Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition

Delian Ruan[1,2], Yan Yan[1]*, Shenqi Lai[2], Zhenhua Chai[2], Chunhua Shen[3], Hanzi Wang[1]
[1]Xiamen University, China [2]Vision Intelligence Center, Meituan [3]University of Adelaide

Abstract

In this paper, we propose a novel Feature Decomposition and Reconstruction Learning (FDRL) method for effective facial expression recognition. We view the expression information as the combination of the shared information (expression similarities) across different expressions and the unique information (expression-specific variations) for each expression. More specifically, FDRL mainly consists of two crucial networks: a Feature Decomposition Network (FDN) and a Feature Reconstruction Network (FRN). In particular, FDN first decomposes the basic features extracted from a backbone network into a set of facial action-aware latent features to model expression similarities. Then, FRN captures the intra-feature and inter-feature relationships for latent features to characterize expression-specific variations, and reconstructs the expression feature. To this end, two modules including an intra-feature relation modeling module and an inter-feature relation modeling module are developed in FRN. Experimental results on both the in-the-lab databases (including CK+, MMI, and Oulu-CASIA) and the in-the-wild databases (including RAF-DB and SFEW) show that the proposed FDRL method consistently achieves higher recognition accuracy than several state-of-the-art methods. This clearly highlights the benefit of feature decomposition and reconstruction for classifying expressions.

1. Introduction

Facial expression is one of the most natural and universal signals for human beings to express their inner states and intentions [4]. Over the past few decades, Facial Expression Recognition (FER) has received much attention in computer vision, due to its various applications including virtual reality, intelligent tutoring systems, health-care, *etc.* [29]. According to psychological studies [9], the FER task is to classify an input facial image into one of the following seven categories: angry (AN), disgust (DI), fear (FE), happy (HA), sad (SA), surprise (SU), and neutral (NE).

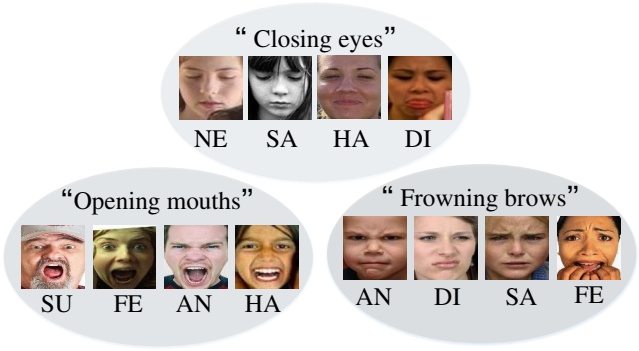

Figure 1 – The images in each group show a similar facial action, but they are from different expressions. Images are from the RAF-DB database [13].

A variety of FER methods [3, 13, 20, 26] have been proposed to learn holistic expression features by disentangling the disturbance caused by various disturbing factors, such as pose, identity, illumination, and so on. However, these methods neglect the fact that the extracted expression features corresponding to some expressions may still not be easily distinguishable, mainly because of high similarities across different expressions.

An example is shown in Figure 1. We can observe that some facial images corresponding to the NE, SA, HA, and DI expressions exhibit closing eyes. The facial images corresponding to the SU, FE, AN, and HA expressions all show opening mouths, while those corresponding to the AN, DI, SA, and FE expressions show frowning brows. The images from different facial expressions in each group give a similar facial action, where the distinctions between some expressions are subtle. Therefore, how to learn effective fine-grained expression features to identify subtle differences in expressions by considering similar facial actions is of great importance.

The expression information is composed of the shared information (expression similarities) across different expressions and the unique information (expression-specific variations) for each expression. The expression similari-

*Corresponding author (email: yanyan@xmu.edu.cn).

ties can be characterized by shared latent features between different expressions, while the expression-specific variations can be reflected by importance weights for latent features. Therefore, the expression features can be represented by combining a set of latent features associated with their corresponding importance weights. Traditional FER methods [15, 18, 31, 5] adopt Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) to extract eigenvectors (corresponding to latent features) and eigenvalues (corresponding to importance weights). However, these eigenvectors only capture holistic structural information rather than fine-grained semantic information of facial images, which is critical for FER.

Motivated by the success of deep learning in various vision tasks, here we propose a novel Feature Decomposition and Reconstruction Learning (FDRL) method for effective FER. FDRL is mainly comprised of two crucial networks, including a Feature Decomposition Network (FDN) and a Feature Reconstruction Network (FRN). The two networks are tightly combined and jointly trained in an end-to-end manner.

Specifically, a backbone convolutional neural network is first used to extract basic features. Then, FDN decomposes the basic feature into a set of facial action-aware latent features, which effectively encode expression similarities across different expressions. In particular, a compactness loss is developed to obtain compact latent feature representations. Next, FRN, which includes an Intra-feature Relation Modeling module (Intra-RM) and an Inter-feature Relation Modeling module (Inter-RM), models expression-specific variations and reconstructs the expression feature. Finally, an expression prediction network is employed for expression classification.

In summary, our main contributions are summarized as follows.

