Joint Training on Multiple Datasets with Inconsistent Labeling Criteria for Facial Expression Recognition

Chengyan Yu*, Dong Zhang*, Wei Zou, Ming Li†

Abstract—One potential way to enhance the performance of facial expression recognition (FER) is to augment the training set by increasing the number of samples. By incorporating multiple FER datasets, deep learning models can extract more discriminative features. However, the inconsistent labeling criteria and subjective biases found in annotated FER datasets can significantly hinder the recognition accuracy of deep learning models when handling mixed datasets. Effectively perform joint training on multiple datasets remains a challenging task. In this study, we propose a joint training method for training an FER model using multiple FER datasets. Our method consists of four steps: (1) selecting a subset from the additional dataset, (2) generating pseudo-continuous labels for the target dataset, (3) refining the labels of different datasets using continuous label mapping and discrete label relabeling according to the labeling criteria of the target dataset, and (4) jointly training the model using multi-task learning. We conduct joint training experiments on two popular in-the-wild FER benchmark databases, RAF-DB and CAER-S, while utilizing the AffectNet dataset as an additional dataset. The experimental results demonstrate that our proposed method outperforms the direct merging of different FER datasets into a single training set and achieves state-of-the-art performance on RAF-DB and CAER-S with accuracies of 92.24% and 94.57%, respectively.

Index Terms—facial expression recogniton, deep convolutional neural networks, continuous label mapping, joint training.

1 INTRODUCTION

FACIAL Expression Recognition (FER) aims to recognize the discrete expression categories (e.g., anger, fear, happiness, sadness, disgust, surprise and neutral) or the continuous levels (e.g., valence, arousal) from still images or videos. Based on different sample-collecting scenarios, FER datasets can be grouped into lab-controlled FER datasets [1], [2], [3], [4], [5] and in-the-wild FER datasets [6], [7], [8], [9], [10], [11]. Compared with the FER task for lab-controlled datasets, recognizing facial expression from the in-the-wild environmental conditions is more practical and challenging.

As a typical task of pattern recognition, in-the-wild FER is expected to yield improved accuracy with an increased number of training samples. However, existing large-scale in-the-wild FER datasets often suffer from inconsistent annotation standards due to uncertainties arising from annotators' subjectivity and the inherent ambiguity of facial images captured in real-world settings. Annotators with different culture backgrounds may interpret discrete emotions in distinct ways [12], [13]. Because of the annotation bias between different datasets, directly merging multiple FER datasets to expand the training set may not necessarily enhance the performance of FER models.

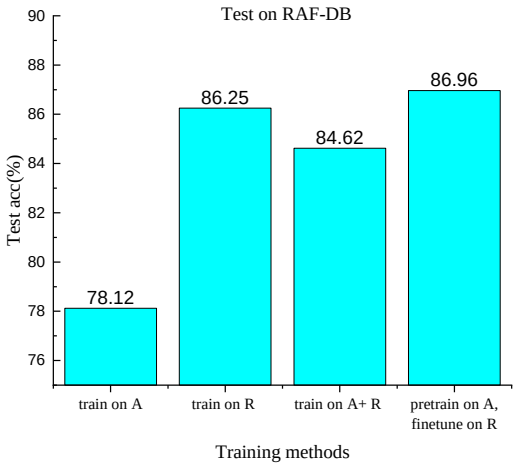

Fig. 1. The performance of discrete label classification on the RAF-DB test set with different training strategies. A denotes training set of AffectNet while R denotes training set of RAF-DB.

As shown in Fig.1, when we use the training set of the RAF-DB dataset [6], [7] for model training, the recognition accuracy on the test set of the RAF-DB dataset is 86.25%. In the case that we only use the training set of the AffectNet or directly merge the training set of RAF-DB and AffectNet for model training, the recognition accuracy decreases by 8.13% and 1.63%, respectively. In the context

Chengyan Yu is with the School of Electronics and Information Technology, Sun Yat-sen University, China, 510006 and the Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Data Science Research Center, Duke Kunshan University, Kunshan, China, 215316. E-mail: yuchy33@mail2.sysu.edu.cn
Dong Zhang and Wei Zou are with the School of Electronics and Information Technology, Sun Yat-sen University, China, 510006. E-mail: zhangd@mail.sysu.edu.cn, zouw23@mail3.sysu.edu.cn
Ming Li is with the Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Data Science Research Center, Duke Kunshan University, Kunshan, China, 215316 and the School of Computer Science, Wuhan University, Wuhan, China, 430072. E-mail: ming.li369@dukekunshan.edu.cn
* *Equal Contributions.*
† *Corresponding author: Ming Li*

of transfer learning, if we pretrain the FER model on the AffectNet dataset and subsequently finetune it on the RAF-DB dataset, the recognition accuracy increases by only 0.71% in comparison to training on RAF-DB and testing on RAF-DB. These experimental findings clearly demonstrate that directly employing transfer learning or merging different FER datasets into a single training set, without unifying the annotation criteria, generally results in limited performance improvement or even degraded performance.

To address this issue, many methods have been proposed in the context of discrete category-based emotion classification. Zeng et al. [12] first consider the problem of inconsistent annotation among different facial expression recognition datasets. They propose an uncertainty learning-based framework that assigns each sample with two labels, including model prediction and human annotation. Then they use the EM algorithm with CNN to discover the latent truth based on the pseudo labels. Wang et al. [13] focus on suppressing these uncertainties to learn better facial expression features. They propose a Self-Cure Network to find the confidence weight of each sample and use a relabel mechanism to modify the labels with low confidence. The aforementioned methods target to minimize the annotation bias among different FER datasets and obtain promising recognition accuracy when merging multiple FER datasets with categorical labels.

In addition to discrete emotion category classification, certain studies consider Facial Expression Recognition (FER) as a task involving the regression of continuous emotion levels [14], [15]. Continuous measurements of emotion typically involve two dimensions: valence, which indicates the negativity or positivity of the emotional display, and arousal, which reflects the calming or exciting nature of the emotional display [14]. When representing emotions on a two-dimensional plane, images labeled as 'happy' tend to be positioned to the right of the vertical axis. Images labeled as neutral typically exhibit relatively lower absolute values of valence and arousal and are situated near the origin on the two-dimensional plane. In summary, images with the same discrete label exhibit a distribution in continuous space when represented as continuous-scale emotion measures.

Fig.2 illustrates the existence of a mapping relationship between discrete and continuous labels. This observation serves as inspiration for quantifying the mismatch in labeling standards for a specific emotion category between two datasets labeled with discrete categories, using statistical measures such as mean value and variance of the continuous labels. These measures allow us to assess the disparity between the datasets in a unified two-dimensional valence-arousal plane. This characteristic presents an opportunity to leverage multi-task learning and effectively merge multiple FER datasets that include both continuous and discrete emotion labels, thereby improving efficiency and performance.

Based on the mapping relationship between the discrete and continuous labels, in this paper, we propose a new noisy label learning method named the Discrete and Continuous labels Joint Training(DCJT) framework for in-the-wild FER. In DCJT, from a large-scale dataset B with both continuous and discrete labels, we first select a subset of samples from B that has consistent annotation criteria with the samples in another discrete dataset A. Then we use the selected subset

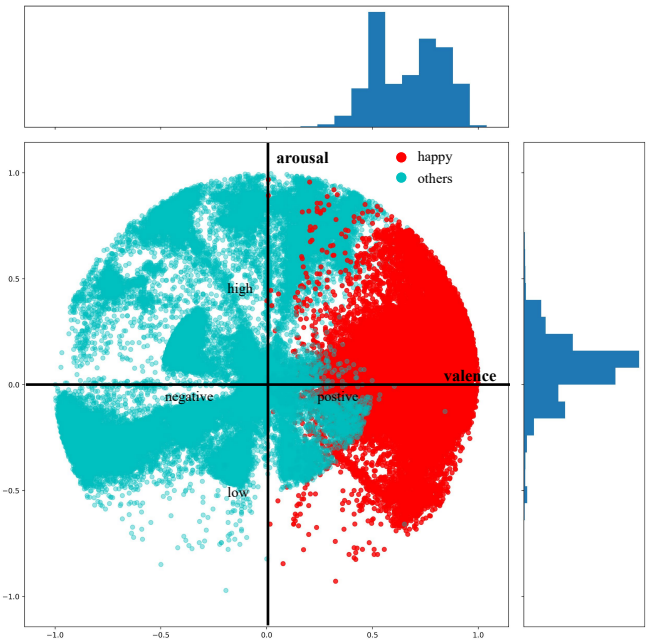

Fig. 2. Visualization of the continuous labels of AffectNet in the two-dimensional valence and arousal space. The histograms show that annotators from AffectNet labeled images with positive valence and small absolute arousal values as happy images.

of B to train a continuous-scale regressor and generate pseudo-continuous labels for all samples in A. Then, we adopt a continuous label mapping to unify the labeling standards according to A's continuous labels.

