Sensors and Materials, Vol. 36, No. 11 (2024) 4631–4649
MYU Tokyo

S & M 3821

Self-cure Dual-branch Network for Facial Expression Recognition Based on Visual Sensors

Dongsheng Wu,[1] Yifan Chen,[1,2*] Yuting Lin,[1] Pengfei Xu,[1] and Dongxu Gao[2**]

[1]School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China
[2]School of Computing, University of Portsmouth PO13HE, UK

(Received April 4, 2024; accepted May 31, 2024)

Keywords: visual sensors, self-cure network, two-branch method, facial expression recognition

With the rapid development of sensors and sensor technology, facial expression recognition (FER) systems can be developed and applied to real-world scenarios. Vision scan sensors and ambient light sensors capture clear and noise-free images of faces. However, in the real world, annotating large facial expressions is challenging owing to inconsistent labels, which are caused by the annotators' subjectivity and the facial expressions' ambiguity. Moreover, current studies present limitations when addressing facial expression differences due to the gender gap. We not only rely on visual sensors for FER but also utilize nonvisual sensors. Therefore, in this paper, we propose a self-cure dual-branch network (SC-DBN) for FER, which automatically prevents deep networks from overfitting ambiguous samples. First, on the basis of SC-DBN, a two-branch training method is designed, taking full advantage of the gender information. Furthermore, a self-attention mechanism highlights the essential samples and weights, each with a regular weighting. Finally, a relabeling module is used to modify the labels of these samples in inconsistent labels. Many experiments on public datasets show that SC-DBN can effectively integrate gendered information and self-cure networks to improve performance.

1. Introduction

Facial expression is one of the most direct, powerful, and universal signals for human beings to convey their intentions and emotional states.[1,2] A range of sensors such as cameras, eye trackers, electrocardiogram (ECG), electromyogram (EMG), and electroencephalogram (EEG) are used in different facial expression recognition (FER) systems. Cameras are the most popular sensors owing to their simplicity of use and relatively low cost. As shown in Fig. 1, our emotion recognition human–computer interaction system uses a visual sensor to collect facial expression information and transmit it to the expression recognition model on the computer through a wireless network. The deep learning model outputs the results, and then the robot can make corresponding actions.[3] However, using vision sensors alone has limitations. For example, people of different genders may have distinctive ways of expressing their feelings.[4] Furthermore, recent psychological research found that there are different expression styles

*Corresponding author: e-mail: yifan.chen@port.ac.uk
**Corresponding author: e-mail: dongxu.gao@port.ac.uk
https://doi.org/10.18494/SAM5064

ISSN 0914-4935
https://myukk.org/

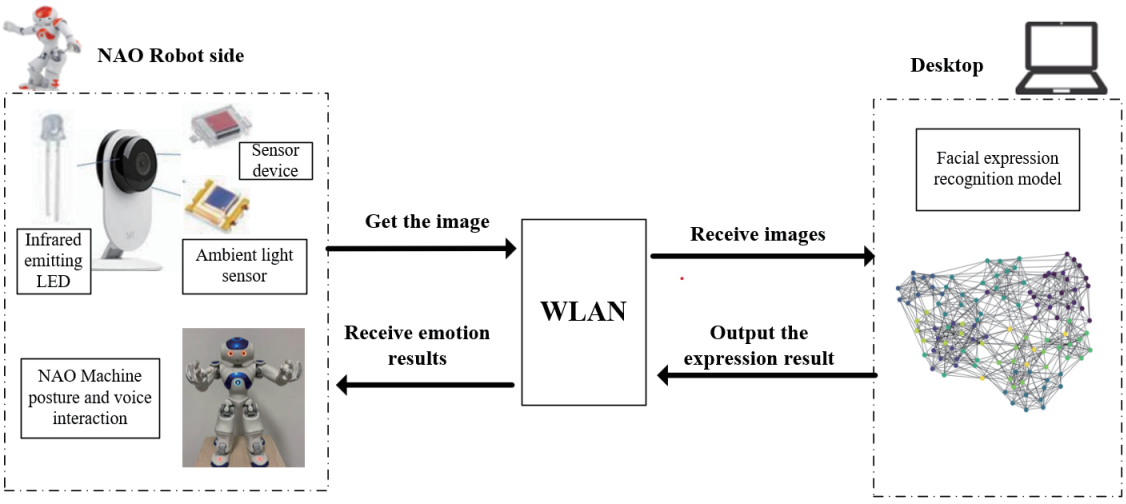


Fig. 1. (Color online) Expression recognition human–computer interaction based on visual sensors.

among subjects of different genders. For example, males tend to express their emotions through their mouth, whereas females often convey their feelings through their eyes.[5] Inspired by this research, in this study, we utilized a two-branch method to leverage the gender knowledge to capture discriminative features. It would be intuitive to investigate the effects of gender differences on the recognition of facial expressions using multimodal sensing technology to obtain discriminative features.

Although the field of visual sensors is developing rapidly, there are still many problems that should be solved in real-world facial expression datasets. Owing to different lighting conditions and sensors, the illumination changes in the expression recognition dataset are large, making it difficult to recognize the deep learning model, which cause the problem of inconsistent labels. Generally, inconsistent labels will lead to the following problems: First, they may result in overfitting on uncertain samples. Moreover, the model may learn useless facial expression features instead of helpful information, resulting in insufficient accuracy. The development of robustness in face recognition algorithms is still a difficult task.

To resolve these issues, we propose a self-cure dual-branch network (SC-DBN) using multimodal sensors. In this study, we divided the training into two branches, the gender recognition branch [branch (a)] and the face recognition branch [branch (b)]. In branch (a), gender is used as the classification condition to capture the distinguishing features in the model so that the parameters in the model are adjusted. In branch (b), the attention relearning process guided by the expression category and the weight regularization and relabeling modules are added so that the model can be self-cured in this process.

The main contributions of this paper are summarized as follows.

- We proposed a nonvisual sensing technology method, including a bilinear fusion module, which integrates gender-related and expression-related features in a framework for improving the classification performance of FER.

- We designed a simple and effective weight regularization module for learning proper importance weights, which also provides a reference for the relabeling module to enhance the robustness of the network.
- We proposed the inconsistency problem in FER, and visual sensors collect data in the wild. The SC-DBN framework is designed to reduce the impact of inconsistencies, which are end-to-end in the backbone, and puts no additional burden on inference.