- A novel FDRL method is proposed to perform FER. In FDRL, FDN and FRN are respectively developed to explicitly model expression similarities and expression-specific variations, enabling the extraction of fine-grained expression features. Thus, the subtle differences between facial expressions can be accurately identified.
- Intra-RM and Inter-RM are elaborately designed to learn an intra-feature relation weight and an inter-feature relation weight for each latent feature, respectively. Therefore, the intra-feature and inter-feature relationships between latent features are effectively captured to obtain discriminative expression features.
- Our FDRL method is extensively evaluated on both the in-the-lab and the in-the-wild FER databases. Experimental results show that our method consistently outperforms several state-of-the-art FER methods. In particular, FDRL achieves 89.47% and 62.16% recognition accuracy on the RAF-DB and SFEW databases, respectively. This convincingly shows the great potentials of feature decomposition and reconstruction for FER.

2. Related work

With the rapid development of deep learning, extensive efforts have been made to perform FER. State-of-the-art deep learning-based FER methods mainly focus on two aspects: 1) disturbance disentangling, and 2) expression feature extraction.

2.1. Disturbance Disentangling

Many FER methods have been proposed to predict expressions by disentangling the disturbance caused by various disturbing factors, such as pose, identity, illumination, and so on. Wang *et al.* [22] propose an adversarial feature learning method to tackle the disturbance caused by facial identity and pose variations. Ruan *et al.* [20] propose a novel Disturbance-Disentangled Learning (DDL) method to simultaneously disentangle multiple disturbing factors. Note that the above methods depend largely on the label information of disturbing factors. A few methods address the occlusion problem of FER. Wang and Peng [24] propose a novel Region Attention Network (RAN) to adaptively adjust the importance of facial regions to mitigate the problems of occlusion and variant poses for FER.

Recently, some methods are concerned with the noisy label problem in the FER databases. Zeng *et al.* [28] propose an Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) method to deal with the problem of inconsistency in different FER databases. Wang *et al.* [23] introduce a Self-Cure Network (SCN) to prevent the trained model from over-fitting uncertain facial images.

The above methods perform FER by alleviating the influence caused by disturbing factors or noisy labels. However, they do not take into account subtle differences between different facial expressions. In this paper, we formulate the FER problem from the perspective of feature decomposition and reconstruction, which successfully models expression similarities and expression-specific variations. Therefore, high-level semantic information can be effectively encoded to classify facial expressions.

2.2. Expression Feature Extraction

Some FER methods design effective network architectures and loss functions to reduce inter-class similarities and enhance intra-class compactness for expression feature extraction. Li *et al.* [13] propose a deep locality-preserving loss based method, which extracts discriminative expression features by preserving the locality closeness. Cai *et*

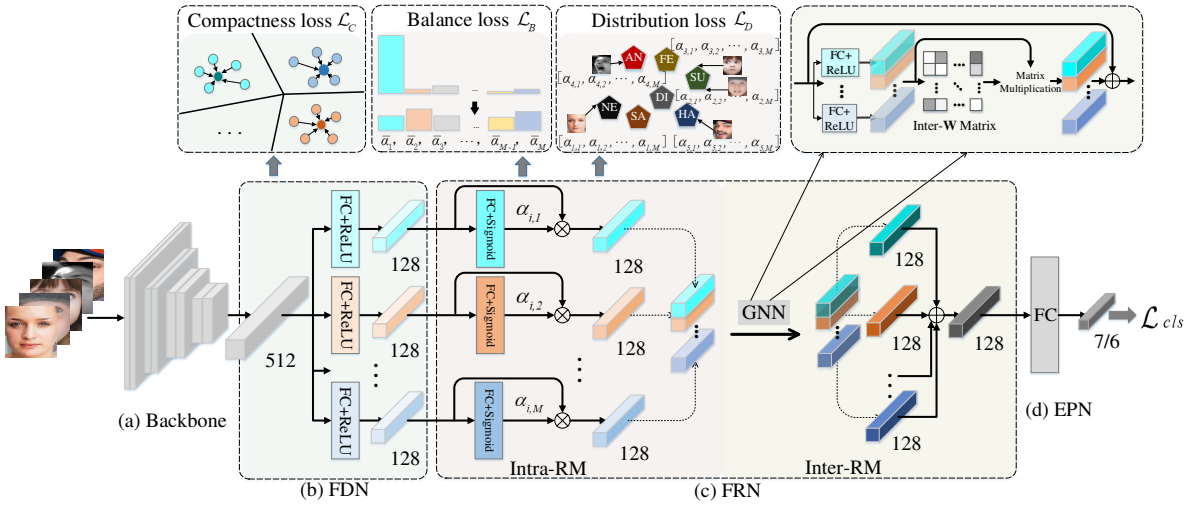

Figure 2 – Overview of our proposed FDRL method. (a) The backbone network (ResNet-18) that extracts basic CNN features; (b) A Feature Decomposition Network (FDN) that decomposes the basic feature into a set of facial action-aware latent features; (c) A Feature Reconstruction Network (FRN) that learns an intra-feature relation weight and an inter-feature relation weight for each latent feature, and reconstructs the expression feature. FRN contains two modules: an Intra-feature Relation Modeling module (Intra-RM) and an Inter-feature Relation Modeling module (Inter-RM); (d) An Expression Prediction Network (EPN) that predicts an expression label.

al. [3] design a novel island loss to simultaneously increase inter-class separability and intra-class compactness.

