IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 22, 2021, accepted August 20, 2021, date of publication August 30, 2021, date of current version September 3, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3108838*

# A Novel Multi-Feature Joint Learning Ensemble Framework for Multi-Label Facial Expression Recognition

**WANZHAO LI[1,2], MINGYUAN LUO[3], PENG ZHANG[4], AND WEI HUANG[1,2]**

[1]Department of Computer Science, School of Information Engineering, Nanchang University, Nanchang 330031, China
[2]China Mobile-NCU AI&IOT Jointed Laboratory, Informatization Office, Nanchang University, Nanchang 330031, China
[3]Laboratory of Medical UltraSound Image Computing (MUSIC), School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China
[4]School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Wei Huang (n060101@e.ntu.edu.sg)

This work was supported in part by the National Natural Science Foundation of China under Grant 61862043 and Grant 61971352, in part by the Natural Science Foundation of Jiangxi Province under Grant 20204BCJ22011, and in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JM6015.

**ABSTRACT** The facial expression is one of the most common ways to reflect human emotions. And understand different classes of facial expressions is an important method in analyzing human perceived and affective states. In the past few decades, facial expression analysis (FEA) has been extensively studied. It illustrates few of the facial expressions are exactly individual of the predefined affective states but are blends of several basic expressions. Some researchers have realized that facial expression recognition can be treated as a multi-label task, but they are still troubled by the inaccurate recognition of multi-label expressions. To overcome this challenge, a novel multi-feature joint learning ensemble framework, called MF-JLE framework, is proposed. The proposed framework combines global features with several different local key features to consider the multiple labels of expressions embodied in many facial action units. The ensemble learning is introduced into the framework, combines the global module and the local module on the loss, and carries out the joint iterative optimization. The ensemble of the whole framework improves the accuracy of multi-label recognition of different modules as weak classifiers. In addition, the traditional multi-classifier cross-entropy loss has been replaced by the binary cross-entropy loss for a better ensemble. The proposed framework is evaluated on the real-world affective faces (RAF-ML) dataset. The experimental results show that the proposed model is better than other methods in both quantitative and qualitative aspects, whether compared with traditional shallow learning methods or recent deep learning methods.

**INDEX TERMS** Multi-label, facial expression recognition, ResNet-18, deep learning.

## I. INTRODUCTION

In recent years, people have become increasingly interested in improving all aspects of human-computer interaction. Facial expressions, as an indispensable way of human communication, can convey abundant information about human emotions. Facial expressions are most commonly used in daily communication between people of each other. For instance, the smile to show greeting, the frown corresponding confuse, and open their mouths when surprised. The fact which we comprehend emotions and how to react to other people's expressions abundantly enriches the interaction. Researchers attempt to analyze facial expressions, which try to comprehend and classify these emotions.

However, most previous research on recognizing emotion through facial expressions was originated by the 1970s [2]. Researchers attempt to depict each facial picture with one of the predefined affective labels on account of extensive studies, such as six basic expressions namely happiness, sadness, anger, surprise, disgust, and fear [4]. Because of the evolution of theory, most of the previous studies on the discrete categorical emotion analysis have regarded the problem of emotion recognition as a binary classification problem, which let researchers turned to classify facial expressions into one of those six categories or seven categories (include neutral). The research [5] was described almost single-label

The associate editor coordinating the review of this manuscript and approving it for publication was Varuna De Silva.

classification studies, and introduced the performance in these studies [6].

Nevertheless, when analyzing the natural interaction of human beings, we cannot hope for everyone to perform pure emotional substance. Therefore, researchers realized the humans' facial expressions are not always pure individuals, they will mix different emotions of humans, and there are also many kinds of research focus on these relationships in recent years [7]–[10]. These studies indicate that the expression of people often combinations, blends, or compounds of different basic emotions. For instance, someone will perform both disgust and fear concurrently when he suffers some depressing accidents. Therefore, facial expression analysis is not a simple single-label classification problem, we have to treat this problem with a more practical method.

There are many researchers provide their approaches to solve this problem [11]–[14]. Some researchers address this complex emotion problem via facial action coding system (FACS) analysis which became the standard scheme for facial expression research. Unfortunately, FACS coding requires professionally trained coders to annotate and is extremely time-consuming especially in unconstrained real-world conditions. There are also very few existing datasets that are available for corresponding the combinations of action unit and specific emotion categories, such as CK+ [29] and GEMEP-FERA(FERA) [50]. In contrast, other researchers focus on multi-label expression recognition, which is a more practical approach. This method can intuitively recognize blended expressions and perform the result of recognition via multiple emotion labels. But these related studies were applied on few small-scale lab-controlled databases, which may not have enough generalization capability. Then, to address this problem, some new databases were proposed such as FER+ [15] and RAF-ML [1].

With the development of deep learning in recent years, it has been the state-of-art method in many classified tasks. There are also have some methods of deep learning that have been applied in the area of multi-label facial expressions. For instance, VGG [15] was applied in FER+, which confirms the distribution information of label is useful in facial expression recognition but lacked accuracy. Deep bi-manifold CNN (DBM-CNN) [1] comprehensive utilizes crowd-sourced label information and feature compactness in the low-dimensional manifolds, which was applied by bi-manifold loss. But, it is not an efficient method, which is complex and ignores the relation of each label.

In general, facial expression analysis cannot attribute as an individual issue. Therefore, multi-label facial expression recognition will be an effective direction to handle the issue of facial expression analysis. And many researchers have been explored this area with many methods [34], [36], [38], [40]. With the developed of deep learning, it has been achieved huge success in many areas include single label facial expression recognition. But, there is barely systematic research in the area of multi-label facial expression recognition by deep learning method until the appearance of DBM-CNN [1] which also has many limitations. So, it is a novel and valuable direction which improves the method of deep learning in the area of multi-label facial expression recognition.

To address above problems, a novel multi-feature joint learning ensemble (MF-JLE) framework is proposed to effectively address the difficulty of multi-label facial expression recognition. The main contribution is three-fold. Firstly, the issue of multi-label expression recognition has been handled by using global features and local key features. Then ensemble learning is used to combine different modules to improve the overall recognition performance. Secondly, the traditional multi-classifier cross-entropy loss is replaced by binary cross-entropy loss in multi-label facial expression recognition. Finally, multi-feature joint learning ensemble (MF-JLE) framework achieves better performance and outperforms some state-of-the-art multi-label facial expression recognition methods in most criteria by RAF-ML dataset [1] and JAFFE dataset [31]. Consequently, the study in this article display that multi-label facial expression recognition by deep learning method also have many space to improve. The MF-JLE which be proposed in this study also points out some directions to enhance the deep learning model in the area of multi-label facial expression recognition, such as the loss function and ensemble learning.

