Facial Expression Recognition Through Cross-Modality Attention Fusion

Rongrong Ni, Biao Yang, *Member, IEEE*, Xu Zhou, *Graduate Student Member, IEEE*, Angelo Cangelosi, *Senior Member, IEEE*, and Xiaofeng Liu, *Member, IEEE*

***Abstract*—Facial expressions are generally recognized based on handcrafted and deep-learning-based features extracted from RGB facial images. However, such recognition methods suffer from illumination/pose variations. In particular, they fail to recognize these expressions with weak emotion intensities. In this work, we propose a cross-modality attention-based convolutional neural network (CM-CNN) for facial expression recognition. We extract expression-related features from complementary facial images (gray-scale, local binary pattern, and depth images) to handle the illumination/pose variations and to capture appearance details that describe expressions with weak emotion intensities. Rather than directly concatenating the complementary features, we propose a novel cross-modality attention fusion network to enhance spatial correlations between any two types of facial images. Finally, the CM-CNN is optimized with an improved focal loss, which pays more attention to facial expressions with weak emotion intensities. The average classification accuracies on VT-KFER, BU-3DFE(P1), BU-3DFE(P2), and Bosphorus are 93.86%, 88.91%, 87.28%, and 85.16%, respectively. Evaluations on these databases demonstrate that our approach is competitive to state-of-the-art algorithms.**

***Index Terms*—Convolutional neural network (CNN), cross-modality attention fusion, facial depth images, facial expression recognition (FER), focal loss.**

I. Introduction

THE CAPABILITY of humans to recognize and show emotions is an essential skill for social and cognitive development [1]–[4]. Studies on the development of children's understanding of emotion contribute to a greater understanding of the cognitive and affective systems involved [5]. Thus, effective social interaction between people, and between people and robots, depending on the capability to perceive emotions and infer the states and desires of the other agents. Moreover, many developmental disorders are related to cognitive impairment resulting from poor emotional development, such as autism, depressive disorder, and antisocial behavior [6]–[11]. Studies have demonstrated that niche targeting emotion cognition interventions for these people are beneficial to their symptoms. For example, the cognitive mental deficit and the stereotyped behaviors of autistic patients can be modified by boosting the cognitive ability of emotion recognition after good training [12], [13]. The depressive disorder can be treated by cognitive behavior therapy (CBT) [14].

Facial expression recognition (FER) is a crucial technology in emotion perception. It provides a possible way to explore the emotion cognition systems of humans and is related to the perceived quality of human–computer interactions in the rehabilitation system [15]. FER aims to predict disparate facial expressions of human beings from their facial images. Then, it could infer human emotions from their facial expressions and then assist the computer in making appropriate responses. FER has recently gained increasing attention in cognitive psychology and artificial intelligence [16]–[18] due to its potential applications in human–robot interactions, health management, abnormal behavior detection, autonomous driving, and related areas.

FER can be achieved by detecting facial regions [19], extracting expression-related features, and recognizing different expressions [20], [21]. Among different stages, extracting expression-related features, including geometric and appearance features, plays the most important role. For geometric features, facial landmark points are detected and then encoded into informative feature vectors [22]. For appearance features, holistic spatial analysis [23] is used to capture appearance information. Convolutional neural networks (CNNs) could automatically extract expression-related features [24]. Recently, temporal variations of facial image sequences are used to capture dynamic expression changes [25].

Despite the increasing attention on FER, it remains challenging due to various disturbances, such as illumination/pose variations and head deflections. These disturbances hamper the researchers from extracting robust expression-related features from facial RGB images. Therefore, the recognition accuracy

Manuscript received 23 July 2021; revised 22 December 2021; accepted 3 February 2022. Date of publication 9 February 2022; date of current version 13 March 2023. This work was supported in part by the National Key Research and Development Program under Grant 2018AAA0100800; in part by the Key Research and Development Program of Jiangsu under Grant BK20192004 and Grant BE2018004-04; in part by the International Cooperation and Exchanges of Changzhou under Grant CZ20200035; in part by the National Postdoctoral General Fund under Grant 2021M701042; in part by the Postdoctoral Foundation of Jiangsu Province under Grant 2021K187B; in part by the Changzhou Sci&Tech Program under Grant CJ20210052; and in part by the State Key Laboratory of Integrated Management of Pest Insects and Rodents under Grant IPM1914. The work of Angelo Cangelosi was supported in part by the H2020 Projects PERSEO, TRAINCREASE, and eLADDA. *(Corresponding author: Xiaofeng Liu.)*

Rongrong Ni, Xu Zhou, and Xiaofeng Liu are with the College of IoT Engineering, Hohai University, Changzhou 213000, China (e-mail: xfliu@hhu.edu.cn).

Biao Yang is with the College of IoT Engineering, Hohai University, Changzhou 213000, China, and also with the School of Microelectronics and Control Engineering, Changzhou University, Changzhou 213000, China.

Angelo Cangelosi is with the Cognitive Robotics Laboratory, The University of Manchester, Manchester M13 9PL, U.K.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2022.3150019.

Digital Object Identifier 10.1109/TCDS.2022.3150019

is low because of the unfaithful features. Other facial modalities, e.g., facial depth images, are robust to illumination/pose variations. However, current fusion strategies of multimodality facial images are rough and could not fully utilize their complementarity. Except for the problems mentioned above, it is also difficult to recognize facial expressions with weak emotion intensities or indistinct visual appearances.

In order to solve the above problems, we propose a cross-modality attention-based CNN (CM-CNN) based on facial RGB and depth images. Facial depth images are introduced to handle illumination/pose variations. Besides, the local binary pattern (LBP) components are calculated from facial RGB images to capture facial details, which could handle indistinct visual appearances. Compared with other commonly used textural features like the histogram of oriented gradients (HOG) [26] and Gabor texture [27], LBP could better capture facial profiles, thus is more suitable for tasks, such as face recognition and FER. Then, three specially designed feature extraction networks (FENs) are presented to perform automatic feature engineering from different facial modalities. Rather than directly concatenating extracted features, a novel cross-modality attention fusion network (CMFN) is proposed by enhancing the spatial correlations between different facial modalities. Finally, CM-CNN is optimized with an improved focal loss, which pays more attention to facial expressions with weak emotion intensities. In conclusion, our main contributions are threefold as follows.

1) Three complementary modalities of facial images are introduced to perform precise FER. In addition to facial RGB images, facial depth images are introduced to handle illumination/pose variations. Facial LBP images are calculated to capture facial details. Besides, three specially designed FENs are proposed to perform automatic feature extraction from different modalities.
2) Rather than directly concatenating extracted features, a novel CMFN is proposed to fuse features extracted by different FENs. A spatial attention mechanism is introduced to enhance spatial correlations between different facial modalities. Then, the complementarity among different modalities could be fully utilized to perform robust and precise FER.
3) To pay more attention to facial expressions with weak emotion intensities, CM-CNN is optimized with a focal loss. Rather than using the original focal loss which forces the classifier to focus on samples with significant training losses, the focal loss is improved by using the soft label.

RGB
depth image
LBP

Evaluations are performed on the VT-KFER, BU-3DFE, and Bosphorus databases. State-of-the-art performance is achieved in most cases. Besides, we propose a new database that consists of facial RGB&Depth (RGBD) images, namely, the CCZU-FER, to test our method in laboratory conditions. All images in the new database are captured by a Kinect 2.0. Fig. 1 shows the six basic facial expressions of two subjects in the database. It is challenging to recognize different expressions shown in this figure using the facial depth images only because they lack details. However, facial RGBD images comprise complementary information, thus can be beneficial for FER. For each given subject, the expression class is labeled by ten well-educated adults through voting. Then, we use the class that gets the most votes.