A few FER methods employ attention mechanisms [26] to improve the discriminative ability of expression features. Xie *et al.* [26] design an attention layer to focus on salient regions of a facial expression. Wang *et al.* [24] determine the importance of different facial regions by leveraging an attention network.

The above methods enhance the discriminative capability of expression features by designing different loss functions or attention mechanisms. These methods consider the expression features as holistic features. In contrast, we decompose the basic features into a set of facial action-aware latent features and then model the intra-feature and inter-feature relationships for latent features. Compared with holistic features used in traditional methods, the latent feature representations developed in our method are more fine-grained and facial action-aware. Such a manner is beneficial to learn expression features for identifying subtle differences between facial expressions.

3. Our Method

Overview The proposed FDRL method consists of a backbone network, a Feature Decomposition Network (FDN), a Feature Reconstruction Network (FRN), and an Expression Prediction Network (EPN). An overview of the proposed method is shown in Figure 2.

Given a batch of facial images, we first feed them into a backbone network (in this paper, we use ResNet-18 [11] as the backbone) to extract basic CNN features. Then, FDN decomposes the basic features into a set of facial action-aware latent features, where a compactness loss is designed to extract compact feature representations. Next, FRN learns an intra-feature relation weight and an inter-feature relation weight for each latent feature, and reconstructs the expression feature. Finally, EPN (a simple linear fully-connected layer) predicts a facial expression label.

In particular, FRN consists of two modules: an Intra-feature Relation Modeling module (Intra-RM) and an Inter-feature Relation Modeling module (Inter-RM). To be specific, Intra-RM is first introduced to assign an intra-feature relation weight to each latent feature according to the importance of the feature, and thus an intra-aware feature is obtained. To ensure similar distributions of intra-feature relation weights for facial images from the same expression category, a distribution loss and a balance loss are employed in Intra-RM. Then, Inter-RM computes an inter-feature relation weight by investigating the relationship between intra-aware features, and thus an inter-aware feature is extracted. At last, the expression feature is represented by a combination of the intra-aware feature and the inter-aware feature. FRN exploits both the contribution of each latent feature and the correlations between intra-aware features, enabling the extraction of discriminative expression features.

3.1. Feature Decomposition Network (FDN)

Given the i-th facial image, the basic feature extracted by the backbone network is denoted as $\mathbf{x}_i \in \mathbf{R}^P$, where P is the dimension of the basic feature. As men-

tioned previously, FDN decomposes the basic feature into a set of facial action-aware latent features. Let $\mathbf{L}_i = [\mathbf{l}_{i,1}, \mathbf{l}_{i,2}, \cdots, \mathbf{l}_{i,M}] \in \mathbf{R}^{D \times M}$ denote a facial action-aware latent feature matrix, where $\mathbf{l}_{i,j} \in \mathbf{R}^D$ represents the j-th latent feature for the i-th facial image. D and M represent the dimension of each latent feature and the number of latent features, respectively.

Specifically, to extract the j-th latent feature, we employ a linear Fully-Connected (FC) layer and a ReLU activation function, which can be formulated as:

$$\mathbf{l}_{i,j} = \sigma_1(\mathbf{W}_{d_j}^{\mathrm{T}} \mathbf{x}_i) \quad \text{for } j = 1, 2, \cdots, M, \tag{1}$$

where $\mathbf{W}_{d_j}$ denotes the parameters of the FC layer used for extracting the j-th latent feature and σ_1 represents the ReLU function.

Compactness Loss. Since different facial expressions share the same set of latent features, it is expected that a set of compact latent feature representations are extracted. In other words, the j-th latent feature extracted from one basic feature should be similar to that extracted from another basic feature. To achieve this, inspired by the center loss [25], we develop a compactness loss. The compactness loss $\mathcal{L}_C$ learns a center for the same latent features and penalizes the distances between the latent features and their corresponding centers, which can be formulated as:

$$\mathcal{L}_C = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \| \mathbf{l}_{i,j} - \mathbf{c}_j \|_2^2, \tag{2}$$

where N denotes the number of images in a mini-batch. $\| \cdot \|_2$ indicates the L_2 norm. $\mathbf{c}_j \in \mathbf{R}^D$ denotes the center of the j-th latent features, and is updated based on a mini-batch. Thus, the intra-latent variations are minimized and a set of compact latent features are effectively learned.

To visually demonstrate the interpretation of latent features, we collect a group of images that corresponds to the highest intra-feature relation weight (see Section 3.2) of the same latent feature and then visualize them. In Figure 3, we can observe that the images from each group show a specific facial action. The images from the nine groups show the facial actions of "Neutral", "Lip Corner Puller", "Staring", "Opening Mouths", "Lips Part", "Closing Eyes", "Grinning", "Frowning Brows", and "Lip Corner Depressor", respectively. Therefore, the latent features obtained by FDN are fine-grained and facial action-aware features, which can be useful for subsequent expression feature extraction.

