Neurocomputing 432 (2021) 159–169

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

LBAN-IL: A novel method of high discriminative representation for facial expression recognition

Hangyu Li [a], Nannan Wang [a,*], Yi Yu [b], Xi Yang [a], Xinbo Gao [c]

[a] State Key Laboratory of Intergrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an, Shannxi 710071, PR China
[b] Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan
[c] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, PR China

ARTICLE INFO

Article history:
Received 21 April 2020
Revised 30 November 2020
Accepted 12 December 2020
Available online 28 December 2020
Communicated by Steven Hoi

Keywords:
Facial expression recognition
Local binary attention network
Islets loss
High discriminative representation

ABSTRACT

Existing facial expression recognition (FER) works have achieved significant progress on constrained datasets. However, these methods only consider the sample distribution and achieve limited performance on unconstrained datasets. Facial expressions in the wild are influenced by various factors, e.g. illumination and partial occlusion, providing great challenge for model design and putting forward the higher requirement for feature discrimination. In this paper, we propose a novel LBAN-IL for FER in the wild, including local binary attention network (LBAN) and islets loss (IL). LBAN is based on two operations, local binary standard layer and encoder-decoder module. The former is derived from local binary convolution, so as to prevent excessive sparseness of feature maps and reduce the number of learnable parameters. The purpose of the latter is to generate attention-aware features and accurately discover local changes in the face. The proposed IL aims to enhance the discrimination of expression features by increasing the amplitude of vectors. Experimental results on RAF-DB, SFEW 2.0, FER-2013 and ExpW datasets validate the effectiveness of LBAN-IL and perform over some state-of-the-art methods.

1. Introduction

Facial expression is a form of nonverbal communication consisting of one or more muscle motions beneath facial skin for humans to convey the emotional state of an individual to observers. Numerous studies on FER have been done in the computer vision community, due to its practical importance in many applications such as medical diagnosis, public safety, and many other human–computer interaction techniques. These works can be generally categorized into two groups: traditional hand-crafted methods and deep learning-based methods.

Early works on FER mainly focus on recognizing expressions from static images or dynamic image sequences in tightly controlled environments, such as CK+ [1] dataset. The traditional methods have used pre-designed features or shallow learning. Gabor transform [2] and Local Binary Patterns (LBP) [3] have been demonstrated to be the most successful hand-crafted features in static FER. In addition, 2D features can be extended to three-dimensional space. LBP from three orthogonal planes (LBP-TOP) [4] solved the texture analysis of facial expression sequence and is more robust in processing the illumination variation problem. Though the above methods effectively extract spatial and temporal information, they achieve inferior performance on in-the-wild datasets with noise and clutter backgrounds, and cannot be applied to real-world scenarios. Therefore, it is worth to study an automatic facial expression analysis system in the wild.

Recently, deep convolutional neural networks (CNNs) have been applied to many visual tasks and achieved state-of-the-art performance, due to the manually designed architectures, such as AlexNet [5], VGGNet [6], ResNet [7] and DenseNet [8]. In the past, affected by the scale of facial expression datasets, deep learning could not be applied in FER efficiently. Up to now, there have been several relatively large-scale facial expression databases, such as FER-2013 [9], RAF-DB [10] and ExpW [11]. Inspired by the success of CNNs and the above FER data, there exists many deep learning-based works on FER [12–14] and their applications [15,16]. Most CNNs-based approaches adopt the softmax loss function to supervise CNN models in the purpose of forcing deep features from different image classes to stay apart, but they cannot effectively distinguish different facial expressions with only slight variants. Therefore, many works are devoted to the study of loss functions, in order to obtain more discriminative deep features.

* Corresponding author.
E-mail addresses: hangyuli.xidian@gmail.com (H. Li), nnwang@xidian.edu.cn (N. Wang), yiyu@nii.ac.jp (Y. Yu), yangx@xidian.edu.cn (X. Yang), gaoxb@cqupt.edu.cn (X. Gao).

https://doi.org/10.1016/j.neucom.2020.12.076
0925-2312/© 2020 Elsevier B.V. All rights reserved.

In [17], the center loss was proposed to minimize the intra-class distance of deep features for deep face recognition, but only pulled deep representation in the same class to a center point without processing different center points. In [10,12], the Deep Locality-Preserving CNN (DLP-CNN) method was proposed to enhance the discriminative power of deep FER features, which can be viewed as the generalization of center loss. Cai et al. [18] introduced the island loss for FER, reducing the intra-class variation while enlarging the inter-class difference, which can be viewed as an improvement on the center loss. Specifically, the island loss is a combination of the center loss and the cosine similarity between center points of different classes. In other words, the closer the cosine similarity is to -1, the greater the angle between two centers points in certain space is. ***However, once the cosine similarity of two centers reaches -1, that is, when two centers are in the same line, the distance can still be enlarged, which is not considered in [18].***

Given the above analysis, as illustrated in Fig. 1, the combination of different loss functions has a great effect on the deep feature distribution. When only softmax loss function controls feature distribution, the deep features are not only overlapped but also scattered. In view of the disadvantage of scattered features, the introduction of center loss can reduce the intra-class distance, but the effect of feature overlapping remains unsolved. Then, the island loss tries to decrease cosine similarity between center points. Here, three classes are taken as an example. Three centers constitute an equilateral trangle in order to minimize the mutual cosine similarity, however a small number of samples are still misclassified. Therefore, it is necessary to further improve existing methods to make deep features from the same category better gather together and deep features from different categories as far as possible.

In addition to the above issue, the attention mechanism is also vital for FER architecture. Bahdanau et al. [19] first utilized attention mechanism in neural machine translation and jointly learned to aligned and translate. Mnih *et al.* [20] presented the attention-based RNN model to adaptively select regions or locations for image classification. Wang *et al.* [21] proposed Residual Attention Network (RAN) and achieved satisfactory object recognition performance. Afterwards, many attention-based methods [22–24] are proposed for the FER task. Sun *et al.* [22] proposed the Visual Attention method for facial expression recognition, and integrated it into a single CNN model. Li *et al.* [23] designed the patch-based attention network and perceived the occlusion facial regions. The patches are cropped from the facial area based on facial landmarks and fed into an attention architecture. These attention models allow salient features to appear dynamically as needed. This is especially helpful when some facial expressions are difficult to observe. Compared with existing attention FER models, we hope to design a straightforward and easy pixel-based attention model. Especially, without the facial landmarks, the attention information can guide the model to focus on important facial pixels.

The above analysis illustrates that all the aforementioned algorithms do not get deep facial expression features well, and can not achieve better performance. Therefore, in order to recognize facial expressions more accurately, it is of great significance to study a novel network and an efficient loss function for high discriminative representation.

In this paper, in order to propose a novel method of high discriminative representation for FER in the wild, one of problems to be solved is to design an efficient network (LBAN). Specifically, according to the concept of local binary convolution proposed by [25], we extract deep features from images by combining local binary convolution and standard convolution at the aim of extracting proper sparse expression feature. Furthermore, we introduce the concept of attention mechanism in LBAN. In addition, in view of the unsatisfactory discrimination performance with the supervision of the existing loss functions, we further propose a novel islets loss (IL), aiming to enhance the distance between center points in all classes while minimizing the intra-class scatters. Jointly trained with the classical softmax loss which forces different classes to stay apart, islets loss further enhances the distance between samples from different classes and the discriminative power of learned representation greatly.