Furthermore, for those remaining samples in B, we relabel their discrete labels by calculating the Euclidean distance to the class center of each emotion category in the inferred two-dimensional continuous label space of A. Finally, we use the technique of multi-task learning to perform joint training with multiple datasets (e.g., A+B) with both discrete and continuous labels.

Our contribution can be summarized as follows:

We propose a novel noisy label learning method named Discrete and Continuous labels Joint Training (DCJT) for the FER task. In our framework, we leverage continuous emotion labels to map the labels of multiple FER datasets towards the target dataset's standard. We quantify the mismatch of annotation standards between datasets on a unified valence-arousal plane and use multi-task learning, discrete label relabeling and continuous label mapping to perform robust joint training with adapted discrete and continuous labels. We conduct several experiments on the popular in-the-wild facial expression recognition datasets, including RAF-DB and CAER-S. The experimental results show that our method achieves better or comparable performance compared to the state-of-the-art approaches.

2 RELATED WORKS

Facial expression recognition is a popular research topic in computer vision. Before the broad application of deep learning, FER was mainly focused on designing hand-crafted patterns [16], including Local Binary Patterns (LBP) [17], [18], Gabor wavelet coefficients [19], Scale-Invariant

Feature Transform (SIFT) [20], and Histogram of Oriented Gradient (HOG) [21] for features extraction. The performance of these hand-crafted patterns is easily affected by variations in illumination, rotation and occlusion, which limits the application of hand-crafted patterns on in-the-wild FER tasks [22]. With the development of deep learning and computational resource, many neural network-based feature learning methods have been proposed and achieve superior performance on FER [23], [24], [25], [26], [27]. An intuitive idea to further improve the performance of FER method is to merge multiple datasets and enlarge the training set. However, the inconsistent labeling standards or annotation bias among different FER datasets with only discrete labels sometimes results in a degraded performance if we simply combine those two FER datasets together.

2.1 Learning with Noisy Labels

Deep learning has achieved remarkable success with the help of large-scale datasets [28]. However, the low quality of the labels remains an issue. Unreliable labels may lead to overfitting and poor generalization [29]. To remedy this, various methods have been proposed [30], [31], [32], [33].

Similar to other computer vision tasks, the performance of in-the-wild FER is also challenged with ambiguous labels and inconsistent annotations. Due to different settings for data collection and the subjectiveness of annotation, data bias and inconsistent annotations are common among various facial expression datasets [22]. Enlarging the training dataset by directly merging multiple datasets can hardly help the model to improve its performance from training [12].

Many works have been proposed to handle this problem based on the uncertainty learning framework. A common approach involves estimating the latent truth and reweighting the training samples, focusing on high-quality samples and disregarding less reliable ones. Zeng et al. [12] propose the Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) framework, which assigns each sample multiple labels using human annotations or model predictions. IPA2LT utilizes an end-to-end trainable LTNet that discovers the latent truth from inconsistent pseudo labels and input face images [34]. Subsequently, the LTNet is used to train a final end-to-end Facial Expression Recognition (FER) model. Chen et al. [35] propose Label Distribution Learning on Auxiliary Label Space Graphs (LDL-ALSG) to leverage topological information from related tasks and guide label distribution learning for Facial Expression Recognition. They utilize topological information from related tasks, such as action unit recognition and facial landmark detection, to guide label distribution learning in Facial Expression Recognition datasets. The method leverages the deviation between images and predictions of their neighboring images to train the backbone network. It also employs prior knowledge to transform logical labels into discretized bivariate Gaussian label distributions. Wang et al. [13] present a new self-cure network (SCN) for robust feature learning in Facial Expression Recognition by mitigating uncertainties in both synthetic and real-world datasets. They use self-attention and a ranking regularization module to assign weights to each image and split them into high and low-confidence groups. By refining labels in the low-confidence group, the model can robustly learn features while mitigating uncertainties in datasets. She et al. [25] propose DMUE, a solution to address annotation ambiguity in Facial Expression Recognition caused by subjective annotation and inherent inter-class similarity. DMUE employs an uncertainty estimation module to assign confidence scores to samples based on the statistics of their relationships. Weighted training in the target branch is then performed using these confidence scores.

Another typical way is the ones adopting sample selection, namely selecting true-labeled examples from noisy datasets [36], [37]. Veit et al. [38] divide the dataset into subsets with clean and noisy labels. They use a multi-task network to learn a mapping from noisy labels to clean ones. Wang et al. [39] propose Emotion Ambiguity-Sensitive (EASE) cooperative networks to addresses the challenges in learning with noisy label in facial expression recognition. EASE consists of two components: an ambiguity-sensitive learning module that distinguishes between clean, noisy, and ambiguous labels and a diversity enhancing module that enhances the cooperative intelligence of the two networks.

2.2 Multi-task Learning

Multi-task learning is an efficient technique of machine learning in which multiple relevant learning tasks are solved simultaneously while exploiting commonalities and differences across tasks [40]. This can result in improved learning efficiency and prediction accuracy for some task-specific models compared to training the models separately [41], [42], [43].

Numerous existing networks for FER focus on a single task. However, in the real world, FER is affected by various factors, such as head pose, illumination, and subject identity (facial morphology). To address this issue, multi-task learning is introduced to transfer knowledge from other relevant tasks and to disentangle nuisance factors. Several works [44], [45] suggest that simultaneously conducting FER with additional tasks, such as facial landmark localization and facial AU [46] detection, can jointly improve FER performance. Zhang et at. [11] train a multi-signal CNN (MSCNN) under the supervision of both FER and face verification tasks, force the model to focus on expression information. Toisoul et al. [15] propose a deep neutral network that integrates face alignment and jointly estimates both categorical and continuous labels in a single pass.

The aforementioned approaches suggest that there is a potential to benefit from both discrete and continuous labels by exploiting multi-task learning in FER tasks. However, most existing works based on multi-task learning focus on training on a single FER dataset. Moreover, existing methods to remedy noisy labels ignore the potential of exploiting both discrete and continuous labels. Therefore, we propose our Discrete and Continuous labels Joint Learning (DCJT) framework to train a FER model from multiple inconsistently labeled datasets. The proposed framework integrates sample selection and label correction to deal with inconsistent labels. Meanwhile, DCJT benefits from both kinds of labels by means of multi-task learning. The details of our framework are introduced in Section 3.

