Applied Intelligence (2021) 51:549–559
https://doi.org/10.1007/s10489-020-01855-5

E-FCNN for tiny facial expression recognition

Jie Shao[1] · **Qiyu Cheng**[1]

Published online: 20 August 2020

Abstract
As a hot issue in recent years, facial expression recognition(FER) has been widely applied in many fields, but it still faces great challenges in tiny facial expression recognition. Currently, most of the FER networks only consider images of ideal sizes. Their recognition accuracy would significantly decrease as the image resolution decreases. However, images captured by surveillance cameras often contain tiny faces with low-resolution. This paper proposes an edge-aware feedback convolutional neural network(E-FCNN) for tiny FER, which associates image super-resolution and facial expression recognition together. To effectively leverage the texture information of faces, we design a novel three-stream super-resolution network, which is embedded with an edge-enhancement block as one branch. The other two branches are the up-sampling branch and SR(Super-Resolution) primary branch. Specifically, visual features are extracted from tiny images based on a hierarchical strategy, and then put into a feedback block with fused results of the three branches. Experiments are performed on down-sampled images in four facial expression datasets: CK+, FER2013, BU-3DFE, RAF-DB. The results demonstrate the favorable performance of our network.

Keywords Facial expression recognition · Super-resolution · Tiny face · Residual network · Edge extraction · Feedback networks

1 Introduction

Facial expression recognition has been widely used in human-computer interaction, communication field, robot manufacturing [1], transportation [2], facial nerve analysis, clinical psychology [3], and so on. However, none of the remaining open challenges is recognizing expressions from tiny faces. Facial images with Normal resolution (at least 48×48 pixels) are fundamental properties of almost all current facial expression recognition systems. The most advanced facial expression datasets have facial images no less than normal size, such as 48×48px in CK+ and 256×256px in BU-3DFE. But from a practical perspective, sensors would capture faces of all kinds of scales. Faces less than 40px tall often appear in surveillance videos. Many recent works in facial expression recognition made use of scale normalization as preprocessing. For example, Mollahosseini et al. [4] resized the input images to 48×48px. Huang et al. [5] resized the input images to 224×224px size for their Densely Connected Convolutional Networks. However, the sharpnesses of input images would be blurry and the facial expression details would be lost with size normalization. Furthermore, if the original images are too tiny, the performances of traditional FER methods would be unsatisfactory after image magnification.

Super-resolution algorithms are used to restore large-sized HR(high-resolution) images from small-sized LR(low-resolution) images, so they are suitable for enlarging tiny faces. Most recent deep learning-based super-resolution algorithms would only include a single-cycle CNN(convolutional neural network) such as [6], [7], [8]. As the network depth increases, the number of parameters increases as well. Large-capacity networks will consume huge storage resources and face overfitting problems. Therefore, we choose to add multiple cycles from the horizontal cycle instead of deepening the network.

This paper proposes a new tiny FER method called edge-aware feedback convolutional neural network(E-FCNN). The overview of our recognition pipeline is shown in Fig. 1.

✉ Qiyu Cheng
chengqy@mail.shiep.edu.cn

Jie Shao
shaojie@shiep.edu.cn

[1] Department of Electronic and Information Engineering, Shanghai University of Electric Power, Shanghai 200120, China

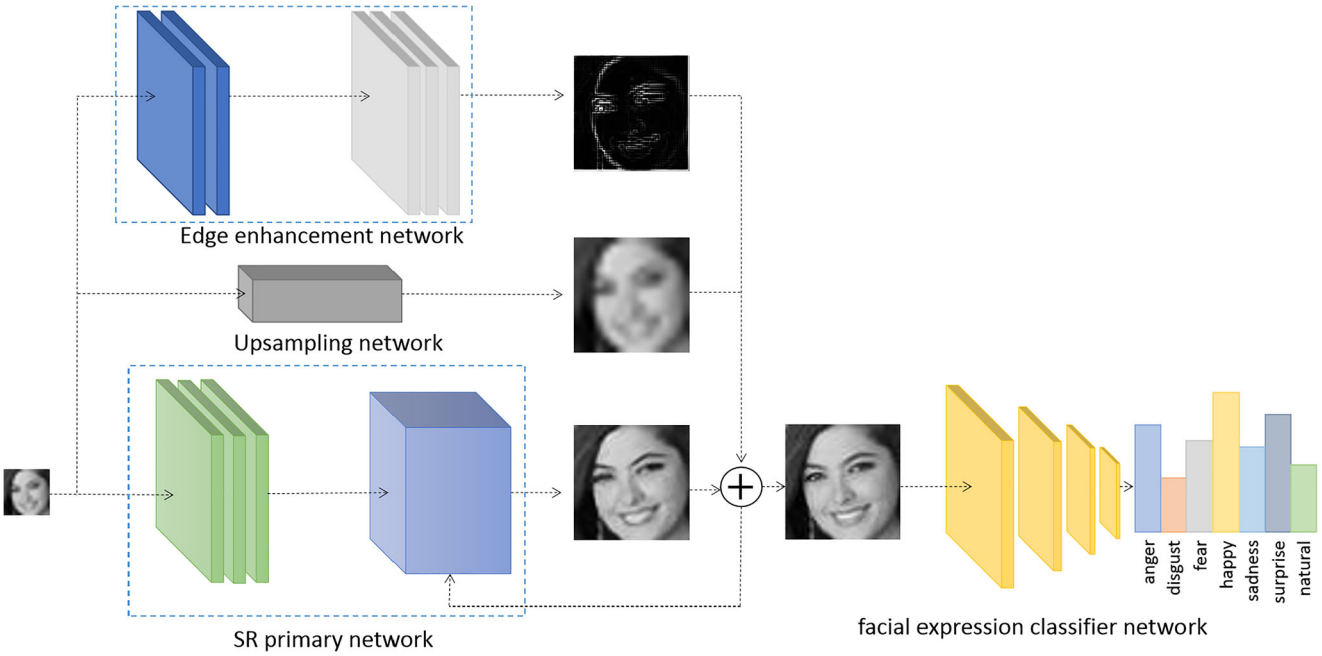


Fig. 1 The overview of our recognition pipeline with E-FCNN

Small faces are enlarged with the super-resolution strategy, and then recognized by a facial expression classifier.