The main contributions of this paper are shown as follows:

(1) ***Local binary attention network.*** Unlike LBCNN in [25], we combine local binary convolution (LBC) layer with standard convolution layer to propose the local binary standard layer (LBSL) at the aim of alleviating the influence of excessive sparsity caused by LBC only. In addition, we introduce an encoder-decoder module in our model (LBAN), which effectively extracts local important expression feature.

(2) ***Islets loss.*** We propose a novel loss function, which makes learned deep representation more discriminant, and makes up for the disadvantage that cosine similarity ignores the amplitude of vectors.

(3) ***Superior performance.*** Experimental results on RAF-DB, SFEW 2.0, FER-2013 and ExpW datasets demonstrate our proposed LBAN-IL can achieve impressive results than previous methods and largely mitigate the impact in the wild.

The rest of this paper is organized as follows. Our framework is introduced in Section 2, including the proposed local binary attention network and the functions of each part in detail. Section 3 introduces the proposed islets loss function and we apply it to

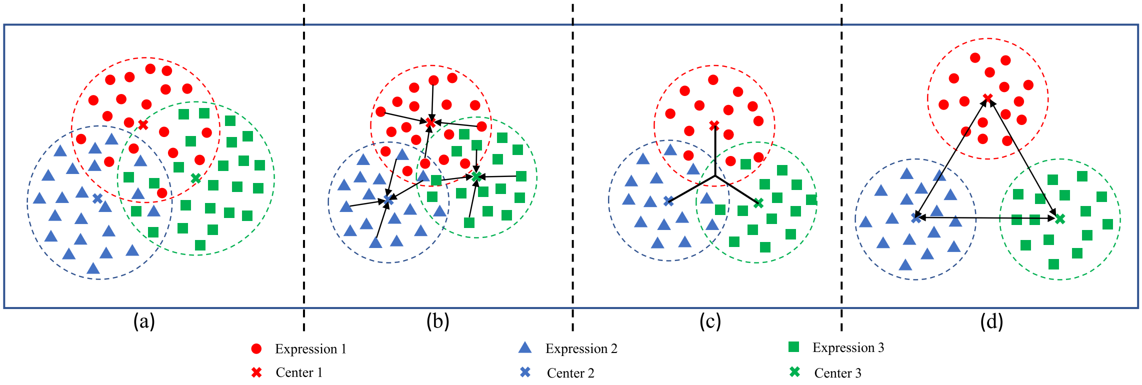


Fig. 1. An illustration of learned deep features by different combinations of loss functions, (a) the softmax loss, (b) the softmax loss + the center loss, (c) the softmax loss + the island loss, and (d) the softmax loss + the islets loss. In order to simplify the distribution, we only give examples of three classes, and the same is true for multiple classes..

the proposed model that is combined with softmax loss function for training. Section 4 provides the experimental results and analysis. Finally, the conclusion and future work are given in Section 5.

2. Our framework

In general, the CNN model directly affects the performance of the artificial intelligence system. In particular, FER is a very challenging problem and facial expressions are highly complex and similar, which put forward higher requirements for model design. To address this issue, we propose a novel method of high discriminative representation, i.e, an end-to-end local binary attention network (LBAN) with islets loss for FER in the wild as shown in Fig. 2. Our LBAN is constructed by stacking three local binary attention blocks (LBAB), each of which is designed according to local binary standard layer (LBSL) and encoder-decoder module. LBSL based on local binary convolution and standard convolution is used to extract proper sparse expression representation. In addition, the encoder-decoder network based on attention mechanism makes our framework more suitable for FER in the wild.

Given an expression image, a standard convolutional layer with the kernel of 3 × 3 is used to extract low-level expression features. Before expression features are input into each local binary attention block, there is a LBSL, which not only realizes the extraction of deep features, but also plays an important role in reducing the expression size. Subsequently, the activation function for the last block is ReLU, and the output size is further reduced through the average pooling. Finally, the last part of the architecture is two fully connected layers, in which the first one outputs a 128 dimension feature for calculating the proposed islets loss, and the second one outputs a seven dimension feature for calculating softmax loss. The details of LBSL and LBAB and the islets loss are given in the following sections.

2.1. Local binary standard layer

At this stage, the sparse feature is of great significance for FER model performance. According to the sparse binary convolution filter defined in advance, LBC obtains sparse features. However, excessive sparsity can easily lead to the unclear pattern of the image. Especially for facial expression recognition, over-sparse facial features cannot highlight the important information of expressions. Therefore, we introduce the traditional convolution to alleviate the influence of sparsity to a certain extent and extract the expression feature map better.

We now give more details about the smallest unit in our model as shown in Fig. 2. Unlike the residual block consisting of LBC layers entirely in [25], there are an LBC layer and a standard convolutional layer in LBSL. Among them, the LBC layer extracts sparse features of facial expression and consists of two convolutional layers according to LBCNN [25]. The first one is a standard convolutional layer with filters in the size of 3 × 3 without back propagation, only randomly assigning parameters to 1 or −1 according to the Bernoulli distribution and then assigning some parameters to 0 according to the sparsity level. In our LBSL, the sparsity level is set to 0.5, that is, half of all parameters are randomly assigned to 0, and the other half are assigned to 1 or −1. According to sparse parameters, the second one with the kernel of 1 × 1 and back propagation carries out information interaction and integration across channels. Besides, there is a weight size in the LBC layer, which is the transition channels between the first convolutional layer and the second one. It is worth noting that the sigmoid function was used as the activation function in LBCNN. However, in order to keep consistent with the activation layer of standard convolution, the ReLU function is used as the nonlinear activation in our method. Following the LBC layer, standard convolutional layer with a filter in the size of 3 × 3 plays the same role as in CNNs, and further improve the nonlinear representation of this network. At this point, the number of channel in this standard convolution is the same as that of the second convolution with filter size of 1 × 1 in LBC layer, which is also different from the operation in LBCNN. In addition, the input features of this unit are added to features processed by LBC layer and standard convolutional layer as the input of next layer.

2.2. Local binary attention block

In this section, we introduce the proposed local binary attention block (LBAB), based on the local binary standard layer and attention operation, to ensure that global features are learned while combining local attention features to classify facial expression in the wild.

In order to achieve better facial expression recognition, we further put forward the LBSL by combining local binary convolution with standard convolution. While extracting the sparse expression features, LBSL improves the nonlinear representation of neural networks, reasonably prunes unnecessary parameters and reduces the complexity of the model. However, extracting global features of facial expression by LBSL merely cannot effectively solve this problem. For FER in the wild especially, it is clear that not all areas of the face are involved in expression detection. In general, an expression is only accompanied by the change in only one or several areas of the face. Therefore, we need to pay more attention to a specific area to obtain the real emotion hidden under the face, i.e. to obtain