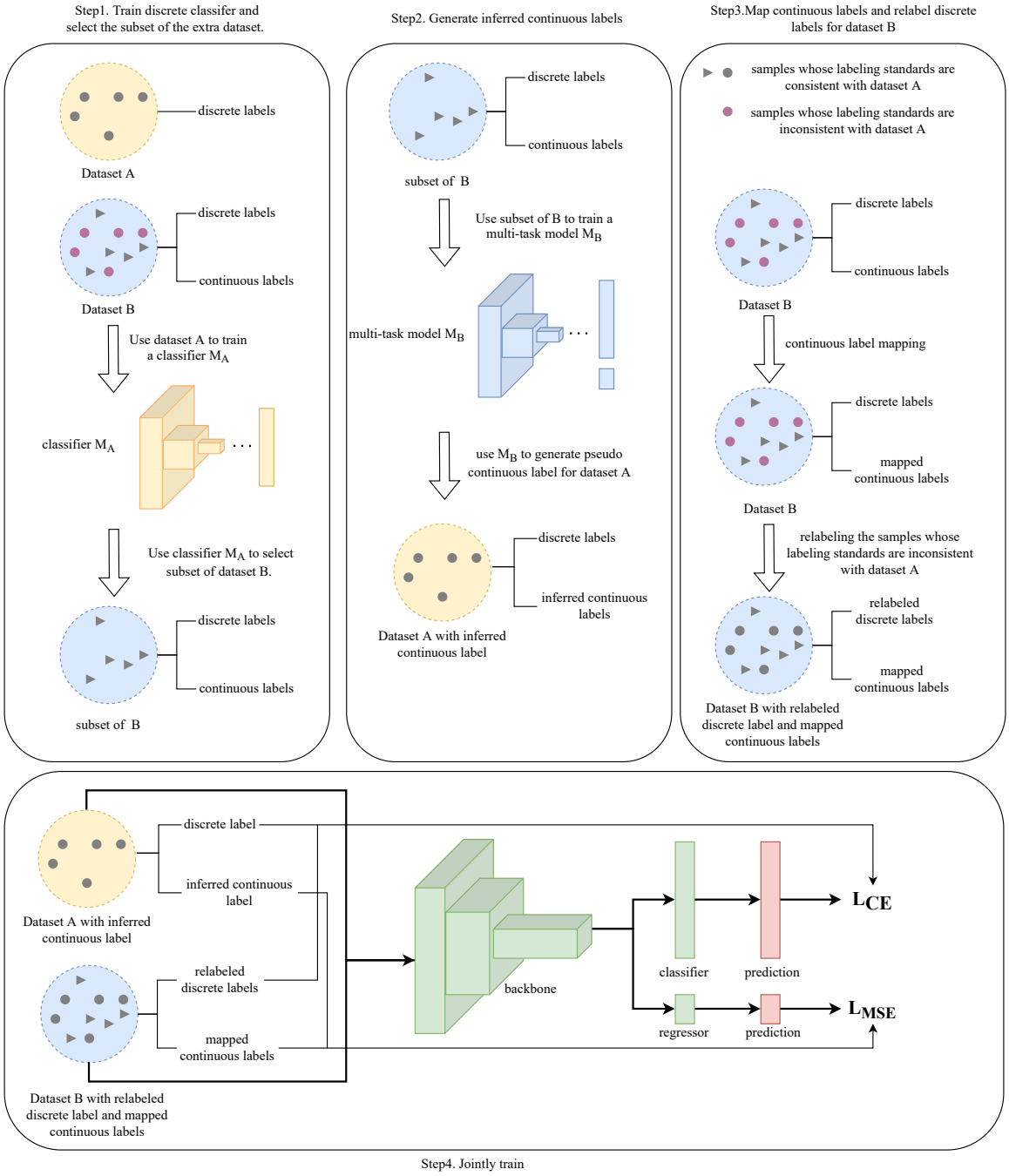

Fig. 3. Overview of our proposed DCJT joint training framework. The framework consists of four steps: training the discrete classifier and selecting the subset of the extra dataset, generating inferred pseudo-continuous labels, mapping continuous labels and relabeling discrete labels, and joint training.

3 THE PROPOSED METHOD

We propose a Discrete and Continuous labels Joint Learning (DCJT) framework for training a Facial Expression Recognition (FER) model using multiple inconsistently labeled datasets. This section provides an overview of our DCJT framework, followed by a detailed description of each step.

3.1 Overview

In order to tackle the disparity in labeling standards, we leverage the mapping relationship between discrete and continuous labels. We assume the continuous labels of the samples in each specific emotion category in any dataset follow a two-dimensional Gaussian distribution. By calculating the mean and covariance of these distributions, we can characterize the labeling standards. We utilize a linear transformation to align the labeling standards across datasets. This linear transformation maps the continuous labels of a specific discrete emotion category from multiple datasets to the distribution of same discrete emotion in the target dataset. This mapping ensures consistency

and enables us to leverage the knowledge from multiple datasets. Furthermore, we incorporate multi-task learning to take advantage of the mapping relationship between discrete and continuous labels. By simultaneously solving both learning tasks, we enhance the learning efficiency and improve the prediction accuracy of the proposed methods. This joint learning approach allows us to extract the underlying patterns and correlations between discrete and continuous labels. Through our proposed method, we aim to leverage the mapping relationship between discrete and continuous labels to address labeling standard discrepancies and enhance the overall performance of facial expression recognition models.

Fig.3 presents an overview of the 4-step DCJT framework. We consider two facial expression recognition datasets: Dataset A, which contains only discrete labels, and Dataset B, which includes both discrete and continuous labels. At this stage, all labels are manually annotated.

Step 1 of the framework, as illustrated in Fig. 3, trains a classifier M_A using the data with discrete labels in the dataset A. By utilizing model M_A, we select a subset from the dataset B and ensure each predicted category aligns with the corresponding discrete label. This selection process aims to align the labeling criteria between dataset A and the selected subset of B. Step 2, the selected subset of dataset B is used to train a multi-task FER model, denoted as M_B. This model is designed to jointly estimate both discrete and continuous labels. Subsequently, M_B is utilized to generate pseudo-continuous labels for dataset A. At this stage, dataset A comprises both discrete and inferred pseudo-continuous labels. Step 3 focuses on aligning the continuous labels of the two datasets. We calculate the values of mean and covariance for the pseudo-continuous labels of the dataset A and the original continuous labels of the dataset B. To achieve consistency, we employ a linear transformation based on the calculated mean and covariance. This transformation maps the distribution of dataset B's continuous labels to match the distribution of dataset A's continuous labels. Consequently, after the transformation, the continuous labels of B are aligned with the continuous labels of A with a similar labeling criteria. Furthermore, the discrete labels of dataset B, excluding the subset selected in Step 1, are adjusted to adhere to the labeling criteria of dataset A. Consequently, dataset A now possesses discrete and inferred continuous labels, while dataset B has relabeled discrete and mapped continuous labels. Lastly, in Step 4, multi-task learning is utilized to train an end-to-end Discrete Continuous Net (DCN). Through this process, the DCN model is trained to to estimate both discrete and continuous emotion labels simultaneously. During the inference phase, the trained DCN can be utilized to predict both the discrete and continuous labels for facial expressions.

With the 4-step DCJT framework, we aim to address the labeling standard discrepancies between datasets A and B, enabling effective training of a facial expression recognition model that can estimate both discrete and continuous emotion labels.

3.2 Generate Pseudo Continous Labels

To generate the pseudo-continuous labels for dataset A, we utilize model M_A, which is trained using samples from dataset A. In Step 1, a subset of samples is selected from dataset B based on the consistency between the predicted labels by model M_A, denoted as $\hat{y}_x^D$, and the discrete human annotations, denoted as y_x^D, for each image x. If the predicted discrete label $\hat{y}_x^D$ matches the discrete human annotation y_x^D, we consider the image and the associated label to be more likely in the same domain of dataset A. Therefore, using these selected samples, our regressor M_B can generate higher-quality continuous labels for samples in dataset A.

After selecting the subset from dataset B, we train a multi-task model, denoted as M_B, using the chosen samples. This model is designed to jointly estimate both discrete and continuous labels. Next, for each image x in dataset A, we utilize model M_B to predict the inferred valence and arousal, denoted as $(\tilde{v}_x, \tilde{a}_x)$. These predicted values are regarded as the pseudo-continuous labels for dataset A.

In summary, through the DCJT framework, the pseudo-continuous labels for dataset A are generated by training model M_A on dataset A and selecting a subset of samples from dataset B based on the consistency between the predicted and annotated discrete labels. Then, model M_B, trained using the selected subset, is used to estimate the continuous labels for dataset A, which are treated as the pseudo-continuous labels.

3.3 Continuous Label Mapping

After completing Step 1 and Step 2, both datasets A and B now have discrete and continuous labels. By utilizing the continuous labels, we can directly quantify the inconsistency and divergence in labeling criteria across different facial expression recognition datasets. This quantification is achieved by calculating the mean and variance of the continuous labels within each dataset.

We assume the continuous labels of the datasets A and B follow two bivariate normal distributions, $\mathcal{N}_\mathcal{A}$ and $\mathcal{N}_\mathcal{B}$, respectively. We denote μ_A and Σ_A the mean and covariance for $\mathcal{N}_\mathcal{A}$, and μ_B and Σ_B the mean and covariance for $\mathcal{N}_\mathcal{B}$.