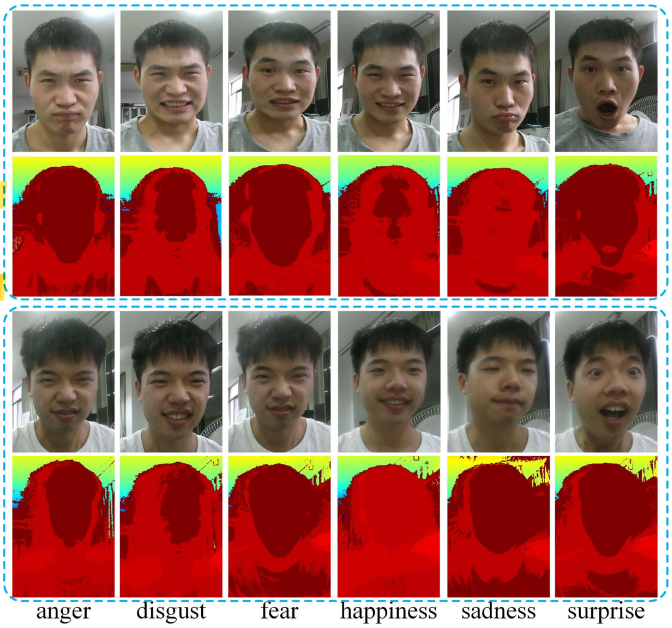


Fig. 1. Six basic facial expressions of two subjects in the proposed database. Expressions belonging to the same subject are enclosed by a blue rectangle. In each rectangle, the first and second rows represent facial RGB and depth images, respectively. A thermodynamic diagram is used to display facial depth images clearly.

The remainder of this work is organized as follows. Section II reviews representative FER approaches. Section III discusses the proposed method. The evaluation results and analysis are discussed in Section IV. Section V presents the conclusions.

II. Related Work

A. Expression-Related Feature Engineering

Feature engineering is crucial in FER. Early works typically utilize handcrafted features, such as pixel intensity [28], LBP [29], Gabor texture [27], and HOG [26]. However, these features are insensitive to variations in local facial regions, which are critical for precise FER. Therefore, different geometric features [30], [31], as well as facial action units (AUs) [32], are extracted via encoding locations of facial landmark points [33].

With the recent success of deep learning techniques in various fields, deep neural networks (DNNs) have increasingly been leveraged to learn discriminative representations for automatic FER [34]–[36]. In the early stage, multilayer perceptrons (MLPs) [37], [38] are used to extract expression-related features and perform expression recognition. Then, CNNs [39], [40] are commonly used to extract features from facial images automatically. Multiple CNNs are simultaneously trained, for robust FER, and the recognition results are output in a weighted manner [41]. Mollahosseini *et al.* [42] used CNNs with inception layers to extract deeper features. CNNs are also jointly used with AUs, namely, AU-inspired DNNs, to explore subtle facial expressions [43]. Recently, Sanchez *et al.* [44] introduced

a stochastic autoencoder to extract expression-related features. Zhang *et al.* [45] proposed a weakly supervised local–global relation network for FER. Siqueira *et al.* [46] explored the effects of wide ensemble-based CNNs for efficient feature engineering. Xie *et al.* [47] integrated adversarial learning and graph representation to extract expression-related features suitable for adaptation. Wang *et al.* [48] presented phase space reconstruction to extract spatiotemporal features and then performed dynamic FER. However, all these methods extract expression-related features from facial RGB images sensitive to illumination/pose variations.

B. Multimodality Fusion

Facial RGB images are sensitive to illumination/pose variations. Besides, different emotional intensities may lead to indistinct visual appearances of facial expressions. Therefore, many researchers introduce other facial modalities, e.g., facial depth, LBP [49], and Gabor [27] images. Aly *et al.* [50] presented a multimodality feature fusion framework using dual kernel discriminant analysis. Fu *et al.* [51] combined facial 2-D and 3-D data into a 4-D tensor and then recognized different expressions through low-rank tensor completion.

For deep-learning-based fusion, multichannel CNNs are commonly used to fuse features automatically extracted from different facial modalities [52]. For example, Li *et al.* [53] generated 2-D texture and 3-D geometry descriptors around detected 2-D and 3-D facial landmarks. Then, they fused 2-D and 3-D descriptors based on both feature-level and score-level fusion strategies. Li *et al.* [54] generated six 2-D facial attribute maps from a 3-D scan and then fed all attribute maps into a multichannel CNN. Zeng *et al.* [55] proposed a deep fusion CNN to learn from four local regions: 1) the eyebrows; 2) eyes; 3) mouth; and 4) nose, from facial gray-scale and depth images. Expression recognition is achieved by using a nonlinear SVM. Deng *et al.* [56] integrated spatial and temporal streams to perform both macro- and micro-FER. Zhao *et al.* [57] proposed an end-to-end visual–audio attention network to recognize emotion from videos. Behzad *et al.* [58] computed several attributes from 4-D data of face scans. Then, they generate cross-domain dynamic images via rank pooling that encapsulates facial deformations over time to perform accurate FER. In conclusion, FER based on multimodality fusion is becoming a mainstream method since there are abundant complementarities among different modalities [54]. However, most works perform multimodality fusion by directly concatenating feature vectors extracted from different modalities, thus not fully utilizing their complementarities.

C. Loss Functions

The cross-entropy loss is a commonly used loss function for DNNs. It pays equal attention to all samples. Many studies have attempted to improve the cross-entropy loss. For example, large margin loss is proposed to enhance the discriminative learning of features and then improves the classification performance [59]. The noisy robust cross-entropy loss is used to handle noisy labels [60]. The center loss is presented to encourage the discriminative features by constraining the distances between the learned features and the corresponding class centers [61]. The triplet loss is used in face recognition [62]. It attempts to reduce the intraclass variance while increasing the interclass variance.

The focal loss is an improved version of the cross-entropy loss. It is originally designed to handle the one-stage object detection problem in which extreme imbalance between foreground and background classes exists [63]. The focal loss could pay more attention to samples that are hard to recognize by adjusting the hyperparameter. Therefore, we employ the focal loss and improve it by using soft labels.

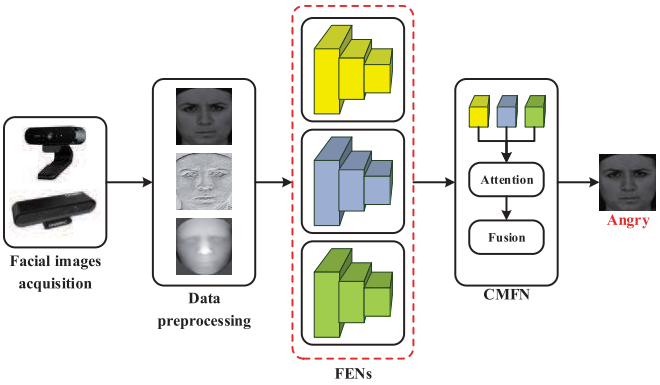


Fig. 2. Flowchart of the proposed CM-CNN approach.

III. Proposed Method

A. Flowchart

Fig. 2 illustrates the flowchart of the proposed method. Input data are facial RGB and the corresponding depth images (or 3-D scans) captured by regular or 3-D cameras. Then, data preprocessing is conducted to prepare the data fed into the FEN. Specifically, we initially conduct face detection and rotation rectification in facial RGB images. Then, we convert facial RGB images into gray-scale versions and generate facial LBP images in detected facial regions. Besides, we calculate the corresponding facial depth images in these regions if 3-D scans are provided. Subsequently, we extract expression-related features from three types of facial images with specially designed CNN-based FENs. Later, a CMFN is proposed to fuse features extracted from different types of facial images. CMFN enhances the spatial correlations between any two types of facial images by introducing the spatial attention mechanism. Finally, expression recognition is performed by optimizing the CM-CNN with an improved focal loss.