2. Related Work

2.1 Visual sensors in FER

In the face detection stage, facial sensors are often used,[6] which detect small dynamic changes in facial components, such as eye trackers. Moreover, it can help distinguish background noise from facial features. A depth camera is a sensor that distributes the intensity of depth image pixels based on camera distance. Depth images can provide more powerful features than RGB images, so they are widely used in computer vision, such as face recognition. However, the problem is that it should be more conducive for us to collect expression samples efficiently and rapidly. The second is nonvisual sensors,[7,8] such as audio and EEG sensors, which provide additional information beyond the visual dimension. Thus, nonvisual sensors can improve recognition reliability. Nonvisual sensors such as audio are often used in FER, and facial expressions have been shown to have a reasonable correlation with body and voice.[9] The third type is a target-focusing sensor, such as an infrared thermal sensor, which can help filter out information unrelated to expressions, thereby more accurately helping the deep learning model recognize expressions.

It may be that subtle changes in external factors, such as illumination, occlusion, and pixel, make FER challenging.[10] Environmental effects challenge FER, so combining facial expressions and physiological signals improves the reliability and accuracy of FER systems in real-world situations. The commonly used sensors at this time are ECG, EMG, and EEG sensors.[11] Hassan *et al.* designed an expression recognition model that fuses physiological signals such as EEG and zEMG to accurately and reliably identify different emotions.[12]

2.2 Nonvisual sensors in FER

Recent findings[4] in psychology suggest subtle variations in expressive style among subjects of different genders. Therefore, it would be intuitive to investigate the effects of gender on FER using discriminative features found by computer vision algorithms. Cai *et al.*[13] proposed an identity-free generative adversarial network to reduce the effect of gender and generate an average gender face image. Zhang *et al.*[14] proposed an identity–expression dual-branch network (IE-DBN). First, identity and expression-related features are learned from the same input facial expression image by two branches. Then, those two features are aggregated with a fusion module. Fan *et al.*[15] proposed a two-stage training approach. In one stage, the attention mechanism is demographic aware, forcing the network to focus on informative regions of a face

based on demographic knowledge. In the other stage, the expression of category-driven attention is incorporated to strengthen the discriminative capabilities of selected features. In general, the main idea of these methods is to achieve gender recognition, which means that the effect of gender on expression recognition is weakened as much as possible. In this study, gender information is used as a reference rather than reducing its usefulness.

2.3 Methods of resolving label ambiguity and label noise

The uncertainties in the FER task mainly originate from ambiguous facial expressions and low-pixel facial images. At the same time, facial expression labels originate from the subjectivity of annotators, while basic expressions mix different facial expressions, leading to inconsistent labels. Thus, expressions can be interpreted in many ways, creating label ambiguity and making FER challenging. The early solutions mainly originated from the crowdsourcing community[16] to leverage a small set of clean data that can be used to assess the quality of the labels during training or to estimate the noise distribution or to train the feature extractor.[17] However, these methods ignore the subjectivity of facial expressions; in fact, the label inconsistency is caused by the subjectivity that annotating errors cannot solve. Zeng *et al.*[18] first considered the problem of label inconsistency between FER datasets and proposed to leverage these uncertainties to improve the performance. We should not directly treat inconsistent labels as noise but as a reference. Recently, Chen *et al.*[19] have proposed to treat the inconsistent labels as noise and labels that can describe the image to a certain degree to improve the accuracy of FER.

3. Materials and Methods

3.1 Contactless visual sensing system

In this paper, we introduce a visual-sensor-based expression recognition system utilizing deep learning algorithms, as shown in Fig. 2. During the system development stage, vision scan and ambient light sensors are used to capture comprehensive and transparent face images. Vision scan sensors can capture nearly 3D visual images, helping our algorithm to achieve its due performance; ambient light sensors can help in dark situations. Accurate samples were collected under bright light. Finally, after weighing the recognition effect and cost, we decided on a robot with two vision sensors to collect data.

3.2 Multimodal sensing system in expression recognition

The SC-DBN consists of two parts. Branch (a) leverages the gender information to capture gender-related features. In contrast, expression-related features are generated by branch (b), and then the two features form hybrid features, which directly concentrate on expression categories in a task-driven manner. The backbone CNN is used to extract facial features. The weight of the self-attention importance module for each input picture is obtained by the sigmoid function and fully connected (FC) layer. The obtained weights are multiplied by the logits for a sample

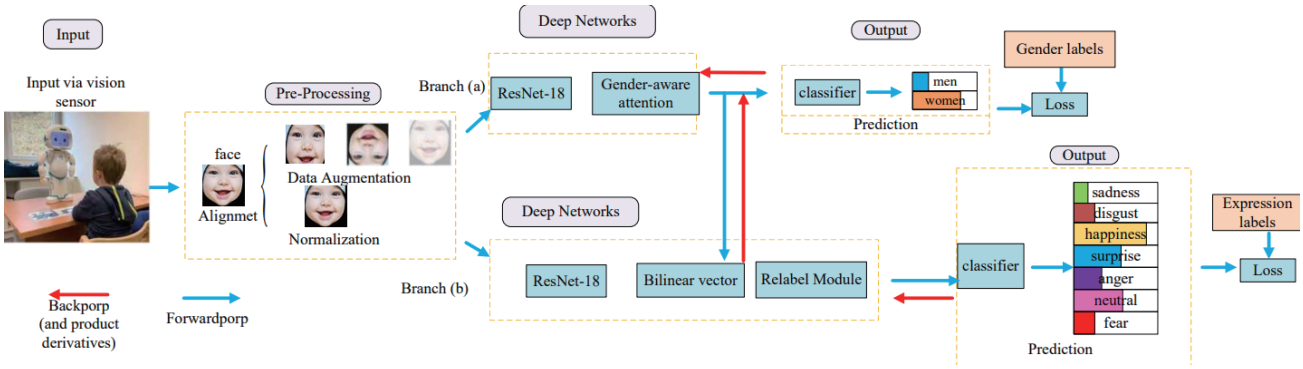


Fig. 2. (Color online) Illustration of proposed network. First, face images are collected through the visual and ambient light sensors; second, the collected face images are preprocessed; finally, a bilinear fusion module is used simultaneously to form hybrid features to improve the FER performance in branch (b).

reweighting scheme. The reliable annotations are assigned high-importance weights. Correspondingly, ambiguous facial images are posted with low-importance weights.

Furthermore, to reduce the importance of ambiguous samples, the obtained importance weights are sorted from high to low by the weight regularization module, which splits them into two groups; the high-importance group accounts for α times the total. This regularization is implemented with a loss function, termed weight regularization loss (WR-Loss), highlighting specific samples and suppressing uncertain samples. The relabeling module is used to relabel these samples from the bottom group. An example is assigned to a pseudolabel if the maximum predicted chance is greater than the probability of the original label with a margin threshold.

Finally, the refined features are used to improve the accuracy of FER. Note that the two branches are optimized by gender and expression categories. The gender information in branch (a) is divided into male and female. The categories in branch (b) are seven basic expressions.