The following sections are organized as follows. Related works are discussed in Section II. In Section III, the proposed framework will be described in detail. Then, the particular progression of experiments has been illustrated in Section IV. Finally, the conclusions have been summarized in Section V.

## II. RELATED WORKS

In this section, several works related to multi-label facial expression recognition methods and network structures have been briefly introduced, which is significant preliminary knowledge that can help to understand the technical details presented in the proposed work.

### A. EXPRESSION RECOGNITION AND MULTI-LABEL FACIAL EXPRESSION

Since the 20th century, recognizing facial expressions has been an active research topic especially in recent decades. Many previous works focus on the hand-crafted feature, such as Gabor wavelets [17], local binary patterns on three orthogonal planes (LBP-TOP) [18], pyramid histogram of oriented gradients (PHOG) [19] and local quantized patterns (LPQ) [20]. Then, with deep learning has been widely applied in image classification tasks, the deep learning methods (e.g., DCNN) were also transplanted to facial expression recognition as a method and achieved a huge breakthrough.

A subjective emotion is often non-exclusive, many psychological studies and cognitive sciences support the theory that the ability of a face often combines more than one emotional component at a given moment, which can be observed even in still facial pictures. Silvan [25] investigated how to combine various emotions, and gave an example to illustrate that in certain parenting styles, children may experience a mixed

emotional state of fear and shame. In [26], the self-report research shows that people usually experience a variety of emotional blends. If subjects are required to make a fancy that they are set in a fearful environment, then they are likely to have fear, plus surprise or distress. Experiments conducted by Nummenmaa [27] testified that it is possible for people who will express pleasure, surprise, hate, fear, and sorrow, and pairwise combinations of these. The still image is also proved to be useful for this particular objective. Izard [9] also enumerated a lot of normal patterns or combinations of effects. For instance, anxiety can be defined as an emotion that combines a mixture of sadness, fear, and anger. Individual changes in the patterns of basic emotions can produce different kinds of anxiety. In recently, Donahue *et al.* [28] proposed a composite facial expression consisting of a combination of two basic emotion categories and identified 15 composite expressions that are consistently produced across cultures. There is much research in other relative areas that also testify to this theory. In [21], it discovered that more than one emotion will be evoked by the music at the same time, and compared four multi-label classification methods to solve the problem. In [22], the authors applied the annotation of emotion mixtures to speech recognition, and there is plenty of work that follows similar techniques in speech emotion recognition [23], [24].

All of these discussions indicate that the single individual emotions can be blended or synthesized to form new emotions and these prototype emotions should prove useful in delineating the precise blended composition of mental representations.

### B. FACIAL EXPRESSION DATASET

There are several widely used facial expression databases: CK, CK+, JAFFE, MMI, Belfast, Multi-PIE, SFEW, AFEW, LFW, FER+, and RAF-ML [29]–[31]. Most existent datasets only provide images attached with one expression category. And only three datasets provide expression data with multilabel information, such as the lab-controlled database JAFFE, FER+, and in-the-wild database RAF-ML [1].

JAFFE dataset includes 213 facial images of ten Japanese female expressers posing the six basic expressions plus neutral expression. Each of the subjects poses three to four examples per expression to make a total of 213 gray-scaled images in the size of 256 $\times$ 256 pixels. The images were captured under controlled environment in terms of pose and illumination, but it is worth mentioning that besides a single label representing the predominant expression of each image, semantic ratings of the expressions are provided as well, which represent the intensity on each emotion. A five-level scale was used for each of the 6 adjectives (5 represents highest emotion intensity, while 1 represents lowest emotion intensity).

FER+ dataset asked crowd taggers to label the image into one of 8 emotion types: neutral, happiness, surprise, sadness, anger, disgust, fear, and contempt. The taggers are required to choose one single emotion for each image and the gold standard method has been adopted to ensure the tagging quality. In a first attempt, tagging was stopped as long as two taggers agreed upon a single emotion, but the obtained quality was unsatisfactory. In the end, asked 10 taggers to label each image, thus obtaining a distribution of emotions for each face image. With 10 annotators for each face image, they generate a probability distribution of emotion capture by the facial expression, which enables them to experiment with multiple schemes during training.

Real-world affective faces (RAF-ML) dataset is a multi-label facial expression dataset with around 5K great-diverse facial images downloaded from the Internet with blended emotions and variability in subjects' identity, head poses, lighting conditions and occlusions. During annotation, 315 well-trained annotators are employed to ensure each image can be annotated enough independent times. And images with multi-peak label distribution are selected out to constitute the RAF-ML. RAF-ML provide 4908 real-world images with blended emotions, 6-dimensional expression distribution vector for each image, 5 accurate landmark locations and 37 automatic landmark locations, and baseline classifier outputs for multi-label emotion recognition.

Although the three datasets which were mentioned upside offered multilabel information, but the database of JAFFE and FER+ have obvious drawbacks. For instance, JAFFE only has 213 facial images, which is not enough especially in the multi-label recognition area. The image in the database of FER+ which just annotated by 10 annotators, so between each label of the image which in the FER+ lacked enough information of distribution. The RAF-ML dataset was established to solve these flaws that were mentioned in the upside. Therefore, the benefit of the proposed framework will be illustrated by the RAF-ML dataset. Then as the comparison, the proposed framework will also be applied by the JAFFE dataset.

### C. MULTI-LABEL LEARNING IN FACIAL EXPRESSION RECOGNITION

Over the past few decades, a great deal of progress has been made in the multi-label classification learning paradigm. Now, researchers can reference these works that have detailed definition, evaluation metrics, and representative multi-label learning algorithms [32], [33]. Nevertheless, few multi-label facial expression models for facial expression analysis have been developed in recent years.

In [34], the change among different expressions is presented as the evolution of the posterior probability of the six basic paths depend on a probabilistic model that can recognize mixture expressions. In [36], a novel implicit multi-emotion video tagging method is proposed, which balances the relationship between multiple facial expressions, and the relationship between expressions and emotions.