Assume T_A is a two-dimensional random vector and $T_A \sim \mathcal{N}(\mu_A, \Sigma_A)$, T_B is a two-dimensional random vector and $T_B \sim \mathcal{N}(\mu_B, \Sigma_B)$. We can get the linear transformation from $\mathcal{N}_B$ to $\mathcal{N}_A$ using Equation (1), where P and Q are the cholesky decomposition of Σ_A and Σ_B.

$$T_A = PQ^{-1}(T_B - \mu_B) + \mu_A \tag{1}$$

We introduce three mechanisms for the combination of two datasets

- Combining two datasets without the continuous label mapping: We combine two the datasets A and B without using any mapping on the continuous labels of these datasets.
- Combining two datasets by mapping global continuous labels: We assume the continuous labels of each dataset follow a single bivariate normal distribution. We calculate the global mean and covariance matrix on each dataset for label mapping.
- Combining two datasets by mapping emotion dependent continuous labels: When we combine two FER datasets, we perform continuous label mapping separately for each emotion category.

Algorithm 1 Global Continuous Label Mapping

Input: Dataset A with its estimated continuous labels Y_A^C; Dataset B with its original continuous labels Y_B^C. B can be devided into two parts, the selected subset of B $(subB)$ and the remaining samples in B other than the selected subset $(B - subB)$. The corresponding continuous labels are Y_{subB}^C and Y_{B-subB}^C, respectively.

Output: The mapped continuous labels $\tilde{Y}_{B-subB}^C$

1: Compute the mean μ_A and covariance Σ_A of Y_A^C.
2: Compute the mean μ_B and covariance Σ_B of Y_B^C.
3: Compute the cholesky decomposition of Σ_A: $PP^T = \Sigma_A$
4: Compute the cholesky decomposition of Σ_B: $QQ^T = \Sigma_B$
5: Process the global continuous label mapping $\tilde{Y}_{B-subB}^C = PQ^{-1}(Y_{B-subB}^C - \mu_B) + \mu_A$

Algorithm 2 Emotion Dependent Continuous Label Mapping

Input: Dataset A with its estimated continuous labels Y_A^C; Dataset B with its original continuous labels Y_B^C. B can be devided into two parts, the selected subset of B $(subB)$ and the remaining samples in B other than the selected subset $(B - subB)$. The corresponding continuous labels are Y_{subB}^C and Y_{B-subB}^C, respectively.

Output: The mapped continuous labels $\tilde{Y}_{B-subB}^C$

1: **for** the i^{th} emotion category **do**
2: Compute the mean μ_A^i and covariance Σ_A^i of Y_A^{iC}.
3: Compute the mean μ_B^i and covariance Σ_B^i of Y_B^{iC}
4: Compute the cholesky decomposition of Σ_A^i: $P^i(P^i)^T = \Sigma_A^i$
5: Compute the cholesky decomposition of Σ_B^i: $Q^i(Q^i)^T = \Sigma_B^i$
6: Process the i^{th} emotion dependent continuous label mapping $\tilde{Y}_{B-subB}^{iC} = P^i(Q^i)^{-1}(Y_{B-subB}^{iC} - \mu_B^i) + \mu_A^i$
7: **end for**

A detailed description of global continuous mapping and emotion dependent continuous mapping is presented in Algorithms 1 and 2.

3.4 Relabeling Discrete Labels

After completing the continuous label mapping in Step 3, a question arises about the discrete labels of the remaining samples in dataset B, excluding the selected subset, namely $(B - subB)$. In the DCJT framework, we utilize a simple mechanism to relabel the discrete labels of these samples. This involves computing the Euclidean distance between each sample and the class center of each emotion category in a two-dimensional continuous label space.

Once continuous label mapping is performed, the remaining subset of dataset B, denoted as $(B - subB)$, has its mapped continuous labels $\tilde{Y}_{B-subB}^C$. To compute the class center of each emotion category, we use the inferred pseudo-continuous labels of dataset A. For the emotion category i, the class center, CC^i, can be obtained by calculating the mean of the inferred pseudo-continuous labels associated with the i_{th} class, as shown in Eq.(2).

$$CC^i = \frac{1}{N_i} \sum_{j \in i_{th} class} \tilde{Y}_j^C \tag{2}$$

Then the new label of sample j from the remaining set $B - subB$ is given by

$$\tilde{y}_j^D = \arg\min_i (\tilde{y}_j^C - CC^i)^2. \tag{3}$$

3.5 Joint Training

Multi-task learning has already been used in FER task and obtained promising results [14], [15], [44], [45]. To leverage the advantages of both types of labels, we introduce our joint training framework based on multi-task learning. After completing Step1, 2 and 3, the dataset A has two kinds of labels, i.e., the original discrete labels and the inferred pseudo-continuous labels. The dataset B also has its relabeled discrete labels and the mapped continuous labels. In Step. 4, we train our model to jointly estimate both discrete and continuous labels of input samples. A detailed description of mini-batch training is presented in Algorithm 3. We specify several input hyperparameters, including the maximum epoch, warm-up epoch, and joint training frequency. During the warm-up epoch, which is the initial phase of training, only images from dataset A with their original discrete labels Y_A^D and its inferred continuous labels $\tilde{Y}_A^C$ are used for training. Within each batch, the model is jointly trained with both discrete and continuous labels. We employ a loss function that consists of two components: a cross-entropy loss for the categorical loss (discrete emotion classes) and a mean squared error (MSE) loss for the continuous loss where $\hat{Y}_A^D$ and $\hat{Y}_A^C$ denote the predicted discrete and continuous labels.

$$\begin{aligned} L_{MSE}(\tilde{Y}_A^C, \hat{Y}_A^C) &= MSE(\tilde{Y}_A^C, \hat{Y}_A^C) \\ L_{CE}(Y_A^D, \hat{Y}_A^D) &= CrossEntropy(Y_A^D, \hat{Y}_A^D) \\ &= -\Sigma_{i=1}^n \hat{y_i} log(y_i) \end{aligned} \tag{4}$$

So the overall loss function minimized by our model is given by Equation (5):

$$L(Y_A, \hat{Y}_A) = L_{CE}(Y_A^D, \hat{Y}_A^D) + \lambda L_{MSE}(\tilde{Y}_A^C, \hat{Y}_A^C) \tag{5}$$

where λ is a hypeparameter to balance two components. If the epoch number is greater than the warm-up epoch and is divisible by the joint training frequency, we include additional images from dataset B with their relabeled discrete labels $\tilde{Y}_B^D$ and mapped continuous labels $\tilde{Y}_B^C$ for training the model. In such cases, the training process consists of two stages. The first stage involves training the model using the images from dataset A and jointly considering both discrete and continuous labels as shown in Equation (5). In the second stage, we use images from the extra dataset B.

$$L(\tilde{Y}_B, \hat{Y}_B) = L_{CE}(\tilde{Y}_B^D, \hat{Y}_B^D) + \lambda L_{MSE}(\tilde{Y}_B^C, \hat{Y}_B^C) \tag{6}$$

Algorithm 3 Joint Training

Input: Dataset A, with training images X_A,
discrete labels Y_A^D and estimated continuous labels $\tilde{Y}_A^C$;
Dataset B with training images X_B,
relabeled discrete labels $\tilde{Y}_B^D$ and mapped continuous labels $\tilde{Y}_B^C$;
Max Epoch;
Warm-up Epoch;
joint training frequency;
Output: Trained model with classifer branch θ^c and regressor branch θ^r

Initialize θ^c and θ^r with random values
while epoch<Max Epoch **do**
  if epoch<Warm-up epoch **then**
    From $(X_A, Y_A^D, \tilde{Y}_A^C)$, sample a batch;
    Compute $L = L_{CE}(Y_A^D, \hat{Y}_A^D) + \lambda L_{MSE}(\tilde{Y}_A^C, \hat{Y}_A^C)$
    Update classifer branch θ^c and regressor branch θ^r
  else
    if epoch%joint training frequency == 0 **then**
      From $(X_A, Y_A^D, \tilde{Y}_A^C)$, sample a batch;
      Compute $L = L_{CE}(Y_A^D, \hat{Y}_A^D) + \lambda L_{MSE}(\tilde{Y}_A^C, \hat{Y}_A^C)$
      Update classifer branch θ^c and regressor branch θ^r
      From $(X_B, \tilde{Y}_B^D, \tilde{Y}_B^C)$ sample a batch;
      Compute $L = L_{CE}(\tilde{Y}_B^D, \hat{Y}_B^D) + \lambda L_{MSE}(\tilde{Y}_B^C, \hat{Y}_B^C)$
      Update classifer branch θ^c and regressor branch θ^r
    else
      From $(X_A, Y_A^D, \tilde{Y}_A^C)$, sample a batch;
      Compute $L = L_{CE}(Y_A^D, \hat{Y}_A^D) + \lambda L_{MSE}(\tilde{Y}_A^C, \hat{Y}_A^C)$
      Update classifer branch θ^c and regressor branch θ^r
    end if
  end if
end while

4 EXPERIMENTS

4.1 Datasets

In this study, we evaluate our proposed method on three widely used in-the-wild facial expression recgonition datasets, namely the RAF-DB [6], [7] dataset, the CAER-S [9] dataset and the AffectNet dataset [8].