3.2. Feature Reconstruction Network (FRN)

In this section, FRN, which models expression-specific variations, is carefully designed to obtain discriminative expression features. FRN contains two modules: Intra-RM and Inter-RM.

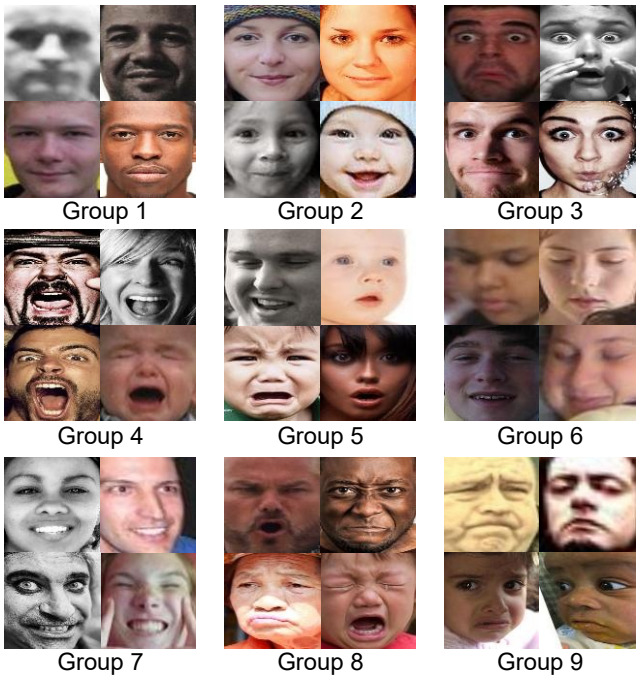

Figure 3 – Visualization of the image groups from the RAF-DB database when M is set to 9. Each group corresponds to the highest intra-feature relation weight of the same latent feature.

Intra-feature Relation Modeling Module (Intra-RM). Intra-RM consists of multiple intra-feature relation modeling blocks, where each block is designed to model the intra-feature relationship between feature elements.

To be specific, each block is composed of an FC layer and a sigmoid activation function, that is:

$$\boldsymbol{\alpha}_{i,j} = \sigma_2(\mathbf{W}_{s_j}^{\mathrm{T}} \mathbf{l}_{i,j}) \quad \text{for } j = 1, 2, \cdots, M, \tag{3}$$

where $\boldsymbol{\alpha}_{i,j} \in \mathbf{R}^D$ denotes the importance weights for the j-th latent feature corresponding to the i-th facial image, $\mathbf{W}_{s_j}$ represents the parameters of the FC layer, and σ_2 indicates the sigmoid function.

With Eq. (3), we compute the L_1 norm of $\boldsymbol{\alpha}_{i,j}$ as the Intra-feature relation Weight (Intra-W) to determine the importance of the j-th latent feature, that is:

$$\alpha_{i,j} = \| \boldsymbol{\alpha}_{i,j} \|_1 \quad \text{for } j = 1, 2, \cdots, M, \tag{4}$$

where $\| \cdot \|_1$ denotes the L_1 norm.

It is desirable that the distributions of Intra-Ws corresponding to different images from the same expression category are as close as possible. Therefore, similarly to the compactness loss, a distribution loss is used to learn a center for each expression category and penalize the distances between the Intra-Ws from one class and the corresponding center. Hence, the variations caused by different disturbing factors are alleviated.

Suppose that the i-th facial image belongs to the k_i-th expression category. Mathematically, the distribution loss $\mathcal{L}_D$ is formulated as:

$$\mathcal{L}_D = \frac{1}{N}\sum_{i=1}^{N} \parallel \mathbf{w}_i - \mathbf{w}_{k_i} \parallel_2^2, \tag{5}$$

where $\mathbf{w}_i = [\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,M}]^{\mathrm{T}} \in \mathbf{R}^M$ represents the Intra-W vector for the i-th facial image. $\mathbf{w}_{k_i} \in \mathbf{R}^M$ denotes the class center corresponding to the k_i-th expression category.

By optimizing the distribution loss, the Intra-W vectors corresponding to different images from the same expression category are closely distributed. Thus, they are able to focus on expression-specific variations.

In practice, as shown in Figure 2, some Intra-Ws (corresponding to one or two latent features) usually show much higher values than the other Intra-Ws in the Intra-W vector for each image, since these Intra-Ws are individually computed. To alleviate this problem, we further design a balance loss to balance the distributions of elements in each Intra-W vector as:

$$\mathcal{L}_B = \parallel \overline{\mathbf{w}} - \mathbf{w}_u \parallel_1, \tag{6}$$

where $\overline{\mathbf{w}} = [\overline{\alpha}_1, \overline{\alpha}_2, \cdots, \overline{\alpha}_M]^{\mathrm{T}} \in \mathbf{R}^M$ represents the mean Intra-W vector for a batch of samples (i.e., $\overline{\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_i$). $\mathbf{w}_u = [\frac{1}{M}, \frac{1}{M}, \cdots, \frac{1}{M}]^{\mathrm{T}} \in \mathbf{R}^M$ denotes a uniformly-distributed weight vector.

