Tsinghua Science and Technology
Doi: 10.26599/TST.2026.90100028

Research Article

Expression Recognition based on Emotion-Wheel-Guided Label Distribution Learning

Jingyang Zhou, Jinyao Liu, and Jing Wang (✉)

Abstract Facial expression recognition (FER) is a critical component in many fields, such as human-computer interaction and affective computing. However, existing FER methods face several key challenges such as label ambiguity, mixed emotional expressions, and class imbalance. To address these issues, we propose a novel framework based on Label Distribution Learning (LDL) that captures the complex and compound nature of real-world emotions. Our approach introduces MixFeature, a feature-level augmentation strategy that synthesizes new samples with label distribution by mixing single-label ones. Such process is guided by the Robert Plutchik's Emotion Wheel to ensure semantic consistency in the generated label distribution. By modeling emotions as label distribution, our method provides a more nuanced representation of blended emotional states. Extensive experiments on widely used datasets demonstrate that our framework significantly outperforms existing single-label and LDL methods in recognition accuracy and robustness, particularly in handling ambiguous and mixed emotions and addressing class imbalance.

Keywords facial expression recognition, label distribution learning, feature mixing, data augmentation, affective computing

1 Introduction

With the advancement of natural language processing (NLP) technologies and the emergence of large language models, human interaction with intelligent systems has become increasingly prevalent [1]. These innovations have enabled machines to perform tasks, such as voice search and automated question answering, significantly improving the naturalness and efficiency of human-computer interaction (HCI) [2]. Meanwhile, as users demand more intuitive and emotionally aware interfaces, emotion recognition has emerged as a necessary research direction across multiple domains, such as HCI and affective computing [3, 4].

In particular, facial expression plays a pivotal role in conveying human emotions and social intentions. For example, the earlier study by Mehrabian [5] highlighted that facial cues often surpass verbal ones in communicating affective states. Facial expression can reveal not only transient emotions but also subtle psychological shifts [4]. As a result, it helps interpret one's emotions, which is essential for building responsive intelligent systems [6, 7].

Recently, with the advent of deep learning, particularly convolutional neural networks (CNNs), has improved the performance of facial expression recognition (FER) systems with end-to-end learning from facial images [4, 8, 9]. This line of work outperforms traditional approaches that relied on manual feature extraction. However, existing CNN-based FER frameworks roughly treat FER as a single-label learning problem, which fails to model blended emotional states [10, 11]. For example, a facial image with anger emotion generally blends other emotions, such as disgust and sadness [4]. Moreover, conventional data augmentation strategies, such as random flipping and cropping, fail to capture the complexity of real-world emotions, and thereby, suffer from noise [12, 13]. Popular datasets, such as AVEC 2013 [14], AVEC 2014 [15], and AffectNet [16], often suffer from low-resolution images and inaccurate labels, which could damage model generalization [17]. Although recent developments in lightweight architectures and attention mechanisms have improved the performance to some extent, they still struggle to handle blended emotions [4].

To address the above issues, we introduce label distribution learning (LDL) [18] into FER. LDL is a novel learning paradigm that assigns each instance with label distribution.

• Jingyang Zhou and Jing Wang are with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China. E-mail: zhoujingyang@seu.edu.cn; wangjing91@seu.edu.cn.
• Jinyao Liu is with the Chien-Shiung Wu College, Southeast University, Nanjing 210096, China. E-mail: 213220701@seu.edu.cn.
Manuscript received: 2025-04-23; revised: 2025-08-28; accepted: 2026-02-06

In particular, label distribution is a multi-variate vector, each element of which describes the relevance of the corresponding label [18, 19]. We model blended emotions as label distribution over all emotions rather than a single deterministic one. That enables a more nuanced representation of emotion states, capturing subtle and blended expressions, which could address the ambiguity in compound emotions.

In this paper, we propose a novel feature-level mixing strategy, called **MixFeature**, which linearly combines single-label samples to synthesize new samples with (emotion) label distribution. **MixFeature** enhances the diversity of training data and improves robustness to class imbalance and annotation uncertainty. Moreover, we incorporate the Plutchik Emotion Wheel [20] to guide the synthetic process, which helps ensure the semantic consistency of the synthesized samples with label distribution. Then, we train a deep network to learn the original single-label samples, as well as the synthesized ones with label distribution. We conduct experiments on several widely used FER datasets, and the experimental results demonstrates the superior performance of our method across both constrained and real-world datasets, particularly in recognizing compound emotions.

Our contributions can be summarized as follows:

(1) We propose a novel feature-level augmentation method, called **MixFeature**, which can synthesize new samples with label distribution by mixing single-label samples.

(2) We introduce the Robert Plutchik's Emotion Wheel to guide the semantic consistency of the generated label distribution.

(3) We conduct extensive experiments on widely used datasets and justify the advantages of **MixFeature** for recognizing compound emotions.

We organize the rest of the paper as follows. First, Section 2 reviews the related works. Second, Section 3 presents the method. Third, Section 4 reports the experiments. Finally, Section 5 concludes this paper.

2 Related work

This paper is related to two lines of works, i.e., facial expression recognition and label distribution, as discussed below.

2.1 Facial expression recognition

Emotion recognition, particularly FER, is a pivotal topic in the field of affective computing [4]. Early approaches primarily relied on hand-crafted features such as edge detectors and texture descriptors, combined with machine learning algorithms, such as linear discriminant analysis [21], and Support Vector Machines (SVM) [22]. These works achieved reasonable performance under controlled conditions but demonstrated limited robustness when applied to real-world scenarios [23].

The advent of deep learning has significantly improved FER. Convolutional neural networks (CNN) enable end-to-end learning by automatically extracting hierarchical feature representations from large-scale datasets, thus eliminating the need for hand-crafted features [24]. Advanced deep learning methods further considered background [25], illumination [26], and head pose [27] for FER, which could improve the model robustness in real-world scenarios [28]. Moreover, given the scarcity of labeled data, researchers have adopted data augmentation for FER [29], such as Mixup [30], CutMix [31], which could increase sample diversity and improve the model performance.

However, the above works roughly treated FER as a single-label learning problem, that is, each instance is associated with one emotion, which fail to model complex and blended emotions. In contrast, our work introduces LDL to solve this issue. Moreover, although existing data augmentation methods, such as Mixup and CutMix, can increase the diversity of training data, they can not solve the ambiguity in FER, for example, compound emotions. In comparison with these works, **MixFeature** is novel augmentation method to synthesize label-distribution samples, which not only enhance the diversity of training data but also address the ambiguity in FER.