3.3 Bilinear fusion module

As shown in Fig. 3, bilinear fusion module is proposed to establish the correlations between the gender-related and expression-related feature maps. $\mathcal{G}_{\theta_1}(l, X)$ is the gender descriptor at location l from gender feature F_{gen}. At the same time, $\mathcal{E}_{\theta_2}(l, X)$ is the expression descriptor at location l from expression feature F_{exp}, which can be formulated as Eqs. (1) and (2).

$$\mathcal{B}\left(\mathcal{G}_{\theta_1}(X), \mathcal{E}_{\theta_2}(X)\right) = vec\left(\mathcal{G}_{\theta_1}(l, X) \otimes \mathcal{E}_{\theta_2}(l, X)\right) \tag{1}$$

$$F = pooling_{l \in \mathcal{L}}\left\{\mathcal{G}\left(\mathcal{X}_{\theta_1}(X), \mathcal{E}_{\theta_2}(X)\right)\right\} = \sum_{l \in \mathcal{L}} \mathcal{G}\left(\mathcal{X}_{\theta_1}(X), \mathcal{E}_{\theta_2}(X)\right) \tag{2}$$

Here, $\otimes$ calculates the outer product of two vectors, *vec* (.) turns the matrix into a vector, and $\mathcal{L}$ is the set of all spatial locations on the output feature map.

The bilinear fusion module establishes the correlations between gender and expression

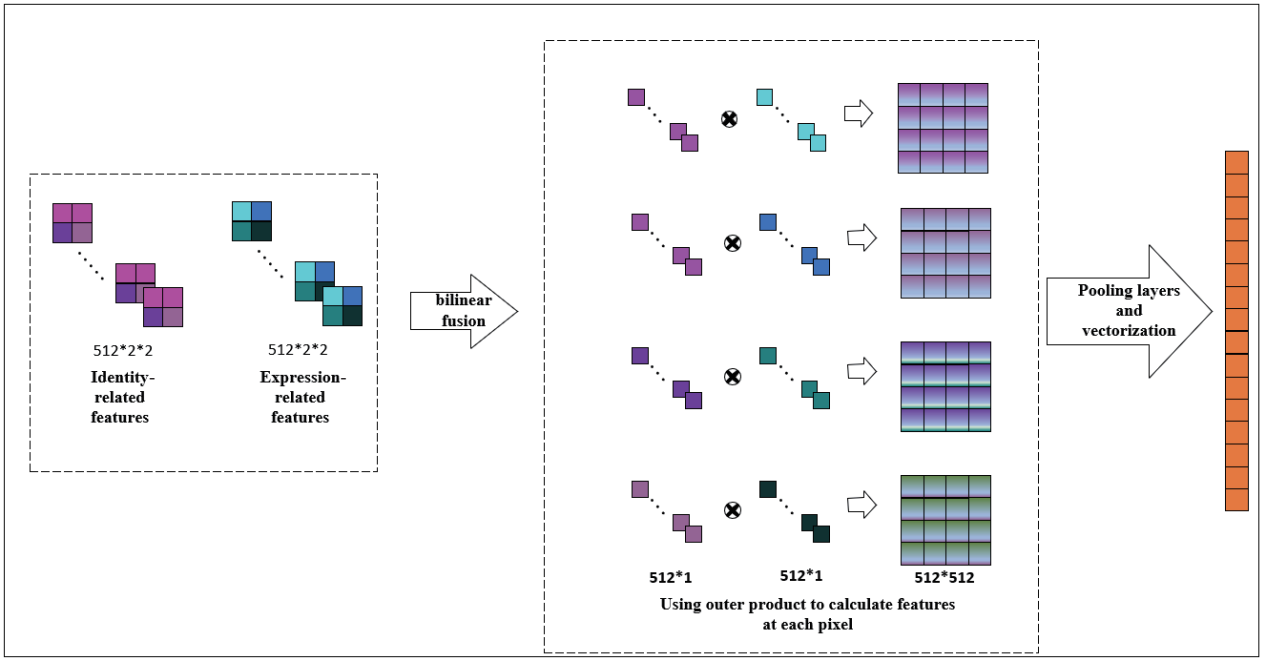


Fig. 3. (Color online) Schematic diagram of feature fusion.

feature maps at each spatial location. For example, the pixel of each channel of gender-related feature maps is multiplied by the pixel of each channel at the same spatial location of expression-related feature maps; then, a fusion matrix is calculated.

3.4 Self-attention importance weighting module

The self-attention importance weighting module is introduced to capture the importance of samples. The FC layer and sigmoid are used to roughly estimate the importance of the weight. $F = [x_1, x_2, x_{n-1}, x_n]$ donates the bilinear features of n images, the self-attention importance weighting module takes F as input. The high correlation will obtain the high-importance weight, and the low correlation will obtain the low-importance weight. It is expected that this module will enhance the current useful information and suppress the previously less helpful information, which can be formulated as Eq. (3).

$$\mu_i = f(\theta_0^\top x_i) \tag{3}$$

Here, f is the sigmoid function, $\theta_0^\top$ is the parameter of the fully connected layer used for attention, x_i is the input feature of a specific image, and μ_i is the importance weight obtained from the i-th sample, which is also used as a reference later. Generally, the attention block in our network carries out standard backpropagation with end-to-end training to guide the learning process and changes adaptively without constraints. The self-attention importance weighting module is only used in branch (b).

3.5 Multiclass cross-entropy loss

In this study, the weights are learned from CNN, which is also end-to-end. Thus, the logit weighting has been chosen.[20] In the multiclass cross-entropy loss, we call the weighted loss logit, and according to our two-branch method model, we improve the common cross-entropy loss, which is formulated as Eq. (4).

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\log\frac{e^{\mu_i\theta_{y_i}^{\top}\boldsymbol{x}_i}}{\sum_{l=1}^{c}e^{\mu_i\theta_l^{\top}\boldsymbol{x}_i}} \tag{4}$$

Here, c is the number of tag categories. In branch (a), only male and female are distinguished, so the value is two. In branch (b), the expression category is divided into seven categories, so the value is seven. The batch size is n, that is, how many samples there are in a batch. The input feature of the i-th sample is correspondingly defined as x_i, and its corresponding label variable is y_i. $\theta_l^{\top}$ is the classifier, μ_i is the attention weight assigned to the i-th sample, and $\mathcal{L}$ has a positive correlation with μ_i.