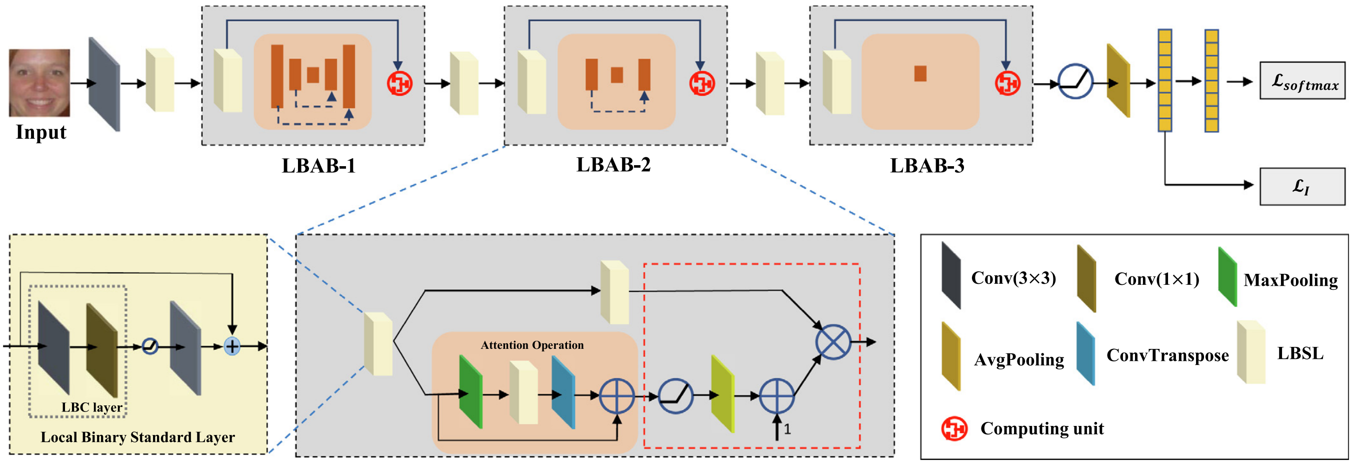


Fig. 2. The pipeline of our proposed LBAN for facial expression recognition in the wild. Since the internal principles of three blocks are similar, we only describe the internal structure of the second block in detail here.

local expression features through the attention mechanism and improve the performance of FER system.

Motivated by the attention mechanism in DBN [26] and RAN [21], our LBAB contains two streams, including forward LBSL and attention operation, as illustrated in Fig. 2. First, an LBSL processes sparse characteristics which are sampled down. Forward LBSL extracts the global expression features and attention operation is composed of an encoder-decoder network to extract the local expression features. The decoder is symmetric with the encoder, and the max-pooling realizes down-sampling and increases the range of the sensory field, while the deconvolution realizes upper sampling. It is worth noting that an LBSL processes features in the sampling processing to further obtain sparsity features. Skip-connections connect corresponding layer of encoder and decoder at the same level. The decoder outputs attention features of the same resolution as inputs. Based on provided attention and global features, the main feature is contained in the computing unit, which can improve the global information by attention information and better extract key features. In other words, a supervision is realized by combining the original global information with the expanded attention features. The attention features are connected by a ReLU layer and a standard convolution layer with filter size of 1 × 1. Now, one special thing to note is that the activation function for the convolution of size in 1 × 1 is not ReLU, but the sigmoid layer. The reason is that the sigmoid function normalizes the output range to [0, 1] and the output is used as a weight to facilitate the product of global features extracted above. Finally, attention features through activation and convolution and global features are added and multiplied to serve as the input of next layers. Since the functions of three blocks are similar, the main difference lies in channels in three blocks, which are respectively 128, 256 and 512. In addition, the times of sampling in three blocks are different, namely 2, 1 and 0. As shown in LBAB-2, the feature map goes through one down-sampling and one up-sampling.

3. Islets loss function for facial expression recognition

Here, we introduce the proposed loss function and its optimization process in detail.

For the center loss, minimizing it tends to reduce intra-class distances of deep features, while it is possible that clusters corresponding to different classes may be overlapped with each other. In view of this problem, the island loss was proposed to increase inter-class distances while reducing intra-class distances, which is actually the summation of the center loss and the cosine similarity between all the center points [18]. The cosine similarity reflects the correlation between two vectors, but ignores their amplitude. However, when the cosine similarity of two points is -1, the Euclidean distance between them can be further increased to make the discrimination of learned deep features stronger. Therefore, we propose the islets loss as $\mathcal{L}_I$ for FER to further increase the inter-class distances and reduce the intra-class differences.

In order to optimize the process, namely minimize the objective function, we define a new additive penalty term $\mathcal{L}_p$ shown in the Eq. (1) to further enhance the inter-class discrepancy. In this term, the L2 norm of differences between any two different class centers is calculated. In order to facilitate the optimization, the norm is calculated exponentially after taking a negative value. As shown in the Eq. (2), the islets loss function is defined as the summation of the island loss and the penalty term.

$$\mathcal{L}_p = exp\left(-\frac{1}{2}\sum_{\substack{\mathbf{c}_j \in N \\ \mathbf{c}_k \in N}} \left\|\mathbf{c}_j - \mathbf{c}_k\right\|_2\right) \tag{1}$$

$$\mathcal{L}_I = \mathcal{L}_{Island} + \lambda_2 \mathcal{L}_p \tag{2}$$

where $j \neq k$, $\mathbf{c}_j$ and $\mathbf{c}_k$ denote the j^{th} and k^{th} centers respectively and N is the set of facial expression centers in databases. By minimizing the islets loss, samples from the same classes get closer to each other and those from different classes are separated as far as possible.

The gradients of $\mathcal{L}_I$ with respect to the feature $\mathbf{x}_i$ and update equation of $\mathbf{c}_{y_i}$ are computed as:

$$\frac{\partial \mathcal{L}_I}{\partial \mathbf{x}_i} = \mathbf{x}_i - \mathbf{c}_{y_i} \tag{3}$$

$$\begin{aligned}\Delta \mathbf{c}_j = &\frac{\sum_{i=1}^{m}\delta(y_i = j)\cdot(\mathbf{c}_j - \mathbf{x}_i)}{1 + \sum_{i=1}^{m}\delta(y_i = j)} + \frac{\lambda_1}{|N| - 1}\sum_{\substack{\mathbf{c}_k \in N \\ j \neq k}} \frac{\mathbf{c}_k}{\|\mathbf{c}_j\|_2\|\mathbf{c}_k\|_2} \\ &- \left(\frac{\mathbf{c}_j \cdot \mathbf{c}_k}{\|\mathbf{c}_j\|_2^3\|\mathbf{c}_k\|_2}\right)\mathbf{c}_j - \frac{\lambda_2}{2|N|}\sum_{\mathbf{c}_k \in N}\frac{\mathbf{c}_j - \mathbf{c}_k}{\|\mathbf{c}_j - \mathbf{c}_k\|_2} exp\left(-\frac{1}{2}\sum_{\substack{\mathbf{c}_j \in N \\ \mathbf{c}_k \in N}}\|\mathbf{c}_j - \mathbf{c}_k\|_2\right)\end{aligned} \tag{4}$$

where $|N|$ is the total number of expressions, m is the number of samples in the mini-batch, $\mathbf{x}_i$ represents the i^{th} sample, y_i is the class label of the i^{th} sample, $\mathbf{c}_{y_i}$ denotes the y_i^{th} class center of deep features. The Eq. (3) is actually the same as the gradients of center loss, and in the Eq. (4) $\delta(condition)$ = 1 if the condition is satisfied and 0 otherwise. Finally, in our proposed method, the loss function in the whole training process is defined in the Eq. (5), where $\mathcal{L}_{softmax}$ denotes the softmax loss function. In addition, the hyperparameter λ is used to balance two loss functions. In Algorithm 1, we introduce the learning details with the joint loss function.