2.2 Label distribution learning for FER

Recently, LDL has emerged as an effective learning paradigm to address label ambiguity [32]. Unlike conventional single-label or multi-label learning, LDL assigns each instance with label distribution over all possible labels categories [18]. Each element of label distribution reflects the degree to which the corresponding label describes one instance [18, 19]. Formally, the label distribution for an instance x is expressed as a vector:

$$\boldsymbol{d} = [d_x^{c_1}, d_x^{c_2}, \ldots, d_x^{c_m}],$$

where $d_x^{c_j}$ denotes the relevance of c_i to x and satisfies $d_x^{c_j} \in [0, 1]$ and $\sum_j d_x^{c_j} = 1$.

LDL captures the nuanced representation of labeling information. For FER, label distribution could represent blended emotions by explicitly defining the description degrees of all emotions [33]. As a result, researchers have adopted LDL to solve compound FER problem [34]. For example, Zhou et al. [33] adopted label distribution to model the compound emotions and proposed emotion distribution learning. Similarly, Le et al. [35] used neighborhood information to adaptively construct emotion distribution, which solves the

ambiguity in FER. Li and Deng [36] applied crownsourcing to annotate each facial image with an emotion distribution. Then, they regarded FER as an LDL problem. Compared with single-label methods, LDL could significantly improve FER performance [34].

However, these works assumed that facial images are annotated with label distribution [33, 35, 37] or applied costly crowdsourcing to obtain label distribution. In comparison with them, our proposal applies emotion wheel to guide generating label distribution, which is more effective.

3 The MixFeature method

As illustrated in Fig. 1, the proposed **MixFeature** method introduces two modules implemented through a dual-path architecture. The first module generates hybrid emotion distributions guided by emotion wheel. The second module employs adaptive LDL to jointly optimize mixed feature embedding and the corresponding label representations, enabling dynamic preservation of compound emotional patterns during training. By integrating feature learning and emotion wheel-guided label distribution generation, the framework could effectively capture nuanced emotional states in real-world scenarios, where multiple emotion cues coexist. This synergistic approach addresses the limitations of conventional single-label emotion recognition through a unified optimization of complementary feature and label distribution learning.

3.1 Label distribution samples synthesis

MixFeature can synthesize realistic label distributions from single-label facial expression datasets. The objective is to leverage label distribution to enhance the generalization of the model for recognizing complex and blended emotions in real-world scenarios.

Let $S = \{(x_1, \boldsymbol{y}_1), (x_2, \boldsymbol{y}_2), \cdots, (x_n, \boldsymbol{y}_n)\}$ stand for a training set, where x_i denotes the ith facial image and $\boldsymbol{y}_i$ its one-hot label vector. We first extract the semantic feature vector $\boldsymbol{v}_i$ for x_i using a CNN. The extracted semantic vectors represent the high-level emotional features embedded in facial expressions.

Conventional approaches generally treat emotion recognition as a single-label classification problem. These methods directly map features to discrete labels via a fully connected layer followed by a softmax function. Instead, the proposed **MixFeature** approach aims to synthesize new feature-distribution pairs that attain (emotion) label distributions. Specifically, we first randomly selected K samples and then linearly combine their corresponding feature vectors $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K)$ using weights $(a_1, a_2, \ldots, a_K)$ that satisfy $\sum a_i = 1$. The synthesized vector is as

$$\boldsymbol{v} = \sum_{i=1}^{K} a_i \boldsymbol{v}_i. \tag{1}$$

Likewise, the corresponding label distribution $\boldsymbol{d}$ is generated by applying the same weights to the original one-hot labels, which is denoted by

$$\boldsymbol{d} = \sum_{i=1}^{K} a_i \boldsymbol{y}_i. \tag{2}$$

For example, we could use uniform weights, i.e., $\alpha_i = 1/K$. However, uniform weights failed to consider the semantic of expressions. In the next subsection, we will apply emotion wheel to guide such process that considers the semantic of expressions.

MixFeature enables the transformation of single-label samples into new ones with label distributions. Thus, it could provide richer information for learning emotion representations.

3.2 Emotion-wheel guided synthesis

To ensure the semantic consistency of the synthesized label distributions, **MixFeature** incorporates the emotion distance metrics as defined in the Plutchik's emotion wheel model [38]. This model defines eight primary emotions, including joy, anger, surprise, fear, disgust, sadness, neutral, and blended states. It arranges these emotions in a circular structure that reflects their emotional proximity.

Fig. 2 visualizes the emotion wheel and the relative positions of eight basic emotions. To better align with existing facial expression datasets, this paper has added the neutral emotion on the basis of the fundamental emotions in the wheel. The distance between adjacent expressions equals one unit, while the distance between the neutral emotion and each of the basic emotions is three units. The closer two emotions are in distance, the more likely they are to blended. For instance, the likelihood of mixing anger and sadness is much greater than that of anger and happiness, as these adjacent emotions are spatially closer in the wheel.

As in Eq. (2), we linearly synthesize K single-label samples into label-distribution samples. First, we randomly select one emotion from that of these K single-label samples. We treat this emotion as the primary emotion and employ the emotion wheel to guide the synthesis process. In details, for the ith sample (x_i, y_i), it contributes to the primary emotion according to the its emotional distance from the primary one in the wheel. The closer the distance, the greater the weight