3.6 Weight regularization module

In the self-attention weighting module, each picture is assigned a different weight between 0 and 1. In the weight regularization module, attention weights are ranked in descending order. Then, each batch will be divided into two groups: high importance and low importance. The high-importance group accounts for α times the total, so $n\alpha$ is the number of high-importance groups; correspondingly, $(1-\alpha)n$ is the number of low-importance groups. To ensure that the average attention weight of the high-importance group is higher than that of the low-importance group with a margin, a direct constraint is required as follows.

$$\mathcal{L}_{WR} = \max\left\{0, v - \left(\mu_{max} - \mu_{min}\right)\right\} \tag{5}$$

Here, v is the margin, μ_{max} is the average of the high-importance group, and μ_{min} is the average of the low-importance group. The total loss of training is shown as

$$\mathcal{L}_{all} = \lambda\mathcal{L}_{WR} + (1-\lambda)\mathcal{L}\,. \tag{6}$$

.

Among them, λ is the trade-off ratio, which weighs the size of the two loss function values. One of them is more important, and λ is used to allocate the ratio.

3.7 Self-cure module

The importance weight of uncertain samples will be relatively low in the weight regularization module. Thus, the idea is to relabel these low-importance samples, and the relabeling module is

performed on Softmax probabilities. A pseudolabel is generated to a sample if the highest probability is greater than the probability corresponding to the original label with a threshold. Mathematically, it can be defined as

$$Y_{max} = \operatorname{argmax} \frac{e^{\mu_i \theta^T x_i}}{\sum_{l=1}^{c} e^{\theta_l^T x_i}} \tag{7}$$

with

$$Y = Y_{max} \text{ if } P_{max} > P_0 + \tau \,. \tag{8}$$

Here, Y is the expression label, Y_{max} is the label with the highest predicted probability, τ is the threshold, P_0 is the predicted probability of the original label, and P_{max} is the maximum probability. It is judged whether to replace the label if the label of the maximum probability predicted by Softmax is not the original label. This module will start after 10 epochs.

In the system, the importance weights of the uncertain samples are expected to be low, so the low-importance weight is increased as much as possible by reweighting to suppress their negative effect. In fact, it can be corrected as certain samples by relabeling the module to suppress the uncertainty to gradually improve the performance. The parameters in the network are learned from the gender information in branch (a) and the expression information in branch (b). Finally, the prediction result originates from branch (b) obtained through Softmax.

4. Experiments

4.1 Datasets

4.1.1 RAF-DB[21]

The Real-world Affective Faces database (RAF-DB) is a large-scale facial expression dataset in the wild. The image faces in this dataset are extremely diverse in terms of age, gender and race, head posture, lighting conditions, occlusions (such as glasses, facial hair or self-occlusion), postprocessing operations (such as various filters and special effects), and so forth. Not only that, the dataset has been divided into a test set, a verification set, and a training set. The specific performance is as follows: 12271 expression data are used for model training, 3068 samples are used for verification, and 3068 samples are used for testing.

4.1.2 Extended Cohn–Kanade (CK+) database[22]

The Extended Cohn–Kanade (CK+) database is a database containing 593 video sequences. The videos were from 123 different subjects, ranging in age from 18 to 50 years old, with different genders and ancestries. In this database, samples are labeled as one of the seven expression categories: anger, contempt, disgust, fear, happiness, sadness, and surprise. The CK+

database is widely considered to be the most widely used laboratory-manipulated facial expression classification database currently and is used in major facial expression classification methods.

4.1.3 Oulu-CASIA[23]

The Oulu-CASIA NIR&VIS facial expression database (Oulu-CASIA) is a facial expression database that consists of 80 individuals, ranging in age from 23 to 58 years, and 73.8% of the subjects are male. The database contains six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) and was shot under three types of lighting (dark, strong light, and low light). The image resolution is 320 × 240 pixels.

4.1.4 FER2013[24]

The Facial Expression Recognition 2013 (FER2013) dataset contains approximately 30,000 facial RGB images of different expressions, and the size of the images is 48×48 pixels. The samples are all images obtained from the wild, so the sample images will be affected by lighting or facial occlusion. The main annotations in this dataset can be divided into seven types. Disgust expression has the smallest number of images at only 600, whereas each of the other types has nearly 5,000 samples.

4.2 Implementation

For fair comparison with other state-of-the-art methods, we trained SC-DBN using the training part of the public dataset RAF-DB.[25] We trained 10 times, and the final accuracy was the average of the 10 running results. Test models were evaluated using the official validation set. The parameters were optimized by the stochastic gradient descent (SGD) method. The momentum was 0.9, the weights decayed to 0.0004, and the learning rate was initialized to 0.1 and further divided by 10 after 15 and 30 epochs. Our baseline was a ResNet-18 model trained directly using a modified cross-entropy loss function. α was set to 0.7, and the average difference between them was at least v, set to 0.07. $\mathcal{L}$ and $\mathcal{L}_{WR}$ are combined to jointly optimize the entire network. The relabeling margin τ was set to 0.2 by default.

4.3 Data preprocessing

First, the multitask convolutional neural network was used to implement face detection and remove irrelevant background information.[26] The detected faces were then rotated and aligned using a face alignment algorithm and further resized to uniformly sized pixels. Finally, the data was enhanced by rotating, flipping, changing pixel values, cropping, and adding Gaussian noise. The SC-DBN network was implemented in Pytorch and GPU for end-to-end training. The parameters were initialized using backbone network pretraining on ImageNet. The last fully connected layer of the backbone network was removed and information was extracted from its last pooling layer.

4.4 Experimental results analysis

4.4.2 Results on RAF database

In our comparative analysis detailed in Table 1, SC-DBN demonstrates superior performance over advanced feature learning methods by achieving a notable accuracy improvement of 7.44% over VGG-face. This comparison highlights the efficacy of incorporating both a self-cure network and gender information into the learning process. SC-DBN's enhanced performance, when compared with frameworks such as MRE-CNN and FSN that also integrate gender data, underscores the unique advantage of the self-cure mechanism in enriching the model with more discriminative information for FER.

Further internal analysis comparing variations of SC-DBN—with and without gender information, and with and without the self-cure network—illustrates the positive impact of gender information on network efficacy. At the same time, because the self-healing network can correct wrong labels, it will not be disturbed by noise labels. Specifically, when gender information is added, the accuracy of expression recognition increases to a certain extent, which further promotes the ability of the self-curing network to modify wrong labels, because it needs to be built on a strong recognition model. The stronger the original model, the stronger the cure ability of the self-curing network, and there is a positive relationship between them.

Overall, the finding not only bolsters the argument for the inclusion of gender data but also emphasizes the critical role of the self-cure network in SC-DBN's superior performance, suggesting its potential as a valuable feature for improving accuracy in FER tasks.

After the model proposed in this article was trained for basic emotion classification, we provided a more detailed confusion matrix to observe the classification of specific categories, as shown in Fig. 4. The diagonal entries of the confusion matrix represent the accuracy of each label being correctly identified. It can be clearly seen that the SC-DBN model performs better in

Table 1
Performance of different methods on RAF-DB.