After computing an Intra-W for each latent feature, we assign this weight to the corresponding feature and obtain an intra-aware feature for the i-th facial image as:

$$\mathbf{f}_{i,j} = \alpha_{i,j}\mathbf{l}_{i,j} \quad \text{for } j = 1, 2, \cdots, M, \tag{7}$$

where $\mathbf{f}_{i,j} \in \mathbf{R}^D$ represents the j-th intra-aware feature for the i-th facial image.

Inter-feature Relation Modeling Module (Inter-RM). Intra-RM learns an Intra-W for each individual latent feature. However, these Intra-Ws are independently extracted. Although the distribution loss imposes consistency regularization on the Intra-W, it does not fully consider the inter-relationship between latent features. In fact, for each facial expression, different kinds of facial actions usually simultaneously appear. For example, the FE expression often involves the facial actions of frowning brows and opening mouths. The HA expression contains the facial actions of stretching brows, closing eyes, and opening mouths. Therefore, it is critical to exploit the correlations between different facial action-aware latent features. To achieve this, we further introduce Inter-RM to learn an Inter-feature Relation Weight (Inter-W) between intra-aware features based on Graph Neural Network (GNN) [2, 21].

Inter-RM learns a set of relation messages and estimates the Inter-Ws between these messages. Specifically, for each $\mathbf{f}_{i,j}$, it is first fed into a message network for feature encoding. In this paper, the message network is composed of an FC layer and a ReLU activation function, which is:

$$\mathbf{g}_{i,j} = \sigma_1(\mathbf{W}_{e_j}^{\mathrm{T}}\mathbf{f}_{i,j}) \quad \text{for } j = 1, 2, \cdots, M, \tag{8}$$

where $\mathbf{W}_{e_j}$ denotes the parameters of the FC layer used for feature encoding and σ_1 represents the ReLU function. $\mathbf{g}_{i,j} \in \mathbf{R}^D$ denotes the j-th relation message for the i-th facial image.

Then, a relation message matrix $\mathbf{G}_i = [\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \cdots, \mathbf{g}_{i,M}] \in \mathbf{R}^{D\times M}$ is represented as nodes in the graph $G(\mathbf{G}_i, E)$. In our formulation, G is an undirected complete graph and E represents the set of relationships between different relation messages. $\omega_i(j,m)$ is the Inter-W which denotes the relation importance between the node $\mathbf{g}_{i,j}$ and the node $\mathbf{g}_{i,m}$. It can be calculated as:

$$\omega_i(j,m) = \begin{cases} \sigma_3(S(\mathbf{g}_{i,j}, \mathbf{g}_{i,m})) & j \neq m \\ 0 & j = m \end{cases}, \tag{9}$$

where $\mathbf{g}_{i,j}$ and $\mathbf{g}_{i,m}$ are the j-th and the m-th relation messages for the i-th facial image, respectively. S is a distance function, which estimates the similarity score between $\mathbf{g}_{i,j}$ and $\mathbf{g}_{i,m}$. In our paper, we use the Euclidean distance function. Since the results of $S(\cdot)$ are all positive, we further adopt the tanh activation function σ_3 to normalize the positive distance value to [0,1). The purpose of setting $\omega_i(j,j)$ to 0 is to avoid self-enhancing. According to Eq. (9), an Inter-W matrix $\mathbf{W}_i = \{\omega_i(j,m)\} \in \mathbf{R}^{M\times M}$ can be obtained.

Hence, the j-th inter-aware feature $\hat{\mathbf{f}}_{i,j} \in \mathbf{R}^D$ for the i-th facial image can be formulated as:

$$\hat{\mathbf{f}}_{i,j} = \sum_{m=1}^{M} \omega_i(j,m)\mathbf{g}_{i,m} \quad \text{for } j = 1, 2, \cdots, M. \tag{10}$$

By combining the j-th intra-aware feature and the j-th inter-aware feature, the j-th importance-aware feature $\mathbf{y}_{i,j} \in \mathbf{R}^D$ for the i-th facial image is calculated as:

$$\mathbf{y}_{i,j} = \delta\mathbf{f}_{i,j} + (1-\delta)\hat{\mathbf{f}}_{i,j} \quad \text{for } j = 1, 2, \cdots, M, \tag{11}$$

where δ represents the regularization parameter that balances the intra-aware feature and the inter-aware feature.

Finally, a set of importance-aware features are added to obtain the final expression feature, that is,

$$\mathbf{y}_i = \sum_{j=1}^{M} \mathbf{y}_{i,j}, \tag{12}$$

where $\mathbf{y}_i \in \mathbf{R}^D$ represents the expression feature for the i-th facial image.

3.3. Joint Loss Function

In the proposed FDRL, the backbone network, FDN, FRN, and EPN are jointly trained in an end-to-end manner. The whole network minimizes the following joint loss function:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_C + \lambda_2 \mathcal{L}_B + \lambda_3 \mathcal{L}_D, \quad (13)$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_C$, $\mathcal{L}_B$, and $\mathcal{L}_D$ represent the classification loss, the compactness loss, the balance loss, and the distribution loss, respectively. In this paper, we use the cross-entropy loss as the classification loss. λ_1, λ_2, and λ_3 denote the regularization parameters. By optimizing the joint loss, FDRL is able to extract discriminative fine-grained expression features for FER.