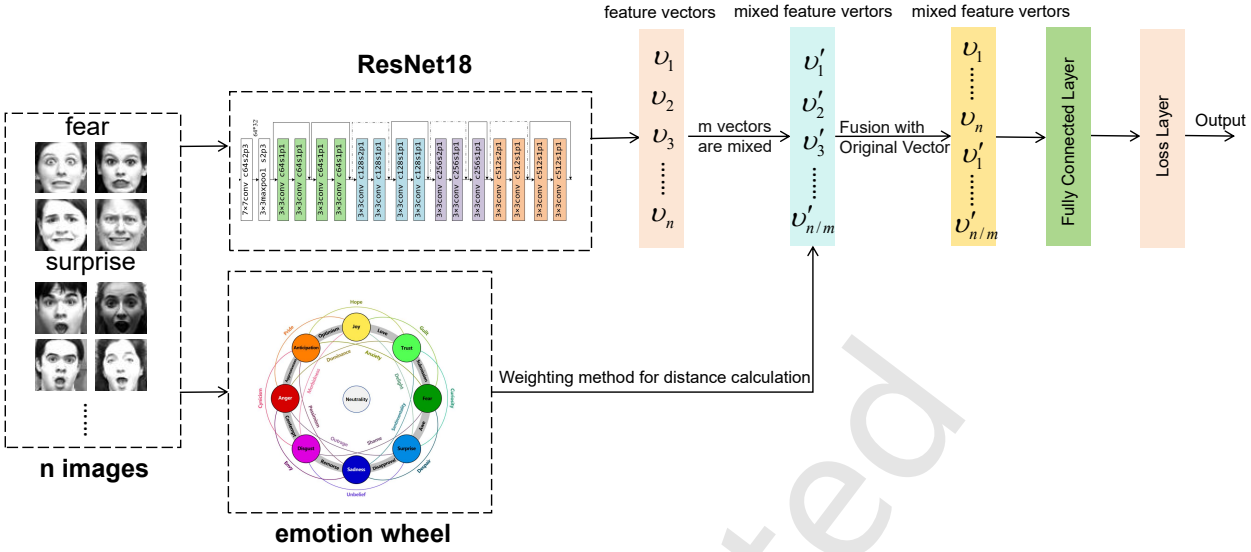

Fig. 1 The framework of the proposed method. First, we generate feature vectors from facial expression images through ResNet-18. Then, we synthesize blended samples (i.e., a feature vector and its corresponding label distribution) guided by the emoton wheel. Finally, we train a hybrid neural network with cross-entropy loss and Kullback-Leibler divergence to learn the original samples and synthesized ones, respectively.

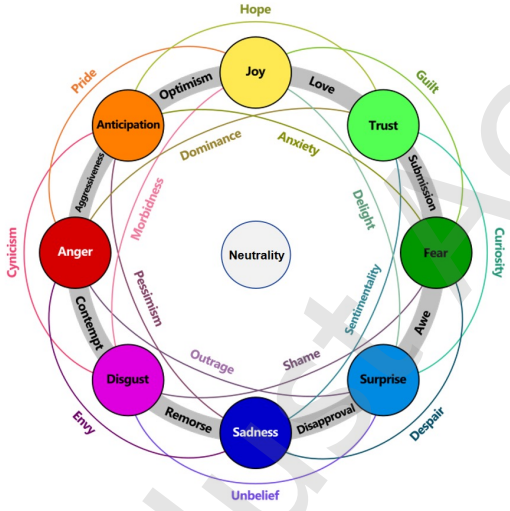

Fig. 2 An illustration of the Plutchik emotion wheel that defines the relative distances between primary emotions [20]. We have added a new emotion, ie., neutrality, to the emotion wheel. The distance between two adjacent expressions equals one unit, while the distance between the neutral emotion and each basic emotions is three units.

is. We define the emotional weight for the i-th sample by:

$$a_i = \frac{\exp(-d_{i,\text{primary}})}{\sum_{j=1}^{K} \exp(-d_{j,\text{primary}})}, \tag{3}$$

where $d_{i,\text{primary}}$ denotes the distance between the emotion label of x_i and the primary emotion on the wheel. That is, the samples having adjacent emotions with the primary one are assigned with larger weights. That favors emotionally similar samples in the mixing process, ensuring that the synthesized label distributions remain semantic consistence. Algorithm 1 summarizes the emotion-wheel guided synthesis.

Algorithm 1 Emotion-wheel-guided synthesis

Parameter: K

Input: K random samples $\{(\boldsymbol{v}_1, \boldsymbol{y}_1), (\boldsymbol{v}_2, \boldsymbol{y}_2), \cdots, (\boldsymbol{v}_K, \boldsymbol{y}_K)\}$

Output: one synthesized sample $(\boldsymbol{v}, \boldsymbol{d})$

1: Randomly select one primary emotion from K samples;
2: Calculate weights $a_1, a_2, \cdots, a_K$ according to Eq. (3)
3: Synthesize $\boldsymbol{v}$ and $\boldsymbol{d}$ by $\boldsymbol{v} = \sum_i a_i \boldsymbol{v}_i$ and $\boldsymbol{d} = \sum_i a_i \boldsymbol{y}_i$
4: **return** $(\boldsymbol{v}, \boldsymbol{d})$

3.3 Training with label distribution

The proposed framework integrates **MixFeature** and an LDL pipeline to train deep models, which is capable of recognizing ambiguous and mixed emotional expressions. The network consists of a backbone for feature extraction and an LDL module to learn both original and synthesized samples.

For a mini-batch $s = \{(x_1, \boldsymbol{y}_1), \ldots, (x_p, \boldsymbol{y}_p)\}$ of size p, the CNN backbone extracts feature vectors $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$. **MixFeature** then generates additional synthesized features and the corresponding label distribution. We randomly select K single-label samples without replacement to synthesize one label-distribution sample. As a result, there are p/K synthesized samples in one mini-batch. The original samples are concatenated with the synthesized ones, that is, both the

original and mixed samples are used for training.The fused samples are as

$$s_{\text{fusion}} = \{(\boldsymbol{v}_1, y_1), \ldots, (\boldsymbol{v}_p, y_p), (\boldsymbol{v}_{p+1}, \boldsymbol{d}_{p+1}), \ldots, (\boldsymbol{v}_{p+m}, \boldsymbol{d}_{p+m})\}, \tag{4}$$

where $\boldsymbol{d}_{p+l}$ is the label distribution for $\boldsymbol{v}_{p+l}$, and $m = {}^{p}/_{K}$ is the number of synthesized samples with label distribution.

Let $C = \{c_1, c_2, \cdots, c_m\}$ denote m emotion labels. We apply softmax function and define the outputs as $\{q_1, q_2, \cdots, q_m\}$. We apply two loss functions:

- **Cross-entropy loss** to learn original single-label samples:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{p}\sum_{i,j}\mathbb{I}(y_{i,j} = c_j)\log q_j^{(i)}, \tag{5}$$

  where $y_{i,j}$ is the jth element of $\boldsymbol{y}_i$, $q_j^{(i)}$ is the jth output for $\boldsymbol{v}_i$.

- **Kullback-Leibler (KL) divergence loss** to learn the synthesized samples with label distribution:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{m}\sum_{i,j} d_{p+i}^{c_j}\log\frac{d_{p+i}^{c_j}}{q_j^{(p+i)}}, \tag{6}$$

  where $d_{p+i}^{c_j}$ is the label description degree of c_j to $\boldsymbol{v}_{p+i}$,

To adaptively weight the contribution of synthesized samples, we introduce a factor β based on the entropy of label distribution:

$$\beta = -\sum_{i,j} d_{p+i}^{c_j}\log d_{p+i}^{c_j} \tag{7}$$

The final loss function is defined as a weighted combination of cross-entropy and KL divergence losses:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{CE}} + (1-\lambda)\beta\mathcal{L}_{\text{KL}} \tag{8}$$

This learning paradigm enables the model to capture subtle and overlapping emotional cues, thus improving the recognition of complex facial expressions in real-world environments.