Method	Accuracy
DLP-CNN[27]	84.22
IPA2LT(LTNet)[18]	83.80
Baseline[18]	81.81
BbaseDCNN[27]	82.66
Center Loss[27]	82.86
VGG-FACE[28]	79.16
MRE-VGG[28]	82.63
PAT-ResNet[29]	84.19
ARM[30]	85.28
RAN[31]	85.58
SC-DBN	85.96
SC-DBN (without gender information)	85.04
SC-DBN (without self-cure network)	84.65

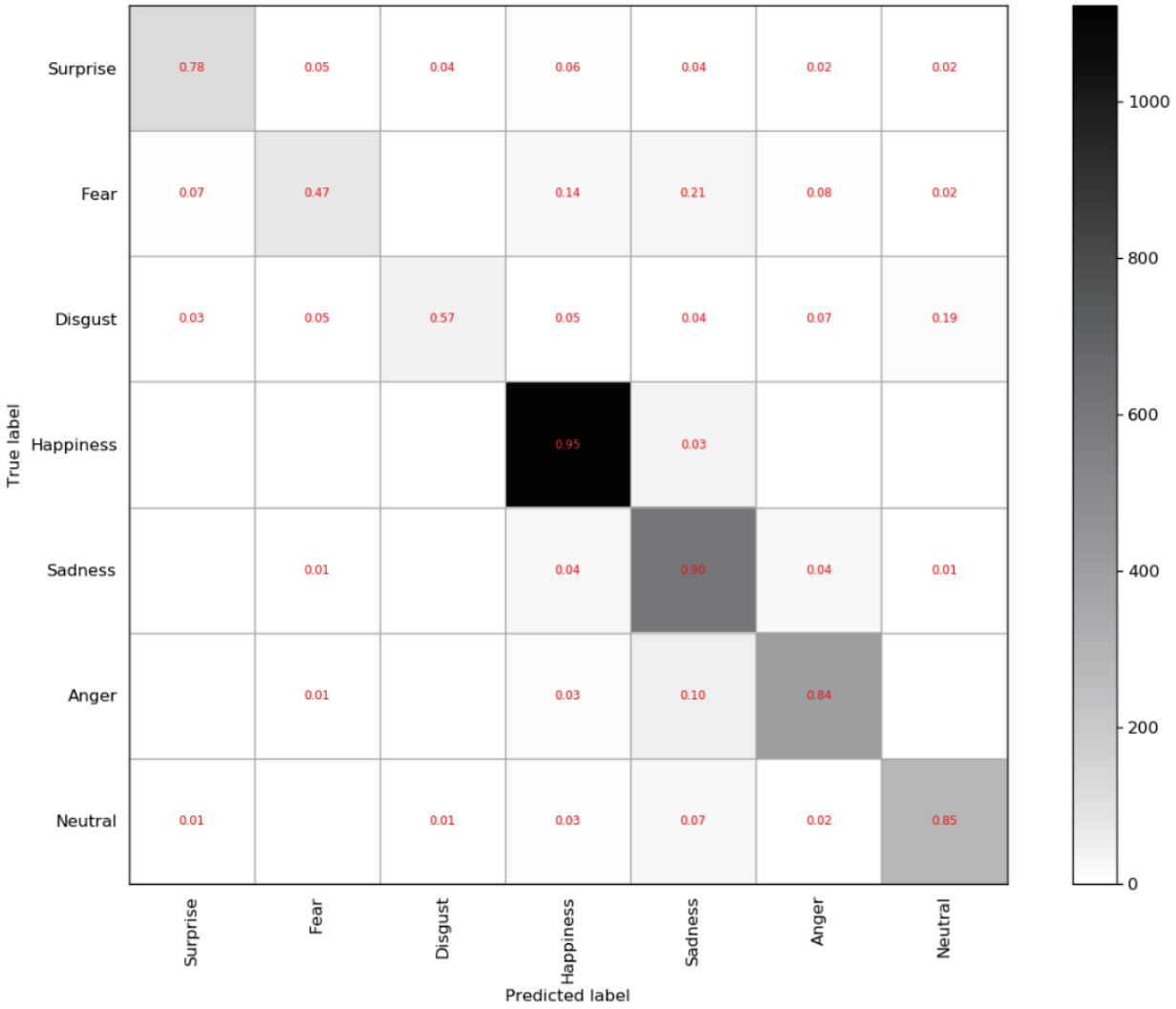


Fig. 4. (Color online) Confusion matrices computed from the true and predicted expression labels of RAF-DB (angry, disgust, fear, happy, sad, surprise, and neutral). The darker the color, the higher the accuracy.

identifying positive expressions, such as happiness and surprise, with happiness having the highest accuracy. However, we found that the recognition rate of a negative expression such as disgust is very low. We conclude that for expression recognition, first, negative expressions are often difficult to detect, whereas positive expressions are often exaggerated. Second, expression data are usually collected from social media, and photos uploaded on social media are usually of happy moments. This leads to the problem of less negative expression data, resulting in the low recognition accuracy of negative expressions.

4.4.3 Evaluation of different modules

In this study, we meticulously evaluated the effects of three distinct modules—bilinear fusion, weight regularization, and relabeling—on the SC-DBN's efficacy in FER within the

RAF-DB. Table 2 illustrates a nuanced finding: integrating the relabeling module into the baseline model paradoxically diminishes accuracy. This counterintuitive outcome underscores the complexity of relabeling operations, necessitating a judicious approach to when labels should be adjusted. Our methodology, which selectively targets low-importance groups for relabeling through weight regularization, introduces a pragmatic constraint that mitigates indiscriminate label modifications, thus fostering the model's self-corrective capability. The data further reveals that omitting the relabeling module slightly reduces accuracy, attesting to its value under a constrained application in bolstering model performance.