In [38], multi-label group Lasso regularized maximum margin classifier (GLMM) and group Lasso regularized regression (GLR) algorithms are proposed which can model FER jointly with multiple outputs. In [40], an emotion

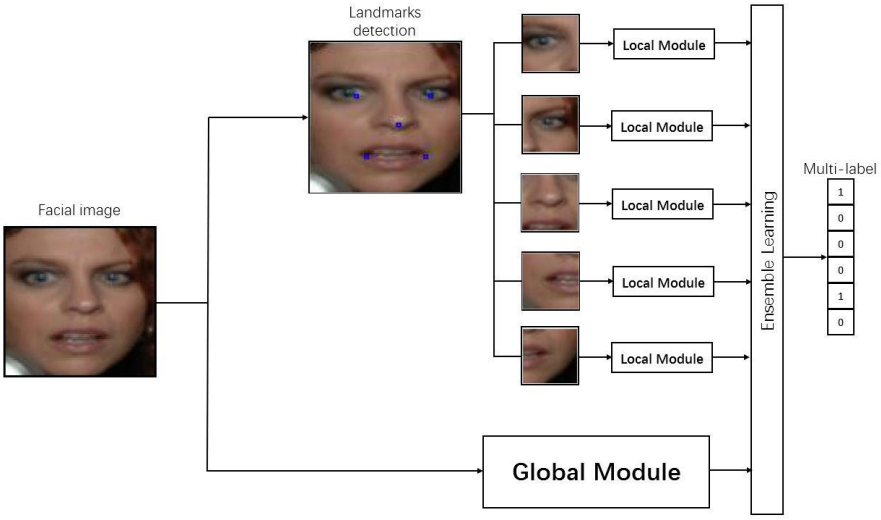



**FIGURE 1.** The pipeline of the multi-feature joint learning ensemble (MF-JLE) framework.

distribution learning (EDL) algorithm has been created that learns the definite description degrees of all six basic emotions and maps the given expression image to the emotion distributions. In [41], an additive weighted function regression from a statistical standpoint, logistic boosting regression (LogitBoost), is used to create two label distribution learning (LDL) algorithms named LDLogitBoost and AOSO-LDLogitBoost, which can cause better performances on expression recognition.

Different from these methods that were conducted on small-scaled laboratory-controlled facial expression datasets and are shallow-learned. VGG [15] was applied in FER+, which confirms that distribution information of label is useful in facial expression recognition, but it is not an efficient method in the area of multi-label facial recognition. Li [1] propose a new deep manifold feature learning based framework, deep bi-manifold CNN (DBM-CNN), which simultaneously and efficiently considers crowd-sourced label information and feature compactness in the low-dimensional manifolds by adding a new loss layer, bi-manifold loss. Jointly trained with the cross-entropy loss which forces images with different labels to stay apart, the bi-manifold loss drives the locally neighboring faces sharing the similar intensity distribution to become coherent, and thus the discriminative power of the deeply learned features can be highly enhanced. Nevertheless, it also has quite a few drawbacks. For example, it uses cross-entropy as its loss function which completely ignores the relation of each label. The structure of the model is complex and the network is too deep, which cannot make the loss convergence swiftly.

## III. METHOD

### A. MULTI-FEATURE JOINT LEARNING ENSEMBLE FRAMEWORK

A novel multi-feature joint learning ensemble (MF-JLE) framework is proposed to overcome the issue that a single facial image contains multiple expression labels. The schematic illustration of the MF-JLE framework as shown in Fig. 1. As can be seen that the framework contains a global feature learning module and several local feature learning modules. The original facial image is fed to the global feature learning module to learn global features. At the same time, since the fusion expression is usually reflected in the tiny details of the face (action unit), multiple local images are extracted according to the given mark points in the image, and each local image is input into different local feature learning modules, in order to learn the local features at different mark points. In addition, the commonly used multi-label cross-entropy loss is replaced by the binary cross-entropy loss, and the loss of different modules is combined for ensemble learning. The proposed framework structure and its training process are described in detail in the following sections.

### B. GLOBAL AND LOCAL MODULE

The facial expression of a human face can be described as a combination of different action units, and different action units have clear and different meanings. Previous research has shown that these action units can be used in any higher-order decision-making process, including the recognition of basic emotions [51]–[53]. Based on this advantage, we introduced action units to multi-label facial expression recognition, and propose a novel and robust ensemble framework combining global and local features for the first time to cope with the complexity of multiple facial expressions. Specifically, parallel models have been applied to analyze images at the global scale and several different local scales, and integrate all models for optimization iteration to handle the subtle features of facial expressions and address the multi-label problem.

Fig. 1 shows the overall framework of the MF-JLE framework. As can be seen that the framework is composed of a global module and multiple local modules, and each module

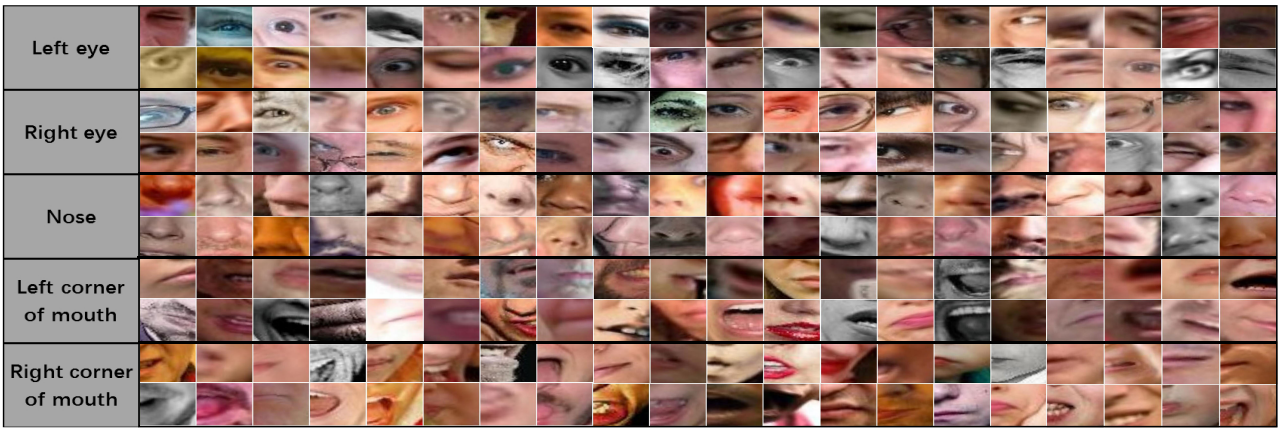



**FIGURE 2.** The samples of five key landmarks.

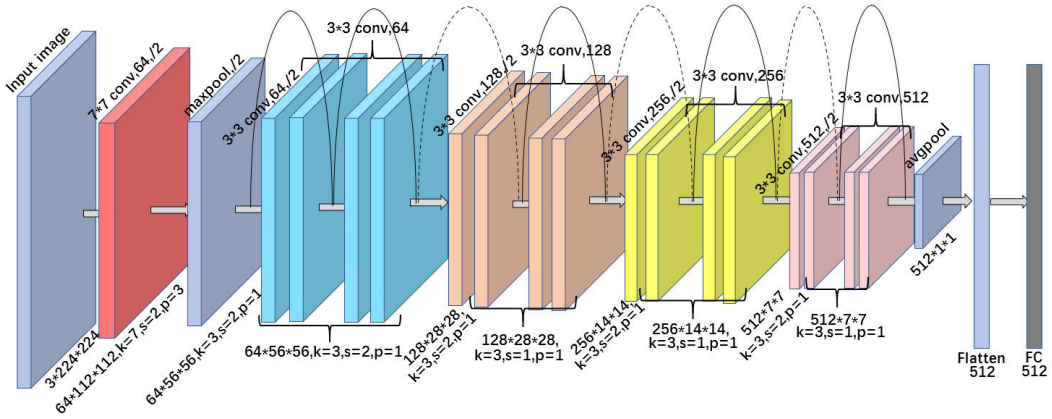



**FIGURE 3.** The structure details of global/local module, every convolution contains ReLU activation function and a layer of batch normalization.