4 Experiments

In this section, we conduct experiments to validate the performance of **MixFeature**. We will elaborate on the experimental settings, result analysis, ablation studies, and parameter sensitivity evaluations

4.1 Methodology

4.1.1 Data sets

We use three widely recognized facial expression datasets, each of which is characterized by distinct features and annotation protocols:

- **JAFFE Dataset**: The JAFFE dataset consists of 213 grayscale images of facial expressions, collected from 10 Japanese female subjects. Each subject displays seven expressions: six basic emotions (happiness, sadness, surprise, fear, anger, disgust) and a neutral expression. All images are standardized to 256×256 pixels, with intensities rated on a scale of 1 to 5 [39].
- **CK+ Dataset**: CK+ contains 593 video sequences from 123 individuals aged 18–50 [40]. From these sequences, 981 peak expression frames are extracted and annotated with seven emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise.
- **FER+ Dataset**: An enhanced version of FER 2013 [27], comprising 28,709 training, 3,589 validation, and 3,589 test images, all resized to 48×48 pixels and relabeled into eight emotions: happiness, sadness, surprise, fear, anger, disgust, contempt, and neutral [17]. For extended analysis, we further create **FER+_argument**, an augmented variant incorporating rule-based facial action unit annotations generated through OpenFace 2.0 toolkit [41], which provides pseudo-annotations for compound facial movements to facilitate learning of blended emotional expressions.

Fig. 3 Example images from JAFFE, FER+, CK+ datasets.

Figure 3 illustrates three sample images from the three datasets, showcasing differences in resolution, annotation, and expression diversity. Table 1 summarizes the distribution of emotion categories. CK+ and FER+ are dominated by positive emotions (e.g., joy, surprise), whereas negative emotions (e.g., sadness, anger) are underrepresented. Our **MixFeature** method synthesizes additional samples for minority classes and assigns probability-based label distributions, enabling the network to learn balanced representations across all emotions.

4.1.2 Implementation details

All experiments were conducted using the PyTorch deep learning framework, with training performed from scratch. We run all experiments on an NVIDIA RTX 4090 GPU.

A standardized preprocessing pipeline was used across the datasets. JAFFE and CK+ images were resized to 256×256 pixels, while FER+ images retained their native resolution of 48×48 pixels. Image normalization was applied to map pixel intensities to [0, 1]. Data augmentation techniques, including horizontal flipping and random cropping, were applied to the JAFFE and CK+ datasets, while FER+ underwent additional

Table 1 Class distribution of emotions in the JAFFE, CK+, and FER+ datasets. JAFFE does not contain contempt emotion whose distribution is denoted by "-".

Emotion	JAFFE	CK+	FER+
Happiness (HAP)	14.5%	25.1%	29.10%
Surprise (SUR)	14.1%	30.2%	12.54%
Sadness (SAD)	14.5%	10.2%	12.06%
Neutral (NEU)	14.1%	14.9%	34.85%
Fear (FEA)	15.0%	9.1%	2.14%
Disgust (DIS)	13.6%	21.5%	0.46%
Anger (ANG)	14.1%	16.4%	8.38%
Contempt (CON)	-	6.6%	0.48%

transformations, such as random rotations and brightness adjustments.

The network architectures were chosen based on the complexity of the datasets. For JAFFE, a compact three-layer CNN was used to prevent overfitting. Besides, for CK+ and FER+, we employed ResNet-18 [42] as the backbone.

Training was performed using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 5×10^{-4}. Learning rate schedules were customized per dataset:

- JAFFE: we use an initial learning rate of 0.001 for 100 epochs;
- CK+: we start with 0.01, decaying by 0.9 every two epochs after epoch 10, for 30 epochs;
- FER+: we use an initial learning rate of 0.01, decaying by 0.9 every five epochs after epoch 50, for 200 epochs.

4.1.3 Evaluation metrics

To assess the performance of **MixFeature**, several evaluation metrics are employed. We apply classification accuracy to indicate the proportion of correctly classified samples. Since JAFFE and CK+ do not provide predefined test splits, 10-fold cross-validation was adopted for a more robust evaluation. However, classification accuracy may not fully capture model performance, especially for imbalanced datasets. Thus, we also report the confusion matrix with average per-class accuracy, offering a more balanced assessment. Additionally, we also evaluate the Kullback-Leibler (KL) divergence, which is widely used in LDL. For two label distribution $\boldsymbol{p}$ and $\boldsymbol{q}$, the KL divergence equals

$$\mathrm{KL}(\boldsymbol{p}, \boldsymbol{q}) = -\sum_j p_i \log \frac{p_i}{q_i} \tag{9}$$

where p_i and q_i are the ith elements of $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. Lower KL divergence values indicate improved probabilistic modeling of expression variations.

Table 2 The comparison of accuracy (%) between the single-label baseline and MixFeature.

Dataset	Baseline	MixFeature	Acc. Gain
JAFFE	82.22	88.89	+6.63
CK+	90.79	94.06	+3.27
FER+	82.26	82.51	+0.22

4.2 FER accuracy

To validate the effectiveness of **MixFeature** in addressing class imbalance and expression ambiguity, we conducted systematic evaluations across the three datasets.

First, we design a single-label learning baseline model for comparison. For the JAFFE dataset, the baseline network adopts a simple three-layer convolutional neural network, while in the CK+ and FER+ datasets, the baseline network is based on ResNet-18. The baseline simply treats FER as a single-label learning problem and learns the single-label. Table 2 shows the comparison results between **MixFeature** and the baseline. From Table 2, **MixFeature** improves accuracy from 82.22% to 88.89% on **JAFFE** , from 90.79% to 94.06% on **CK+**, and from 82.26% to 82.51% on **FER+**. The consistent improvement across datasets demonstrates the effectiveness of the proposed feature-level augmentation strategy.

Next, we plot the confusion matrix of **MixFeature** on the JAFFE, CK+, FER+, and FER+_argument datasets in Fig. 4. Generally, the networks trained on imbalanced datasets tend to perform better on positive emotions, such as joy and surprise, with much higher accuracy compared to negative emotions like sadness and disgust. However, after incorporating **MixFeature**, the recognition accuracy for negative emotions, such as sadness, anger, and contempt, shows a significant improvement. **MixFeature** not only augments the dataset by mixing images but also effectively enhances the underrepresented negative emotion samples. As a result, the network is trained with a larger and more balanced data, particularly for the negative emotions. Additionally, **MixFeature** generates mixed images with label distributions. Even though the primary emotion of a mixed image might be a positive one, such as joy and happiness, the network also learns from the mixed distribution, which includes a small portion of negative emotions. The network focuses on the label distribution, even when the mixed image predominantly contains positive emotions, which helps improve its ability to recognize negative emotions that are typically underrepresented in the data sets.