- RAF-DB: The RAF-DB [6], [7] dataset contains 30,000 facial images annotated with basic or compound expressions. For our experiments, we focus on images with basic emotion categories, specifically using 12,271 training samples and 3,068 test samples. The annotations in RAF-DB are obtained through a combination of 40 human coders and crowdsourcing techniques.
- CAER-S: The CAER-S [9] dataset is derived from the CAER dataset. The CAER-S dataset contains 65,983 images. It is divided into a training set with 44,996 samples and a test set with 20,987 samples. Each image in CAER-S is assigned with one of seven discrete expressions: neutral, happiness, sadness, surprise, fear, disgust, and anger.
- AffectNet: The AffectNet [8] dataset comprises over one million images and provides annotations for both discrete emotional labels and continuous labels. The images in AffectNet are obtained from the Internet by querying major search engines using 1,250 emotion-related keywords. Among the annotated images, 450,000 were manually labeled with 11 discrete emotion labels. For our experiments, we focus on the seven expression categories and their continuous labels. AffectNet serves as an additional dataset with both discrete and continuous labels.

4.2 Implementation Details

For the RAF-DB dataset, we use aligned images with seven basic discrete labels. Furthermore, images from RAF-DB are resized to 224×224 pixels. For the CAER-S dataset, we first detect and align all faces using similarity transformation. Then we resize them to 224×224 pixels. In our experiments, we incorporate AffectNet as an additional FER dataset. We begin with cropping face images based on the provided bounding boxes from the AffectNet annotations. Subsequently, we perform similarity transformation to align the faces and resize them to 224×224 pixels. To augment the data, we apply random horizontal flipping and random erasing.

To generate pseudo-continuous labels for RAF-DB and CAER-S, we employ a multi-task model trained on AffectNet. When training the multi-task model with the entire AffectNet, we utilize all 283,901 images comprising seven discrete expression categories. For the multi-task models trained with selected subsets of AffectNet, we initially train a classifier on RAF-DB and CAER-S. Subsequently, we employ the trained classifier to select samples from AffectNet. Finally, using the selected samples, we train the multi-task models and generate pseudo labels for RAF-DB and CAER-S.

In the case of the RAF-DB dataset, our network is trained using a batch size of 128, an initial learning rate of 0.001, and ADAM as the optimizer. The hyperparameter λ is set to 15, the warm-up epoch is set to 30, and the joint frequency is set to 10. Regarding the CAER-S dataset, our network is trained using a batch size of 128, an initial learning rate of 0.01, and ADAM [47] as the optimizer. Additionally, we set the hyperparameter λ to 100, the warm-up epoch to 100, and the joint frequency to 5. We implement our approach using the PyTorch toolbox on the GeForce RTX 1080Ti platform.

4.3 Ablation Study

To assess the effectiveness of each step in our DCJT framework, we conduct an ablation study on RAF-DB using two different backbones: Resnet18 [48] and ARM [27]. Note that for experiments requiring an additional dataset, we utilize the 7-class AffectNet training set as an additional training set. For this ablation study, we conduct the following step-by-step experiments.

- Expriment I: In the first experiment, we utilize the RAF-DB dataset to train a classifier model, denoted

as M_A. Additionally, we employ this trained model M_A to select a subset of the AffectNet dataset. The training set is then formed by the combination of the RAF-DB training set, the selected subset of AffectNet, and their corresponding discrete labels. Finally, we evaluate the performance of the model on the RAF-DB test set. The objective of this experiment is to demonstrate the effectiveness of incorporating a selected subset of data from AffectNet, which shares similar annotation criteria, in improving the overall performance of the classifier.

- Expriment II: For the second experiment, we train a multi-task model by incorporating the generated pseudo-continuous labels derived from RAF-DB, along with their original discrete labels. The objective of this experiment is to demonstrate the potential improvements achieved through the utilization of additional pseudo-continuous labels in conjunction with multi-task learning.
- Expriment III: The results obtained from the two previous experiments provide evidence for the following observations: 1) the addition of the selected subset of AffectNet to RAF-DB can improve the performance of our model in the discrete emotion classification task, and 2) the generation of pseudo-continuous labels for RAF-DB and the utilization of multi-task learning also contribute to enhanced performance.
  Based on these findings, it is natural to shift our focus towards joint training, involving RAF-DB and the selected subset of AffectNet, leveraging the benefits of multi-task learning. Thus, in the third experiment, we train a multi-task model and evaluate its performance on the RAF-DB test set. The model is trained on RAF-DB, which includes both discrete and inferred pseudo-continuous labels, and the selected subset of AffectNet, which consists of discrete and continuous labels. The aim of this experiment is to investigate whether our model can derive advantages from the additional data provided by both the selected subset of AffectNet and the pseudo-continuous labels.
- Expriment IV: The experimental findings indicate that our model experiences advantages when incorporating additional data from the selected subset of AffectNet and utilizing pseudo-continuous labels. However, our experiments have solely involved the utilization of RAF-DB and the selected subset of AffectNet. Nevertheless, there is a portion of AffectNet data that remains untapped, specifically the remaining data in dataset B, excluding the selected subset. Consequently, we proceed with the following set of experiments: we employ RAF-DB, which comprises discrete labels and inferred continuous labels, in conjunction with the entire AffectNet dataset, wherein the discrete labels have been relabeled and the continuous labels have been mapped. This particular experiment yields the most optimal performance for our model.

The results of experiment I, II and III are shown in Table 1. Table 2 shows the results of experimet IV.

4.3.1 Effectiveness of Selecting Subset of Extra Dataset

Table 1 demonstrates the enhanced performance of our model through the combination of RAF-DB and the selected subset of AffectNet. When employing Resnet18 [48] as the backbone, the inclusion of additional data leads to a notable improvement of 1.20% compared to training with the RAF-DB dataset alone. Similarly, when utilizing ARM [27] as the backbone, the incorporation of extra data results in a performance boost of 0.25% compared to using only the RAF-DB dataset. These results suggest that the selected subset of the AffectNet is labeled with a similar criteria to the RAF-DB dataset.

4.3.2 Effectiveness of Generating Pseudo Continuous Label

We conducted experiments to validate the effectiveness of generating pseudo-continuous labels for RAF-DB and to investigate the utilization of a selected subset for this purpose. As shown in Table 1, the results indicate that incorporating pseudo-continuous labels to train a multi-task model improves performance compared to using only RAF-DB with its discrete annotation for classifier training. This implies that multi-task models can derive benefits from the pseudo-continuous labels generated by our DCJT approach. Additionally, the group that employs the selected subset of AffectNet for generating pseudo-continuous labels (DL & CLS) for RAF-DB outperforms the group using the entire AffectNet dataset for generating pseudo-continuous labels (DL & CLW) for RAF-DB. When using Resnet18 as the backbone, compared to those using pseudo-continuous labels generated by the whole AffectNet, using the selected subset to generate pseudo-continuous labels brings an improvement of 0.45%. The group that generates pseudo-continuous labels using selected subset also performs better when using ARM as the backbone.

In Experiment III, we augment the training set by incorporating a selected subset of AffectNet. The experimental results presented in Table 1 demonstrate that our model achieves the highest performance when trained using both RAF-DB and the selected subset of AffectNet, utilizing both discrete and continuous labels. This suggests that our model can derive advantages not only from the inclusion of additional selected samples but also from the utilization of pseudo-continuous labels. Unsurprisingly, the group that generated pseudo-continuous labels using the selected subset of AffectNet achieved superior performance.