is composed of an 18-layer ResNet [44] which is applied to extracts features. The ResNet-18 contains 17 convolution layers and 1 full connected layer (FC). And it includes 8 basic residual blocks and 4 residual constructions. Every residual construction stack 2 basic residual blocks. The basic residual block contains 2 convolution layers and a skip connection, each convolution layer will be followed by batch normalization layer and ReLU activation function. Between each residual block, the different stride will be applied to fit the channel when the channel is changing. The first layer of ResNet-18 is a convolution layer. After the first layer, a maxpooling layer and several deep residual blocks are applied. Then, there is an average pooling layer after all residual blocks propagate, and only one full connected layer (FC) is applied in the end. The Fig 3 displays these details. Let the input facial image as $x$. Firstly, the MTCNN [49] have been applied to extract the five key landmarks on the input facial image $x$, including left eye, right eye, noes, light mouth corner, right mouth corner, as shown in Fig. 2. Then, five local images as $\{x_i|i = 1, 2, \cdots, 5\}$ are cropped according to the corresponding five key landmarks. After that, the five local images are input into the corresponding local modules to learn the local image features and predict the probability of different expressions. At the same time, the whole image $x$ is input into the global module to learn rough global features on the whole face and predict the probability that the face belongs to different expressions. Finally, inspired by ensemble learning, we combine the prediction probabilities of all modules for different expressions to get the final predicted multi-label expressions. The main idea behind the introduction of ensemble learning into the framework is that the multi-label of expression requires the framework to pay more attention to the global and local subtle differences which namely action units. The framework can distinguish the subtle differences between different expressions, only when the action units are fully considered.

$$L = L_{global} + \sum_{i} \lambda_i L_{local,i} \tag{1}$$

Among them, $\lambda_i$ represents the weights of $i$-th local modules. Multi-label of expressions are merged into a single probability vector through softmax to replace the traditional multi-classifier method based on a single probability.

On this basis, the global module and all local modules are combined for optimization. Specifically, let the multi-expression label corresponding to image $x$ be $y = [y_1, y_2, \cdots, y_N]$, where $y_i$ is a binary expression label and $N$ (we set 5) represent the total number of expression categories. When image $x$ is represent to the $j$-th expression category, the corresponding $y_j$ equal to 1 and 0 otherwise. Therefore, for any module, the loss function uses binary cross-entropy loss instead of the traditional multi-class cross-entropy loss, which can be expressed as Eq. 2.

$$L_{global/local} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log p_i + (1-\log p_i)\log(1-y_i)) \tag{2}$$

where $p_i$ represents the softmax of the output of $i$-th local module. Thus, the framework iteration algorithm is shown in Algorithm 1, which summarizes the optimization process for the framework.

**Algorithm 1** Optimization Algorithm

**Input:** training data $\phi = \{(x, y)_i\}_{i=1}^{n}$, $n$ is the batchsize, learning rate $\mu$, hyper parameter $\lambda$.

**Output:** model $M$ including parameters $\Theta$

**Initialize:** set the number of iteration $t \leftarrow 0$, initialize model parameters $\Theta^0$.

1: **repeat**
2: $t \leftarrow t + 1$;
3: Attract and corp the local feature at $\phi = \{(x, y)_i\}_{i=1}^{n}$;
4: Fed model $M$ with image $x$;
5: Compute the global loss $L_{global}$ and the local loss $L_{local}$ by Eq. 2;
6: Compute the joint loss $L$ by Eq. 1;
7: Calculate the gradient of the joint loss $L$ with respect to the parameters $\Theta^t$;
8: Update the parameters $\Theta^t$ of the model $M$.
9: **until** Model $M$ convergence

### C. OPTIMIZATION WITH ENSEMBLE LEARNING

On the other hand, ensemble learning has been proved to improve the performance of weak classifiers. Its main idea is to combine multiple weak classifiers to vote. In the MF-JLE framework, the final probability of classification is obtained by summing all probability of classification from the global module and local modules according to certain weights, which can be expressed as Eq. 1. Then the predicted label will be obtained by the process that is the activation function of Sigmoid have been applied to the probability of classification, and the probability of classification is the result of the model which has been pre-trained by the MF-JLE framework.

## IV. EXPERIMENTS

In this section, we conduct some experiments on RAF-ML datasets to evaluate the proposed framework. The data preprocessing, indicators, compared methods, and experimental results will be introduced successively.

### A. DATASET AND DATA PRE-PROCESSING

To verify the effectiveness of the proposed ensemble framework, we conducted a series of experiments on the RAF-ML dataset and the JAFFE dataset. The RAF-ML dataset is a multi-label facial expression dataset. It contains 4,908 real-world images of blended emotions annotated by 315 well-trained annotators, with other details described in Section II. The JAFFE dataset also contains multi-label information of facial expression. For the comparability between DBM-CNN [1] and MF-JLE on the JAFFE dataset, we applied the same method to set the multi-label of the JAFFE dataset. We applied a threshold of 3 to obtain the label set of each image according to the five-scale (1-5) intensity principle: relevant emotions whose value is greater than 3 are set as 1 and the irrelevant emotions are set as 0. And the details of the global and local key features are displayed in Fig 4.

During the experiment, all datasets have been divided into three-part, including train set, validate set, and test set, which respectively contained 60%, 20%, and 20% of all images, to ensure that results can be accurately reproduced. In addition, all facial images are cropped to the 224 × 224 size and transformed to RGB images.

### B. TRAINING DETAILS AND EVALUATION CRITERIONS

All input images have been normalized in the same way, mini-batches of 3-channel RGB images of shape $(3 \times H \times W)$, where $H$ and $W$ were assigned as 224. The images have been loaded in to a range of $[0, 1]$ and then normalized by using mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225]. The learning rate is initially set to $10^{-4}$ and the step adjustment is used to decrease the learning rate by a factor of 0.9 at every 50 epoch, and training will be finished at 218 epochs (approximate 8k iterations). In addition, the Adam has been chosen as optimizer with $\beta$ is assigned as [0.9, 0.999] and mini-batch with 128 samples is applied in the training process.