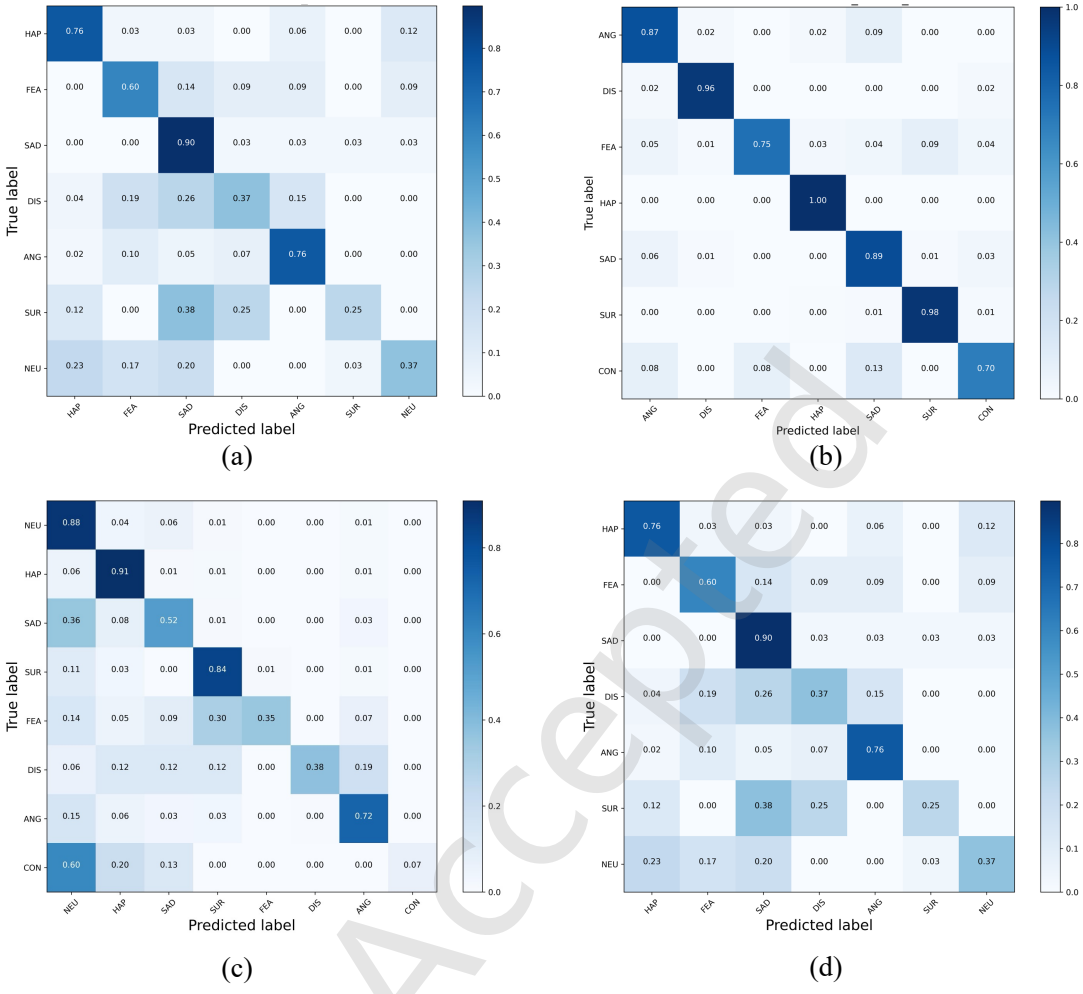

Fig. 4 The confusion matrices of **MixFeature** on (a) JAFFE, (b) CK+, (c) FER+, and (d) FER+_argument.

4.3 Comparison with LDL methods on JAFFE

Notice that we can treat the mean ratings of JAFFE and FER+_argument as label distribution, as conducted by Geng [18]. As a result, we also conduct comparison with LDL methods on JAFFE and FER+_argument. We compare **MixFeature** with four LDL methods, including AA-kNN, SA-BFGS [18], EDL [33], and LDL-LDML [43]. AA-kNN and SA-BFGS are two representative methods that apply the kNN and maximum entropy model to learn label distribution [18]. EDL is the first LDL method for FER that designs a weighted Jeffery's divergence to learn emotion distribution. LDL-LDML is a specially designed LDL method for classification. Moreover, we also compare **MixFeature** with a label enhancement method GLLE which recovers label distribution from logical labels [44]. For GLLE, we first run it on JAFFE and FER+_argument to enhance single-label into label distribution, and then SA-BFGS on the enhanced label distribution. For these five LDL methods, we treat the highest emotion in the predicted label distribution as the prediction.

Table 3 tabulates the comparison results of **MixFeature** against LDL methods. **MixFeature** significantly outperforms AA-kNN, SA-BFGS, EDL, and GLLE that learn the label distribution. In comparison, **MixFeature** jointly learns single-label and (synthesized) label-distribution samples, which bring better FER accuracy. Although LDL-LDML improves the classification accuracy of these four LDL methods, it still has inferior performance compared with **MixFeature**. The reason may lie in that **MixFeature** augments the training samples with better performance.

4.4 An example of compound emotions

Fig. 5 illustrates an example image with compound emotions from the JAFFE dataset. The expression depicted in the image primarily conveys a sense of *surprise*, while also exhibiting a noticeable component of *happiness*, as indicated in the ground-truth label distribution. We also draw the ground-

Table 3 Comparison results against LDL methods on JAFFE. The best results are highlighted in boldface.

Method	Accuracy (%)	
	JAFFE	FER+_argument
AA-kNN	79.32	78.33
SA-BFGS	80.81	79.91
EDL	80.88	81.05
GLLE	79.37	79.25
LDL-LDML	82.31	82.39
MixFeature	**88.89**	**87.75**

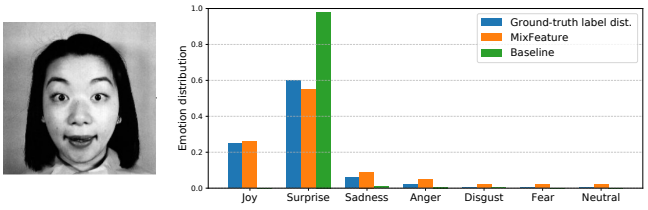

Fig. 5 Example of a compound emotion with ground-truth label distribution and the prediction results.

truth and predicted label distribution by **MixFeature**, as well as the prediction by the baseline.