$$\mathcal{L} = \mathcal{L}_{softmax} + \lambda \mathcal{L}_I \tag{5}$$

$$\mathcal{L}_{softmax} = -\frac{1}{m}\sum_{i=1}^{m} log \frac{e^{\theta_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{|N|} e^{\theta_j^T x_i + b_j}} \tag{6}$$

where θ_j denotes the j^{th} column of the last fully connected parameter θ and b_j is the bias term.

```
**Algorithm 1**: Optimization Algorithm of Our Method
Input: Training data $\{\mathbf{x}_i\}_{i=1}^{m}$;
Given: mini-batch size $m$, the maximum number of iterations
    T, hyperparameters $\alpha$, $\lambda$, $\lambda_1$ and $\lambda_2$, and learning rate $\mu$;
Initialize: t = 0, network layer parameters $W$, softmax loss
    parameters $\theta$, and islets loss parameters $c_j$;
Repeat:
Step 1: t = t + 1
Step 2: Calculate the joint loss:
$\mathcal{L} = \mathcal{L}_{softmax} + \lambda \mathcal{L}_I$
Step 3: Update the backprogation error:
$\frac{\partial \mathcal{L}^t}{\partial x_i^t} = \frac{\partial \mathcal{L}_{softmax}^t}{\partial x_i^t} + \lambda \frac{\partial \mathcal{L}_I^t}{\partial x_i^t}$
    Step 4: Update the softmax loss parameters:
$\theta^{t+1} = \theta^t - \mu \frac{\partial \mathcal{L}_{softmax}^t}{\partial \theta^t}$
Step 5: Update the class centers:
$\mathbf{c}_j^{t+1} = \mathbf{c}_j^t - \alpha \Delta \mathbf{c}_j^t$
Step 6: Update the network layer parameters:
$W^{t+1} = W^t - \mu \frac{\partial \mathcal{L}^t}{\partial W^t} = W^t - \mu \frac{\partial \mathcal{L}^t}{\partial x_i^t}\frac{\partial x_i^t}{\partial W^t}$
Until convergence or t $\geqslant$T
Output: Network layer parameters $W$, softmax loss
    parameters $\theta$, and islets loss parameters $c_j$.
```

4. Experimental results and analysis

In this section, we conduct experiments to validate the effectiveness of our proposed method. We compare its performance to some architectures and loss functions on four widely used databases, RAF-DB, SFEW 2.0, FER-2013 and ExpW, of which some examples are shown in Fig. 3. The following subsections show details of our experimental results.

4.1. Datasets

RAF-DB [10] is a large-scale facial expression database with nearly 30 K images, which contains two different subsets, basic and compound expressions. In our experiment, images with basic expressions were used basically, including 12,271 images as training data and 3,068 images as testing data. The compound subset is only compared to existing methods. **SFEW 2.0** [27] is separated into three sets, namely training (with 958 images), validation (with 436 images) and testing (with 372 images) sets. **FER2013** [9] is created automatically by the Google image search API and consists of 28,709 training images, 3,589 validation images and 3,589 testing images. **ExpW** [11] database contains 91,793 original web images manually labeled with six expressions (angry, disgust, fear, happy, sad and surprise) and neutral faces. The number of images in this database is much larger and face variations are more diverse than others databases. However, the training samples and testing samples are not given clearly in the ExpW database.

Data Preprocessing: In our experiments, all facial images are firstly aligned by MTCNN [28]. In addition, we adapt the strategy of data augmentation to increase the number of training data. Specifically, for the RAF-DB and SFEW 2.0, 80×80 images are randomly cropped from the aligned 100 × 100 facial images, and the cropped images are randomly flipped horizontally as the input of our model. During the testing process, in order to keep the size of testing images consistent with that of training images, we crop the middle and four corners of images in the size of 80 × 80, and get ten facial images after horizontal flip. Finally, the average value of testing results of ten images is taken as the testing result of the original image. Similar to the above operation, we adapt the same strategy in FER2013. The only noticeable one is that 40×40 images are randomly cropped from the aligned 48 × 48 facial expression images. For ExpW, we choose 41,748 images with the size of 100 × 100 or more in the facial area as training or testing samples which are resized into 100 × 100, and the sample number of each category is 2,280, 1,950, 627, 12,453, 6,158, 3,924 and 14,356. In our experiment, we adapt the 10-fold cross-validation strategy, that is, we divide all images into 10 groups and nine groups are selected as training samples and the rest is used for testing in each time.

4.2. Implementation details

In the whole process of our experiments, we implement our proposed method with the PyTorch library. It is worth noting that we adopt the same data augmentation operation in all experiments. Furthermore, stochastic gradient descent with a momentum rate of 0.9, a mini-batch size fixed to 64, and a weight decay parameter of 0.05 are set during the whole training process. The learning rate μ is set to 0.001 at the beginning, and decreased by 0.9 every five epoches since the 20^{th} epoch. We adapt the dropout layer before the last two fully connected layers with the rate of 0.5, that is to say, assigning the output of the neuron to zero with a probability of 50%. In practice, we set $\lambda = 0.6, \lambda_1 = 0.2$ and $\lambda_2 = 0.9$.

4.3. Evaluation Metric

The performance metrics we employ are Average and Accuracy rates. The Average rate (Ave) means the average of all categories. The Accuracy rate (Acc) refers to the overall accuracy on seven facial expression categories (i.e., six basic expressions plus neural face).

4.4. Ablation study

We conduct experiments to verify the effectiveness of each component in the proposed framework on both RAF-DB and FER-2013 databases. We report the results with only LBSL, only encoder-decoder module (attention operation) and LBAN as shown in Table 1. It is worth noting that the loss function used in all structures is the islets loss we present.

Comparing the results on both databases, we can observe that LBSL and attention operation can get better performance than basic structure separately. Furthermore, the combination of them can improve the recognition performance of the structure to a greater extent. The results of this part can verify that the combination of all components works best for FER in the wild.

4.5. Comparison of different architectures and learnable parameters

Considering that our proposed LBSL can alleviate the influence of excessive sparsity caused by LBC only, we compare the effects of each module (LBC, standard convolution and skip connection) in LBSL on RAF-DB and FER-2013. Furthermore, we compare results

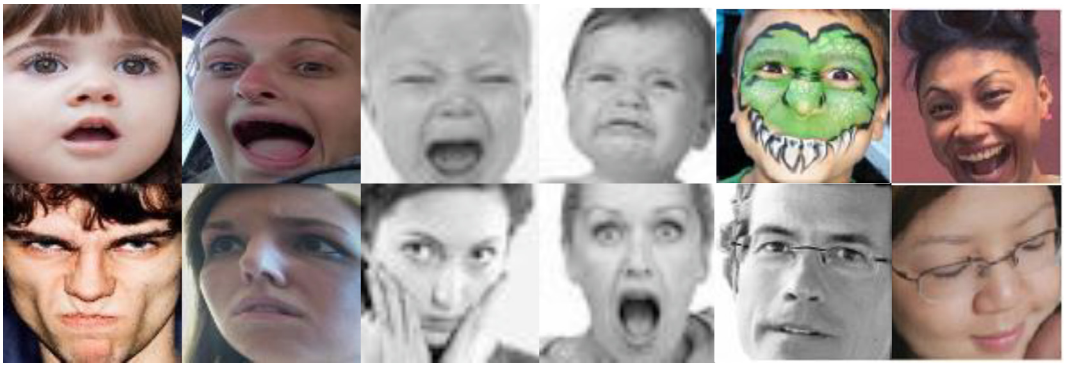


Fig. 3. Example photos from the FER databases: the first two columns are form the RAF database [10], the first of which is from the single-label subset and the second of which is from two-tab subset, the middle two columns are from the FER-2013 database [9] and the last two columns are from the ExpW database [11], which are extracted by us from original images.