B. Data Preprocessing

1) Face Detection: FER is performed on images that contain faces and other elements (e.g., bodies and different background items). Thus, it is important to localize facial regions from background regions, which are uncorrelated to expression recognition. The performance of expression recognition depends on the face detection accuracy. Face detection is not reviewed thoroughly in this study for the sake of brevity. Nevertheless, readers can refer to [64]. We use the

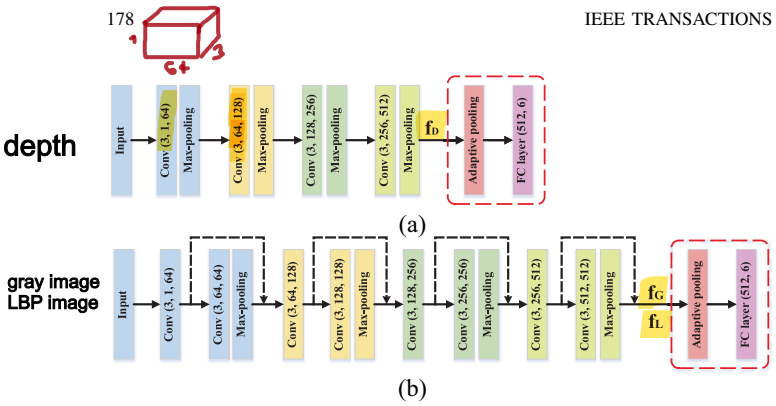


Fig. 3. Structure of the specially designed FENs.

V-J detector [65] to detect facial regions. Such regions are used in corresponding facial LBP and depth images. The proposed FENs only extract expression-related features from facial regions to avoid background disturbance.

2) Rotation Rectification: Detected facial regions may vary in angles. These variations in angles are uncorrelated to expression changes and may influence recognition performance. To solve this problem, all detected facial regions are aligned by using rotation rectification via the following rotation transformation matrix:

$$(x', y', 1) = [x, y, 1] \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where (x, y) and (x', y') refer to the original and after rotation rectification coordinates, respectively. θ denotes the rotation angle between the line connecting two eye centers and the horizontal axis. DRMF [66] is used to localize both eyes.

3) LBP Components Calculation: LBP [67] is a commonly used descriptor to explore the textural details. A traditional LBP descriptor of a given pixel is a 0/1 sequence. In each bit, the value is calculated by comparing the values of the given and adjacent pixels. LBP components could be calculated as follows:

$$\text{LBP} = \sum_{n=1}^{N} S(\boldsymbol{g_n} - \boldsymbol{g_c}) \times \boldsymbol{2^n} \tag{2}$$

where $S(\cdot)$ represents the sign function, and $\boldsymbol{N}$ denotes the number of adjacent pixels. $\boldsymbol{g_c}$ and $\boldsymbol{g_n}$ refer to the values of the given and adjacent pixels, respectively.

4) Facial Depth Images Calculation: The facial depth images should be calculated in the detected facial regions if facial 3-D scans are provided. In this work, the facial depth image is generated by fitting a surface from the 3-D point cloud (provided by the facial 3-D scans) in the form of $z(x,y)$ using the gridfit algorithm [68]. To process the facial RGBD images captured by Kinect 2.0, we only introduce the facial depth images whereas many other methods introduce facial azimuth, elevation, and curvature maps, which are calculated based on 3-D scans.

C. FENs Structures

FENs structures for different facial modalities are specially designed based on their properties. Details of the FENs, including the channel numbers and kernel sizes, are illustrated in Fig. 3.

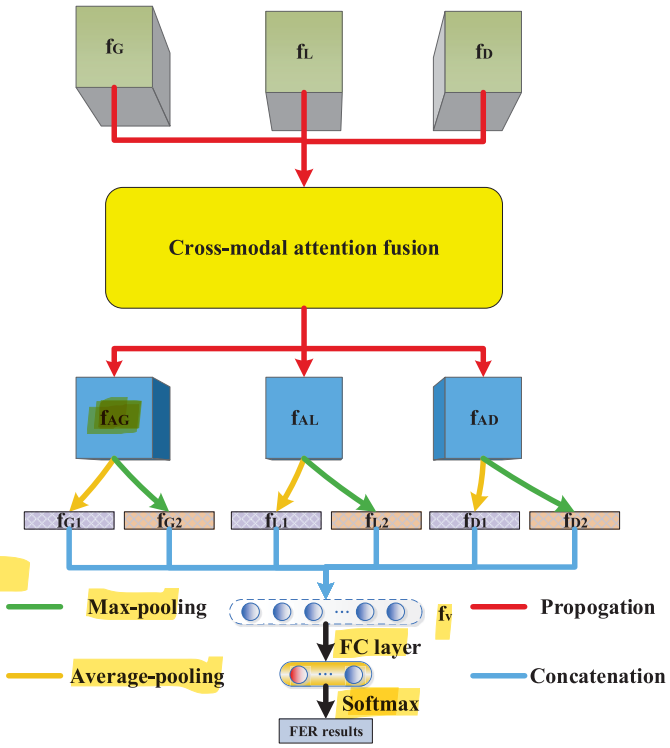


Fig. 4. Structure of the CMFN. $\boldsymbol{f_G}$, $\boldsymbol{f_L}$, and $\boldsymbol{f_D}$ represent the features extracted by FENs from gray, LBP, and depth facial images/regions, respectively. $\boldsymbol{f_{AG}}$, $\boldsymbol{f_{AL}}$, and $\boldsymbol{f_{AD}}$ represent the fused feature of gray, LBP, and depth facial images/regions, respectively.

The FEN structure for facial depth is similar to DeepID [69], a shallow CNN-based face detection framework. As shown in Fig. 3(a), the network contains four convolutional layers, and each convolutional layer is followed by batch normalization, PReLU activation, and max pooling. In each layer, Conv($\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$) represents the convolutional operation. $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, and $\boldsymbol{x_3}$ represent the convolutional kernel size, the number of input channels, and the number of output channels, respectively. FC layer($\boldsymbol{x_4}$, $\boldsymbol{x_5}$) means the fully connected operation. $\boldsymbol{x_4}$ and $\boldsymbol{x_5}$ represent the number of input neurons and the number of output neurons, respectively. A shallow network is introduced to extract information from facial depth images since there exist fewer details compared with the other two modalities. For facial gray-scale and LBP images, a deeper network is presented to extract texture details. As shown in Fig. 3(b), the network is deepened by adding four convolutional layers. Skip connections are used to avoid the problem of vanishing gradient. Notably, the classification heads in red rectangles are used to determine the parameters of different FENs. In the classification head, the output of the last max-pooling layer ($\boldsymbol{f_G}$, $\boldsymbol{f_L}$, or $\boldsymbol{f_D}$) is fed into an adaptive pooling layer and then is fed into the fully connected (FC) layer to perform FER from a single facial modality. Then, FENs' parameters (mainly channel numbers) are determined by cross-validation on different benchmarks.

D. CMFN Structure

Fig. 4 illustrates the structure of the proposed CMFN. CMFN fuses features extracted from three FENs and outputs

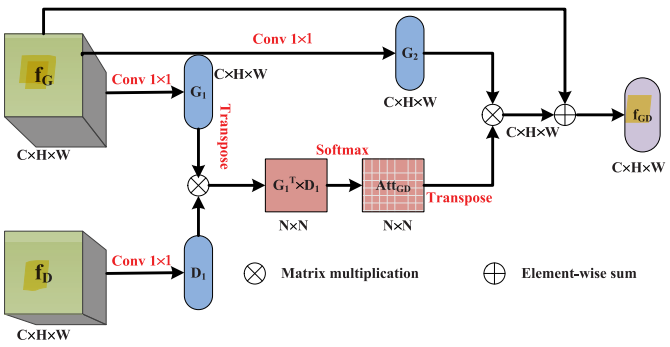


Fig. 5. Details of the cross-modality attention fusion module.

the final recognition results. To better utilize the complementary facial images, we propose a CMFN instead of directly fusing features by concatenation. CMFN is inspired by SAGAN [70], which performed self-attention for improved performance. Unlike SAGAN, cross-modal attention is calculated based on different modalities to enhance their spatial correlations. Details of the cross-modal attention fusion network could be found in Fig. 5. CMFN is beneficial to multimodal FER based on the following finding: we assume that some local facial regions are highly related to expression changes. However, these regions may present different activation degrees after convolutions due to the different properties of input data, such as facial gray-scale image's sensitivity to illumination changes. Without cross-modal attention, these regions may be neglected by the model since expressionless activations may cover up the weak activations. However, these regions could be enhanced with the cross-modal attention since strong activations could be observed in another complementary modality. The cross-modal attention fusion network outputs three embeddings of different modalities. Then, both max-pooling and average-pooling are used to vectorize the three embeddings. Such an operation is inspired by CBAM [71], which claims that using both max-pooling and average-pooling could significantly improve the representation power of networks rather than using each independently. Finally, the vectorized results are concatenated into the final embedding f_v, which is used for classification. Notably, we perform hyperparameter tuning on benchmarks, such as BU-3DFE and Bosphorus.