Moreover, as presented in Table 2, regardless of the backbone model used, the results obtained using pseudo-continuous labels generated by the entire AffectNet consistently exhibit poorer performance compared to those obtained using pseudo-continuous labels generated by the selected subset of the AffectNet. One possible explanation for this observation is that the selected subset shares similar labeling criteria with the samples in the RAF-DB dataset. Therefore, the model trained with the selected subset can give higher quality continuous labels in terms of performance on RAF-DB compared to the model trained with the entire AffectNet.

TABLE 1
Evaluation of pseudo-continuous label generation and subset selection in our DCJT on discrete FER dataset RAF-DB using Resnet18, ARM as backbone pre-trained on ImageNet. DL, CLE and CLS denote discrete labels, continuous labels generated using the entire AffectNet, and continuous labels generated using a selected subset of AffectNet.

Backbones	Training Datasets		Training Labels			Acc.(%)
	RAF-DB	RAF-DB and selected subset of AffectNet	DL	DL & CLE	DL & CLS	
Resnet18	✓		✓			86.25
	✓			✓		87.03
	✓				✓	87.48
		✓	✓			87.45
		✓		✓		87.19
		✓			✓	87.58
ARM	✓		✓			90.42
	✓			✓		90.15
	✓				✓	90.83
		✓	✓			90.67
		✓		✓		90.93
		✓			✓	91.72

TABLE 2
Evaluation of continuous label mapping and discrete label relabeling in our DCJT system on the RAF-DB test set using Resnet18, ARM as the backbone pre-trained on ImageNet. RAF-DB training set and the entire AffectNet training set are used as the training set. The DL, CLE, CLS, GCLM, EDCLM denote discrete labels, continuous labels generated using the entire AffectNet, continuous labels generated using the selected subset of AffectNet, global continuous label mapping and emotion dependent continuous label mapping. * Denotes the results are pretrained on MSCeleb [49].

backbone	Training set								Acc.(%)
	RAF-DB		selected subset of AffectNet	remaining samples in Affectnet					
	labels		labels	discrete labels		continuous labels			
	DL & CLE	DL & CLS	DL & CL	original DL	relabeled DL	original CL	GCLM	EDCLM	
Resnet18	✓		✓	✓		✓			86.27
	✓		✓	✓			✓		86.83
	✓		✓	✓				✓	86.89
		✓	✓	✓		✓			87.15
		✓	✓	✓			✓		87.28
		✓	✓	✓				✓	87.39
	✓		✓		✓	✓			87.05
	✓		✓		✓		✓		86.92
	✓		✓		✓			✓	86.93
		✓	✓		✓	✓			87.31
		✓	✓		✓		✓		87.50
		✓	✓		✓			✓	87.74(**88.48***)
ARM	✓		✓	✓		✓			90.61
	✓		✓	✓			✓		91.39
	✓		✓	✓				✓	91.41
		✓	✓	✓		✓			91.49
		✓	✓	✓			✓		91.57
		✓	✓	✓				✓	91.81
	✓		✓		✓	✓			91.36
	✓		✓		✓		✓		91.37
	✓		✓		✓			✓	91.27
		✓	✓		✓	✓			91.76
		✓	✓		✓		✓		92.01
		✓	✓		✓			✓	**92.24**

4.3.3 Effectiveness of Continuous Labels Mapping and Discrete Labels Relabeling

Based on the findings of our previous experiments, we have established that our model can derive advantages from the incorporation of the selected subset of AffectNet and the utilization of pseudo-continuous labels. Consequently,

we intend to explore strategies to harness the untapped potential of the remaining data in AffectNet, which extends beyond the selected subset. We do not use this data in the previous experiment for the samples possess different labeling criteria compared to those present in RAF-DB. Hence, we propose the utilization of continuous label mapping and discrete label relabeling methods to effectively harness this segment of the data and enhance the performance of our multi-task model. For experiment IV, we utilize the RAF-DB training set in conjunction with the complete 7-class AffectNet training set as the training dataset. Specifically, the selected subset of AffectNet is employed with its original continuous and discrete labels, while the remaining AffectNet data undergoes relabeling of the discrete labels and mapping of the continuous labels.

Now we want to figure out which continuous label mapping methods perform best. As shown in Table 2, the group utilizing emotion-dependent continuous label mapping (EDCLM) achieves a performance of 87.74% when using Resnet18 as the backbone (pretrained on ImageNet [50]). This group shows improvements of 0.43% and 0.24% compared to the systems without continuous label mapping (Without CLM) and with global continuous label mapping (GCLM), respectively. The group utilizing emotion-dependent continuous label mapping (EDCLM) achieves the highest performance of 92.24% accuracy when using ARM as the backbone. Therefore, these experiments demonstrate the effectiveness of our proposed approach of emotion-dependent continuous label mapping.

We then turn to discrete label relabeling. The results in Table 2 demonstrate that utilizing the original discrete labels of AffectNet and RAF-DB leads to a decrease in model performance. The Resnet18 and ARM backbones achieve the highest recognition accuracies of 87.39% and 91.81%, respectively, when using the original discrete labels, which are lower than the performance (87.74% and 92.24%) obtained using the relabeled discrete labels. In conclusion, the discrete label relabeling mechanism proves to be advantageous when incorporating the remaining data beyond the selected subset in AffectNet.

4.4 Comparison with Other Joint Training Methods

To tackle the problem of annotation inconsistency in discrete FER tasks, numerous studies have explored uncertainty learning techniques [13], [39], [51]. In contrast to our framework, previous approaches primarily emphasize joint training using discrete labels. In order to compare the performance of joint training with discrete labels and our method, we conduct experiments using two backbones on two discrete FER datasets.

We conducted experiments using four different training methods and selected the 7-class AffectNet training set as an additional FER dataset.

We compare various joint training methods on the discrete FER datasets, RAF-DB and CAER-S, and use Resnet18 [48] and ARM [27] as backbone CNNs. The results of each joint training method are presented in Tables 3 and 4. When evaluating on the test set of RAF-DB, merging the two FER datasets directly and performing joint training with discrete labels result in degraded performance of 84.63% and 89.27% using Resnet18 and ARM as backbone architectures. SCN based joint training method yields performance of 88.14% and 84.31% when using Resnet18 and ARM as backbone architectures, respectively. SCN improves the recognition accuracy by 1.89% when Resnet18 is used as the backbone, compared to training solely on RAF-DB. However, when ARM is employed as the backbone, SCN results in a degradation of recognition accuracy by 0.36%. This suggests that SCN lacks generalization across different backbones. Joint training the model using DCJT results in performance of 88.48% and 92.24% when using Resnet18 and ARM as backbone architectures, respectively. These results highlight that our joint training approach achieves the highest accuracy on both backbones. This suggests that joint training with both discrete and continuous labels, along with multi-task learning, yields superior performance compared to joint

TABLE 3
Comparison of different joint training methods on the RAF-DB test set based on the ARM, Resnet18 backbones. * denotes these results are reproduced by us.

Backbone	Joint-learning method	Acc.(%)
Resnet18	Without extra datasets	86.25
	Combination straightly	84.62
	SCN [13]	88.14
	DCJT	**88.48**
ARM	Without extra datasets	90.42
	Combination straightly	89.27
	SCN [13]	91.06*
	DCJT	**92.24**

Without extra dataset means we use only RAF-DB dataset and the associated discrete labels.
Combination straightly means combining two datasets' images without any other operation.
SCN [13] is a method for joint training using discrete labels. We borrow its key idea and using in different backbones.
DCJT is our proposed framework for multi-task learning on multiple different FER datasets using continuous label mapping and discrete label relabeling.

TABLE 4
Comparison of different joint training methods on the CAER-S test set based on the ARM, Resnet18 backbones. * denotes these results are reproduced by us.

Backbone	Joint-learning method	Acc.(%)
Resnet18	Without extra datasets	84.67
	Combination straightly	81.45
	SCN [13]	84.31*
	DCJT	**86.39**
ARM	Without extra datasets	91.54
	Combination straightly	84.36
	SCN [13]	90.39*
	DCJT	**94.57**

Without extra dataset means we use only RAF-DB dataset and the associated discrete labels.
Combination straightly means combining two datasets' images without any other operation.
SCN [13] is a method for joint training using discrete labels. We borrow its key idea and using in different backbones.
DCJT is our proposed framework for multi-task learning on multiple different FER datasets using continuous label mapping and discrete label relabeling.