First of all, we introduce the feedback mechanism from the super-resolution feedback network(SRFBN) [9]. Then we insert residual block [8] between the LR feature extraction block and the feedback block to build a hierarchical feature extraction structure. They constitute the basic structure of our image SR(super-resolution) network, namely the SR primary network. After comprehensively analyzing current FER methods, we find that although edge information accounts for no more than 20% in facial images, it is the most effective information in facial expression recognition. Therefore, we embeded an edge enhancement block to strengthen the sharpness in face images. Besides, an up-sampling block is exploited to enlarge the face images as well, which complements low-frequency information. The super-resolution results are fused by all the above multiple sources, and then repeatedly fed to the feedback block. The SR images are obtained after several iterations, and taken as the input of facial expression classifier.

In summary, our main contributions are as follows:

- A new end-to-end network structure is proposed for tiny FER, which effectively encapsulates multiple facial attributes with a feedback mechanism.
- Texture details are taken into account in our facial super-resolution(FSR) network for the first time. An edge enhancement module is designed, which can efficiently promote differentiation of tiny facial expression.
- The improved residual blocks are leveraged to extract deep features, without increasing computational cost in inference.
- Extensive experiments are performed on several challenging datasets for tiny FER. Our E-FCNN achieves favorable performances.

2 Related work

2.1 Facial expression recognition

The development of machine learning and the advent of deep learning have significantly improved the research of FER. It has the following three aspects. Firstly, the networks become deeper and deeper in order to deal with more complex classifications. For example, Zhang et al. [10] proposed a DNN (depth neural network) with SIFT(Scale Invariant Feature Transform) features and experimented on the BU-3DFE dataset. Secondly, local features attract more attention. To this end, attention mechanisms are introduced in many networks. For example, Sharma et al. [11] proposed a facial expression action recognition network with the visual attention mechanism. Finally, facial expression recognition in the natural state has also made some progress. Shao et al. [12] proposed three CNN facial expression recognition networks for multiple uncertain interference factors in the wild environment. However, most researches deal with facial images of standard sizes. Their performances may degrade in processing tiny faces. Cheng

et. al. [13] developed a novel framework embedded with a decoder for low-resolution facial expression recognition in videos which is a fresh try for LR facial expression recognition.

2.2 Super-resolution

Super-resolution techniques are usually explored to generate HR images from LR face images, which is normally called face hallucination. Chen et al. [14] proposed an SR network related to facial prior information. They constructed a super-resolution network to get a rough HR image, and then put the image features obtained by the fine SR encoder and the facial landmarks in the prior information into the SR decoder. Dogan et al. [15] upsampled the image by combining the LR features with information from other images of the same face. They tried to reduce the ambiguity in FSR by additional information. However, instead of FSR, our interests are the accuracy of FER. As a result, we also refer to a lot of super-resolution works in other fields. For example, Zhang et al. [16] proposed an RDN(Residual Dense Network) network which associated global residual connections with dense skip connections. However, dense-skip-connections was a bottom-up approach. That was to say, the lower levels only accepted information from previous levels, which led to a small receptive field and lack of contextual information. To resolve this issue, Li et al. [9] introduced a super-resolution network in the form of feedback, namely SRFBN (super-resolution feedback network). High-level information is put into the low-level through a feedback module to correct low-level information so that sufficient context information could be obtained. We are inspired by the feedback concept. Nevertheless, we find that recognition accuracy is not as good as we expect. Hence edge detection strategy is considered to strengthen the texture information of the face.

2.3 Edge detection

In facial expression recognition, texture information is always leveraged as an important feature. In recent years, edge detection algorithms have been greatly improved by deep learning. Yang et al. [17] argued that the edge information of high-resolution images could be inferred by combining low-resolution images and their edge information. Thus, they proposed a recursive residual learning method to reduce the loss of edge information during amplification, so that the texture information in low-resolution images was enriched and the similarities between HR and SR images were increased. Some works have tried to add texture related information to facial super-resolution. Chen et al. [18] speculated that adding face priors to the network could effectively reduce information loss in facial super-resolution networks. Therefore, they used geometric priors (landmark heat maps and analytical maps) in the facial super-resolution network. Liu et al. [19] proposed a novel multi-cycle CNN. Each cycle extracted different features and then fused them to minimize information loss. Experiments showed that the edge information extracted by the proposed edge extraction branches had a great promotion effect on the super-resolution effect. Inspired by [19], we added an edge enhancement block to the super-resolution network. It is composed of an edge extraction block [20] and an enhancement block. The specific introduction is in Section 3.4.

3 The proposed method

The main structure of our edge-aware feedback convolutional neural network is shown in Fig. 2. It consists of four sub-networks, namely the SR primary network, the upsampling network, the edge enhancement network, and the facial expression recognition network. The main function of

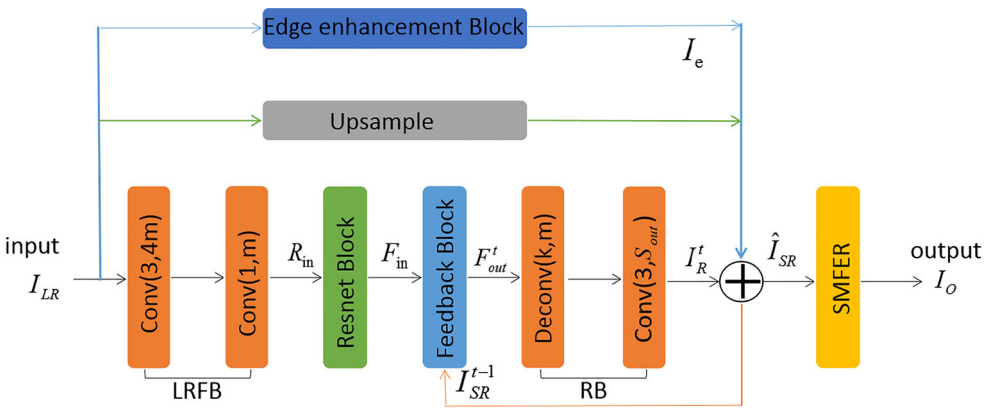


Fig. 2 The architecture of edge-aware feedback convolutional neural network(E-FCNN)