4. Experiments

We first briefly introduce five public FER databases. Then, we describe the implementation details, and perform ablation studies with qualitative and quantitative results to show the importance of each component in FDRL. Finally, we compare FDRL with state-of-the-art FER methods.

4.1. Databases

CK+ [14] contains 327 video sequences, which are captured in controlled lab environments. We choose the three peak expression frames from each expression sequence to construct the training set and the test set, thus resulting in a total of 981 images. **MMI [19]** is also a lab-controlled database, containing 205 video sequences with six basic expressions. We choose the three peak frames from each sequence to construct the training set and the test set, thus resulting in a total of 615 images. **Oulu-CASIA [30]** contains videos captured in controlled lab conditions. We select the last three frames in each sequence captured with the visible light and strong illumination to construct the training set and the test set (consisting of 1,440 images in total). Similarly to [8, 17, 27, 32], we employ the subject-independent ten-fold cross-validation protocol for evaluation on all the three in-the-lab databases.

RAF-DB [13] is a real-world FER database, which contains 30,000 images labeled with basic or compound expressions by 40 trained human labelers. The images with six basic expressions and one neutral expression are used in our experiment. RAF-DB involves 12,271 images for training and 3,068 images for testing. **SFEW [6]** is created by selecting static frames from Acted Facial Expressions in the Wild (AFEW) [7]. The images in SFEW are labeled with six basic expressions and one neutral expression. We use 958 images for training and 436 images for testing.

4.2. Implementation Details

For each database, all the facial images are detected and cropped according to eye positions, and the cropped images are further resized to the size of 256×256. During the training process, the facial images are randomly cropped to the size of 224×224, and then a random horizontal flip is applied for data augmentation. During the test process, the input image is center cropped to the size of 224×224 and then fed into the trained model. The FDRL method is implemented with the Pytorch toolbox and the backbone network is a lightweight ResNet-18 model [11]. Similarly to [23], the ResNet-18 is pre-trained on the MS-Celeb-1M face recognition database [10].

The dimension of the basic feature is 512. The dimensions of both the latent feature and the expression feature are 128. The value of δ in Eq. (11) is empirically set to 0.5. We train our FDRL in an end-to-end manner with a single TITAN X GPU for 40 epochs, and the batch size for all the databases is set to 64. The model is trained using the Adam algorithm [12] with the initial learning rate of 0.0001, $\beta_1 = 0.500$, and $\beta_2 = 0.999$. The learning rate is further divided by 10 after 10, 18, 25, and 32 epochs.

4.3. Ablation Studies

To show the effectiveness of our method, we perform ablation studies to evaluate the influence of key parameters and components on the final performance. For all the experiments, we use one in-the-lab database (MMI) and one in-the-wild database (RAF-DB) to evaluate the performance.

Influence of the number of latent features. As shown in Figure 4, we can see that the proposed method achieves the best recognition accuracy when the number of latent features is set to 9. On one hand, when a small number of latent features are used, the expression similarities cannot be effectively modeled. On the other hand, when a large number of latent features are used, there exist redundancy and noise among latent features, thus leading to a performance decrease. In the following experiments, we set the number of latent features to 9.

Influence of the parameters. We evaluate the recognition performance of the proposed method with the different values of λ_1, λ_2, and λ_3 in Eq. (13), as shown in Table 1.

Specifically, we first fix $\lambda_2 = 1.0$ and $\lambda_3 = 0.0001$, and set the value of λ_1 from 0 to 0.01. Experimental results are given in Table 1 (a). We can observe that our method achieves the best performance when the value of λ_1 is set to 0.0001. When $\lambda_1 = 0$, our method is trained without using the compactness loss, and the performance decreases. Table 1 (b) shows the recognition performance obtained by our method, when the values of λ_1 and λ_3 are both set to 0.0001, and the value of λ_2 varies from 0 to 2.0. When the value of λ_2 is set to 1.0, our method achieves the top performance. Then, we fix $\lambda_1 = 0.0001$ and $\lambda_2 = 1.0$, and set the value

Table 1 – Ablation studies for the different values of λ_1, λ_2, and λ_3 (represent the balance parameters for compactness loss, balance loss, and distribution loss, respectively) on MMI and RAF-DB. The recognition accuracy (%) is used for performance evaluation.

(a) Influence of λ_1.

λ_1	MMI	RAF-DB
0	84.64	88.75
0.00001	85.02	89.02
0.0001	**85.23**	**89.47**
0.001	82.67	88.82
0.01	82.24	88.63

(b) Influence of λ_2.

λ_2	MMI	RAF-DB
0	82.66	88.23
0.5	83.68	88.89
1.0	**85.23**	**89.47**
1.5	84.94	88.92
2.0	83.23	88.63

(c) Influence of λ_3.