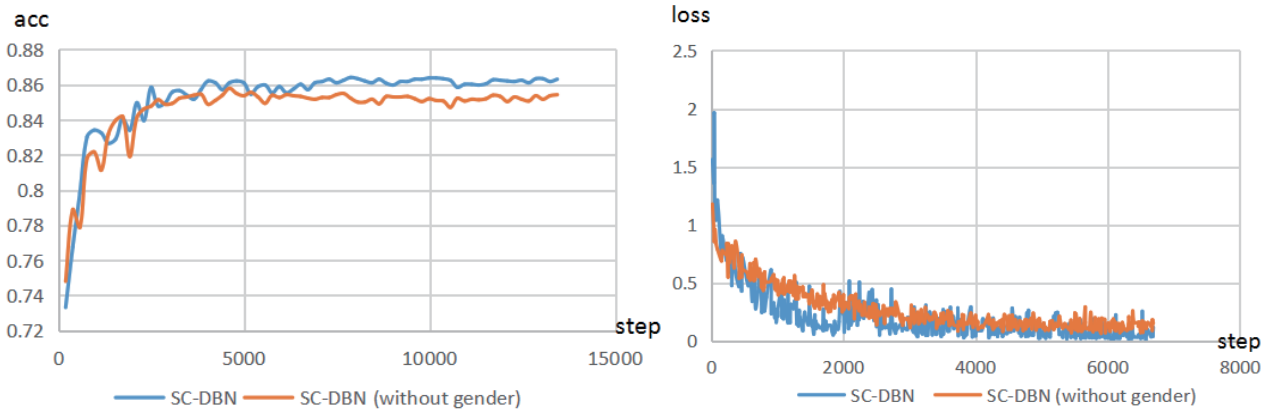
Moreover, the incorporation of the bilinear fusion module, which integrates gender characteristics, significantly enhances SC-DBN's performance as shown in Fig. 5. Despite similar convergence times between configurations with and without gender information, the absence of gender insights leads to a notable decline in average accuracy. This phenomenon highlights the pivotal role of gender information in refining SC-DBN's performance, by directing the model's focus towards more discriminative weight groups, and thereby optimizing the self-curing network's efficacy.

4.4.4 Results on CK+ database

In the evaluation conducted on the CK+ database, as detailed in Table 3, our SC-DBN demonstrates performance sperior to that of contemporary methodologies, including feature

Table 2
Evaluation of different modules in SC-DBN.

Bilinear	Weight	Relabel	RAF-DB
√			76.37
√		√	83.38
√	√		85.79
√	√	√	85.96
		√	60.04
	√		73.66
	√	√	85.04



(a) (b)

Fig. 5. (Color online) The two columns display the (a) testing accuracy and (b) testing loss of SC-DBN. The testing processes were performed using RAF-DB.

Table 3
Performance comparisons with other approaches on CK+ database.

Method	Accuracy
Baseline9[18]	88.99
SCAN-CCI[32]	97.31
IF-GAN[13]	97.52
DLP-CNN[27]	95.78
Vit+SE[33]	99.80
Center Loss[34]	92.26
IL-CNN[34]	94.35
FAN[35]	99.70
DeepEmotion[36]	90.63
SC-DBN	99.83
SC-DBN (without gender information)	98.39
SC-DBN (without self-cure network)	97.62

fusion methods and CNN-based models. Specifically, SC-DBN surpasses the SCAN-CCI method by a margin of 2.52% and exhibits a 0.09% improvement over the facial deep representation Learning. Furthermore, it outperforms the IF-GAN method, which is designed to address dataset noise, by 2.31% and achieves a 0.03% higher accuracy than Vit+SE. These comparisons not only underscore SC-DBN's effectiveness but also highlight the incremental benefits of incorporating a dual-branch strategy that leverages gender information and a self-healing network to enhance performance.

4.4.5 Results on OULU database

As shown in Table 4, when augmented solely with gender information, our SC-DBN surpasses established feature fusion methodologies such as DTAGN (Joint) and LOMo, underscoring the critical importance of gender considerations in the accurate identification of expressions.

Furthermore, when juxtaposed against methods designed to mitigate dataset noise such as FN2EN, Center Loss, and IL-CNN, SC-DBN consistently demonstrates superior performance. This consistent outperformance across various noise-adaptive methodologies provides evidence of the self-curing network's effectiveness in navigating the complexities of datasets afflicted by background noise. Such findings not only affirm the value of integrating gender information but also showcase the self-curing network's adeptness at enhancing recognition accuracy in environments compromised by lighting variations or background disturbances.

4.4.6 Results on FER2013 database

In our detailed examination of the FER2013 "in-the-wild" database, known for its imbalanced expression categories and a notable scarcity of negative expressions,[42] the SC-DBN method has been meticulously compared with current state-of-the-art approaches. Our findings, as

Table 4
Performance comparisons with other approaches on the Oulu-CASIA database.

Method	Accuracy
HOG 3D[37]	70.63
DTAGN(Joint)[38]	81.46
Baseline[18]	59.35
FN2EN[39]	87.71
Center Loss[34]	75.63
IL-CNN[34]	77.29
DeRL[40]	88.00
SC-DBN	95.40
SC-DBN (without gender information)	93.48
SC-DBN (without self-cure network)	90.23

illustrated in Table 5, reveal that SC-DBN significantly enhances the handling of samples with complex backgrounds and label ambiguities. Impressively, SC-DBN achieves a performance improvement of 1.54% over the best existing method specifically in scenarios characterized by label noise and complex sample backgrounds. Furthermore, even in configurations where gender information is not utilized and the model relies solely on its self-healing module to correct ambiguous or erroneous labels, SC-DBN still manages to surpass the leading method by 0.47%.

4.4.7 Visualization analysis

The visual representation derived from testing on the challenging RAF-DB is depicted in Fig. 6. Initially, as showcased in Fig. 6(a), the embedding space before model training presents a scenario where data clusters are markedly unbalanced, with a disordered distribution across various categories.[43] This visual prelude starkly contrasts with the post-training landscape illustrated in Fig. 6(b), where the SC-DBN model's effect manifests through significantly mitigated category overlap and overfitting issues.

This separation is not merely a dispersion of data points but evolves into the formation of compact clusters, each cohesively organized around their respective labels. Such visual evidence not only confirms the model's adeptness in feature extraction and recognition within complex datasets but also highlights the structured learning process that underpins the SC-DBN model's ability to refine and enhance expression categorization, thereby contributing to the broader discourse on effective FER methodologies.