Table 1
The comparison of components on RAF-DB and FER-2013 databases.

LBSL	×	×	✓	✓
Encoder-decoder module	×	✓	×	✓
Ave(RAF-DB)	72.98	75.65	76.24	**77.80**
Acc(FER-2013)	69.91	71.86	71.91	**73.11**

with classical backbones and LBAN without LBC to highlight reduction of learnable parameters. To make a fair comparison, we implement each model with the proposed islets loss. In Table 2, LBCNN denotes LBAN where LBSL without standard convolution. LBC is based on LBCNN and lacks the skip connection operation. The SC denotes LBC in which the convolution operation is replaced with standard convolutions.

Table 2 illustrates the quantitative results and reveals several observations: 1) Despite the 1.8 M difference in the number of learnable parameters (the sixth and last row), the Ave rate on RAF-DB varies by nearly 6 percent. This directly proves that our strategy of combining LBC with standard convolution is effective and alleviates the influence of excessive sparsity. 2) In order to compare the impact of learnable parameters, LBAN† replaces the LBC layer in LBAN with the combination of two standard convolutions (the kernel size 3 × 3 and 1 × 1 respectively). Similarly, LBAN‡ removes skip connections in LBSL based on LBAN†. Comparing the results (the last three rows), our LBAN can obtain better classification performance. This can be explained by the reason that the parameters in LBAN† and LBAN‡ increase sharply, reducing the generalization performance. 3) The first 3 × 3 convolution in LBSL is not involved in training process, which greatly reduces the calculation and effectively improves training speed. Especially compared with LBAN†, LBAN can reduce training time by 10 s.

Next, our focus is to introduce the difference in the number of parameters between LBAN† and LBAN. As mentioned in the explanation of local binary residual block, the LBC layer includes two standard convolutional layers, but the first convolutional layer does not carry out back propagation, so we conduct two standard convolutional layers in LBAN† in accordance with this difference. The result of this is the difference in learnable parameters between them, i.e., the latter has H × W × 3 × 3 more parameters than the former, where H and W represent the input and output channels respectively. As shown in Fig. 2, we illustrate the reduction of parameters in the first LBSL, where the input and output channels in the LBC layer are 64 and 128, and the weight size is 128, i.e., compared with the standard convolution, it reduces 64 × 128 × 3 × 3 parameters. In addition, it is worth noting is that when the skip connection is deleted, the parameter decreases partly. This is due to the difference in the input and output channels of the LBSL before each LBAB.

Table 2
The comparison of the learnable parameters, the Ave rate on RAF-DB and the Acc rate on FER-2013 by different modules. We give the parameters of our proposed LBAN excluding the parameters without back propagation, namely the fixed parameters. The SC is obtained by replacing LBC with standard convolution. The STD Conv and SK CONN denote the standard convolution and the skip connection respectively.

Model	Params	Ave	Acc
VGG19 + softmax	20.3 M	68.86	67.96
ResNet18 + softmax	11.7 M	66.84	68.93
VGG19	20.3 M	74.24	71.75
ResNet18	11.7 M	74.07	72.81
LBC	16.7 M	72.06	71.13
LBCNN(LBC + SK CONN)	18.0 M	72.21	71.61
LBAN‡(SC + STD Conv)	29.4 M	75.06	69.66
LBAN†(SC + STD Conv + SK CONN)	30.7 M	75.67	71.72
LBAN(LBC + STD Conv + SK CONN)	19.8 M	**77.80**	**73.11**

4.6. Comparison with state-of-the-art methods

Our proposed LBAN-IL is compared with the state-of-the-art methods evaluated on four facial expression datasets, such as CNNs-based methods (DLP-CNN [10,12], IL-CNN [18], DeepEmo [29], Boosting-POOF [30], Visual Attention [22], MDLBP + BAE [31] and IPA2LT [32]) on RAF-DB and SFEW 2.0 and methods (Bag of Visual Words [33], GoogleNet [34], Deep-Emotion [35], etc) on FER-2013. Furthermore, the confusion matrices based on our proposed method are presented for SFEW 2.0, FER-2013 and ExpW datasets respectively. In order to further compare the performance of various loss functions, we conduct related experiments on ExpW database.

(1) Results on the RAF-DB: The comparative results of each category and the Ave results are shown in Table 3. To highlight the advantage of islets loss, we compare the performance of three loss functions in Table 4. The comparison of the total Acc rate is given in Table 5(a).
From Table 5(a), we can intuitively observe that our method can exceed all existing methods in terms of the Acc rate. In addition, from the results in Table 3, we use the result of the VGG19 and ResNet18 networks under the supervision of the softmax loss as the baseline and have the following

Table 3
Comparisons of different methods on RAF-DB in terms of the Ave rates. * means the results are produced by our implementation.

Method	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral	Ave	Compound
VGG (baseline 1)	84.19	37.84	38.75	92.66	80.13	66.67	81.79	68.86	42.44
ResNet (baseline 2)	72.34	37.84	38.12	95.36	79.71	62.35	82.15	66.84	37.95
DLP-CNN [12]	81.16	62.16	52.15	92.83	80.13	71.60	80.29	74.20	44.55
DeepEmo [29]	–	–	–	–	–	–	–	68.20	–
Boosting-POOF [30]	80	**64**	**57**	89	74	73	76	73.19	–
Visual Attention* [22]	83.89	47.29	45.00	93.75	**84.51**	72.22	**86.91**	73.37	–
LBAN-IL	**86.63**	62.16	56.25	**95.36**	82.85	**79.01**	82.35	**77.80**	**45.77**

Table 4
Comparisons of different loss functions produced by us with LBAN on RAF-DB.

Loss	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral	Ave
Softmax loss	78.72	47.30	41.88	92.49	**85.36**	67.90	**87.21**	71.55
Center loss	84.80	48.65	48.75	94.68	81.59	75.31	84.41	74.03
Island loss	84.50	54.05	48.13	94.18	83.47	76.54	86.62	75.36
Islets loss	**86.63**	**62.16**	**56.25**	**95.36**	82.85	**79.01**	82.35	**77.80**