TABLE 5
Comparison with state-of-the-art method on RAF-DB

Methods	Acc.(%)	
IPA2LT [12]	86.77	With extra data
RAN [51]	86.90	
SCN [13]	88.14	With extra data
MANET [24]	88.40	
DMUE [25]	88.76	
EASE [39]	89.56	
EAC [52]	89.99	
ARM [27]	90.42	
ARM+DCJT	**92.24**	With extra data

TABLE 6
Comparison with state-of-the-art method on CAER-S. * denotes the result is reproduced by us.

Methods	Acc.(%)	
Resnet18 [48]	84.67	
Resnet50 [48]	84.81	
Res2Net [53]	85.35	
CAER-NET-S [9]	73.51	
MANET [24]	88.42	
EASE [39]	90.95	
EAC [52]	91.33*	
ARM [27]	91.54	
ARM+DCJT	**94.57**	With extra data

training with discrete labels alone.

A similar result is observed in another dataset, CAER-S. When evaluating on CAER-S, merging the two FER datasets directly and performing joint training result in degraded performance of 81.45% and 84.36% when using Resnet18 and ARM as backbone architectures. Interestingly, when evaluating on CAER-S, employing the SCN joint training method yields recognition rates of 84.31% and 90.39% on both backbone architectures. The performance of the SCN joint training method surpasses that of directly combining both datasets; however, it does not match the performance achieved by training solely on the CAER-S dataset. One possible explanation for this is that, despite SCN's correction of some discrete labels with partial labeling standard mismatch, the issue of label inconsistency still exists on CAER-S. Now, focusing on our DCJT joint training method, the results demonstrate that our approach achieves the highest performance of 86.39% and 94.57% using Resnet18 and ARM as backbone architectures, respectively. These results suggest that our joint training method is superior to the SCN method.

4.5 Comparison with Other State-of-the-art Methods

Table 5 compares our method to several state-of-the-art methods on the test set of RAF-DB. RAN [51] and MA-NET [24] focus on attention methods. RAN utilizes face regions and original faces with a cascade attention network. MA-NET uses attention methods of different scales and fuses features of different scales. The accuracy of our method is 5.34% and 3.84% higher than RAN and MA-Net. DMUE [25] makes use of the latent distribution in the label space and estimates the ambiguity extent in the instance space. Nonetheless, DMUE does not use additional data for joint training. IPA2LT [12] and SCN [13] provide joint training methods using discrete labels. IPA2LT introduces the idea of latent ground truth for training with inconsistent annotations across different FER datasets. SCN uses self-attention to weight the training samples and relabel the samples in the low-quality group using the weights. Our method demonstrates accuracy improvements of 5.47% and 4.10% over IPA2LT and SCN, respectively. EASE [39] and EAC [52] aim to address the problem of noisy labels in training FER models. However, these methods only take into account discrete labels during training. In comparison, our method achieves accuracy improvements of 2.68% and 2.25% over EASE and EAC, respectively. The results from Table 5 indicate that joint training using continuous labels is potentially more effective by utilizing the relationship between discrete and continuous labels. Therefore, our method outperforms different methods in Table 5.

Table 6 compares our method with multiple state-of-the-art methods on the CAER-S test set. The results in Table 6 demonstrate that our DCJT method outperforms the state-of-the-art methods, even when compared with deeper networks, such as Resnet50 and Res2Net50.

4.6 Visualization of The Learned Continuous Labels

To further evaluate the effectiveness of continuous label mapping and discrete label relabeling, we conducted a visualization analysis on the RAF-DB training set and the seven-class AffectNet training set. In Fig. 4, we present the visualizations of sample images with predicted continuous labels and relabeled discrete labels. In the first row of the visualization, the pseudo-continuous labels generated by the selected subset of the AffectNet exhibit meaningful patterns. These labels provide confirmation of the relationship between discrete and continuous labels, reinforcing their alignment. Moving to the second row, we compare the original continuous labels of the images in AffectNet with the mapped continuous labels. The visualization clearly indicates that after applying continuous label mapping, the continuous labels of the AffectNet images become more closely aligned with the standard continuous labels of the RAF-DB dataset. This demonstrates the effectiveness of the continuous label mapping technique in achieving better alignment between datasets. Finally, in the third row of the visualization, we examine the original discrete labels of the images in comparison to the relabeled discrete labels. It is evident that after the discrete label relabeling process, these images are assigned new discrete labels that are more rational and in line with the discrete labeling criteria of the RAF-DB dataset. This further validates the effectiveness of the discrete label relabeling approach. Overall, these visualizations provide compelling evidence of the successful integration of continuous label mapping and discrete label relabeling techniques, highlighting their ability to improve the consistency and alignment of labels across datasets.

4.7 Discussion on The Generalization Capabilities

In our method, we could utilize additional samples with adapted labels that match the labeling standard of the target

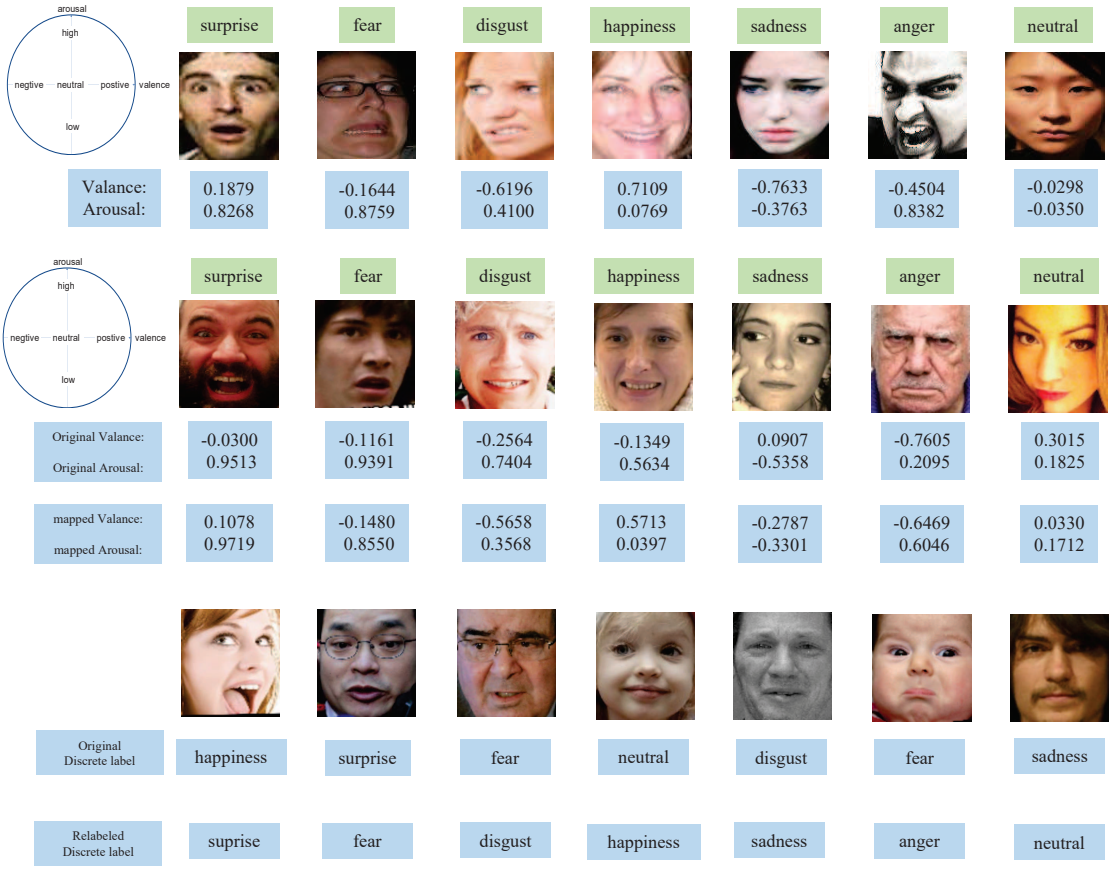

Fig. 4. Visualization of the predicted labels on the RAF-DB training set and the AffectNet training set. The first row shows the predicted continuous labels for RAF-DB. The second row shows the original and mapped continuous labels of AffectNet. The third row shows the original and relabeled discrete labels of AffectNet.

set A, hence the performance in the test set of A is improved. To further validate the generalization capabilities of the final resulted model towards other new datasets, we use the trained model to perform finetuning on the training set of FERPlus and test on the test set of FERPlus. The results are shown in Table 7.