Fig. 5 depicts the details of the cross-modal attention fusion network. Given an input $f_G \in \mathbb{R}^{C\times H\times W}$ (where C, H, and W represent the channel number, the height, and the width of feature maps, respectively), we first feed it into two 1×1 convolutional layers to generate G_1 and G_2, respectively, where $\{G_1, G_2\} \in \mathbb{R}^{C\times H\times W}$. We then feed another input $f_D \in \mathbb{R}^{C\times H\times W}$ into another 1×1 convolutional layer to generate $D_1 \in \mathbb{R}^{C\times H\times W}$. We reshape $\{D_1, G_1, G_2\}$ to $\mathbb{R}^{C\times N}$, where $N = H \times W$ is the number of pixels. We conduct a matrix multiplication between G_1^T and D_1, and the Softmax function $\sigma(\cdot)$ is introduced to calculate the attention weight $Att_{GD} \in \mathbb{R}^{N\times N}$, which explores the correlations between f_G and f_D as follows:

$$Att_{GD} = \sigma\left(G_1^T \times D_1\right), \text{ where } \sigma(x) = \frac{1}{1+e^{-x}}. \tag{3}$$

Then, we conduct matrix multiplication between G_2 and Att_{GD}^T and reshape the result to $\mathbb{R}^{C\times H\times W}$. Subsequently, we perform an elementwise sum between the result and the input feature f_G to generate the final output $f_{GD} \in \mathbb{R}^{C\times H\times W}$. f_{GD} carries information extracted from facial gray-scale images with the guide of facial depth images. Notably, f_{GD} and f_{DG} carry different information. Similarly, f_{GL} carries information extracted from facial gray-scale images with the guide of facial LBP images. Further, the embedding of facial gray-scale regions f_{AG} is generated via concatenation of f_{GD} and f_{GL}. Other embedding, namely, f_{AL} and f_{AD}, are generated in a similar way. Then, we vectorize the three embeddings by conducting max-pooling and average-pooling as proposed in CBAM [71]. Afterward, all generated vectors are concatenated into the final embedding f_v. Such a concatenation would not weaken the effects of the former fusion strategy since the aggregated feature vectors already contain complementary facial information by performing cross-modal attention fusion. Finally, FER is conducted as follows:

$$F = \sigma(PRelu(Fc1(f_v))) \tag{4}$$

where $Fc1(\cdot)$ represents the fully connected layer that converts f_v into a 6-D vector. PReLU is the activation function.

E. Focal Loss With Soft Label

The categorical cross-entropy loss is a common choice for classification problems such as FER. Such a loss function pays equal attention to all samples. However, there exist samples that are hard to recognize due to different emotional intensities and visual appearances. To handle the problem, we used the focal loss [63], which is defined as follows:

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{5}$$

where p_t refers to the estimated probability distribution of sample t. This work introduces the soft label by replacing the traditional one-hot label Y with a probability distribution p_t to improve the classification performance. Specifically, the probability distribution p_t is a vector of Gaussian distribution with mean μ equals to the hard labels (0–5 represent six basic expressions), and standard deviation σ equals to 0.1. σ is determined through cross-validation on benchmarks. $(1-p_t)^{\gamma}$ represents the modulating factor added to the traditional cross-entropy loss. γ is a tunable focusing parameter used to adjust the weights between samples easy to recognize and samples hard to recognize. Finally, the model is forced to consider more about samples hard to recognize. γ is set to 2 according to [63]. α in original focal loss is used to balance positive and negative samples in the training set and is manually adjusted based on the ratio of positive and negative samples. In our multiclass classification task, there are no positive or negative samples. Hence, α is set to 1.

IV. Experimental Results

A. Databases and Configuration

We evaluate the proposed method on three benchmarking databases. Besides, we test our method on a novel database, namely, CCZU-FER, which consists of facial RGBD images collected by a Kinect 2.0 sensor. The proposed model is built with a Pytorch framework and is trained with an Intel I7 CPU

and an NVIDIA GTX-1080 GPU. Database descriptions are presented as follows.

1) VT-KFER [72]: A Kinect 1.0 sensor is used to capture facial RGBD images in scripted and unscripted scenarios. VT-KFER contains expressions performed in frontal, right, and left poses with 4732 frontal and 5066 nonfrontal frames. According to [72], the emotions of subjects are evoked using verbal, image, and live demonstrations. Following [50], we select 4221 frontal samples, including 2129 and 2092 frames evoked by verbal and image demonstrations, respectively.

2) BU-3DFE [73]: BU-3DFE contains 2500 3-D facial scans of 100 subjects, where 44 and 56 are males and females, respectively. Each subject displays a neutral and six basic facial expressions with four intensity levels. For a fair comparison, we use the same evaluation protocol proposed in [74]. Specifically, 60 subjects are selected and six basic facial expressions of the two highest intensity levels (Protocol I and Protocol II) are chosen for evaluation. For each protocol, $60 \times 6 = 360$ samples are used for training and testing. Notably, all methods used for comparisons employ the same rule.

3) Bosphorus [75]: Bosphorus consists of 4666 textured 3-D facial scans of 105 subjects in various facial expressions, poses, and occlusions. To perform the identity-independent FER fairly, we use Protocol I as used in evaluating the BU-3DFE. Specifically, 60 subjects displaying six basic facial expressions are randomly selected. Then, $60 \times 6 = 360$ samples are used for training and testing. Notably, all methods used for comparisons employ the same rule.

4) CCZU-FER: CCZU-FER is an RGBD FER database collected in the laboratory by using a Kinect 2.0 sensor. RGB images and their corresponding depth images are aligned by the "MapDepthToColorImagePoint" provided by the Kinect software development kit. The motivation to create CCZU-FER is to propose an RGB-D FER data set, which only consists of Asian people, whereas the state-of-the-art data sets contain many other ethnicities. Hence, CCZU-FER would be beneficial to develop 3-D FER systems in China, as well as to boost 3-D FER research by making it publicly available in the future. In this data set, a total of 1960 annotated samples are captured from 15 male and 13 female subjects aged from 18 to 32. For each subject, ten samples are collected for each expression (six basic facial expressions and one neutral) with partial occlusions or head deflections.

For three benchmarking databases, a tenfold cross-validation strategy is used to evaluate the proposed method and methods chosen for comparison. Besides, we train the model on all three benchmarking databases, and test the well-trained model on CCZU-FER to verify its generalization performance. We use the average accuracy as the evaluation criteria.

B. Implementation Details

Each type o facial image is resized to 148×148 pixels and then is normalized to (0, 1). The SGD optimizer with a learning rate, momentum, and weight decay of 0.01, 0.9, and 5×10^{-4} is used to optimize the network, respectively. The learning rate is reduced by 0.5 for every 100 epochs, and 300 epochs are used in each evaluation. The weights of FEN for facial gray-scale and LBP images are pretrained on the publicly available database CASIA-WebFace [76]. The weights of FEN for facial depth images are pretrained on the database proposed by [77].