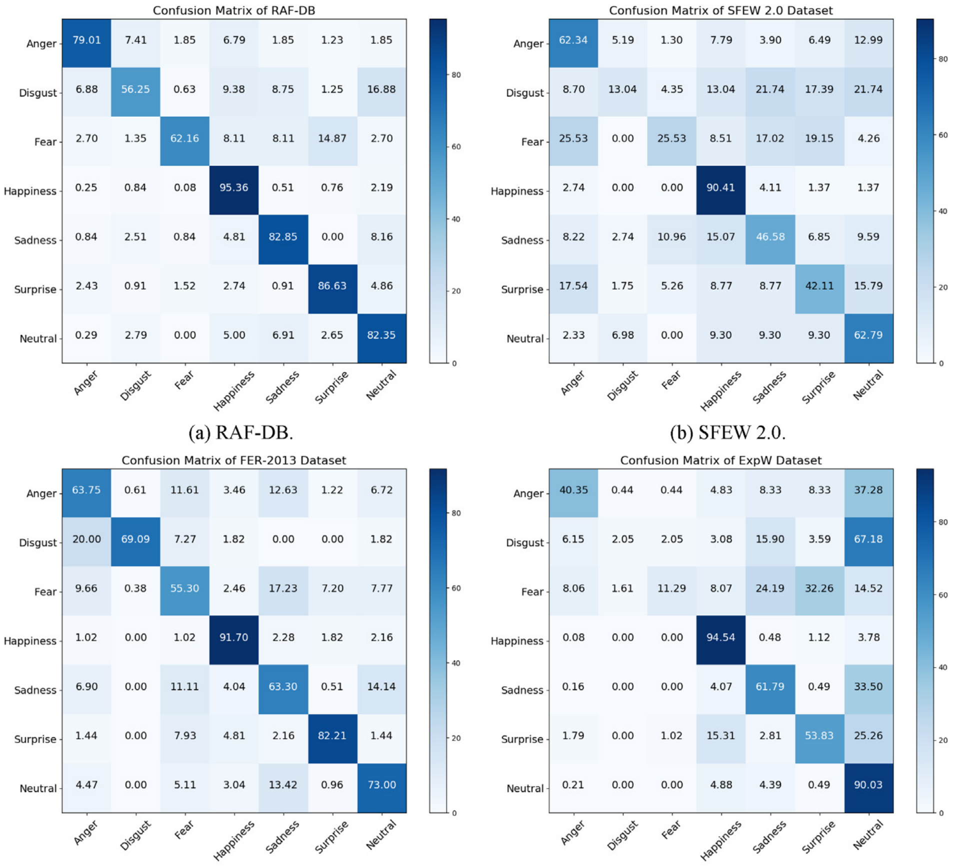
Table 5
Comparisons to the state-of-the-art methods in terms of the Acc rates. * means the results are produced by our implementation.

Method	Acc
(a) Comparison on RAF-DB.	
Occams razor [36]	80
DLP-CNN [12]	84.13
Visual Attention* [22]	84.94
gACNN [23]	85.07
LDL [37]	85.53
LBAN-IL	**85.89**
(b) Comparison on SFEW 2.0.	
Visual Attention [22]	48.3
MDLBP + BAE [31]	49.25
DLP-CNN [12]	51.05
Kim's CNN [38]	52.5
IPA2LT* [32]	53.15
LBAN-IL	**55.28**

observations. First, owing to the uneven distribution of samples, the testing performance of fear and disgust is relatively poor. Second, our proposed method achieves quite outstanding results in most facial expressions and the best performance in terms of the Ave rate. Third, despite the small sample size and complexity of compound expressions, we still achieve the best performance, i.e., the Ave rate 45.77%. Moreover, in order to compare the influence of different loss functions, we implement three existing loss functions in the field of face recognition or FER, and it is obvious that our islets loss performs best, as shown in Table 4. The performance is still the best in five facial expressions, which indicates that the penalty term plays a role of increasing distance in the high dimensional space. Although it is not optimal in the neutral faces, it is only used to distinguish whether there are expressions and does not affect the judgment on the effectiveness.

(2) Results on the SFEW 2.0: We compare our LBAN-IL to several state-of-the-art methods on the SFEW 2.0 dataset in Table 5 (b). The confusion matrix of our method is presented in Fig. 4. We obtain 55.28% on the validation set which is a new state of the art to our knowledge. It is worth noting that Reference [31] is the first to combine a binary feature extractor, a binary unsupervised feature learner and a binary neural network into the FER system. Our proposed local binary standard layer (LBSL) can better extract proper sparse expression features and achieve superior performance. In addition, compared to Visual Attention [22], our LBAN-IL performs better. This suggests that although the regions of interests have been discovered successfully in [22], our method can effectively observe pixel-based facial areas, which are important for discriminant representation.

(3) Results on the FER-2013 database: The comparisons of different methods in terms of Acc results are shown in Table 6. The confusion matrix of our LBAN-IL is presented in Fig. 4. To clarify, the experiments on the FER-2013 dataset are all about the comparison of Acc rates, and the current performance is not very ideal affected by its own defects. According to the comparison in Table 6, we clearly obverse that our islets loss can better monitor the training process. Furthermore, we also use the result of the VGG19 and ResNet18 net-



(a) RAF-DB.

(b) SFEW 2.0.

(c) FER-2013.

(d) ExpW.

Fig. 4. The confusion matrices of our LBAN-IL on the RAF-DB, SFEW 2.0, FER-2013 and ExpW.

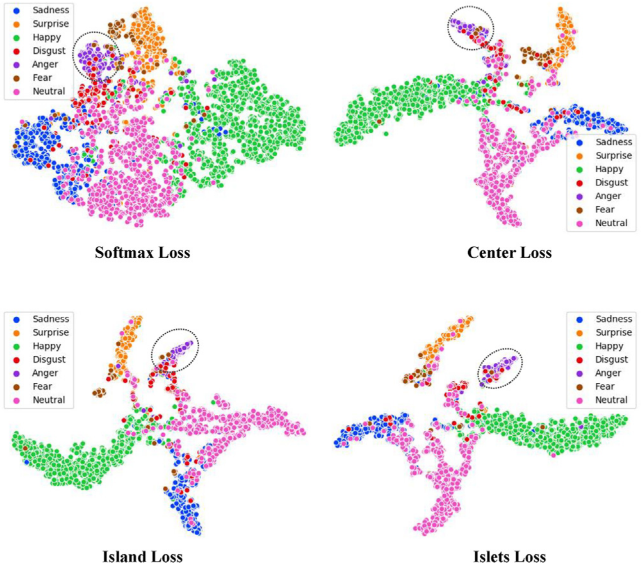
Table 6
Comparisons of different methods on FER-2013 database. * means the results are produced by our implementation. The following three loss functions are implemented with our LBAN.

Method	Acc
VGG (baseline 1)	67.96
ResNet (baseline 2)	68.93
Bag of Visual Words [33]	67.48
GoogleNet [34]	65.2
Mollahosseini et al. [39]	66.4
VGG13 + global SVM [40]	66.31
Deep-Emotion [35]	70.02
Visual Attention* [22]	70.74
DLSVM L2 [41]	71.2
DNNRL [42]	71.33
Softmax loss	71.58
Center loss	71.72
Island loss	71.97
LBAN-IL	**73.11**

works under the supervision of the softmax loss as the baseline and our LBAN-IL is obviously superior to other methods. According to the confusion matrix of FER-2013 shown in Fig. 4, we can obtain that the performance of happiness is over 90%, but it is not particularly good for other categories. In particular, the fear expression is the least effective, with most of expressions falsely classified as sadness because of the high similarity between them. In addition, there is another phenomenon worthy of our attention, that is, the number of other six expressions wrongly identified as disgust is very low, but the disgust expression is still wrongly identified as the other six expressions, especially the anger.

(4) Results on the ExpW database: Table 7 shows the results of 10 crops, providing a comparison for the following experiments on the ExpW database. We give the confusion matrix distribution within crop 5 as shown in Fig. 4.

In this section, we compare our method with several existing loss functions, i.e., softmax loss, center loss and island loss. As shown in Table 7, although we can obtain unsatisfactory results on several crops, but it is caused by the complexity of samples. Importantly, our islets loss achieves the best results on all crops, but it is not much different from other loss functions. This can be explained by the reason that the ExpW database is a challenging task with different sizes of facial expression and even some faces are covered by certain objects. According to the confusion matrix in crop 5, disgust and fear expressions have a poor performance and a large part of them are wrongly classified as neutral faces and surprise expressions, which is largely caused by the similarity between expressions. Therefore, in terms of the ExpW, what's most important is how to separate the fear from the surprise, otherwise the performance on negative emotions can be greatly affected.