4.5 Ablation studies

4.5.1 Evaluation of α

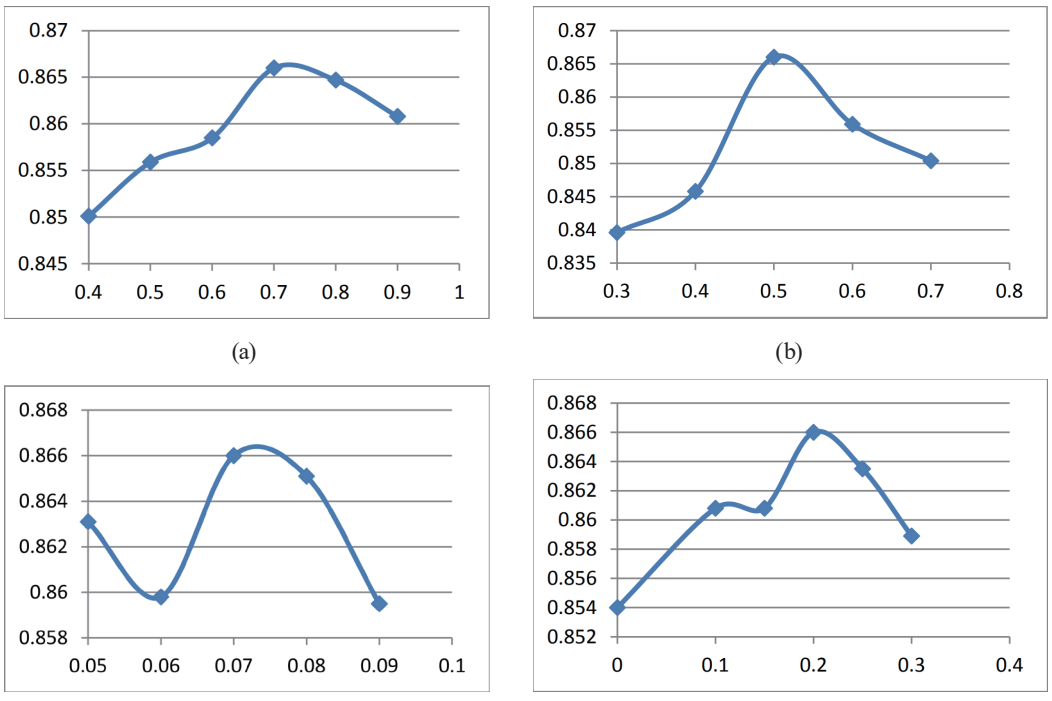
A very large α will degrade the ability to suppress uncertainties as shown in Fig. 7. A very small α will lead to the over-consideration of uncertainties. As shown in the above line chart, the decrease in α has unreasonably reduced the test accuracy. The accuracy of the model only slightly decreaces as α increases. We can conclude that too much consideration of uncertain data

Table 5
Performance of different methods on FER2013.

Method	Accuracy
KDL[27]	71.28
ResNet[27]	72.40
Residual Masking Network[24]	74.14
PASM[41]	73.59
MoVE-CNN[30]	77.70
Attentional ConvNet[36]	70.02
SC-DBN	79.24
SC-DBN (without gender information)	78.32
SC-DBN (without self-cure network)	77.23

(a) (b)

Fig. 6. (Color online) TSNE visualization of the outputs of SC-DBN. There are 15339 samples. (a) Before and (b) after training.



(a) (b)

(c) (d)

Fig. 7. (Color online) Evaluation of (a) α, (b) λ, (c) v, and (d) τ on RAF-DB.

will have a significant impact on accuracy. Thus, α of 0.7 or 0.8 makes the model achieve the best effect.

4.5.2 Evaluation of λ

We evaluated the impact of different ratios between $\mathcal{L}$ and $\mathcal{L}_{WR}$. From our previous experience, the model works best when λ is 0.5; thus, we take a number around 0.5 for comparison. In the line graph, when λ decreases, it will significantly affect the performance, which indicates that $\mathcal{L}$ is more important, and 0.5 is indeed the best.

4.5.3 Evaluation of v

v is an important parameter for determining the size of $\mathcal{L}_{\mathrm{WR}}$ by controlling the average attention scores between the high-importance group and the low-importance group. We will evaluate v between 0 and 0.1. As shown in the graph, the model achieves the best effect when $v = 0.07$. If v is very high, the loss function will always exist, which will lead to the problem of overfitting, and the loss value cannot be reduced. If v is very low, the loss function $\mathcal{L}_{\mathrm{WR}}$ has no meaning.

4.5.4 Evaluation of τ

As shown in Fig. 7, $\tau = 0$ means that the label is modified if the maximum label probability is larger than the probability of the original label, which will severely damage the performance. If τ is very large, it hardly needs to be modified, which is meaningless. It is best when $\tau = 0.2$ is observed from the line chart.

5. Conclusions

In this study, we introduced the SC-DBN, a novel architecture tailored to enhance FER by leveraging gender-specific emotional expressions and mitigating uncertainties within datasets. The network comprises two main branches: one for gender classification and the other for facial expression classification, enhanced with four critical modules (bilinear fusion, self-attention, weight regularization, and relabeling) to reduce feature ambiguities and improve label accuracy. Our findings, validated across standard datasets, demonstrate the SC-DBN's superior performance and robustness, underscoring its potential applicability beyond FER to a broader spectrum of human behavior analysis tasks, including voice, gesture, and posture recognition. This versatility highlights the model's potential as a foundational tool for future research in diverse classification and recognition domains.

Acknowledgments

This work is supported by the Liaoning Provincial Department of Education's Higher Education Institutions Scientific Research Project (JYTZD2023006). This research project is Liaoning Provincial Department of Education Scientific Research Project Plan "Research on Complex Process Set Fault Diagnosis Method Based on Meta Learning Network".

Conflict of Interest

The authors declare no conflict of interest.

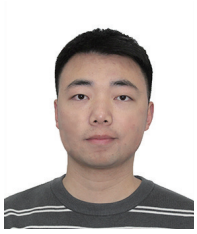
References

1 C. Darwin: The expression of the emotions in man and animals (University of Chicago Press, 2015). https://doi.org/10.1017/CBO9780511694110
2 J. Yu, H. Gao, D. Zhou, J. Liu, Q. Gao, and Z. Ju: IEEE Trans. Cybern. **52** (2021) 13738. https://doi.org/10.1109/TCYB.2021.3114031
3 M. Li, R. Kang, D. T. Branson, and J. S. Dai: IEEE/ASME Trans. Mechatron. **23** (2018) 286 https://doi.org/10.1109/TMECH.2017.2775663
4 C. Chen, C. Crivelli, O. G. Garrod, P. G, Schyns, J. M. Fernández-Dols, and R. E. Jack: PNAS **115** (2018) E10013. https://doi.org/10.1073/pnas.1807862115
5 K. Li, P. Boyd, Y. Zhou, Z. Ju, and H. Liu: IEEE Trans. Autom. Sci. Eng. **16** (2018) 1556. https://doi.org/10.1109/TASE.2018.2882465
6 T. Zhang, H. Lin, Z. Ju, and C. Yang: Int. J. Fuzzy Syst. **22** (2020) 1330. https://doi.org/10.1007/s40815-020-00825-w
7 B. Amos, B. Ludwiczuk, and M. Satyanarayanan: CMU School of Computer Science **6** (2016) 118.
8 J. Yu, H. Gao, Y. Chen, D. Zhou, J. Liu, and Z. Ju: IEEE Trans. Cognit. Dev. Syst. **14** (2021) 1654. https://doi.org/10.1109/TCDS.2021.3131253
9 J. Li, Y. Mi, G. Li, and Z. Ju: Int. J. of Humanoid Rob. **16** (2019) 1941002. https://doi.org/10.1142/S0219843619410020
10 J. Yu, H. Gao, Y. Chen, D. Zhou, J. Liu, and Z. Ju: IEEE Trans. Hum.-Mach. Syst. **52** (2022) 784. https://doi.org/10.1109/THMS.2022.3144951
11 B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran: Expert Syst. Appl. **93** (2018) 143. https://doi.org/10.1016/j.eswa.2017.09.062
12 M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino: Inf. Fusion **51** (2019) 10. https://doi.org/10.1016/j.inffus.2018.10.009
13 J. Cai, Z. Meng, A. S. Khan, J. O'Reilly, Z. Li, S. Han, and Y. Tong: Proc. 2021 IEEE Int. Conf. Image Processing (ICIP) (IEEE, 2021) 1344–1348. https://doi.org/10.1109/ICIP42928.2021.9506593
14 H. Zhang, W. Su, J. Yu, and Z. Wang: IEEE Trans. Cognit. Dev. Syst. **13** (2020) 898. https://doi.org/10.1109/TCDS.2020.3034807
15 Y. Fan, V. Li, and J. C. Lam: IEEE Trans. Affective Comput. **10** (2020) 1057. https://doi.org/10.1109/TAFFC.2020.2988264
16 J. A. Saglia, N. G. Tsagarakis, J. S. Dai, nd D. G. Caldwell: IEEE/ASME Trans. Mechatron. **18** (2013) 1799. https://doi.org/10.1109/TMECH.2012.2214228
17 J. Dai and D. Caldwell: Trends Food Sci. Technol. **21** (2010) 153. https://doi.org/10.1016/j.tifs.2009.10.007
18 J. Zeng, S. Shan, and X. Chen: Proc. European Conf. Computer Vision (ECCV) (2018) 222–237.
19 S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, Y. Rui: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (2020) 13984–13993.
20 W. Hu, Y. Huang, F. Zhang, and R. Li: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (2019) 11887–11896.
21 S. Li and W. Deng: IEEE Trans.n Image Process. **28** (2018) 356. https://doi.org/10.1109/TAFFC.2020.2981446
22 P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews: Proc. 2010 IEEE Computer Society