TABLE I
COMPARISON OF DIFFERENT LOSS FUNCTIONS ON THE BENCHMARKING DATABASES [ACCURACY (%)]. CCE IS THE ABBREVIATION OF CATEGORICAL CROSS ENTROPY. SOFT FOCAL LOSS REPRESENTS THE FOCAL LOSS WITH THE SOFT LABEL. BOLD FONTS INDICATE THE BEST PERFORMANCE

Method	VT-KFER	BU-3DFE(P1/P2)	Bosphorus
CCE loss	91.52	87.88 / 85.50	83.38
Focal loss (γ=1.5)	91.80	87.31 / 85.52	84.22
Soft Focal loss (γ=1.5)	92.50	87.33 / 85.85	84.32
Focal loss (γ=2.5)	91.86	88.12 / 85.61	84.58
Soft Focal loss (γ=2.5)	92.36	88.68 / 86.21	84.98
Focal loss (γ=2.0)	92.52	88.38 / 86.02	84.76
Soft Focal loss (γ=2.0)	**93.86**	**88.91 / 86.28**	**85.16**

C. Evaluation of the Improved Focal Loss Function

As mentioned above, we replace the categorical cross-entropy loss with an improved focal loss to focus more on expressions hard to recognize. We set the hyperparameter $\boldsymbol{\alpha}$ to 1 since we select equal numbers of samples for each expression. Hyperparameter $\boldsymbol{\gamma}$ is initially set to 2 according to [63]. Then, we compare the performance of the categorical cross-entropy loss and our improved focal loss with different $\boldsymbol{\gamma}$ values on three benchmarking databases. As presented in Table I, performance on all three databases is improved by using the focal loss, except for the case of P1 when setting $\boldsymbol{\gamma}$ to 1.5. Compared with the original focal loss, introducing the soft label into the focal loss could boost FER accuracy. A possible reason is that some facial expressions are hard to distinguish. Thus, a soft label may be better than a hard label. Besides, we achieve the best performance by setting $\boldsymbol{\gamma}$ to 2.

D. Computational Time Analysis

The computational time is a critical issue for real-time FER. The time for once forward computation is 15 ms with a 148×148 input facial region. The computational time for the data preprocessing is about 28 ms. Therefore, the model could be run in about 23 frames/s. It could satisfy the needs of many real-time FER applications. After optimization in both time and space, the model could be equipped in real-time edge devices, such as an NVIDIA TX2.

E. Confusion Matrices on Different Databases

We use the confusion matrices to exhibit our method's recognition performance on different databases. For example, the first row in Fig. 6(a) indicates that 96.2%, 2.6%, and 1.2% of anger samples are correctly recognized, identified as fear ones, and considered sadness, respectively. As revealed by the confusion matrices on four databases, CM-CNN performs well in recognizing facial expressions like anger, happiness, and surprise. All these expressions are accompanied by evident appearance changes around the regions of the mouth, eyebrows, and eyes. In contrast, disgust and fear expressions are easily confused. The recognition performance of the sadness expression is the

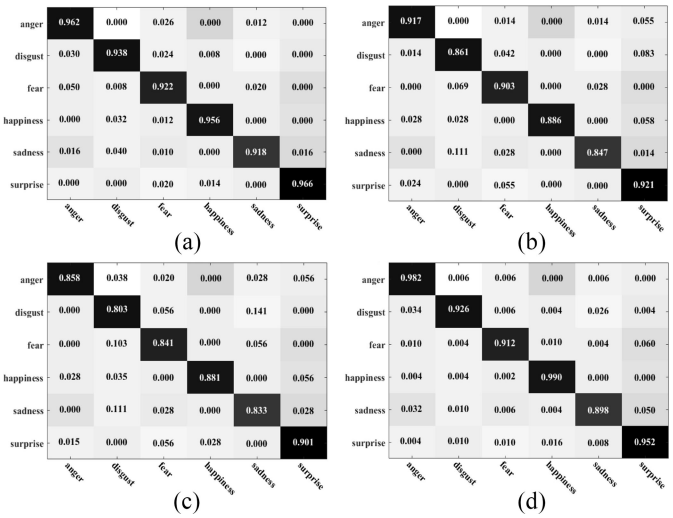


Fig. 6. Confusion matrices on different databases. (a) Confusion matrix of VT-KFER database. (b) Confusion matrix of BU-3DFE database. (c) Confusion matrix of Bosphorus database. (d) Confusion matrix of CCZU-FER database.

TABLE II
ABLATION STUDY ON THE BU-3DFE DATABASE. G REPRESENTS GRAY-SCALE, L REPRESENTS LBP, H REPRESENTS HOG, B REPRESENTS GABOR, AND D REPRESENTS DEPTH

Method	Accuracy (%)	
	P1	P2
G	84.04	82.33
L	78.48	77.14
H	71.32	69.84
B	75.66	74.68
D	80.11	79.02
G + L	84.13	82.36
G + L + CMFN	86.42	84.81
G + D	85.08	82.82
G + D + CMFN	86.65	84.86
L + D	80.93	79.46
L + D + CMFN	81.72	80.18
G + L + D	86.07	84.38
G + H + D + CMFN	84.71	83.38
G + B + D + CMFN	86.86	85.48
G + L + D + CMFN	**88.91**	**86.28**

worst among all six basic expressions. Moreover, it is challenging for CM-CNN to capture subtle expression changes in the sadness expression. Such a finding is consistent with daily experience, that is to say, it is challenging for us to identify whether a person is sad just from their facial expression.

F. Ablation Study

The proposed CM-CNN simultaneously processes three types of facial images to utilize their complementarity for precise FER. To verify its effectiveness, we conduct ablation studies on the BU-3DFE and Bosphorus databases. Specifically, we calculate the average accuracies by feeding the model with different inputs. We also calculate the average accuracies by using CMFN or directly concatenating $\boldsymbol{f_G}$, $\boldsymbol{f_L}$, and $\boldsymbol{f_D}$. Here, G, L, and D represent gray-scale, LBP, and depth, respectively. As presented in Tables II and III, using facial gray-scale images alone outperforms using facial LBP or depth images on both BU-3DFE and Bosphorus. Using LBP as texture features is superior to using HOG or Gabor texture, verifying our point that LBP is better at capturing facial details. Such a point is further verified by comparing the performance of the whole framework but using different texture features. Using more facial modalities always leads to better recognition performance. However, directly concatenating different facial modalities only achieves limited performance gains. Benefitting from the cross-modality attention, CMFN improves the recognition performance of using multiple facial modalities compared with directly concatenating them.

TABLE III
ABLATION STUDY ON THE BOSPHORUS DATABASE. G REPRESENTS GRAY-SCALE, L REPRESENTS LBP, H REPRESENTS HOG, B REPRESENTS GABOR, AND D REPRESENTS DEPTH

Method	Accuracy (%)
G	81.23
L	76.15
H	70.12
B	74.22
D	77.82
G + L	81.86
G + L + CMFN	83.41
G + D	82.00
G + D + CMFN	83.58
L + D	78.21
L + D + CMFN	79.68
G + L + D	83.26
G + H + D + CMFN	83.28
G + B + D + CMFN	84.22
G + L + D + CMFN	**85.16**

G. Visualization of Classification Performance

Fig. 7 presents a 2-D t-distributed stochastic neighbor embedding (tSNE) visualization [78] of the learned high-dimensional embedding fv on BU-3DFE (P1 and P2). The operation to convert the embedding fv into 2-D feature embedding by using tSNE is inspired by [79], which visualizes the last embedding of the learned network since that embedding contains information for classification. Fig. 7(a)–(f) shows the comparisons of different methods in terms of separating distances, which can be used as an indicator of classification performance. As shown in the figure, the method using G + L + D + CMFN has the largest separating distance. The 2-D feature encodings of different expressions are separated from each other, except for angry and sadness expressions. The method using G + L + D has the second-largest separating distance. The separating distances of methods using G + D and G + L both are larger than those of methods using G and D alone, which reveals the effects of complementary facial modalities in FER. All visualization results are consistent with the ablation study results.

H. Visualization of the Attention Maps

In Fig. 8, we visualize the attention maps using the commonly used Grad-CAM [80]. Colors from blue to red represent the increased attention on certain facial regions. Fig. 8(b) and (e) are attention maps generated using facial gray-scale images. Specifically, we only use the FEN structure presented in Fig. 3(b). Attention maps in these two subfigures indicate that the proposed FER could extract expression-related

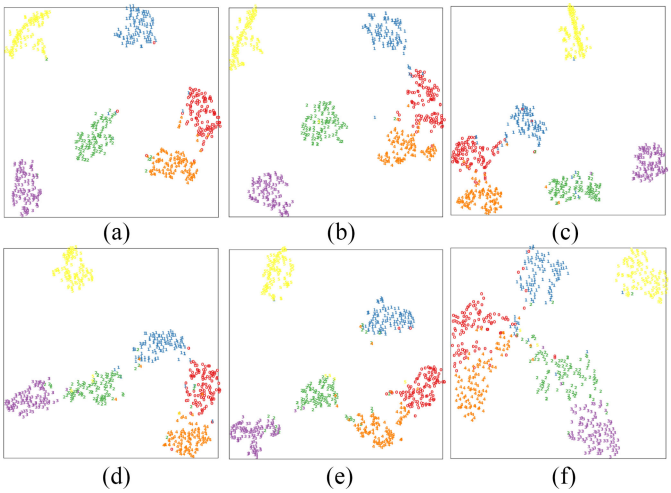


Fig. 7. tSNE visualization of BU-3DFE using (a) G + L + D + CMFN, (b) G + L + D, (c) G + D, (d) G + L, (e) G, and (f) D. Zero to five represent angry (red), disgust (blue), fear (green), happiness (purple), sadness (orange), and surprise (yellow), respectively.