Performance evaluation of the multi-label learning system is different from that of the classical single-label learning system. In this study, Nine widely used evaluation criteria have been adopted to assess the performance of the different methods, including Hamming loss, one error, coverage, label ranking loss, average precision, micro-/macro- F1, and micro-/macro- AUC.

1) Hamming loss: The hamming loss evaluates the degree of disaccord between the predicted results and the ground truth of the sample. It is formally defined as a score of the wrong labels to the total number of labels. The smaller the value of hamming loss the better performance would be.

$$\text{Hamming loss} = \frac{1}{p}\sum_{i=1}^{p}|h(x_i)\Delta Y_i| \tag{3}$$

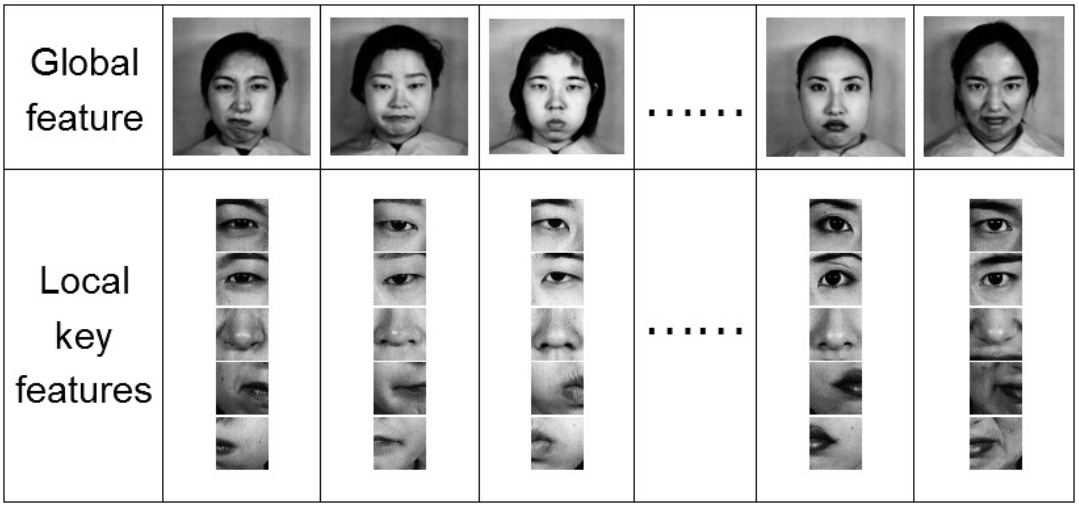



**FIGURE 4.** The global and local key features in JAFFE dataset.

2) One error: The indicator depicts how many times that the top-ranked label is not in the set of correct labels of the example. The smaller the value of one error, the better the performance will be.

$$\text{One error} = \frac{1}{p}\sum_{i=1}^{p}[\text{argmax}_{y\in Y} f(x_i, y)] \notin Y_i \quad (4)$$

3) Coverage: The coverage evaluates how many steps are needed, on average, to move down the ranked label list so as to cover all the relevant labels of the example.

$$\text{coverage} = \frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i}\text{rank}_f(x_i, y) - 1 \quad (5)$$

4) Ranking loss: The ranking loss evaluates the fraction of reversely ordered label pairs, label is ranked higher than a relevant label. The result is perfect when the loss equals 0; the smaller the value, the better the performance would be.

$$\text{Ranking loss} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y||\bar{Y}|}|R| \quad (6)$$

where

$$R = \{(y', y'')|f(x_i, y') \leq f(x_i, y''), (y', y'') \in Y_i \times \bar{Y}\}. \quad (7)$$

5) Average precision: The average precision evaluates the average fraction of relevant labels ranked higher than a particular label $y \in Y_i$.

$$\text{Average}(p) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y|}\sum_{y\in Y_i}\frac{|P|}{\text{rank}_f(x_i, y)} \quad (8)$$

where

$$P = \{y'|\text{rank}_f(x_i, y') \leq \text{rank}_f(x_i, y), y' \in Y_i\} \quad (9)$$

6) Micro-/Macro- AUC: AUC used here generically refers to compute the area under the receiver operating characteristic curve (ROC) from prediction scores. The ROC curve visualizes the trade-off between sensitivity and specificity by plotting both values as a function of a varying classification threshold. And the larger value of AUC is, the better performance of the corresponding classifier is. The Micro-AUC calculates metrics globally by considering each element of the label indicator matrix as a label. The Macro-AUC calculates metrics for each label, and finds their unweighted mean. This does not take label imbalance into account.

7) Micro-/Macro- F1: The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$\begin{cases} \text{Precision} = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \bigcap h(x_i)|}{|h(x_i)|} \\ \text{Recall} = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \bigcap h(x_i)|}{|Y_i|} \\ \text{F1} = \frac{(1+\beta^2)\cdot\text{Precison}\cdot\text{Recall}}{\beta^2\cdot\text{Precison}+\text{Recall}} \end{cases} \quad (10)$$

in which, the most common choice is $\beta = 1$ which leads to the harmonic mean of precision and recall. For micro-averaged method, calculate metrics globally by counting the total true positives, false negatives and false positives. And macro-averaged method calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.

### C. EXPERIMENTAL RESULTS

To investigate the performance of our proposed framework, the MF-JLE framework has been compared with the traditional shallow learning methods LBP, HOG, and the deep

**TABLE 1. Comparisons with different methods on the RAF-ML dataset, the proposed framework get a better performance than others.**

| Comparing models | Evaluation Criterion | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Hamming ↓ | Coverage ↓ | One error ↓ | Ranking ↓ | Average precision ↑ | Micro-$F1$ ↑ | Macro-$F1$ ↑ | Micro-AUC ↑ | Macro-AUC ↑ |
| LBP | 0.251 | 2.487 | 0.229 | 0.218 | 0.785 | 0.623 | 0.542 | 0.812 | 0.809 |
| HOG | 0.230 | 2.383 | 0.211 | 0.201 | 0.804 | 0.653 | 0.586 | 0.826 | 0.815 |
| AlexNet | 0.264 | 2.616 | 0.253 | 0.238 | 0.766 | 0.614 | 0.562 | 0.782 | 0.769 |
| VGG | 0.247 | 2.540 | 0.234 | 0.221 | 0.782 | 0.642 | 0.581 | 0.813 | 0.798 |
| DBM-CNN | 0.167 | 1.954 | **0.119** | 0.116 | **0.879** | 0.769 | 0.745 | 0.903 | 0.894 |
| ResNet-18 | 0.161 | 1.932 | 0.124 | 0.112 | 0.827 | 0.772 | 0.740 | 0.907 | 0.895 |
| MF-JLE | **0.155** | **1.879** | 0.123 | **0.111** | 0.828 | **0.783** | **0.762** | **0.909** | **0.897** |