λ_3	MMI	RAF-DB
0	84.96	89.15
0.00001	85.07	88.89
0.0001	**85.23**	**89.47**
0.001	82.66	88.95
0.01	81.64	88.49

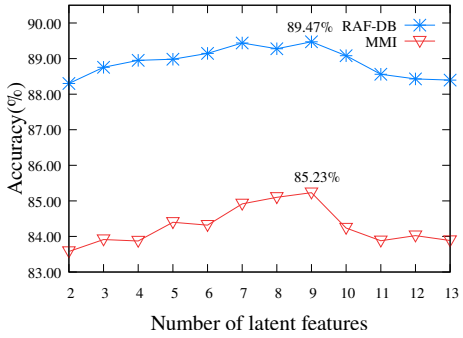

Figure 4 – Ablation studies for the different numbers of latent features on the MMI and RAF-DB databases.

Table 2 – Ablation studies for three key modules of our FDRL on the MMI and RAF-DB databases. The recognition accuracy (%) is used for performance evaluation.

FDN	FRN		MMI	RAF-DB
	Intra-RM	Inter-RM		
×	×	×	79.69	86.93
√	×	×	81.23	87.71
√	×	√	83.44	88.76
√	√	×	84.74	89.34
√	√	√	**85.23**	**89.47**

of λ_3 from 0 to 0.01. Experimental results are given in Table 1 (c). Our method obtains the top performance when $\lambda_3 = 0.0001$. In the following, we set the values of both λ_1 and λ_3 to 0.0001, and set the value of λ_2 to 1.0.

Influence of the key modules. To evaluate the effectiveness of the key modules in FDRL, we perform ablation studies for FDN, Intra-RM, and Inter-RM on the MMI and RAF-DB databases, respectively. Experimental results are reported in Table 2.

We can see that incorporating FDN into the backbone network improves the performance, which shows the importance of FDN. Moreover, by employing Intra-RM or Inter-RM in FRN, we are able to achieve better recognition accuracy than the method combining FDN and the backbone network. This is because the features extracted by FDN are not distinguishable enough to classify different expressions, since FDN does not take expression-specific variations into account. In contrast, Intra-RM and Inter-RM effectively model the intra-feature relationship of each latent feature and the inter-feature relationship between intra-aware features, respectively, leading to performance improvements. Our proposed FDRL method, which combines the backbone network, FDN, and FRN in an integrated network, achieves the best results among all the variants.

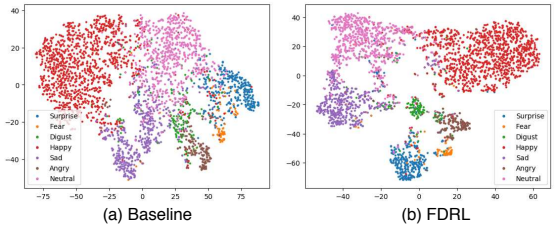

Figure 5 – Visualization of the expression features using t-SNE. Features are extracted from the RAF-DB database.

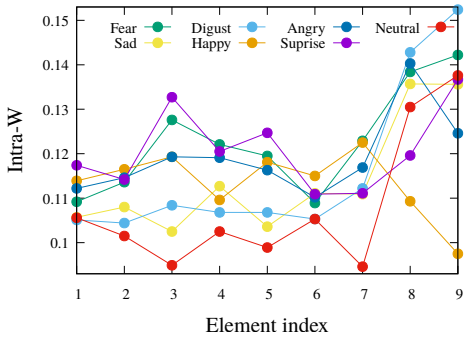

Figure 6 – Visualization of distribution of mean Intra-W vectors for seven basic expression categories on the RAF-DB database.

4.4. Visualization

2D feature visualization. We use t-SNE [16] to visualize the expression features extracted by the baseline method (which only adopts ResNet-18) and the proposed FDRL method on the 2D space, respectively, as shown in Figure 5. We can observe that the expression features extracted from

Table 3 – Performance comparisons among different methods on several public FER databases. The best results are boldfaced. ‡ and † respectively denote that seven expression categories and six expression categories are used in CK+.

(a) Comparisons on the in-the-lab databases.

Methods	Accuracy (%)		
	CK+	MMI	Oulu-CASIA
PPDN [32]	97.30†	-	72.40
IACNN [17]	95.37‡	71.55	-
DLP-CNN [13]	95.78†	78.46	-
IPA2LT [28]	92.45‡	65.61	61.49
DeRL [27]	97.37‡	73.23	88.00
FN2EN [8]	98.60†	-	87.71
DDL [20]	99.16‡	83.67	**88.26**
Baseline	97.15‡	79.69	86.18
FDRL (proposed)	**99.54** ‡	**85.23**	**88.26**

(b) Comparisons on the in-the-wild databases.

Methods	Accuracy (%)	
	RAF-DB	SFEW
IACNN [17]	-	50.98
DLP-CNN [13]	84.13	51.05
IPA2LT [28]	86.77	58.29
SPDNet [1]	87.00	58.14
RAN [24]	86.90	56.40
SCN [23]	87.01	-
DDL [20]	87.71	59.86
Baseline	86.93	58.03
FDRL (proposed)	**89.47**	**62.16**

the baseline method are not easily distinguishable for different facial expressions. In contrast, the features extracted from our proposed method can effectively reduce intra-class differences and enhance inter-class separability for different expressions. Especially, compared with baseline, the differences between fear and surprise, disgust and sadness are more distinct for FDRL.