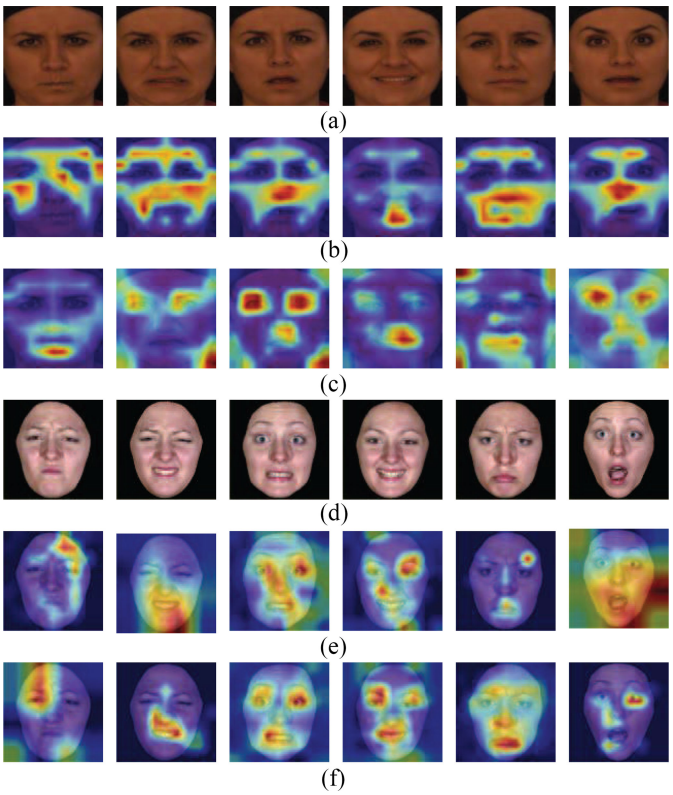


Fig. 8. Visualization of the attention maps generated on samples from the (a) Bosphorus and (d) BU3DFE data sets. Expressions from left to right are anger, disgust, fear, happiness, sadness, and surprise, respectively. Colors from blue to red represent the increased attention on certain facial regions.

features from the mouth, eyes, and nose regions. Although attention maps in Fig. 8(b) and (e) spread over the facial regions, there still exist shortages that they are not concentrated enough. Besides, some key regions, such as the eye regions in Fig. 8(b), are not attended to. The proposed method could solve these problems by effectively fusing complementary facial images. Fig. 8(c) and (f) are attention maps generated by using the proposed method. Attention maps are more concentrated than using facial gray-scale images only.

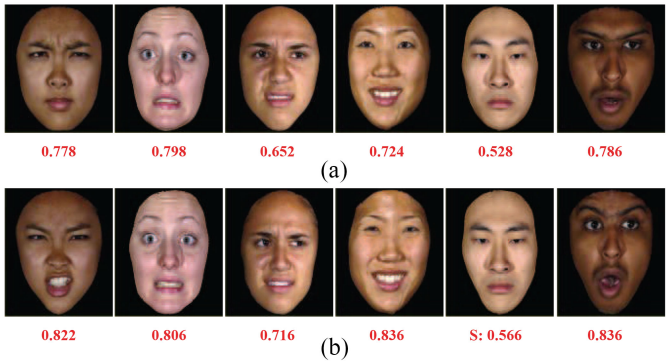


Fig. 9. Samples selected from (a) BU-3DFE(P2) and (b) BU-3DFE(P1) with expressions angry, fear, disgust, happiness, sadness, and surprise. Numbers below the images denote the prediction scores predicted by the proposed method.

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE VT-KFER DATABASE

Method	Feature	Accuracy (%)
Aly *et al.* (2016) [50]	GIST / HOG / LBP	80.00
Rivera *et al.* (2013) [81]	LDN code	82.53
Li *et al.* (2015) [53]	shape / texture	86.94
Lopes *et al.* (2017) [39]	CNN features	88.61
Zhang *et al.* (2015) [31]	CNN features	89.22
CM-CNN	CNN features	**93.86**

We could observe more activation on eyes, mouth, and nose regions, which are significant for FER.

I. Case Study

Fig. 9 shows a case study in which samples with six basic expressions are randomly selected from BU-3DFE (P1) and (P2). According to the officially segmented rules, subjects in Fig. 9(b) have stronger emotional intensity levels than subjects in Fig. 9(a). In Fig. 9, all subjects' facial expressions are correctly recognized and the prediction scores predicted by the proposed method are given in red fonts. It is obvious that the proposed method could predict a higher score if the subject shows a stronger emotional intensity level. The results indicate the effects of emotional intensity levels in FER. Among different expressions, the sadness and disgust expressions have the lowest and the second-lowest prediction scores, which reveal the challenges in recognizing these two expressions. Even a human observer is hard to recognize these expressions precisely. Such results are similar to the confusion matrix results.

J. Comparison With State-of-the-Art Approaches

We compare the proposed method with several state-of-the-art approaches. As presented in Table IV, methods using CNN features significantly outperform handcrafted feature-based approaches [50], [53], [81] on the VT-KFER database. Compared with single-channel CNN-based FER [39], our method achieves better performance since the introduction of complementary facial images. Besides, our method is better than Zhang *et al.* [31], which proposed a multichannel CNN for feature extraction. Such a superiority is achieved due to the CMFN and the improved focal loss, which could better

TABLE V
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE BU-3DFE DATABASE

Method	Feature	Accuracy (%)	
		P1	P2
Zeng *et al.* (2013) [82]	LBP / MCI / CFI	N / A	70.93
Li *et al.* (2012) [83]	normals / LBP	N / A	80.14
Zhen *et al.* (2015) [84]	shape / geometry	84.50	83.20
Yang *et al.* (2015) [85]	scattering features	84.40	82.73
Li *et al.* (2015) [53]	shape / texture	81.22	80.64
Zhang *et al.* (2016) [40]	SIFT / CNN features	80.10	N / A
Qyedotun *et al.* (2017) [86]	CNN features	84.72	N / A
Li *et al.* (2017) [54]	CNN features	86.86	N / A
Chen *et al.* (2018) [87]	CNN features	86.67	85.96
Wei *et al.* (2018) [88]	CNN features	88.03	N / A
Yang *et al.* (2018) [49]	CNN features	84.17	N / A
CM-CNN	CNN features	**88.91**	**86.28**

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE BOSPHORUS DATABASE

Method	**Feature**	**Accuracy (%)**
Ujir *et al.* (2014) [89]	surface normals	63.63
Li *et al.* (2012) [83]	normals / LBP	75.83
Fu *et al.* (2019) [51]	curvatures / LBP	75.93
Li *et al.* (2015) [53]	shape / texture	79.72
Yang *et al.* (2015) [85]	scattering features	77.50
Wei *et al.* (2018) [88]	CNN features	82.50
CM-CNN	CNN features	**85.16**

TABLE VII
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE CCZU-FER DATABASE

Method	**Feature**	**Accuracy (%)**
Fu *et al.* (2019) [51]	curvatures / LBP	88.33
Li *et al.* (2015) [53]	shape / texture	89.62
Wei *et al.* (2018) [88]	CNN features	92.68
CM-CNN	CNN features	**94.33**

utilize the complementary features and pay more attention to samples hard to recognize.

Table V presents the comparisons on the BU-3DFE database. Similar to results in Table IV, CNN features achieve better performance than handcrafted features. Among approaches with CNN features, ours outperforms [88], which introduced unsupervised domain adaptation for feature fusion. Other approaches directly concatenate CNN features, and thus achieve inferior performance. Similar performance could be observed in Table VI, which reposts the comparison results on the Bosphorus database.