Conf. Computer Vision and Pattern Recognition-workshops (IEEE, 2010) 94–101. https://doi.org/10.1109/CVPRW.2010.5543262
23 G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen: Image Vision Comput. **29** (2011) 607. https://doi.org/10.1016/j.imavis.2011.07.002
24 I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, T. Tang, D. Thaler, D.-H. Lee, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio: Proc. Int. Conf. Neural Information Processing (Springer, 2013) 117–124.
25 S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang: Proc. European Conf. Computer Vision (ECCV) (2018) 135–150.
26 W. Zhang, Y. Wang, and Y. Qiao: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (2019) 7373–7382.
27 S. Li, W. Deng, and J. Du: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2017) 2852–2861.
28 Y. Fan, J. C. Lam, and V. O. Li: Proc. Int. Conf. Artificial Neural Networks (Springer, 2018) 84–94.
29 J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong: arXiv (2018). https://doi.org/10.48550/arXiv.1812.07067
30 J. Shi, and S. Zhu: arXiv (2021). https://doi.org/10.48550/arXiv.2103.10189
31 K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao: IEEE Trans. Image Process. **29** (2020) 4057. https://doi.org/10.1109/TIP.2019.2956143
32 D. Gera and S. Balasubramanian: Pattern Recognit. Lett. **145** (2021) 58. https://doi.org/10.1016/j.patrec.2021.01.029
33 M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, R. Seguier: arXiv (2021). https://doi.org/10.48550/arXiv.2107.03107
34 J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong: Proc. 2018 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018) (IEEE, 2018) 302–309. https://doi.org/10.1109/FG.2018.00051
35 D. Meng, X. Peng, K. Wang, and Y. Qiao: Proc. 2019 IEEE Int. Conf. Image Processing (ICIP) (IEEE, 2019) 3866–3870. https://doi.org/10.1109/ICIP.2019.8803603
36 S. Minaee, M. Minaei, and A. Abdolrashidi: Sensors **21** (2021) 3046. https://doi.org/10.3390/s21093046
37 Y. Li, J. Zeng, S. Shan, and X. Chen: IEEE Trans. Image Process. **28** (2018) 2439. https://doi.org/10.1109/TIP.2018.2886767
38 H. Jung, S. Lee, J. Yim, S. Park, and J. Kim: Proc. IEEE Int. Conf. Computer Vision (2015) 2983–2991. https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Jung_Joint_Fine-Tuning_in_ICCV_2015_paper.html
39 H. Ding, S. K. Zhou, R. Chellappa: Proc. 2017 12th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2017) (IEE, 2017) 118–126. https://doi.org/10.1109/FG.2017.23
40 H. Yang, U. Ciftci, and L. Yin: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2018) 2168–2177.
41 P.. Giannopoulos, I. Perikos, and 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. https://doi.org/10.1007/978-3-319-66790-4_1.
42 L. Van der Maaten and G. Hinton: J. Mach. Learn. Res. **9** (2008).
43 L. Berlincioni, L. Cultrera, C. Albisani, L. Cresti, A. Leonardo, S. Picchioni, and A. Del Bimbo: Proc.IEEE/CVF Conf. Computer Vision and Pattern Recognition (2023) 4108–4118. https://openaccess.thecvf.com/content/CVPR2023W/EventVision/html/Berlincioni_Neuromorphic_Event-Based_Facial_Expression_Recognition_CVPRW_2023_paper.html

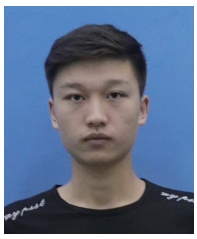
About the Authors

Dongsheng Wu received his B.S. and M.S. degrees from Shenyang Institute of Technology, China, in 1997 and 2005, respectively, and his Ph.D. degree from Changchun University of Science and Technology, China, in 2013. From 2013 to 2019, he was an assistant professor at Shenyang Ligong University, China. Since 2020, he has been a professor at Shenyang Ligong University. His research interests are in intelligent control, fault diagnosis, and sensors. (wuds@sylu.edu.cn)



Yifan Chen received his B.S. degree in automation from Beijing Jiaotong University, Haibin College, China, in 2020. He received his M.S. degree in intelligent systems from Shenyang Ligong University, China, in 2023. He is working on his Ph.D. degree in intelligent robotics at the University of Portsmouth, U.K. His research interests include facial recognition, human motion analysis, and human–robot collaboration. (yifan.chen@port.ac.uk)

Yuting Lin received her B.S. degree from Qingdao Ligong University in 2022. She is currently studying for a master's degree at Shenyang Ligong University. Her research interests are in facial expression recognition and sensors. (linyt1115@163.com)



Pengfei Xu received his B.S. degree from Qingdao Ligong University Qindao College in 2021. He is currently studying for a master's degree at Shenyang Ligong University. His research interests are in facial expression recognition. (xpf_afei@163.com)