Table 7 that combing multiple datasets with our proposed method achieves a more robust model against directly merging in the task of finetuning and testing on other target datasets.

Moreover, in real-world scenarios, sometimes we do not have another dataset B as large as AffectNet. We perform another experiment by setting the size of dataset B as different portions of AffectNet. In this case, we can evaluate the performance of our proposed method against different scale of the additional dataset B. From the results in Table 8, we can observe that our proposed method is still effective even if we have a small scale additional dataset B.

TABLE 7
Comparison of different pre-training methods on the FERPlus dataset using ARM as the backbone. The training set of FERPlus is used for finetuning while the test set of FERPlus is adopted for testing.

Pre-training data & Strategy	Acc.(%)
RAF-DB only	88.09
AffectNet only	88.19
RAF-DB + AffectNet (directly combine)	88.38
RAF-DB + AffectNet (our proposed method)	88.64
RAF-DB + AffectNet + CAER-S (directly combine)	88.45
RAF-DB + AffectNet + CAER-S (our proposed method)	**88.95**

TABLE 8
The results of our proposed method using different proportion of the additional dataset B.

Dataset A and B	Accuracy on RAF-DB test set (%)
RAF-DB only	90.42
RAF-DB + 10% Affectnet	91.10
RAF-DB + 50% Affectnet	91.85
RAF-DB + 100% Affectnet	**92.24**

5 CONCLUSION

In this paper, we present a novel framework called D-CJT (Discrete and Con- tinuous labels Joint Training) for jointly training multiple facial expression recognition (FER) datasets with the goal of addressing the issue of inconsistent labels in in-the-wild FER tasks.

Our proposed framework incorporates three key mechanisms: subset selection, continuous label mapping, and discrete label relabeling. The subset selection mechanism ensures that high-quality pseudo-continuous labels are generated by training regressors on a carefully selected subset of the AffectNet dataset. This helps improve the reliability and accuracy of the continuous labels used in training. The continuous label mapping mechanism plays a crucial role in reducing the mismatch between label standards across different FER datasets. By mapping the continuous labels of images in the AffectNet dataset to be closer to the standard continuous labels of the target RAF-DB dataset, we enhance the alignment between the datasets and improve the overall performance. Additionally, we leverage multi-task learning to enhance the learning efficiency and prediction accuracy of our framework. By jointly training the model on both discrete and continuous labels, we exploit the complementary information between these two label types, leading to improved performance in facial expression recognition tasks.

We conducted extensive experiments on two widely used in-the-wild FER datasets, and the results demonstrate that our DCJT framework achieves state-of-the-art performance. By effectively addressing the issue of inconsistent labels through subset selection, continuous label mapping, and discrete label relabeling, our framework offers a promising approach for improving the accuracy and reliability of facial expression recognition systems in real-world scenarios.

ACKNOWLEDGMENT

This research is funded in part by the Science and Technology Program of Guangzhou City (202007030011), National Natural Science Foundation of China (62173353,62171207) and DKU Interdisciplinary Seed Grant. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

REFERENCES

[1] 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 *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[2] M. Valstar, M. Pantic *et al.*, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. Paris, France., 2010, p. 65.

[3] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

[4] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and vision computing*, vol. 29, no. 9, pp. 607–619, 2011.

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2106–2112.

[6] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[7] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[8] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[9] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 143–10 152.

[10] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.

[11] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

[12] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

[13] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6897–6906.

[14] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE transactions on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[15] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, 2021.

[16] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[17] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognition and Image Analysis*, vol. 17, no. 4, pp. 592–598, 2007.

[18] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[19] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467–476, 2002.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[22] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[23] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, and X. Li, "Facial expression recognition with identity and emotion joint learning," *IEEE Transactions on affective computing*, vol. 12, no. 2, pp. 544–550, 2018.

[24] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021.

[25] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.

[26] W. Zou, D. Zhang, and D.-J. Lee, "A new multi-feature fusion based convolutional neural network for facial expression recognition," *Applied Intelligence*, vol. 52, no. 3, pp. 2918–2929, 2022.

[27] J. Shi, S. Zhu, and Z. Liang, "Learning to amend facial expression representation via de-albino and affinity," *arXiv preprint arXiv:2103.10189*, 2021.

[28] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[29] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.
[30] N. Manwani and P. Sastry, "Noise tolerance under risk minimization," *IEEE transactions on cybernetics*, vol. 43, no. 3, pp. 1146–1151, 2013.
[31] B. Van Rooyen, A. Menon, and R. C. Williamson, "Learning with symmetric label noise: The importance of being unhinged," *Advances in neural information processing systems*, vol. 28, 2015.
[32] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.
[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
[35] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 984–13 993.
[36] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *International conference on machine learning*. PMLR, 2018, pp. 2304–2313.
[37] E. Malach and S. Shalev-Shwartz, "Decoupling "when to update" from" how to update"," *Advances in neural information processing systems*, vol. 30, 2017.
[38] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 839–847.
[39] L. Wang, G. Jia, N. Jiang, H. Wu, and J. Yang, "Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 218–227.
[40] K.-H. Thung and C.-Y. Wee, "A brief review on multi-task learning," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 705–29 725, 2018.
[41] J. Baxter, "A model of inductive bias learning," *Journal of artificial intelligence research*, vol. 12, pp. 149–198, 2000.
[42] S. Thrun, "Is learning the n-th thing any easier than learning the first?" *Advances in neural information processing systems*, vol. 8, 1995.
[43] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
[44] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *2014 Canadian conference on computer and robot vision*. IEEE, 2014, pp. 98–103.
[45] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," *arXiv preprint arXiv:1802.06664*, 2018.
[46] E. L. Rosenberg and P. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 2020.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[49] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.
[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[51] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.
[52] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*. Springer, 2022, pp. 418–434.
[53] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

Chengyan Yu received his Bachelor's degree from Taiyuan University of Technology in 2021. He is currently a postgraduate student of the school of Electronics and Information Technology, Sun Yat-sen University and a research intern in Data Science Research Center at Duke Kunshan University. His research interests include deep learning and facial expression recognition.

Dong Zhang received his B.S.E.E. and M. S. degrees from Nanjing University in 1999 and 2003, respectively, and Ph.D. from Sun Yat-sen University in 2009. He is currently an Associate Professor of the School of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, computer vision, affective computing, and information hiding.

Wei Zou received his bachelor's degree from Xidian University, China in 2019 and master's degree from Sun Yat-sen University, China in 2022 respectively. His research interests include facial expression recognition and disentangled representation learning.

Ming Li (Senior Member, IEEE) received his Ph.D. in Electrical Engineering from University of Southern California in 2013. He is currently an Associate Professor of Electrical and Computer Engineering at Duke Kunshan University. He is also an Adjunct Professor at School of Computer Science in Wuhan University. His research interests are in the areas of audio, speech and language processing as well as multimodal behavior signal processing. He has published more than 180 papers and served as the member of IEEE speech and language technical committee, APSIPA speech and language processing technical committee. He is an area chair at Interspeech 2016, 2018, 2020 and 2024, as well as the technical program co-chair of Odyssey 2022 and ASRU 2023. Works co-authored with his colleagues have won first prize awards at Interspeech Computational Paralinguistic Challenges 2011, 2012 and 2019, ASRU 2019 MGB-5 ADI Challenge, Interspeech 2020 and 2021 Fearless Steps Challenges, VoxSRC 2021, 2022 and 2023 Challenges, ICASSP 2022 M2MeT Challenge, IJCAI 2023 ADD challenge. He received the IBM faculty award in 2016, the ISCA Computer Speech and Language 5-years best journal paper award in 2018 and the youth achievement award of outstanding scientific research achievements of Chinese higher education in 2020. He is a senior member of IEEE.