Besides, we perform the comparisons on the CCZU-FER database as shown in Table VII. For a fair comparison, we compare different methods with the same setting as mentioned in Section IV-A. Our method achieves the best performance on the proposed database. Such a superiority indicates the generalization performance of CM-CNN.

V. CONCLUSION

We propose CM-CNN to perform FER on multiple facial modalities. For each type of facial modality, a specially designed FEN is proposed to automatically extract expression-related features. Features extracted from different facial modalities are fused by a novel CMFN, which could enhance the spatial correlations between different facial modalities. Moreover, an improved focal loss is proposed to enhance the recognition performance by paying more attention to expressions hard to recognize. Evaluations on benchmarking databases indicate the superiority of CM-CNN compared with state-of-the-art approaches. The average recognition accuracies on VT-KFER, BU-3DFE(P1), BU-3DFE(P2), and Bosphorus are 4.64, 0.88, 0.32, and 2.66 better than the second best method, respectively. Tests on a newly collected CCZU-FER verify the generalization performance of CM-CNN in laboratory conditions. All results reveal that CM-CNN is promising as a robust FER method, which is beneficial to emotion recognition. Therefore, the proposed method could boost the study of human cognitive systems, especially emotion perception and generation. Besides, our method could be flexibly extended to more types of facial modalities or other emotional modalities, such as sounds and texts. Our future work focuses on the dynamic FER from facial image sequences. Moreover, we plan to update the CCZU-FER data set by adding the neutral expression, which is essential for practical usefulness.

REFERENCES

[1] R. Thompson, "Infancy and childhood: Emotional development," 2001.
[2] P. D. Ross, L. Polson, and M.-H. Grosbras, "Developmental changes in emotion recognition from full-light and point-light displays of body movement," *PLoS ONE*, vol. 7, no. 9, 2012, Art. no. e44815.
[3] G. Chronaki, J. A. Hadwin, M. Garner, P. Maurage, and E. J. S. Sonuga-Barke , "The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood," *Brit. J. Develop. Psychol.*, vol. 33, no. 2, pp. 218–236, 2015.
[4] A. S. Walker-Andrews, "Emotions and social development: Infants' recognition of emotions in others," *Pediatrics*, vol. 102, no. 5, pp. 1268–1271, 1998.
[5] A. L. Gross and B. Ballif, "Children's understanding of emotion from facial expressions and situations: A review," *Develop. Rev.*, vol. 11, no. 4, pp. 368–398, 1991.
[6] A. Dawel, R. O'Kearney, E. McKone, and R. Palermo, "Not just fear and sadness: Meta-analytic evidence of pervasive emotion recognition deficits for facial and vocal expressions in psychopathy," *Neurosci. Biobehav. Rev.*, vol. 36, no. 10, pp. 2288–2304, 2012.
[7] K. M. Rump, J. L. Giovannelli, N. J. Minshew, and M. S. Strauss, "The development of emotion recognition in individuals with autism," *Child Develop.*, vol. 80, no. 5, pp. 1434–1447, 2009.
[8] M. H. Black *et al.*, "Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography," *Neurosci. Biobehav. Rev.*, vol. 80, pp. 488–515, Sep. 2017.
[9] M. J. West, A. J. Angwin, D. A. Copland, W. L. Arnott, and N. L. Nelson, "Cross-modal emotion recognition and autism-like traits in typically developing children," *J. Exp. Child Psychol.*, vol. 191, Mar. 2020, Art. no. 104737.
[10] D. M. Barch, M. P. Harms, R. Tillman, E. Hawkey, and J. L. Luby, "Early childhood depression, emotion regulation, episodic memory, and hippocampal development," *J. Abnormal Psychol.*, vol. 128, no. 1, pp. 81–95, 2019.
[11] A. A. Marsh and R. Blair, "Deficits in facial affect recognition among antisocial populations: A meta-analysis," *Neurosci. Biobehav. Rev.*, vol. 32, no. 3, pp. 454–465, 2008.
[12] G. Lorenzo, A. Lledó, J. Pomares, and R. Roig, "Design and application of an immersive virtual reality system to enhance emotional skills for children with autism spectrum disorders," *Comput. Educ.*, vol. 98, pp. 192–205, Jul. 2016.
[13] F. Ke and S. Lee, "Virtual reality based collaborative design by children with high-functioning autism: Design-based flexibility, identity, and normconstruction," *Interact. Learn. Environ.*, vol. 24, nos. 5–8, pp. 1511–1533, 2016.
[14] D. F. Tolin, "Is cognitive-behavioral therapy more effective than other therapies? A meta-analytic review," *Clin. Psychol. Rev.*, vol. 30, no. 6, pp. 710–720, 2010.
[15] R. Hortensius, F. Hekele, and E. S. Cross, "The perception of emotion in artificial agents," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 4, pp. 852–864, Dec. 2018.

[16] Y. Zhou, L. Jin, H. Liu, and E. Song, "Color facial expression recognition by quaternion convolutional neural network with gabor attention," *IEEE Trans. Cogn. Devel. Syst.*, vol. 13, no. 4, pp. 969–983, Dec. 2021.

[17] Y. Liu, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via deep action units graph network based on psychological mechanism," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 2, pp. 311–322, Jun. 2020.

[18] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 3, pp. 668–680, Sep. 2018.

[19] C.-R. Chen, W.-S. Wong, and C.-T. Chiu, "A 0.64 mm2 real-time cascade face detection design based on reduced two-field extraction," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 11, pp. 1937–1948, Nov. 2011.

[20] X. Pu, K. Fan, X. Chen, L. Ji, and Z. Zhou, "Facial expression recognition from image sequences using twofold random forest classifier," *Neurocomputing*, vol. 168, pp. 1173–1180, Nov. 2015.

[21] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Trans. Affective Comput.*, vol. 9, no. 1, pp. 38–50, Jan.–Mar. 2018.

[22] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, 2011, pp. 1642–1649.

[23] M. H. Siddiqi, R. Ali, A. Sattar, A. M. Khan, and S. Lee, "Depth camera-based facial expression recognition system using multilayer scheme," *IETE Techn. Rev.*, vol. 31, no. 4, pp. 277–286, 2014.

[24] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5325–5334.

[25] M. Valstar, M. Pantic, and I. Patras, "Motion history for facial action detection in video," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, vol. 1, 2004, pp. 635–640.

[26] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 151–160, Apr./Jun. 2013.

[27] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, pp. 467–476, 2002.

[28] M. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "Pca-based dictionary building for accurate facial expression recognition via sparse representation," *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 1082–1092, 2014.

[29] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[30] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

[31] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, 2015.

[32] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *Int. J. Comput. Vis.*, vol. 126, no. 2-4, pp. 255–271, 2018.

[33] H. Kobayashi and F. Hara, "Facial interaction between animated 3D face robot and human beings," in *Proc. IEEE Int. Conf. Syst. Man Cybern. Comput. Cybern. Simulat.*, vol. 4, 1997, pp. 3732–3737.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[36] S. Li and W. Deng, "Deep facial expression recognition: A survey," 2018, *arXiv:1804.08348*.

[37] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Tech. Rev.*, vol. 32, no. 5, pp. 347–355, 2015.

[38] H. Boughrara, M. Chtourou, C. B. Amar, and L. Chen, "Facial expression recognition based on a mlp neural network using constructive training algorithm," *Multimedia Tools Appl.*, vol. 75, no. 2, pp. 709–731, 2016.

[39] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.

[40] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[41] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.

[42] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016, pp. 1–10.

[43] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015.

[44] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective processes: Stochastic modelling of temporal context for emotion and facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9074–9084.

[45] H. Zhang, W. Su, J. Yu, and Z. Wang, "Weakly supervised local–global relation network for facial expression recognition," in *Proc. IJCAI*, 2021, p. 145.

[46] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 5800–5809.

[47] Y. Xie, T. Chen, T. Pu, H. Wu, and L. Lin, "Adversarial graph representation adaptation for cross-domain facial expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1255–1264.