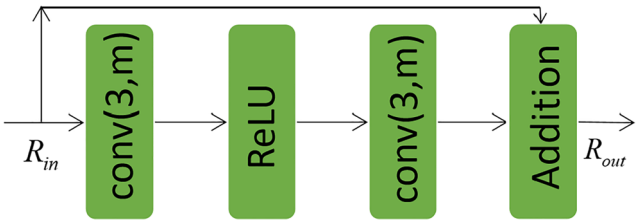


Fig. 3 The Resnet block(ResB)

the SR primary network is to reduce the impact of image loss and increase the resolution of the image while enlarging the input image. The up-sampling network enlarges the original low-resolution image by interpolation. The edge enhancement network first extracts the edge features of the LR image, and then uses SRCNN [7] to enlarge the texture features to enhance the texture details. The outputs of the SR primary network, the up-sampling network and the edge enhancement network are weighted and summarized afterward. Then the weighted sum is reversely put into the feedback block in the SR primary network for T iterations. Finally, the SR result is classified by the FER network, which is designed based on the structure of facial expression recognition using Saliency Maps (SMFER) [21].

3.1 The SR primary network

The SR primary network contains four parts: low-resolution feature extraction block (LRFB), resnet block (ResB), feedback block (FB) and reconstruction block (RB).

As shown in Fig. 2, the $Conv(s, n)$ and $Deconv(s, n)$ stand for the convolution layer and deconvolution layer respectively, the s and n in the formula are the size and number of filters respectively.

First, we put I_{LR} into the LRFB to obtain the shallow features of LR images R_{in}. As we can see in the figure,the LRFB consists of $Conv(3, 4m)$ and $Conv(1, m)$, where m refers to the base number of filters.

$$R_{in} = f_{LRFB}(I_{LR}) \tag{1}$$

$$F_{in} = f_{Res}(R_{in}) \tag{2}$$

Where f_{LRFB} denotes the operations of the LRFB. Its output R_{in} is then put into the resnet block to get deep features. F_{in} represents the output of the resnet block, and it is also used as the input of the feedback block. The output of the feedback block at the t-th cycle can be expressed as:

$$F_{out}^{t} = f_{FB}(I_{SR}^{t-1}, F_{in}) \tag{3}$$

where f_{FB} represents the process of feedback block. We will give specific implementation details in Section 3.3, the specific feedback network is shown in Fig. 4. I_{SR}^{t-1} refers to the super-resolution result of the previous iteration. The iteration procedure runs for T times. The ultimate SR output is represented as $\hat{I}_{SR}$:

$$\hat{I}_{SR} = \alpha I_e + \beta f_{up}(I_{LR}) + \gamma I_R^T \tag{4}$$

where I_e is the SR result after edge detection and enhancement on the tiny image. α, β, γ refer to the weighting factors. Their acquisition method will be given in Section 4. f_{up} denotes the up-sampling process. I_R^T denotes the reconstructed image after T iterations. The reconstruction block(RB) consists of $Deconv(k, m)$ and $Conv(3, s_{out})$. The $Deconv(k, m)$ upscales the LR features F_{out}^{t} to HR features. The $Conv(3, s_{out})$ generates the residual image I_R^T.

$$I_R^t = f_{RB}(F_{out}^{t}) \tag{5}$$

Where f_{RB} refers to the reconstruction function. Finally, the output of this E-FCNN for tiny facial expression recognition network can be expressed as:

$$I_o = f_{SMFER}(\hat{I}_{SR}) \tag{6}$$

where f_{SMFER} refers to the facial expression network [21].

3.2 Resnet block

We exploit the improved resnet block mentioned in the EDSR [8] before the feedback block, shown in Fig. 3.

$$R_{out} = f_{Res}(R_{in}) \tag{7}$$

where R_{in} and R_{out} are the inputs and outputs of the network, respectively. The resnet block is also the input to the feedback block. f_{Res} denotes the residual block.

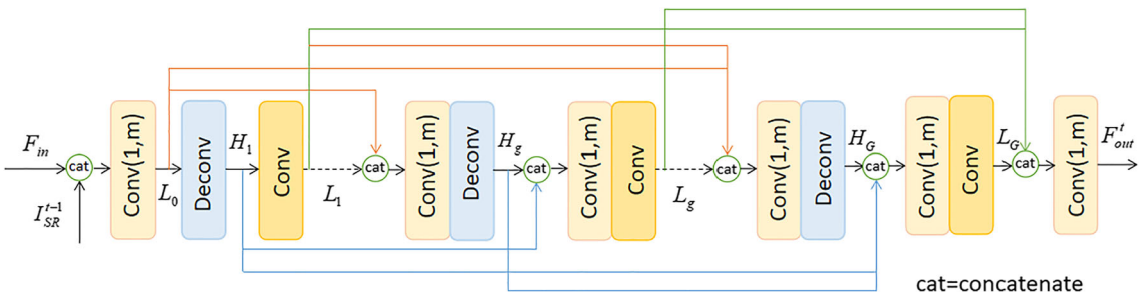


Fig. 4 The Feedback block(FB)

3.3 Feedback block

Figure 4 shows the structure of the feedback block. During the iterations, FB receives feedback information I_{SR}^{t-1} to improve the shallow information in F_{in}. Then F_{out}^{t} is reconstructed after an iteration. Each projection group mainly includes an upsampling operation and a downsampling operation, which can project HR features onto LR features.