**TABLE 2. Comparisons with different methods on the JAFFE dataset, the proposed framework get a better performance than others.**

| Comparing models | Evaluation Criterion | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Hamming ↓ | Coverage ↓ | One error ↓ | Ranking ↓ | Average precision ↑ | Micro-$F1$ ↑ | Macro-$F1$ ↑ | Micro-AUC ↑ | Macro-AUC ↑ |
| LBP | 0.225 | 1.749 | 0.250 | 0.149 | 0.809 | 0.754 | 0.687 | 0.854 | 0.869 |
| HOG | 0.183 | 1.681 | 0.234 | 0.142 | 0.832 | 0.776 | 0.711 | 0.887 | 0.875 |
| AlexNet | 0.211 | 1.854 | 0.223 | 0.142 | 0.833 | 0.771 | 0.715 | 0.892 | 0.884 |
| VGG | 0.186 | 1.718 | 0.271 | 0.153 | 0.817 | 0.752 | 0.691 | 0.873 | 0.862 |
| DBM-CNN | 0.142 | 1.247 | **0.109** | 0.062 | 0.936 | - | - | - | - |
| ResNet-18 | 0.061 | 0.875 | 0.153 | 0.025 | 0.939 | 0.888 | 0.861 | 0.962 | 0.961 |
| MF-JLE | **0.032** | **0.804** | 0.146 | **0.023** | **0.959** | **0.895** | **0.873** | **0.979** | **0.963** |

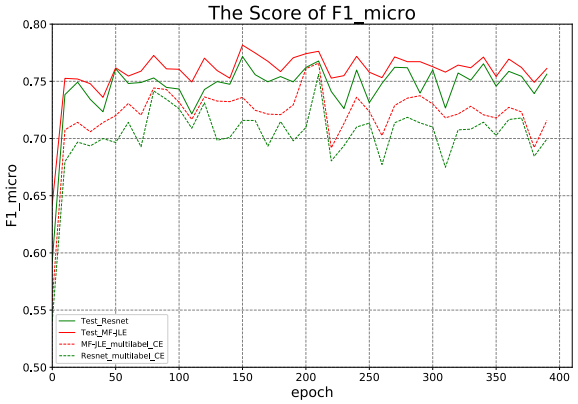



**FIGURE 5. Micro-F1 results of the MF-JLE framework, compared to ablation version (i.e., single ResNet, MF-JLE with multi-label cross-entropy ) during the different epochs of the training process.**

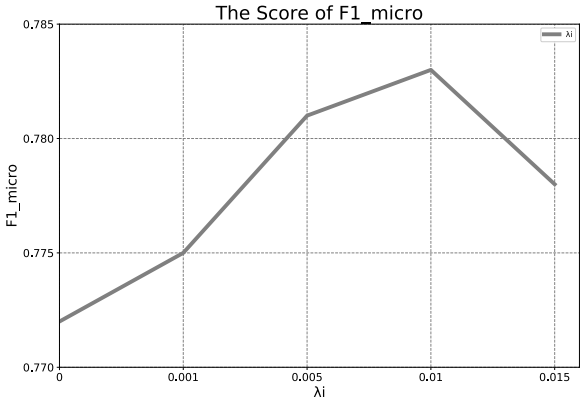



**FIGURE 6. Micro-F1 results of the MF-JLE framework, compared to the different values of $\lambda_i$.**

learning methods AlexNet, VGG, and DBM-CNN. In addition, the proposed MF-JLE framework has been compared with the single ResNet to verify the effectiveness of the model ensemble through an ablation experiment. And another ablation experiment also is applied between multi-label cross-entropy loss and the binary cross-entropy loss to verify the effectiveness of binary cross-entropy loss. The single ResNet has the same structure as each module in the MF-JLE framework and takes the entire facial image as input. All of the models have been trained base on a workstation mainly equipped with an Intel CPU i7-8700K, 16G RAM, an Nvidia GTX 1060 GPU card, and PyTorch 1.7.0 under the Ubuntu OS. In this configuration environment, the running time for each image are 12.13ms, 19.94ms, 10.52ms, and 10.52ms for AlexNet, VGG, ResNet-18, and MF-JLE, respectively.

Table 1 enumerates all performance of the proposed framework and other compared methods on the RAF-ML dataset. Table 2 displays the whole results of the proposed framework and other compared methods on the JAFFE dataset. The ↓ means the smaller the value, the better the performance, and the ↑ means the larger the value, the better the performance. It can be seen that the performance of the early deep learning models AlexNet and VGG is slightly worse than that of the

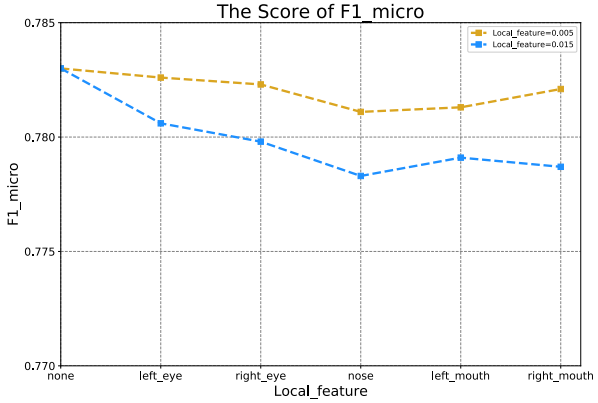


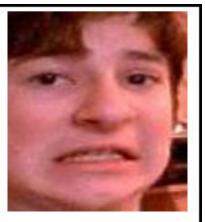

**FIGURE 7.** Micro-F1 results of the MF-JLE framework, compared to the different local feature which set the $\lambda_i$ as 0.005 and 0.015.

traditional shallow learning methods. This shows that for multi-label recognition, traditional learning methods have certain advantages over early deep learning models. However, DBM-CNN, the new deep learning model, has significantly improved its performance compared with AlexNet and VGG. On this basis, the proposed MF-JLE framework achieved the best performance in almost all indicators (except One error and AP, which ranked second). In the RAF-ML dataset, compared with DBM-CNN, the MF-JLE framework is improved by 0.012/0.075/0.005/0.014/0.017/0.006/0.003 on Hamming loss/Coverage/Ranking loss/Micro-F1/Macro-F1/Micro-AUC/Macro-AUC, respectively. In the JAFFE dataset, compared with DBM-CNN, the MF-JLE framework is improved by 0.11/0.443/0.039/0.023 on Hamming loss/Coverage/Ranking loss/Average precision, respectively. The DBM-CNN in [1] is only evaluated by these criteria on the Jaffe dataset and have not released the original code.