From Fig. 5, for the baseline network, it classifies this image as "surprise", which is technically correct from the perspective of traditional single-label tasks. However, it is worth noting that the predicted probability for "surprise" from the baseline model reaches an overwhelming 96%, without reflecting the presence of the secondary emotion. In contrast, the network trained with **MixFeature** and LDL demonstrates a more nuanced understanding of the compound expressions. Due to exposure to a diverse range of mixed-emotion samples during training, the model not only identifies the dominant emotion but also captures the presence of secondary emotions, such as happiness in this case. The result exemplifies the core motivation behind LDL. It models and learns the importance degrees of all relevant emotional cues rather than focusing solely on the dominant label. By doing so, the network achieves a more human-like perception of complex and blended emotional expressions.

4.5 Training dynamics

To validate the effectiveness of *MixFeature*, we analyze the convergence patterns and loss dynamics during the training process. Fig. 6 presents the training process of networks on the FER+ dataset, including the cross-entropy loss and KL divergence. Specifically, Fig. 6(a) illustrates the performance of the baseline network, while Fig. 6(b) shows that of the network trained using **MixFeature**.

As shown in Fig. 6, with the increase in training epochs, the cross-entropy loss on the FER+ validation set decreases rapidly. This is because both the baseline network and the network employing **MixFeature** are optimized using cross-entropy loss as the primary objective. Consequently, a rapid decrease in cross-entropy loss results in a substantial reduction in the overall training loss. However, the KL divergence of the baseline network increases sharply as the cross-entropy loss decreases. In contrast, the network trained with MixFeature exhibits only a slight increase in KL divergence, followed by a dynamic stabilization. This can be attributed to the design of the loss functions. The baseline model relies solely on cross-entropy loss, which pushes the predicted probability for the correct class toward one. As a result, the model becomes increasingly confident in a single label, causing a rise in KL divergence when compared to the actual label distribution. On the other hand, the MixFeature-based network incorporates KL divergence into the overall loss function through the use of mixed images and their corresponding label distributions.

Table 4 Ablation and comparison results on different datasets. The best results are highlighted in boldface.

Method	Accuracy (%)			
	JAFFE	CK+	FER+	FER+_argument
MixFeature	**88.89**	**94.06**	**82.51**	**87.75**
Mixup	84.44	70.89	80.60	87.21
CutMix	84.44	88.91	75.09	87.50
AugMix	86.67	90.10	80.07	87.12
GridMask	82.22	86.93	78.00	85.05
SaliencyMix	86.67	68.61	75.84	87.24
StyleMix	84.44	73.56	59.64	87.43
TokenMix	84.44	81.58	59.32	84.96

4.6 Usefulness of MixFeature

To further validate the effectiveness of **MixFeature**, we compare it with eight representative data augmentation techniques, including Mixup [30], CutMix [31], AugMix [45], GridMask [46], SaliencyMix [47], StyleMix [48], and TokenMix [49]. These methods were evaluated on JAFFE, CK+, and FER+ and FER+_argument. The accuracy are summarized in Table 4.

As shown in Table 4, **MixFeature** consistently outperforms all augmentation strategies across all data sets. In particulary, in the JAFFE dataset, **MixFeature** achieves an accuracy of 88.89%, matching the best performance of Mixup and CutMix, demonstrating its effectiveness in small-scale static facial expression datasets. In the more complex CK+ dataset, **MixFeature** achieves the highest accuracy of 94.06%, surpassing Mixup by 23.17% and TokenMix by 12.48%, which

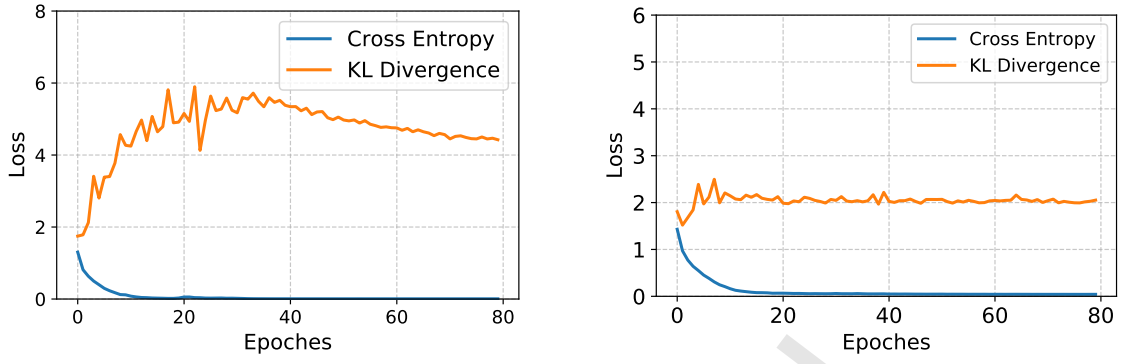

Fig. 6 Training loss of (a) baseline and (b) MixFeature on FER+ dataset.

Table 5 The results of the paired t-test (p-values) between MixFeature and other methods on the JAFFE dataset.

Method	p-value
Mixup	0.0023
CutMix	0.0019
AugMix	0.0031
GridMask	0.0027
SaliencyMix	0.0016
StyleMix	0.0018
TokenMix	0.0015

highlights the superior generalization capability of **MixFeature** in dealing with diverse expression variations. For the FER+ dataset, which involves label distributions with higher ambiguity, **MixFeature** achieves 82. 51% precision, outperforming AugMix (80. 07%) and CutMix (75.09%), and improving over TokenMix by 23.19%. This demonstrates its robustness in capturing label uncertainty by utilizing a feature-level weighted fusion guided by emotion wheel semantics. Finally, on the FER+_argument benchmark, **MixFeature** maintains its advantage, achieving 87.75% accuracy, which is 2.79% higher than TokenMix (84.96%). This results justify its advantages and robustness when applied to more diverse and complex real-world scenarios.

To verify the statistical significance of the improvements, we conduct paired t-tests between MixFeature and the eight comparison methods on the JAFFE dataset. The resulting p-values are reported in Table 5. From Table 5, all p-values are below the 0.05 threshold, demonstrating that the improvements of MixFeature over other methods are statistically significant at the level of 0.05. This confirms the superiority and robustness of **MixFeature**.

In summary, **MixFeature** consistently outperforms traditional augmentation methods and demonstrates strong applicability to complex FER systems.