At the beginning of FB, $Conv(1, m)$ refines the input feature F_{in} through the feedback information I_{SR}^{t-1}, thereby it can generate a re-refined input feature L_0:

$$L_0 = C_0([I_{SR}^{t-1}, F_{in}]) \tag{8}$$

where C_0 refers to the initial concatenation network, $[I_{SR}^{t-1}, F_{in}]$ means to concatenate I_{SR}^{t-1} and F_{in}. H_g and L_g refer to the feature map of HR and LR obtained after the g-th projection in FB, which can be obtained by the following methods:

$$H_g = C_g^{\uparrow}([L_0, L_1, ..., L_{g-1}]) \tag{9}$$

$$L_g = C_g^{\downarrow}([H_1, H_2, ..., H_g]) \tag{10}$$

where $C_g^{\uparrow}$ and $C_g^{\downarrow}$ refer to the upsample and downsample networks that repeatedly use $Deconv(k, m)$ and $Conv(k, m)$ at the g-th group. In addition to the first $Deconv(k, m)$ and $Conv(k, m)$, we add $Conv(1, m)$ before $C_g^{\uparrow}$ and $C_g^{\downarrow}$ in order to improve the calculation efficiency.

To take full advantage of the useful information of each $Deconv(k, m)$ and $Conv(k, m)$ group and the results of this iteration for the next iteration, we perform feature-fusion on the LR features generated by each group (green arrow in Fig. 5) to get the output of FB:

$$F_{out}^{t} = C_{FF}([L_1, L_2, ..., L_G]) \tag{11}$$

where C_{FF} represents the last $Conv(1, m)$.

3.4 Edge enhancement block

In Fig. 5, the first layer of EHB is a 3×3×m convolution. Its output is the input to two branches. One branch is a $Conv(3, m)$ layer following with a $Conv(1, m)$ layer, and the other branch is simply a $Conv(1, m)$ layer. Two edge maps e_1 and e_2 are extracted from two branches. Then they are fused into e_0. The fusion weights are updated during the training process.

$$e_o = \sum(h_1 e_1, h_2 e_2) \tag{12}$$

Where h_1 and h_2 are the fusion weights. Afterward, SRCNN [7] is applied to e_0 to implement edge enhancement and image magnification with super-resolution.

$$I_e = f_{SRCNN}(e_o) \tag{13}$$

Where I_e is the output of edge enhancement block, and f_{SRCNN} denotes the function of SRCNN.

3.5 Learning Strategy

The loss of our proposed network is:

$$L = \lambda_{sr} L_{sr} + \lambda_{fer} L_{fer} \tag{14}$$

where L_{sr} refers to the loss of the super-resolution network, L_{fer} refers to the loss of facial expression recognition network. We choose cross-entropy as the loss function. λ_{fer} and λ_{sr} refer to the regularization parameters. $L1$ regularization is applied to optimize our super-resolution network. So it can be formulated with:

$$L_{sr} = L_{sr1} + \alpha L_{sr2} \tag{15}$$

$$L_{sr1}(\Theta) = \frac{1}{T} \sum W ||I_{HR} - I_{SR}||_1 \tag{16}$$

$$L_{sr2} = Dist(e, e_t) \tag{17}$$

where α has the same value as Eq. 4, L_{sr1} refers to the loss of the SR primary network and the up-sampling network. L_{sr2} refers to the loss of edge-enhance network. T is the iteration time referring to the red arrow in Fig. 2.

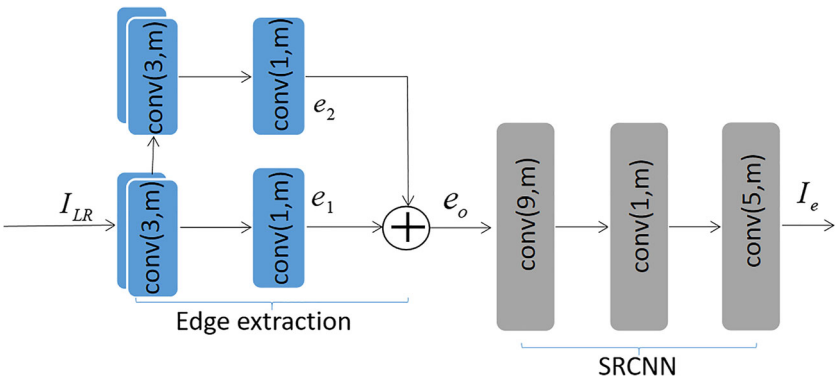


Fig. 5 The Edge enhancement block(EHB). The blue module is the edge extraction module and the gray is the edge enhancement module. We use SRCNN to perform super-resolution enhancement on the edge extraction results

Θ indicates the parameter of the network. I_{SR} refers to the super-resolution image obtained after T iterations in the feedback network, I_{HR} represents the reference image with high-resolution during training. $Dist$ is the difference between the edge enhancement image ideal value(e) and the actual value of the training(e_t), which is expressed as the cross-entropy loss in the network.

4 Experimental results

4.1 Databases and protocols

We use four publicly available expression recognition databases to evaluate the performance of our algorithm. Some samples in the datasets are presented in Fig. 6, which contains seven expressions of CK+, FER2013, BU-3DFE, RAF-DB. All experimental data are down-sampled using bicubic interpolation.

CK+ database: The Extended Cohn-Kanade (CK+) database [22] collects facial expressions of 123 subjects by videos, and a total of 593 facial expression video sequences. The subjects are young people between ages of 18 and 30. It is particularly emphasized that the database acquisition environment is laboratory-controlled. We extract the last three frames of each sequence from these video sequences and obtain 981 facial expression images. Since these images are not only face-images, we crop them to 48×48px only face images. We use 10-fold Cross-validation to allocate the test set and training set.

Fer2013 database: The FER2013 database comes from a facial expression recognition challenge held on Kaggle [23, 24]. The face images in this database are grayscale, and the image size is 48×48px. It has 28709 training images, 3589 validation images, and 3589 testing images. These images are divided into seven categories, which express different emotions: anger, disgust, fear, happy, sad, surprise and natural. Furthermore, these face images source from the Internet with noises and relatively low quality.