[48] S. Wang, H. Shuai, and Q. Liu, "Phase space reconstruction driven spatio-temporal feature learning for dynamic facial expression recognition," *IEEE Trans. Affective Comput.*, early access, Jul. 7, 2020, doi: 10.1109/TAFFC.2020.3007531.

[49] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2168–2177.

[50] S. Aly, A. L. Abbott, and M. Torki, "A multi-modal feature fusion framework for kinect-based facial expression recognition using dual kernel discriminant analysis (dkda)," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016, pp. 1–10.

[51] Y. Fu, Q. Ruan, Z. Luo, Y. Jin, G. An, and J. Wan, "FERLrTc: 2D+3D facial expression recognition via low-rank tensor completion," *Signal Processing*, vol. 161, pp. 74–88, Aug. 2019.

[52] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.

[53] H. Li *et al.*, "An efficient multimodal 2D+3D feature-based approach to automatic facial expression recognition," *Comput. Vis. Image Understand.*, vol. 140, pp. 83–92, Nov. 2015.

[54] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2816–2831, Dec. 2017.

[55] W. Zeng, H. Li, L. Chen, J.-M. Morvan, and X. D. Gu, "Accurate facial parts localization and deep learning for 3D facial expression recognition," in *Proc. 13th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, 2018, pp. 466–472.

[56] D. Deng, Z. Chen, Y. Zhou, and B. Shi, "Mimamo net: Integrating micro-and macro-motion for video emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 2621–2628.

[57] S. Zhao *et al.*, "An end-to-end visual–audio attention network for emotion recognition in user-generated videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 303–311.

[58] M. Behzad, N. Vo, X. Li, and G. Zhao, "Automatic 4D facial expression recognition via collaborative cross-domain dynamic image network," 2019, *arXiv:1905.02319*.

[59] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. ICML*, vol. 2, 2016, p. 7.

[60] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8792–8802.

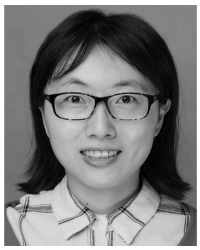
[61] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[62] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[63] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[64] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis. Image Understand.*, vol. 138, pp. 1–24, Sep. 2015.

[65] Y.-Q. Wang, "An analysis of the Viola-Jones face detection algorithm," *Image Process. Line*, vol. 4, pp. 128–148, Jun. 2014.
[66] S. Cheng, A. Asthana, S. Zafeiriou, J. Shen, and M. Pantic, "Real-time generic face tracking in the wild with CUDA," in *Proc. 5th ACM Multimedia Syst. Conf.*, 2014, pp. 148–151.
[67] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, pp. 1657–1663, 2010.
[68] J. D'Errico, "Surface fitting using gridfit," MATLAB Central File Exchange. Accessed: Feb. 15, 2022. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/8998-surface-fitting-using-gridfit
[69] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," 2015, *arXiv:1502.00873*.
[70] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.
[71] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[72] S. Aly, A. Trubanova, L. Abbott, S. White, and A. Youssef, "VT-KFER: A kinect-based rgbd+ time dataset for spontaneous and non-spontaneous facial expression recognition," in *Proc. Int. Conf. Biometrics*, 2015, pp. 90–97.
[73] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, 2006, pp. 211–216.
[74] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3D face by exploring shape deformation," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 569–572.
[75] A. Savran and B. Sankur, "Non-rigid registration based model-free 3D facial expression recognition," *Comput. Vis. Image Understand.*, vol. 162, pp. 146–165, Sep. 2017.
[76] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.
[77] S. Z. Gilani and A. Mian, "Learning from millions of 3D scans for large-scale 3D face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1896–1905.
[78] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[79] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5683–5692.
[80] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[81] A. R. Rivera, J. R. Castillo, and O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, pp. 1740–1752, 2013.
[82] W. Zeng, H. Li, L. Chen, J.-M. Morvan, and X. D. Gu, "An automatic 3D expression recognition framework based on sparse representation of conformal images," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, 2013, pp. 1–8.
[83] H. Li, L. Chen, D. Huang, Y. Wang, and J.-M. Morvan, "3D facial expression recognition via multiple kernel learning of multi-scale local normal patterns," in *Proc. 21st Int. Conf. Pattern Recognition (ICPR)*, 2012, pp. 2577–2580.
[84] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model based automatic 3D facial expression recognition," in *Proc. Int. Conf. MultiMedia Model.*, 2015, pp. 522–533.
[85] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3D facial expression recognition using geometric scattering representation," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, 2015, pp. 1–6.
[86] O. K. Oyedotun, G. Demisse, A. El Rahman Shabayek, D. Aouada, and B. Ottersten, "Facial expression recognition via joint deep learning of RGB-depth map latent representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3161–3168.
[87] Z. Chen, D. Huang, Y. Wang, and L. Chen, "Fast and light manifold CNN based 3D facial expression recognition across pose variations," in *Proc. ACM Multimedia Conf. Multimedia Conf.*, 2018, pp. 229–238.
[88] X. Wei, H. Li, J. Sun, and L. Chen, "Unsupervised domain adaptation with regularized optimal transport for multimodal 2D+3D facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2018, pp. 31–37.
[89] H. Ujir and M. Spann, "Surface normals with modular approach and weighted voting scheme in 3D facial expression classification," *Int. J. Comput. Inf. Technol.*, vol. 3, no. 5, pp. 909–918, 2014.



Rongrong Ni was born in Nantong, China, in 1987. She received the master's degree in instrument science and technology from Southeast University, Nanjing, China, in 2012. She is currently pursuing the Ph.D. degree with the College of IoT Engineering, Hohai University, Changzhou, China.

Biao Yang (Member, IEEE) received the B.S. degree from Nanjing University of Technology, Nanjing, China, in 2009, and the Ph.D. degree in instrument science and technology from Southeast University, Nanjing, in 2014.

He was a Visiting Scholar with the University of California at Berkeley, Berkeley, CA, USA. He currently works with Changzhou University, Changzhou, China. He is also a Postdoctoral researcher in Hohai University, Nanjing. His current research interests include target detection and behavior analysis based on computer vision.

Xu Zhou (Graduate Student Member, IEEE) received the bachelor's degree from Hohai University, Changzhou, China, in 2016, where he is currently pursuing the Ph.D. degree.

His research interests include deep learning, reinforcement learning, and human–robot interaction.

Angelo Cangelosi (Senior Member, IEEE) received the M.S. degree in psychology from the University of Rome La Spaienza, Rome, Italy, in 1991, and the Ph.D. degree in psychology and artificial intelligence from the University of Genoa, Genoa, Italy, in 1997.

He was a Visiting Scholar with the University of California at San Diego, San Diego, CA, USA, and the University of Southampton, Southampton, U.K. His research interests are in developmental robotics, language grounding, human–robot interaction and trust, and robot companions for health and social care.

Xiaofeng Liu (Member, IEEE) received the B.S. degree in electronic engineering and the M.S. degree in computer application from Taiyuan University of Technology, Taiyuan, China, in 1996 and 1999, respectively, and the Ph.D. degree in biomedical engineering from Xian Jiaotong University, Xi'an, China, in 2006.

Then, he joined as an Associate Professor with Shandong University of Science and Technology, Qingdao, China. From 2008 to 2011, he was a Postdoctoral Researcher with the Institute of Artificial Intelligence and Robotics, Xian Jiaotong University. Since 2010, he has been with Hohai University, Changzhou, China, as a Full Professor, where he is currently the Vice Dean of the College of IoT Engineering. He has undertaken more than 20 grants as PI, and more than 16 grants as researcher, including the National High-Tech R&D Program (863) and the National Basic Research Program (973). His current research interests focus on the study of nature-inspired navigation and human–robot interaction.

Prof. Liu has served as the Associate Editor for the *Interaction Studies* and *IET Cognitive Computation and System*, the Guest Editor of several international journals, including *Interaction Studies*, *Complexity*, *Assembly Automation*, *Frontiers in Robotics and AI*, *Advances in Mechanical Engineering*, and *Electronics*, and the Editorial Board Member of *International Journal of Artificial Intelligence and Consciousness*.