In addition, the MF-JLE framework also provides better performance compared to the single ResNet, which fully proves the necessity of ensemble learning. Fig. 5 presents a quantitative evaluation of ablation version (i.e., single

| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 0 | 0 | 0 |
| Fear | 0 | 0 | 0 |
| Disgust | 1 | 1 | 1 |
| Happiness | 0 | 0 | 0 |
| Sadness | 0 | 1 | 1 |
| Anger | 1 | 0 | 0 |

| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 0 | 1 | 1 |
| Fear | 0 | 1 | 1 |
| Disgust | 1 | 1 | 1 |
| Happiness | 1 | 0 | 0 |
| Sadness | 0 | 0 | 0 |
| Anger | 0 | 0 | 0 |

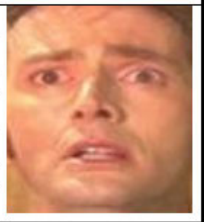


| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 0 | 0 | 0 |
| Fear | 0 | 0 | 0 |
| Disgust | 1 | 1 | 1 |
| Happiness | 0 | 1 | 1 |
| Sadness | 1 | 0 | 0 |
| Anger | 0 | 0 | 0 |

| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 0 | 1 | 1 |
| Fear | 1 | 0 | 0 |
| Disgust | 0 | 0 | 0 |
| Happiness | 0 | 0 | 0 |
| Sadness | 0 | 0 | 0 |
| Anger | 1 | 1 | 1 |

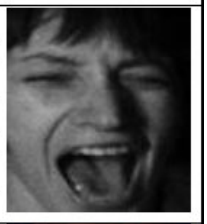
| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 1 | 1 | 1 |
| Fear | 1 | 0 | 0 |
| Disgust | 0 | 0 | 0 |
| Happiness | 0 | 1 | 1 |
| Sadness | 0 | 0 | 0 |
| Anger | 0 | 0 | 0 |

| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 1 | 1 | 1 |
| Fear | 0 | 0 | 0 |
| Disgust | 1 | 0 | 0 |
| Happiness | 0 | 0 | 0 |
| Sadness | 0 | 1 | 1 |
| Anger | 1 | 0 | 0 |



| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 0 | 0 | 0 |
| Fear | 0 | 0 | 0 |
| Disgust | 1 | 0 | 0 |
| Happiness | 1 | 0 | 0 |
| Sadness | 0 | 1 | 1 |
| Anger | 1 | 1 | 1 |

| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 1 | 0 | 0 |
| Fear | 1 | 1 | 1 |
| Disgust | 0 | 0 | 0 |
| Happiness | 0 | 0 | 0 |
| Sadness | 0 | 1 | 1 |
| Anger | 1 | 1 | 1 |



| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 0 | 0 | 0 |
| Fear | 1 | 0 | 0 |
| Disgust | 0 | 1 | 1 |
| Happiness | 1 | 1 | 1 |
| Sadness | 0 | 0 | 0 |
| Anger | 1 | 0 | 0 |

| emotion | Resnet | MF-JLE | True-label |
|---|---|---|---|
| Surprise | 0 | 0 | 0 |
| Fear | 1 | 0 | 0 |
| Disgust | 1 | 1 | 1 |
| Happiness | 0 | 1 | 1 |
| Sadness | 1 | 0 | 0 |
| Anger | 0 | 0 | 0 |

**FIGURE 8.** The examples of multi-label expression recognition.

ResNet) during the training process. It exhibits that ensemble learning improves the result of the single ResNet and gives approximately 0.01 higher Micro-F1. It also points out binary cross-entropy get higher performance than multi-label cross-entropy in the criterion of Micro-F1, which displays the ability of binary cross-entropy superior to multi-label cross-entropy in the area of the multi-label task.

An experiment was also conducted on the multi-label expression recognition task to displays the effects of different values of hyperparameter $\lambda_i$. The $\lambda_i$ was set as the same value in every local feature. The results of Micro-F1 by the MF-JLE model on RAF-ML are shown in Fig. 6. The $\lambda_i$ has been set as 0/0.001/0.005/0.01/0.015, $\lambda_i = 0$ is the case of simply using the global feature. This confirms that the local key features are beneficial to multi-label facial expression recognition. The result shows that the best performance is when the $\lambda_i = 0.01$.

Fig. 7 displays changed results when setting the $\lambda_i$ of a single local feature as 0.005 or 0.015, and the $\lambda_i$ of others stay the same as 0.01. These results reveal the different values of the $\lambda_i$ in different local features will impact the final result. In fact, each local feature doesn't represent a single specific facial expression, it will impact multiple facial expressions which have a similar texture. Therefore, the adjustment of the $\lambda_i$ about the specific local feature should be considering that it will reduce the effect of the weak correlation facial expression in this specific local feature. However, the results of the experiment also prove that it will get a worsening effect if only the $\lambda_i$ of the single local feature has been adjusted. In general, considering the comprehensive relationship between each local feature and facial expression, as well as the constraints of the experiment condition, the value of $\lambda_i$ has been set as 0.01 to each local feature temporarily. The result also shows that set the $\lambda_i$ as 0.01 is the best choice at present.

Fig. 8 shows some examples of multi-label emoticons that the MF-JLE framework can accurately classify. It can be seen that in the difficult examples where the label predicted of other methods are far from the ground truth while the label predicted of the MF-JLE framework is also very close to the ground truth. The quantitative and qualitative results fully prove that the proposed MF-JLE framework with ensemble learning combining global and local modules accurately and robustly addresses the problem of multi-label expression recognition.

## V. DISCUSSIONS AND CONCLUSION

The study in the area of multi-label facial expression recognition by deep learning is still a novel and valuable direction. Although, some studies have been proved that the technology of deep learning is beneficial to multi-label facial expression recognition. But, the MF-JLE which is proposed in this article displays the deep learning technical also having many spaces to be improved. For instance, there are huge differences from the previous studies conducted on the controlled dataset with mixtures emotions, the method in this study, apply a novel multi-feature joint learning ensemble (MF-JLE) framework which effectively addresses the difficulty of multi-label facial expression recognition in the dataset both in the wild and controlled laboratory. And sufficient experimental results confirm the proposed framework can provide more ability to learn discriminative features in a wide range of multi-label facial expression recognition tasks.

Then the details of the training model are elucidated in this article, which indicates some of the directions to improve the deep learning model in multi-label facial expression recognition. Indeed, our proposed MF-JLE framework fully considers the global and local key features and introduces ensemble learning to improve the recognition ability of multi-label expressions. In addition, the proposed framework uses binary cross-entropy loss for multi-label learning.

Finally, it is sincerely hoped this work will make more researchers focus on the area of multi-label facial expression recognition and devote their effort to this area. Then the future work derives from this article will continue, which hopes to make the model more robust and practical in multi-label facial expression recognition.

## REFERENCES

[1] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 884–906, Jun. 2019.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, p. 124, 1971.

[3] P. E. Ekman and W. V. Friesen, "Facial action coding system (FACS)," in *A Human Face*, 2002.