BU-3DFE database: The BU-3DFE database [25] is a database containing three-dimensional models of faces and two-dimensional models of faces. It collects 100 faces from different races, including 56 females and 44 males. 4 images about basic expressions and 1 image about natural expression per person, totaling 2500 images. We take the same Cross-validation of this database and the CK+ database, and choose 2250 pictures as the training set and 250 pictures as the test set. We select the part of the 2D faces in the database. The size of images in the database is 512×512 pixels.

RAF-DB database: The Real-world Affective Faces (RAF-DB) database [26] is a database of about 30,000 facial images downloaded from the Internet. It contains basic emotions database and compound emotions database, and each picture is individually marked by 40 annotators. The basic emotions database contains 15339 facial images in 7 kinds of expressions. We select the basic emotions database as the training object, of which 3068 are the test set and the rest are the training set.

We run our network in a Pytorch3.6 environment with NVIDIA 2080 GPU for training. The batch size of all networks during training is 16. The Adam optimizer is used in the network to optimize the network parameters with an initial learning rate of 0.001, and for every 200 epochs the

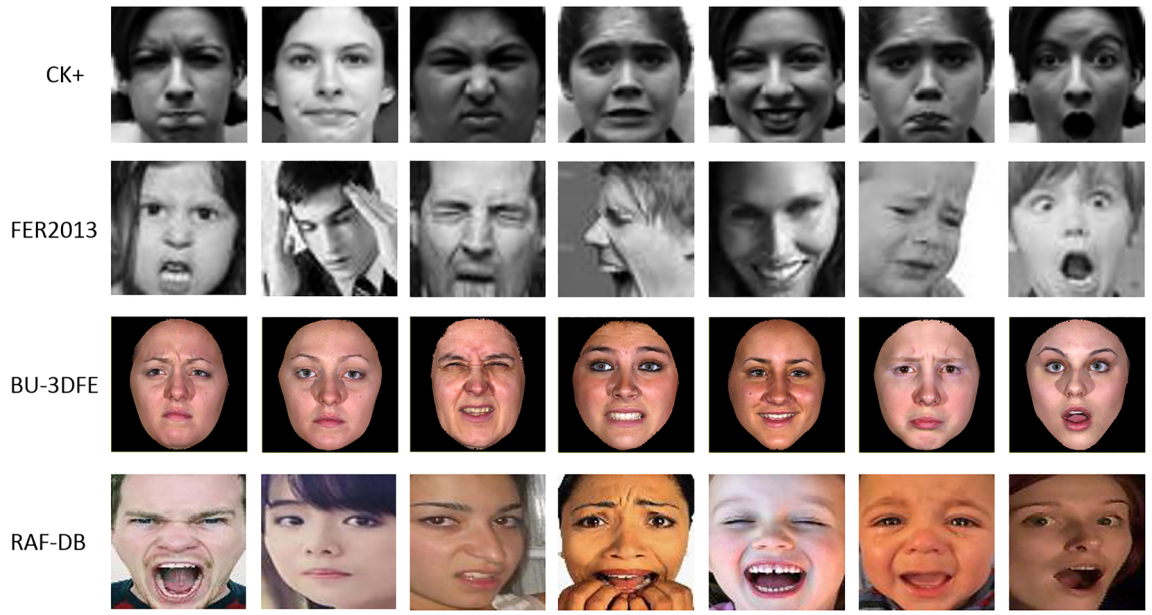


Fig. 6 Image samples in CK+ , FER2013, BU-3DFE, RAF-DB Databases

learning rate is multiplied by 0.5. The average test time of a single image is 0.017 seconds on the condition that the input images are 16 px tall. For small datasets like CK+, our training process includes 100 epochs and the training time is 1400 seconds.

4.2 Discussions on network components

In order to assess the role of each module in our network, ablation experiments are performed on CK+, FER2013, and BU-3DFE respectively. The baseline is called 16FER by reducing some layers from SMFER [21]. To analyze the influence of the feedback network, the resnet block and the edge enhancement block, Figure 7 compares the baseline with methods by adding SRFBN(feedback network), RS(resnet block + SRFBN), ERS(edge extraction block + RS) to it, and our E-FCNN(ERS + edge enhancement block).

All the methods are fed with 16 px tall images. The baseline has the worst performances while testing all three datasets. SRFBN [9] improves the accuracy of facial expression recognition after increasing image resolution. Comparatively, adding resnet block is helpful as RS always outperforms SRFBN. Furthermore, the performance of ERS is not good enough since the edges are extracted without SR enhancement. FER2013 database has more noise and the images are too blurry, so the effect of edge extraction is not ideal. From the experimental results of the CK+ database and the BU-3DFE database, E-FCNN has the best recognition accuracy, because edge enhancement plays an important role in promoting FER. In the FER2013 database, as the original image resolution is much lower than the others, it challenges the edge feature extraction. Therefore, the edge enhancement module doesn't work. In short, based on the ablation analysis, the resnet block and the edge enhancement module both play promoting roles, and their effects are related to the ambiguity of the images and their amplification expectations.

4.3 Discussions on coefficients

There are three hyperparameters α, β, γ in (4), which are used to determine the impact of the SR primary network, the up-sampling network and the edge enhancement network in E-FCNN. We perform experiments on the BU-3DFE database. We use the undetermined coefficient method to keep $\gamma = 1$ constant, and change α and β to observe the FER accuracy. Firstly we set $\alpha = 1$ and gradually increase the value of β. The result is shown in Fig. 8. We can see that when $\beta = 1$, the accuracy of expression recognition is the highest. Then we make $\beta = 1$, and change the value of α. The result is shown in Fig. 9. We can see that the optimal value is reached when $\alpha = 0.95$. In summary, we roughly have the optimal values at $\alpha = 0.95, \beta = 1, \gamma = 1$.

4.4 Algorithm self-evaluation

In order to get a clearer result of the impact of our network on the recognition of each expression, the confusion matrices of our methods on CK+, FER2013 and BU-3DFE databases are shown in Fig. 10. From these three confusion matrices, it can be seen that the accuracies of happy and surprise are higher than the others as the facial movements of these two expressions are larger.