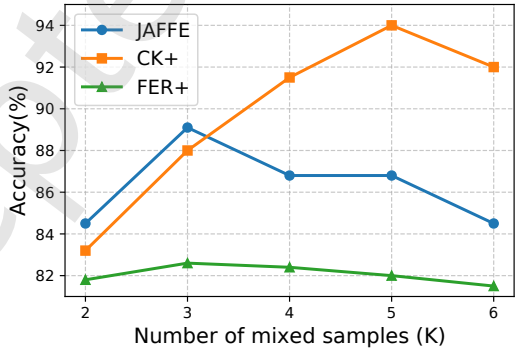

Fig. 7 Sensitivity of MixFeature performance to the number of mixed samples K on the JAFFE, CK+, and FER+ datasets.

4.7 Parameter sensitivity

To further investigate the robustness of **MixFeature**, we conducted a parameter sensitivity analysis with respect to K, i.e., the number of samples mixed for each mini-batch.

In our experiments, we vary K from 2 to 6 and evaluate the corresponding average classification accuracy on the JAFFE, CK+, and FER+ datasets. The results are plotted in Fig. 7, where each curve indicates the accuracy trend on one dataset with varying K. From Figure 7, we observe that:

- **JAFFE**: Accuracy peaks at $K = 3$ (88.89%) and remains above 84% for all values.
- **CK+**: Accuracy increases with K, reaching a maximum of 94.06% at $K = 5$.
- **FER+**: The optimal accuracy occurs at $K = 3$ (82.6%), with a slight decline for larger values.

Overall, the performance of **MixFeature** remains consistently high across a wide range of K values. In particular, accuracy is well-maintained for $3 \leqslant K \leqslant 5$, with the best results observed around $K = 4$. This demonstrates that our method is not overly sensitive to the exact choice of K. Moreover, the stability under varying K confirms the robustness of

MixFeature. In contrast to traditional augmentation methods (e.g., Mixup, CutMix, TokenMix) whose performance often fluctuates with different settings, our method delivers reliable and superior results across all datasets.

To summarize, **MixFeature** demonstrates robustness and reliability in FER. Its stable performance across various datasets and hyperparameter settings highlights its practical advantages.

5 Conclusion

This study introduces a novel framework for facial expression recognition that integrates a feature-level augmentation strategy, called **MixFeature**, with LDL. By synthesizing mixed emotional features and their corresponding label distributions, our approach effectively addresses the inherent ambiguity of facial expressions and mitigates the limitations of single-label annotations. The emotion mixing process is guided by Robert Plutchik's "Wheel of Emotions" model [38], ensuring semantic coherence and avoiding unrealistic emotional combinations. Furthermore, the network is trained through a joint loss function combining weighted cross-entropy and KL divergence, which enhances both classification accuracy and LDL performance. Extensive experiments on widely used benchmarks demonstrate that **MixFeature** consistently outperforms existing data augmentation techniques. The results show significant improvements not only in recognition accuracy but also in the alignment between predicted and ground-truth label distributions, validating its ability to model complex and nuanced emotional expressions.

While the current framework achieves competitive performance, several promising directions deserve further investigation. First, advanced backbone architectures, such as Vision Transformers, could be integrated to enhance the expressive power and scalability of the model. Second, we may apply self-supervised learning techniques to improve feature extraction robustness, particularly for limited data. Third, adaptive emotion representations could be explored, where label distributions are dynamically learned rather than relying on predefined structures.

To better capture human emotions, our future works will investigate multi-modal approaches that combine facial expressions with other modalities [50], such as speech or physiological data. Besides, given the increasing importance of ethical AI, further studies will address potential biases in facial expression datasets and develop fairness-aware mitigation strategies to ensure equitable performance across diverse demographic groups.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 62306073), Natural Science Foundation of Jiangsu Province (Grant No. BK20243012), and China Postdoctoral Science Foundation (Grant No. 2022M720028 and No. 2025T180432).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Reference

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, Language models are few-shot learners, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Virtual Conference, 2020, pp.1877–1901.

[2] P. Hämäläinen, M. Tavast, and A. Kunnari, Evaluating large language models in generating synthetic hci research data: a case study, in *Proc. CHI Conf. Human Factors in Computing Systems*, New York, NY, USA, 2023, pp. 1–19.

[3] M. E. Ayadi, M. S. Kamel, and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recogn.*, vol. 44, no. 3, pp. 572–587, 2011.

[4] S. Li and W. Deng, Deep facial expression recognition: A survey, *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022.

[5] A. Mehrabian, *Silent messages: Implicit communication of emotions and attitudes*, Belmont, CA, USA: Wadsworth, 1971.

[6] P. Faye, S. Sonar, S. Shahare, R. Ambhore, and V. Chafle, Emotion based music recommendation system, *Int. J. Adv. Electr. Comput. Eng.*, vol. 14, pp. 186–191, 2025.

[7] S. Peng, R. Zeng, H. Liu, L. Cao, G. Wang, and J. Xie, Deep broad learning for emotion classification in textual conversations, *Tsinghua Sci. Technol.*, vol. 29, no. 2, pp. 481–491, 2024.

[8] X. Cao, L. Zhai, P. Zhai, F. Li, T. He, and L. He, Deep learning-based depression recognition through facial expression: A systematic review, *Neurocomputing*, vol. 627, p. 129605, 2025.

[9] R. Zeng, H. Liu, S. Peng, L. Cao, A. Yang, C. Zong, and G. Zhou, CNN-based broad learning for cross-domain emotion classification, *Tsinghua Sci. Technol.*, vol. 28, no. 2, pp. 360-369, 2023.

[10] Z. Zhao, Q. Liu, and F. Zhou. Robust lightweight facial expression recognition network with label distribution training, in *Proc. AAAI Conf. Artificial Intelligence*, Virtual Conference, 2021, pp. 3510–3519.

[11] S. Li and W. Deng, Deep facial expression recognition: A survey, *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022.

[12] C. Corneanu, M. Simon, J. Cohn, and S. Guerrero, Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, 2016.

[13] L. Yi and M. Mak, Improving speech emotion recognition with adversarial data augmentation network, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 172–184, 2022.

[14] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, AVEC 2013: the continuous audio/visual emotion and depression recognition challenge, in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Barcelona, Spain, 2013, pp. 3–10.

[15] N. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, AVEC 2014: 3D dimensional affect and depression recognition challenge, in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, New York, NY, USA, 2014, pp. 3—10.

[16] A. Mollahosseini, B. Hasani and M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild, *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2019.

[17] 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 Interaction*, New York, NY, USA, 2016, pp. 279–283.

[18] X. Geng, Label distribution learning, *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, 2016.