[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)*. London, U.K.: Oxford Univ. Press, 1997.

[5] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[6] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, May 2018.

[7] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues*, vol. 10. Sacramento, CA, USA: Ishk, 2003.

[8] R. R. Hassin, H. Aviezer, and S. Bentin, "Inherently ambiguous: Facial expressions of emotions, in context," *Emotion Rev.*, vol. 5, no. 1, pp. 60–65, Jan. 2013.

[9] C. E. Izard, *Human Emotions*. Springer, 2013.

[10] R. Plutchik, *The Emotions*. Lanham, MD, USA: Univ. Press of America, 1991.

[11] X. Ding, W.-S. Chu, F. De La Torre, J. F. Cohn, and Q. Wang, "Facial action unit event detection by cascade of tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2400–2407.

[12] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Multi-conditional latent variable model for joint facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3792–3800.

[13] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[14] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong, "Confidence preserving machine for facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3622–3630.

[15] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[17] Z. Zhang, "Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 13, no. 6, pp. 893–911, Sep. 1999.
[18] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
[19] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. Int. Conf. Image Signal Process.*, Berlin, Germany: Springer, 2008, pp. 236–243.
[20] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image video Retr. (CIVR)*, 2007, pp. 401–408.
[21] K. Trohidis, *Multi-Label Classification of Music Into Emotions*, vol. 8, Izmir, Turkey: ISMIR, 2008.
[22] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Netw.*, vol. 18, no. 4, pp. 407–422, May 2005.
[23] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–8.
[24] T. Sobol-Shikler and P. Robinson, "Classification of complex information: Inference of co-occurring affective states from their expressions in speech," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1284–1297, Jul. 2010.
[25] S. S. Tomkins, *Affect Imagery Consciousness the Negative Affects*, vol. 2, 1963.
[26] P. Ekman, "Expression and the nature of emotion," *Approaches Emotion*, vol. 3, no. 19, p. 344, 1984.
[27] T. Nummenmaa, "The recognition of pure and blended facial expressions of emotion from still photographs," *Scandin. J. Psychol.*, vol. 29, no. 1, pp. 33–47, Mar. 1988.
[28] J. Donahue, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
[29] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.
[30] A. Dhall, O. V. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 423–426.
[31] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.
[32] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.
[33] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[34] Y. Chang, C. Hu, and M. Turk, "Probabilistic expression analysis on manifolds," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2004, p. 2.
[35] W. Huang, S. Zhang, P. Zhang, Y. Zha, Y. Fang, and Y. Zhang, "Identity-aware facial expression recognition via deep metric learning based on synthesized images," *IEEE Trans. Multimedia*, early access, Jul. 9, 2021, doi: 10.1109/TMM.2021.3096068.
[36] S. Wang, Z. Liu, J. Wang, Z. Wang, Y. Li, X. Chen, and Q. Ji, "Exploiting multi-expression dependences for implicit multi-emotion video tagging," *Image Vis. Comput.*, vol. 32, no. 10, pp. 682–691, Oct. 2014.
[37] W. Huang, M. Luo, X. Liu, P. Zhang, H. Ding, W. Xue, and D. Ni, "Arterial spin labeling images synthesis from sMRI using unbalanced deep discriminant learning," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2338–2351, Oct. 2019.
[38] K. Zhao, H. Zhang, Z. Ma, Y.-Z. Song, and J. Guo, "Multi-label learning with prior knowledge for facial expression analysis," *Neurocomputing*, vol. 157, pp. 280–289, Jun. 2015.
[39] W. Huang, H. Ding, and G. Chen, "A novel deep multi-channel residual networks-based metric learning method for moving human localization in video surveillance," *Signal Process.*, vol. 142, pp. 104–113, Jan. 2018.
[40] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1247–1250.
[41] C. Xing, X. Geng, and H. Xue, "Logistic boosting regression for label distribution learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4489–4497.
[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[46] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.
[47] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[48] Y. Wen, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.
[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[50] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 966–979, Aug. 2012.
[51] W.-S. Chu, F. D. L. Torre, and J. F. Cohn, "Learning facial action units with spatiotemporal cues and multi-label sampling," *Image Vis. Comput.*, vol. 81, pp. 1–14, Jan. 2019.
[52] P. Tarnowski, "Emotion recognition using facial expressions," *Procedia Comput. Sci.*, vol. 108, pp. 1175–1184, Jan. 2017.
[53] Y. Li, B. Wu, Y. Zhao, H. Yao, and Q. Ji, "Handling missing labels and class imbalance challenges simultaneously for facial action unit recognition," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 20309–20332, Jul. 2019.

**WANZHAO LI** received the B.Eng. and M.Eng. degrees from Nanchang University, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree, under the supervision of Prof. W. Huang. His research interests include facial expression recognition, machine learning, computer vision, and pattern recognition.

**MINGYUAN LUO** received the B.Eng. and M.Eng. degrees from Nanchang University, under the supervision of Prof. W. Huang. He is currently pursuing the Ph.D. degree with Shenzhen University. He has published several academic papers in well-known international journals and conference proceedings, including IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE ACCESS, *Multimedia Tools and Applications*, MICCAI, and ACM Multimedia. His research interests include medical image processing, machine learning, computer vision, and pattern recognition.

**PENG ZHANG** received the B.E. degree from Xian Jiaotong University, China, in 2001, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2011. He is currently a Full Professor with the School of Computer Science, Northwestern Polytechnical University, China. He is also the Chief Scientist at Mekitec Oy, Finland. He has published more than 80 research papers, including CVPR, ACM Multimedia, *Neurocomputing*, *Signal Processing*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON MEDICAL IMAGING. He has been acting as the PI in three grants of NSFC. His current research interests include computer vision, pattern recognition, and machine learning.

**WEI HUANG** received the B.Eng. and M.Eng. degrees from Harbin Institute of Technology, China, and the Ph.D. degree from Nanyang Technological University, Singapore. Then, he worked at the University of California San Diego, USA, and the Agency for Science Technology and Research, Singapore, as a Postdoctoral Research Fellow. He is currently a Full Professor with the Department of Computer Science and acts as the Head of the Informatization Office, Nanchang University, China. He has been acting as a Principal Investigator in studies supported by nearly 20 national/provincial grants, including five NSF-China projects and four NSF key projects in Jiangxi, China. He has published more than 100 academic journal/conference papers, including IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, MICCAI, and ACM Multimedia. His main research interests include machine learning, pattern recognition, computer vision, and multimedia. He received Jiangxi Provincial Natural Science Award, the Most Interesting Paper Award of ICME-ASMMC, and the Best Paper Award of MICCAI-MLMI. He was designated as the Academic Leader of Jiangxi, in 2020.