4.5 Facial super-resolution comparison

High performance of facial super-resolution could be helpful in tiny facial expression recognition, so we cut the recognition part of the model, and provide the comparison results with the state-of-the-art facial super-resolution algorithms. The experiments are performed in the case of CelebA database, and the PSNR(Peak Signal to Noise Ratio) and SSIM(structural similarity) results obtained by the magnification of 8 times are shown in Table 1.

SRCNN [7] is the first algorithm in image super-resolution reconstruction. Its PSNR and SSIM are 24.56dB

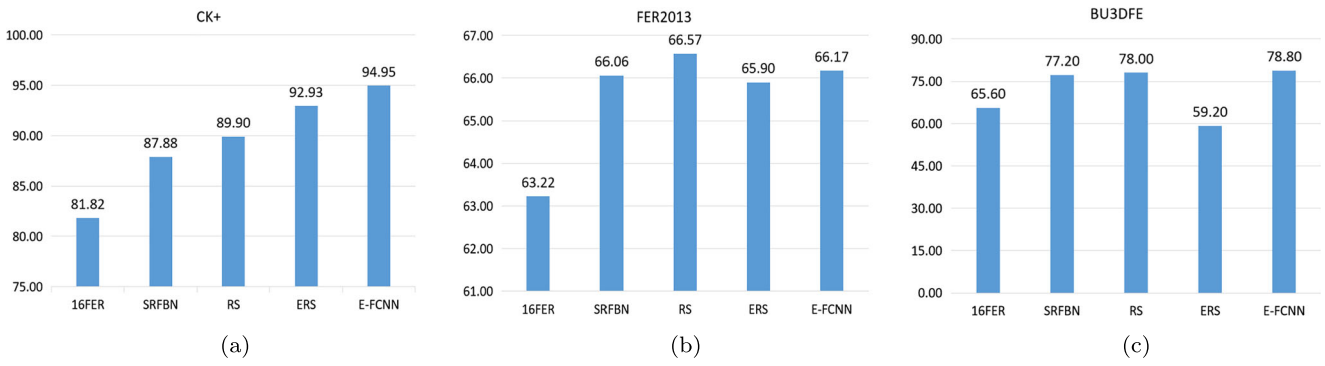


Fig. 7 Experimental results on CK+, FER2013 and BU-3DFE databases for network component analysis

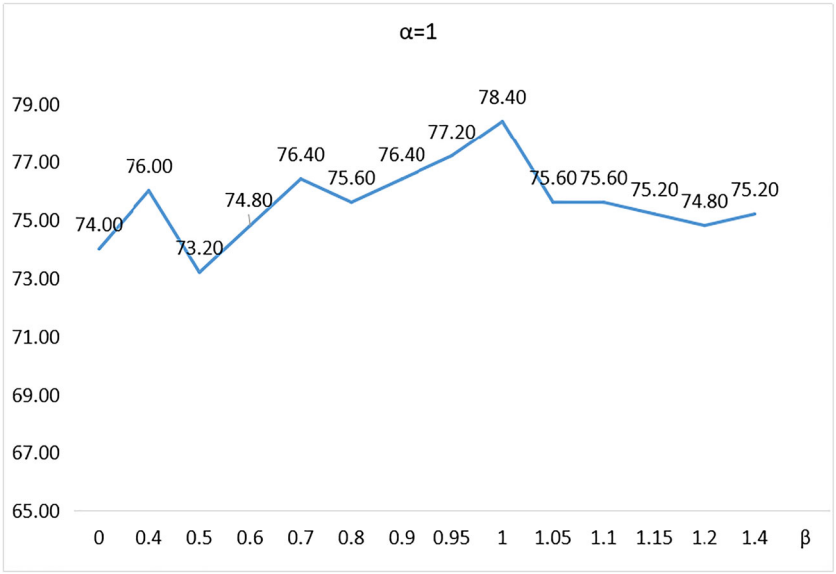


Fig. 8 When $\alpha = 1$ change β,the change of facial expression recognition accuracy

and 0.66. EDSR [8] improves the Resnet block to make the network more flexible, and expands the size of the model to obtain better performance. Its PSNR result is 25.13dB. Yu and Porikli [27] integrates facial structural information into an SR network with two branches. One is a facial super-resolution network, and the other is used to generate landmark heatmaps to improve facial details. Yu et al. [28] is an SR network including facial auxiliary information. Distilled FAN [29] uses an asymptotic method, divides the network into consecutive steps for training, and uses the attention mechanism and facial landmarks in the network at the same time. Our network outperforms all the above algorithms in both PSNR and SSIM comparisons.

4.6 Facial expression recognition comparison

To evaluate the performance of our algorithm in FER, we compare our algorithm with other state-of-the-art algorithms including SVM [30], FIL-CNN [31], Frame [32], GoogleNet [33], ResNet-18 [34] and SMFER [21]. All the algorithms are fed with 16×16px size images. Bicubic interpolation is explored as the preprocessing method to resize the images for the compared algorithms. We train our E-FCNN on CK+, FER2013 and BU-3DFE respectively. The comparison results are shown in Tables 2, 3 and 4. We achieve the accuracy of 94.95%, 66.17% and 78.80% respectively, which outperforms all the other FER methods.