[19] J. Wang, Z. Kou, Y. Jia, J. Lv and X. Geng, Label distribution learning by exploiting fuzzy label correlation, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 8979-8990, 2025.

[20] R. Plutchik, A general psychoevolutionary theory of emotion, in *Theories of Emotion*, New York, NY, USA: Academic Press, 1980, pp. 3–33.

[21] M. Siddiqi, R. Ali, A. Khan, Y. Park, and S. Lee, human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields, *IEEE Trans. Image Process*, vol. 24, no. 4, pp. 1386–1398, 2015.

[22] H. Tsai and Y. Chang, Facial expression recognition using a combination of multiple facial features and support vector machine, *Soft Comput.*, vol. 22, pp. 4389–4405, 2018.

[23] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[24] P. Liu, S. Han, Z. Meng, and Y. Tong, Facial expression recognition via a boosted deep belief network, in *Proc. IEEE Conf. Computer Visual Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1805–1812.

[25] F. Zhang, G. Chen, H. Wang, and C. Zhang, CF-DAN: Facial-expression recognition based on cross-fusion dual-attention network, *Comput. Visual Media*, vol. 10, no. 3, pp. 593–608, 2024.

[26] H. Tao and Q. Duan, Hierarchical attention network with progressive feature fusion for facial expression recognition, *Neural Networks*, vol. 170, pp. 337-348, 2024.

[27] M. Valstar, E. Lozano, J. Cohn, L. Jeni, J. Girard, Z. Zhang, L. Yin, and M. Pantic, FERA 2017 - Addressing head pose in the third facial expression recognition and analysis challenge, in *Proc. 12th IEEE Int. Conf. Automatic Face Gesture Recognition*, Washington, DC, USA, 2017, pp. 839–847.

[28] A. Farzaneh and X. Qi, Facial expression recognition in the wild via deep attentive center loss, in *Proc. IEEE Winter Conf. Applications of Computer Vision*, Waikoloa, HI, USA, 2021, pp. 2401–2410.

[29] S. Hashemifar, A. Marefat, J. Joloudari, and H. Hassanpour, Enhancing face recognition with latent space data augmentation and facial posture reconstruction, *Expert Syst. Appl.*, vol. 238, p. 122266, 2024.

[30] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in *Proc. Int. Conf. Learning Representation*, Vancouver, Canada, 2018, pp. 1–13.

[31] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, CutMix: Regularization strategy to train strong classifiers with localizable features, in *Proc. IEEE Int. Conf. Computer Vision*, Seoul, Korea (South), 2019, pp. 6023–6032.

[32] B. Gao, C. Xing, C. Xie, J. Wu, and X. Geng, Deep label distribution learning with label ambiguity, *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, 2017.

[33] Y. Zhou, H. Xue, and X. Geng, Emotion distribution recognition from facial expressions, in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, Australia, 2015, pp. 1247–1250.

[34] A. Khelifa, H. Ghazouani, and W. Barhoumi, Label distribution learning for compound facial expression recognition in-the-wild: A comparative study, *Expert Syst.*, vol. 42, no. 2, pp. 1–27, 2025.

[35] N. Le, K. Nguyen, Q. Tran, E. Tjiputra, B. Le and A. Nguyen, Uncertainty-aware label distribution learning for facial expression recognition, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, Waikoloa, HI, USA, 2023, pp. 6077–6086.

[36] 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, pp. 884–906, 2019.

[37] X. Jia, X. Zheng, W. Li, C. Zhang and Z. Li, Facial emotion distribution learning by exploiting low-rank label correlations locally, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 9833-9842.

[38] R. Plutchik, The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, *Am. Sci.*, vol. 89, no. 4, pp. 344–350, 2001.

[39] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, Coding facial expressions with Gabor wavelets, in *Proc. 3rd IEEE Int. Conf. Face Gesture Recognition*, Nara, Japan, 1998, pp. 200–205.

[40] 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 Computer Society Conf. Computer Vision Pattern Recognition - Workshops*, San Francisco, CA, USA, 2010, pp. 94–101.

[41] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, OpenFace 2.0: Facial behavior analysis toolkit, in *Proc. 13th IEEE Int. Conf. Automatic Face Gesture Recognition*, Xi'an, China, 2018, pp. 59–66.

[42] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[43] J. Wang and X. Geng, Explaining the better generalization of label distribution learning for classification, *Sci. China Inf. Sci.*, vol. 68, p. 152102, 2025.

[44] N. Xu, Y. Liu and X. Geng, Label enhancement for label distribution learning, *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1632-1643, 2021.

[45] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, AugMix: A simple data processing method to improve robustness and uncertainty, in *Proc. Int. Conf. Learning Representation*, Addis Ababa, Ethiopia, 2020, pp. 1–11.

[46] P. Chen, S. Liu, H. Zhao, X. Wang, and J. Jia, GridMask data augmentation, *arXiv preprint arXiv:2001.04086*, 2020, doi:10.48550/arXiv.2001.04086.

[47] A. F. M. Uddin, M. S. Monira, W. Shin, T. Chung, S. Bae, SaliencyMix: A saliency guided data augmentation strategy for better regularization, in *Proc. Int. Conf. Learning Representation*, Virtual Conference, 2021, pp. 1–12.

[48] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, Domain generalization with MixStyle, in *Proc. Int. Conf. Learning Representation*, Virtual Conference, 2021, pp. 1–12.

[49] J. Liu, B. Liu, H. Zhou, H. Li, and Y. Liu, TokenMix: Rethinking image mixing for data augmentation in vision transformers, in *Proc. European Conf. Computer Vision*, Tel Aviv, Israel, 2022, pp. 455—471.

[50] J. Lv, B. Kim, B.D. Parameshachari, A. Slowik, and K. Li, Large model-driven hyperscale healthcare data fusion analysis in complex multi-sensors, *Inf. Fusion*, vol. 115, p. 102780, 2025.

Author biography

Jingyang Zhou received the BEng and MEng degrees from Nanjing University, Nanjing, China in 2001 and 2004, respectively. He is currently pursuing the PhD degree at Southeast University, Nanjing, China. His research interests include machine learning and pattern recognition.

Jinyao Liu is currently pursuing the BEng degree at Chien-Shiung Wu College, Southeast University, Nanjing, China. His research interests include machine learning and computer vision.

Jing Wang received the BEng degree from Suzhou University of Science and Technology, Suzhou, China, in 2013, the MEng degree from Northeastern University, Shenyang, China, in 2015, and the PhD degree from Southeast University, Nanjing, China, in 2021. He is currently an assistant researcher with the School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include pattern recognition and machine learning.