Softmax Loss

Center Loss

Island Loss

Islets Loss

Fig. 5. A visualization study of the distribution of deep expression features by LBAN using different loss on the RAF-DB. **Best viewed in color.**.

To sum up, our LBAN-IL has achieved the superior performance on the above four datasets, which indicates the effectiveness of our method. While ensuring the reduction of learnable parameters, the performance is improved as far as possible, providing a good direction for future practical FER. In addition, it is also necessary to deal with the sample in ExpW database, that is, the difference of image size is large, and the super-resolution reconstruction method can be adopted to process all face images into the same size with the smallest distortion, which is of great help to the in-the-wild FER.

4.7. Visualization results

In order to verify the effects of the Islets loss and pixel-based attention mechanism in LBAN, we give related visualization results in this section. As shown in Fig. 5, the real expression distribution further demonstrates the validity of toy example in Fig. 1. Based on the feature map shown in Fig. 6, we can obtain how our LBAN observes the important area in facial expressions, and adopts it as important information for the learning of next layers.

To better visualize the expression feature distribution, t-SNE [43] is adopted as the method of feature dimension reduction. As shown in Fig. 5, in order to more intuitively observe the distribution of facial expression features, we display six facial expressions and neutral faces with different colors, for example, the purple point represents the anger expression. According to the above classification results, the overlapping of expression feature

Table 7
Comparisons of different loss functions on ExpW. Three corresponding loss functions are all implemented with our LBAN. The average value (the last row) is the average results obtained after cross validation (the first ten rows).

Method	Softmax loss	Center loss	Island loss	**Islets loss**
Crop1	57.88	59.95	60.33	**60.79**
Crop2	67.79	68.15	68.79	**69.63**
Crop3	71.72	73.61	73.99	**75.02**
Crop4	69.49	70.76	70.90	**72.12**
Crop5	72.27	74.11	74.76	**75.82**
Crop6	71.98	72.53	74.04	**75.10**
Crop7	70.57	71.29	72.48	**73.25**
Crop8	65.39	65.96	66.23	**67.40**
Crop9	55.47	56.45	57.12	**58.77**
Crop10	53.31	56.23	56.38	**57.14**
Average	65.58	66.90	67.50	**68.50**

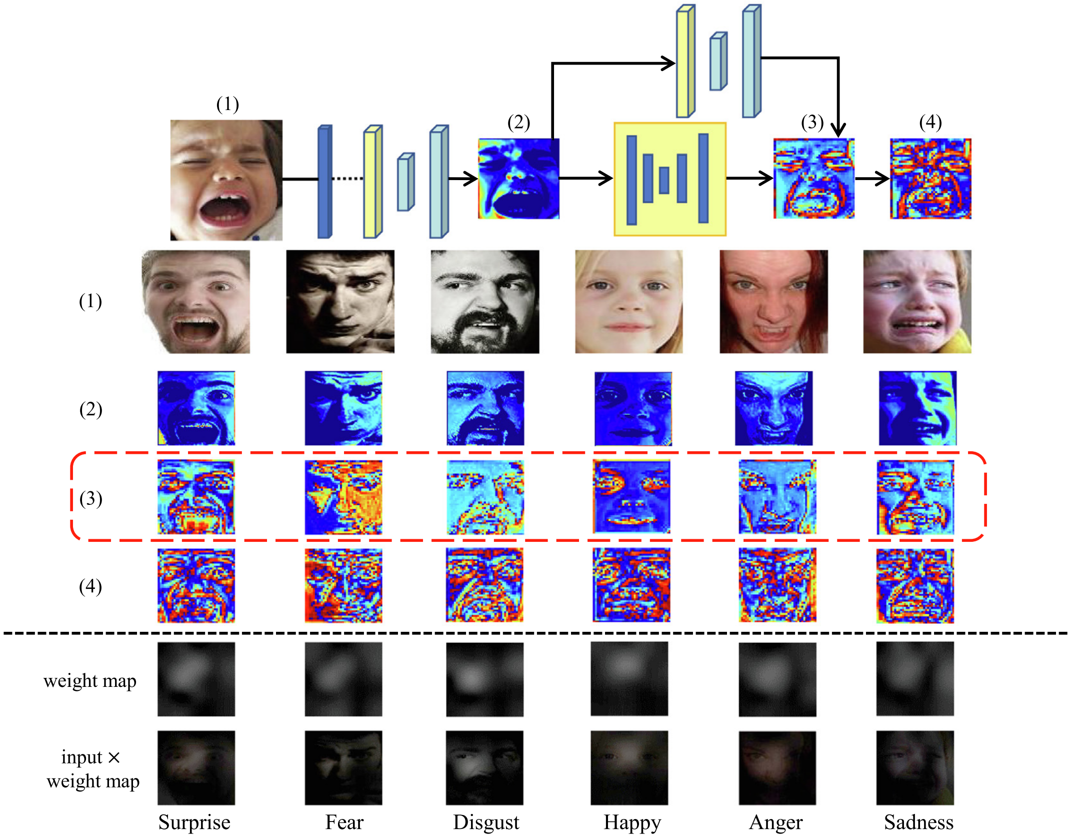


Fig. 6. A visualization of learned deep attention convolution features. The results (the first four rows) are obtained by our LBAB-1. Especially, the third row (in red box) represents the local attention features, which can help to understand how the LBAB pays attention to the important facial areas. The last two lines are reproduced by our implementation based on the reference [22].

distribution in the wild cannot be avoided. Taking the purple points in black dotted box as an example, the first three loss functions cannot separate purple points from others. However, from the visualization result of our islets loss, the obvious distance between purple and other points can be clearly obtained. In addition, the yellow point also acts as an islet, keeping an effective distance from others.

With the help of Emotional FACS [44], we can obtain that the facial expression is one or more action units (AU) motion, such as mouth, noses and eyes. The attention-based method can therefore discover the AU change effectively. As shown in Fig. 6, the Visual Attention [22] (the last two lines) has discovered regions of interests for each type of facial expressions. However, to better locate important facial areas, our LBAN obtains the local attention expression feature through the encoder-decoder network, which is taken as an important information, and combined with the global feature. From the feature maps in Fig. 6 (the first four rows), it can be intuitively observed that the feature map obtained by the attention mechanism can clearly outline multiple organs, and take the internal expression details as the input of the next layer. Especially, the surprise expression (the first column) opens the mouth wide obviously and our LBAB-1 also focuses on the mouth region. Furthermore, our attention operation can observe important eyes and mouth regions in the happy expression (the fourth column).

5. Conclusion

Motivated by the local binary convolution and attention mechanism, in this paper, we proposed the local binary attention network (LBAN) with islets loss for facial expression recognition in the wild. We demonstrated empirically that the LBAN alleviates the influence of excessive sparsity caused by LBC only meanwhile results in a significant reduction of learnable parameters. This provides a good alternative for the large-scale applications of facial expression recognition systems. In addition, the encoder-decoder attention operation can pay more attention to the regions with motions of muscles beneath the skin of face, such as mouth, eyes and nose, providing a choice for the proposed architecture to better extract deep expression features. In order to further enhance the high discrimination of deep representation, we proposed the islets loss combined with the softmax loss to jointly supervise the training process of LBAN. Experimental results on four datasets demonstrated our proposed LBAN-IL have achieved superior performance in the field of deep facial expression recognition.

As we have analyzed about the confusion matrices, it is difficult to identify two kinds of negative expressions, namely the fear and disgust expressions. This is partly due to the uneven distribution of samples and partly due to the high similarity between two expressions and others. In the future, we will consider utilizing metric learning to further learn the facial expression differences and develop a more powerful architecture with better performance on facial expressions recognition in the wild.