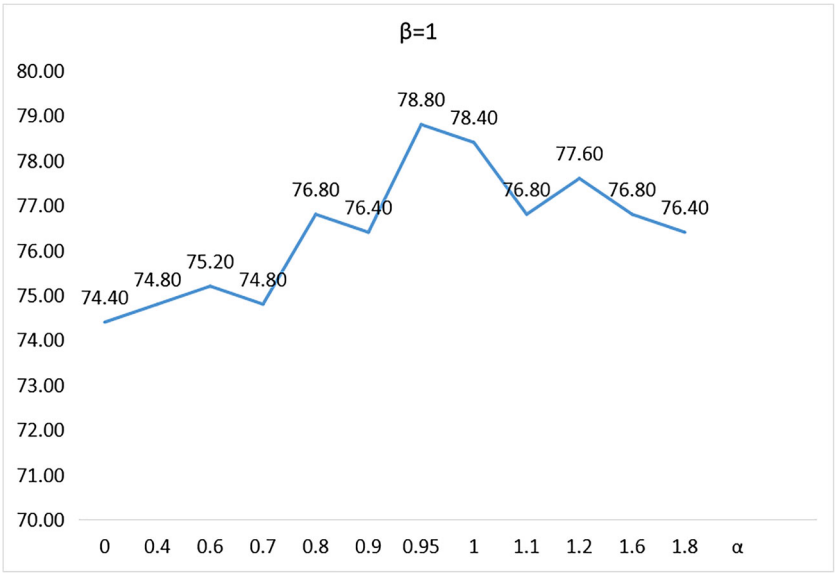
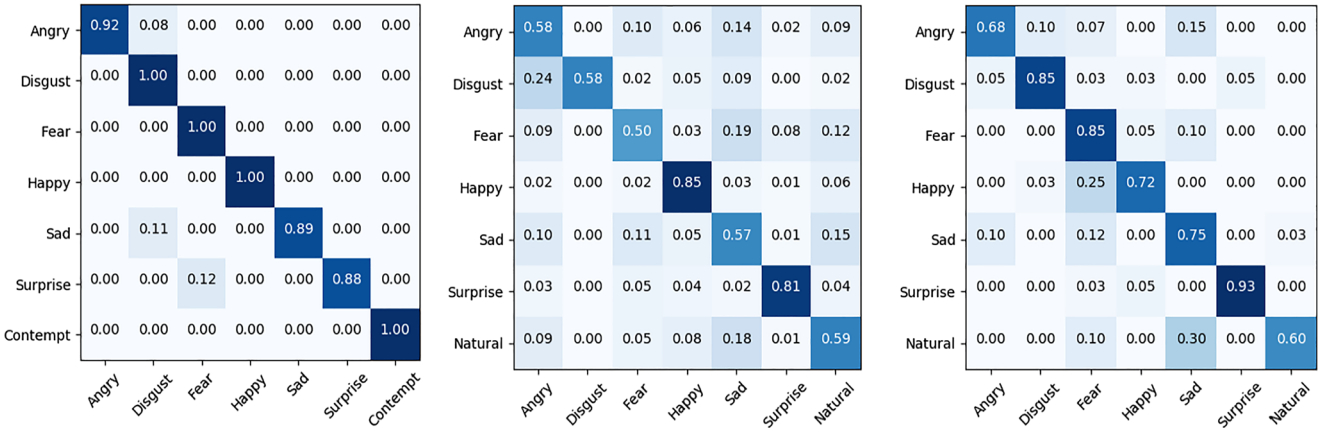


Fig. 9 When $\beta = 1$ change α,the change of facial expression recognition accuracy



(a) Confusion matrix results for CK+ database

(b) Confusion matrix results for FER2013 database

(c) Confusion matrix results for BU-3DFE database

Fig. 10 Confusion matrices of E-FCNN on three expression databases

Table 1 Facial super-resolution comparison (using PSNR / SSIM judgment)

Methods	PSNR(dB)	SSIM
Bicubic	21.53	0.59
SRCNN [7]	24.26	0.66
EDSR [8]	25.13	0.69
Yu and Porikli [27]	23.14	0.68
Yu et al. [28]	21.82	0.62
Distilled FAN [29]	22.66	0.69
Our E-FCNN	**26.02**	**0.76**

Table 2 Facial expression recognition comparison on CK+ database

Methods	Accuracy(%)
SVM [30]	82.70
FIL-CNN [31]	89.98
Frame [32]	84.07
GoogLeNet [33]	90.91
ResNet-18 [34]	82.83
SMFER [21]	82.83
Our E-FCNN	**94.95**

Table 3 Facial expression recognition comparison on FER2013 database

Methods	Accuracy(%)
SVM [30]	44.40
FIL-CNN [31]	55.60
Frame [32]	47.19
GoogLeNet [33]	64.17
ResNet-18 [34]	65.51
SMFER [21]	65.12
Our E-FCNN	**66.17**

Table 4 Facial expression recognition comparison on BU-3DFE database

Methods	Accuracy(%)
SVM [30]	58.80
FIL-CNN [31]	47.60
Frame [32]	63.13
GoogLeNet [33]	74.40
ResNet-18 [34]	74.80
SMFER [21]	76.40
Our E-FCNN	**78.80**

Table 5 Low-resolution facial expression recognition comparison on RAF-DB database

Methods	Scale	Accuracy(%)
FSR-FER [35]	×2	83.21
Our E-FCNN	×2	**84.62**
FSR-FER [35]	×3	81.23
IFSL(SVM) [36]	-	76.90
IFSL(k-NN) [36]	-	72.60
Our E-FCNN	×3	**81.68**
FSR-FER [35]	×4	78.03
Our E-FCNN	×4	**79.04**

4.7 Low-resolution facial expression recognition comparison

Table 5 lists the comparison results between low-resolution facial expression recognition methods and ours. All the experiments are conducted using images from the RAF-DB dataset. FSR-FER [35] proposes a multi-scale super-resolution expression recognition method. IFSL [36] proposes a method to deal with low-resolution facial expression recognition by image filter based subspace learning. Both algorithms use 100*100 px sized images as original data, and down sample them as inputs of their algorithms. As we can see from Table 5, we enlarge the images to ×2, ×3, and ×4 sizes of the inputs, the results show that our performance is superior than the others in all conditions.

5 Conclusion and future works

We propose an edge-aware feedback convolutional neural network(E-FCNN) for tiny facial expression recognition, and evaluate its performance using different analytical methods. Experiments demonstrate that the accuracy of our proposed network can maintain a good level when the input image size is small. However, images after SR reconstruction still look blurry to human eyes, even though the recognition accuracy has achieved a good performance. Therefore, in subsequent research, we will continue to improve our network and strengthen the role of edge enhancement in the network, so as to improve the actual effect of FSR.