Distribution of mean Intra-W vectors. We visualize the distribution of mean Intra-W vectors (corresponding to nine latent features) for seven basic expression categories on the RAF-DB database, as shown in Figure 6. Generally, each expression shows relatively high weights on the latent features associated with facial actions (as shown in Figure 3) closely related to this expression. Nevertheless, we can observe that some latent features (such as 2nd and 6th, 1st and 4th) have similar weights for different expressions. Hence, we further develop Inter-RM to exploit the inter-feature relationship between different intra-aware features.

4.5. Comparison with State-of-the-Art Methods

Table 3 shows the comparison results between our method and several state-of-the-art FER methods on the in-the-lab databases and the in-the-wild databases.

Among all the competing methods, IACNN, DDL, and RAN aim to disentangle the disturbing factors in facial expression images. SCN and IPA2LT are proposed to solve the noise label problem. FN2EN, DTAGN, and SPDNet improve the model performance by designing new network architectures. DLP-CNN alleviates intra-class variations by using a novel loss function. The above methods improve the FER performance by suppressing the influence of different disturbing factors or noise labels, but they ignore large expression similarities among different expressions. In contrast, our method explicitly models expression similarities and expression-specific variations with FDN and FRN, respectively, leading to performance improvements.

PPDN is developed to focus on the differences between expression images. DeRL claims that a facial expression is composed of the expression component and the neutral component. These two methods extract coarse-grained expression features. On the contrary, our proposed FDRL extracts more fine-grained features based on feature decomposition and reconstruction. Such a manner is beneficial to discriminate subtle differences between facial expressions, especially similar expression categories (such as fear and surprise). The above experimental results show the effectiveness of our proposed method.

5. Conclusion

In this paper, we have proposed a novel FDRL method for effective FER. FDRL consists of two main networks: FDN and FRN. FDN effectively models the shared information across different expressions based on a compactness loss. FRN accurately characterizes the unique information for each expression by taking advantage of Intra-RM and Inter-RM, and reconstructs the expression feature. In particular, Intra-RM encodes the intra-feature relationship of each latent feature and obtains an intra-aware feature. Inter-RM exploits the inter-feature relationship between different intra-aware features and extracts an inter-aware feature. The expression feature is represented by combining the intra-aware feature and the inter-aware feature. Experimental results on both the in-the-lab and the in-the-wild databases have shown the superiority of our method to perform FER.

Acknowledgements

This work was in part supported by the National Natural Science Foundation of China under Grants 62071404 and 61872307, by the Natural Science Foundation of Fujian Province under Grant 2020J01001, by the Youth Innovation Foundation of Xiamen City under Grant 3502Z20206046, and by the Beijing Science and Technology Project under Grant Z181100008918018.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 367–374, 2018.

[2] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[3] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *Proceeding of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 302–309, 2018.

[4] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[5] Hong-Bo Deng, Lian-Wen Jin, Li-Xin Zhen, Jian-Cheng Huang, et al. A new facial expression recognition method based on local gabor filter bank and PCA plus LDA. *International Journal of Information Technology*, 11(11):86–96, 2005.

[6] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceeding of IEEE International Conference on Computer Vision Workshops*, pages 2106–2112, 2011.

[7] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, (3):34–41, 2012.

[8] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In *Proceeding of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 118–126, 2017.

[9] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124, 1971.

[10] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 87–102, 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018.

[14] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 94–101, 2010.

[15] Yuan Luo, Cai-ming Wu, and Yi Zhang. Facial expression recognition based on fusion feature of PCA and LBP with SVM. *Optik-International Journal for Light and Electron Optics*, 124(17):2767–2770, 2013.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[17] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *Proceeding of IEEE International Conference on Automatic Face & Gesture Recognition)*, pages 558–565, 2017.

[18] Mohammad Reza Mohammadi, Emad Fatemizadeh, and Mohammad H Mahoor. PCA-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, 25(5):1082–1092, 2014.

[19] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5–15, 2005.

[20] Delian Ruan, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2833–2841, 2020.

[21] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision*, pages 486–504, 2018.

[22] Can Wang, Shangfei Wang, and Guang Liang. Identity-and pose-robust facial expression recognition through adversarial feature learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 238–246, 2019.

[23] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020.

[24] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[25] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 499–515, 2016.

[26] Siyue Xie, Haifeng Hu, and Yongbo Wu. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognition*, 92:177–191, 2019.

[27] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[28] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European Conference on Computer Vision*, pages 222–237, 2018.

[29] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368, 2018.

[30] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[31] Lihong Zhao, Guibin Zhuang, and Xinhe Xu. Facial expression recognition based on PCA and NMF. In *2008 7th World Congress on Intelligent Control and Automation*, pages 6826–6829, 2008.

[32] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *Proceedings of the European Conference on Computer Vision*, pages 425–442, 2016.