CRediT authorship contribution statement

Hangyu Li: Methodology, Software, Writing - original draft. **Nannan Wang:** Conceptualization, Methodology. **Yi Yu:** Supervision, Writing - review & editing. **Xi Yang:** Visualization, Investigation. **Xinbo Gao:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grant 61922066, Grant 61876142, Grant 61772402, Grant 61976166, and Grant 62036007; in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001; in part by the Xidian University Intellifusion Joint Innovation Laboratory of Artificial Intelligence; in part by the Fundamental Research Funds for the Central Universities; and in part by the Innovation Fund of Xidian University.

References

[1] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE, 2010, pp. 94–101.

[2] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans. Image Process. 11 (4) (2002) 467–476.

[3] M. Pietikäinen, A. Hadid, G. Zhao, T. Ahonen, Computer vision using local binary patterns, Vol. 40, Springer Science & Business Media, 2011..

[4] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105..

[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556..

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[8] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[9] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: A report on three machine learning contests, in: International Conference on Neural Information Processing, Springer, 2013, pp. 117–124..

[10] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] Z. Zhang, P. Luo, C.C. Loy, X. Tang, From facial expression recognition to interpersonal relation prediction, Int. J. Comput. Vision 126 (5) (2018) 550–569.

[12] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Trans. Image Process. 28 (1) (2019) 356–370.

[13] J. Shao, Y. Qian, Three convolutional neural network models for facial expression recognition in the wild, Neurocomputing 355 (2019) 82–92.

[14] D.O. Melinte, L. Vladareanu, Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified adam optimizer, Sensors 20 (8) (2020) 2393.

[15] Z. Fei, E. Yang, D.D.-U. Li, S. Butler, W. Ijomah, X. Li, H. Zhou, Deep convolution network based emotion analysis towards mental health care, Neurocomputing 388 (2020) 212–227.

[16] Y. Guo, Y. Xia, J. Wang, H. Yu, R.-C. Chen, Real-time facial affective computing on mobile devices, Sensors 20 (3) (2020) 870.

[17] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 499–515.

[18] J. Cai, Z. Meng, A.S. Khan, Z. Li, J. OReilly, Y. Tong, Island loss for learning discriminative features in facial expression recognition, in: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, IEEE, 2018, pp. 302–309..

[19] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473..

[20] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: Advances in neural information processing systems, 2014, pp. 2204–2212..

[21] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[22] W. Sun, H. Zhao, Z. Jin, A visual attention based roi detection method for facial expression recognition, Neurocomputing 296 (2018) 12–22.

[23] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using cnn with attention mechanism, IEEE Trans. Image Process. 28 (5) (2019) 2439–2450.

[24] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, IEEE Trans. Image Process. 29 (2020) 4057–4069.

[25] F. Juefei-Xu, V.N. Boddeti, M. Savvides, Local binary convolutional neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4284–4293..

[26] H. Larochelle, G.E. Hinton, Learning to combine foveal glimpses with a third-order boltzmann machine, in: Advances in neural information processing systems, 2010, pp. 1243–1251..

[27] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, 2015, pp. 423–426..

[28] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.

[29] W. Deng, J. Hu, S. Zhang, J. Guo, Deepemo: Real-world facial expression analysis via deep learning, in: Visual Communications and Image Processing (VCIP), 2015, IEEE, 2015, pp. 1–4..

[30] Z. Liu, S. Li, W. Deng, Boosting-poof: Boosting part based one vs one feature for facial expression recognition in the wild, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 967–972..

[31] W. Sun, H. Zhao, Z. Jin, An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks, Neurocomputing 267 (2017) 385–395.

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

[33] R.T. Ionescu, M. Popescu, C. Grozea, Local learning to improve bag of visual words model for facial expression recognition, in: Workshop on challenges in representation learning, ICML, 2013..

[34] P. Giannopoulos, I. Perikos, I. Hatzilygeroudis, Deep learning approaches for facial emotion recognition: A case study on fer-2013, in: Advances in Hybridization of Intelligent Methods, Springer, 2018, pp. 1–16.

[35] S. Minaee, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, arXiv preprint arXiv:1902.01019..

[36] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, F. Jurie, An occam's razor view on learning audiovisual emotion recognition with small training sets, in, in: Proceedings of the 2018 on International Conference on Multimodal Interaction, ACM, 2018, pp. 589–593.

[37] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, Y. Rui, Label distribution learning on auxiliary label space graphs for facial expression recognition, in, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13984–13993.

[38] B.-K. Kim, H. Lee, J. Roh, S.-Y. Lee, Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition, in, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 427–434.

[39] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: 2016 IEEE winter conference on applications of computer vision (WACV), IEEE, 2016, pp. 1–10..

[40] M.-I. Georgescu, R.T. Ionescu, M. Popescu, Local learning with deep and handcrafted features for facial expression recognition, arXiv preprint arXiv:1804.10892..

[41] Y. Tang, Deep learning using linear support vector machines, arXiv preprint arXiv:1306.0239..

[42] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, D. Tao, Deep neural networks with relativity learning for facial expression recognition, in: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2016, pp. 1–6.

[43] L.v.d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605..

[44] W.V. Friesen, P. Ekman, et al., Emfacs-7: Emotional facial action coding system, Unpublished manuscript, University of California at San Francisco 2 (36) (1983) 1.

Hangyu Li received the B. Sc degree in electronic and information engineering from Shandong University, Jinan, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Telecommunications Engineering, Xidian University, Xi'an, China. His current research interests include computer vision, machine learning, and affective computing.

Nannan Wang received the B. Sc degree in information and computation science from the Xi'an University of Posts and Telecommunications in 2009 and the Ph.D. degree in information and telecommunications engineering from Xidian University in 2015. From September 2011 to September 2013, he was a Visiting Ph.D. Student with the University of Technology, Sydney, NSW, Australia. He is currently a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published over 100 articles in refereed journals and proceedings, including IEEE T-PAMI, IJCV, NeurIPS, ECCV etc. His current research interests include computer vision, pattern recognition, and machine learning.

Yi Yu is currently an assistant professor with National Institute of Informatics (NII), Japan. Before joining NII, she was a senior research fellow with School of Computing, National University of Singapore. Her research covers large-scale multimedia data mining and pattern analysis, location-based mobile media service and social media analysis. Yu received a Ph.D. in Information and Computer Science from Nara Women's University, Japan.

Xi Yang received the B.Eng. degree in electronic information engineering, and the Ph.D. degree in pattern recognition and intelligence systems from Xidian University, Xi'an, China, in 2010 and 2015, respectively. From 2013 to 2014, she was a visiting Ph.D. student with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX, USA. In 2015, she joined the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, where she is currently an Associate Professor in communications and information systems. Her current research interests include image/video processing, computer vision, and multimedia information retrieval. She has published over 10 papers in refereed journals and proceedings, including the IEEE T-NNLS, T-IP, information sciences, and optics express.

Xinbo Gao received the B.Eng., M.Sc. and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education of P. R. China, a Professor of Pattern Recognition and Intelligent System of Xidian University and a Professor of Computer Science and Technology of Chongqing University of Posts and Telecommunications. His current research interests include Image processing, computer vision, multimedia analysis, machine learning and pattern recognition. He has published six books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and a Fellow of the Chinese Institute of Electronics.