References

1. Weiguo W, Qingmei M, Yu W (2004) Development of the humanoid head portrait robot system with flexible face and expression. In: Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics, pp 757–762. https://doi.org/10.1109/ROBIO.2004.1521877
2. Patil SA, Deore PJ (2016) Local binary pattern based face recognition system for automotive security. In: Proceedings of the International Conference on Signal Processing, Computing and Control, pp 13–17
3. Su MH, Wu CH, Huang KY, Hong QB, Wang HM (2017) Exploring microscopic fluctuation of facial expression for mood disorder classification. In: Proceedings of the International Conference on Orange Technologies, pp 65–69
4. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: Proceedings of the Applications of Computer Vision, pp 1–10
5. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely Connected Convolutional Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
6. Kim J, Lee JK, Lee KM (2016) Deeply-Recursive Convolutional Network for Image Super-Resolution. In: CVPR
7. Dong C, Loy CC, He K, Tang X (2016) Image super-resolution using deep convolutional networks TPAMI
8. Lim B, Son S, Kim H, Nah S, Lee KM (2017) Enhanced deep residual networks for single image Super-Resolution. In: IEEE Computer vision and pattern recognition workshops (CVPRW)
9. Li Z, Yang J, Liu Z, Yang X, Jeon G, Wu W (2019) Feedback network for image Super-Resolution. In: IEEE Conference on computer vision and pattern recognition (CVPR)
10. Zhang T, Zheng W, Cui Z, Zong Y, Yan J, Yan K (2016) A deep neural network driven feature learning method for multi-view facial expression recognition. IEEE Trans Multimed 18(12):2528–2536
11. Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. In: NIPS workshop on Time Series, pp 1–11
12. Shao J, Qian Y (2019) Three convolutional neural network models for facial expression recognition in the wild. In: NEUROCOMPUTING, pp 82–92. https://doi.org/10.1016/j.neucom.2019.05.005
13. Cheng B, Wang Z, Zhang Z, Li Z, Liu D, Yang J, Huang S, Huang T (2017) Robust emotion recognition from low quality and low bit rate video: A deep learning approach. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)
14. Yu C, Tai Y, Liu X, Shen C, Yang J (2018) FSRNet: End-to-end Learning Face Super-Resolution With Facial Priors. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2492–2501
15. Dogan B, Gu S, Timofte R (2019) Exemplar guided face image Super-Resolution without facial landmarks the. IEEE conference on computer vision and pattern recognition (CVPR) workshops
16. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: CVPR
17. Yang W, Feng J, Yang J, Zhao F, Liu J, Guo Z, Yan S et al (2017) Deep edge guided recurrent residual learning for image super-Resolution. IEEE Trans Image Process 26(12):5895–5907. https://doi.org/10.1109/TIP.2017.2750403
18. Chen Y, Tai Y, Liu X, Shen C, Yang J (2018) Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2492–2501
19. Liu B, Ait-Boudaoud D (2019) Effective image super resolution via hierarchical convolutional neural network. In: Neurocomputing. https://doi.org/10.1016/j.neucom.2019.09.035
20. Xie Saining Tu (2017) Zhuowen, Holistically-Nested Edge Detection. In: International Journal of Computer Vision, pp 3–18. https://doi.org/10.1007/s11263-017-1004-z
21. Qin Z, Wu J, Liu Y, Gedeon T (2018) Visual saliency maps can apply to facial expression. Recognition, arXiv:1811.04544
22. Lucey P, Cohn JF, Kanade T, Saragih J (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and

emotion-specified expression. In: Proceedings of the Computer Vision and Pattern Recognition Workshops, pp 94–101
23. Fer K (2013) dataset. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2018-11-10
24. Goodfellow IJ, Erhan D, Luc CP, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee DH (2015) Challenges in representation learning: a report on three machine learning contests. Neural Netw 64:59–63
25. Yin L, Wei X, Sun Y, Wang J, Rosato MJ (2006) A 3D facial expression database for facial behavior research. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition, pp 211–216
26. Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2852–2861
27. Yu X, Fernando B, Ghanem B, Porikli F, Hartley R (2018) Face Super-resolution Guided by Facial Component Heatmaps. The European Conference on Computer Vision (ECCV), pp 217–233
28. Yu X, Fernando B, Hartley R, Porikli F (2018) Super-resolving very Low-Resolution face images with supplementary attributes. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 908–917
29. Kim D, Kim M, Kwon G, Kim D-S Progressive Face Super-Resolution via Attention to Facial Landmark. arXiv:1908.08239 [cs.CV]
30. Liu P, Zhou JT, Tsang WH, Meng Z, Han S, Tong Y (2014) Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In: European conference on computer vision (ECCV), pp 151–166
31. Pramerdorfer C, Martin Kampel E Facial Expression Recognition using Convolutional Neural Networks: State of the Art, arXiv:1612.02903 [cs.CV]
32. Kuo C-M, Lai S-H, Sarkis M (2018) A compact deep learning model for robust facial expression recognition. 2018 IEEE conference on computer vision and pattern recognition
33. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going Deeper with Convolutions. Computer Science
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
35. Jing W, Tian F, Zhang J, Chao K-M, Hong Z, Liu X Feature Super-Resolution Based Facial Expression Recognition for Multi-scale Low-Resolution Faces, arXiv:2004.02234 [cs.CV]
36. Yan Y, Zhang Z, Chen S, Wang H (2020) Low-resolution facial expression recognition: A filter learning perspective. Signal Processing. https://doi.org/10.1016/j.sigpro.2019.107370

Jie Shao received a B.S. and an M.S. degree in the Nanjing University of Aeronautics and Astronautics. Then she got her Ph.D. at Tongji University. At present, she is an associate professor at Shanghai University of Electric Power. Her current research interest includes computer vision, video surveillance, and human emotion analysis.

Qiyu Cheng received her bachelor's degree in electrical engineering and automation from Tongling University in 2018. She is currently a graduate student in the department of electronics and information engineering at Shanghai University of Electric Power. Her research interest includes facial expression recognition and deep learning.