3-D Facial Expression Recognition via Attention-Based Multichannel Data Fusion Network

Yu Gu, *Senior Member, IEEE*, Huan Yan, Xiang Zhang, Zhi Liu, *Senior Member, IEEE*, and Fuji Ren, *Senior Member, IEEE*

***Abstract*— Facial expression has long been recognized as containing meaningful nonverbal affective cues for decoding human emotions. Recently, multimodal 2-D + 3-D fusion method has shown significant potential in facial expression recognition (FER) due to its fine-grained face descriptions in various spatial channels. However, current work mainly relies on feature- or even score-level fusion to find emotion cues spread in different channels and may miss key information due to lack of focus. To this end, we propose an attention-based multichannel data fusion network (AMDFN) to better preserve and find such key facial cues. More specifically, we first map a 3-D face scan into multichannel images and then fuse them in a ResNet18 backbone to get layered emotion features. Second, we leverage a layer attention model to explore the dependencies between features of different layers to learn discriminative affective cues for effective emotion recognition. Our comprehensive experiments on two widely used datasets (i.e., Facescape and Bosphorus) have verified the performance of our approach compared to several state-of-the-art rivals.**

***Index Terms*— Attention, convolutional neural networks (CNNs), data-level fusion, facial expression recognition (FER), multimodal.**

I. Introduction

FACIAL expression recognition (FER), acting as an essential way of human emotional behavior understanding, is rapidly applied in various fields of human–computer interaction in recent years (e.g., human mental health [1], emotional regulation [2], and fatigue detection [3]). With the continuous development of information technology, automatic facial expression recognition (AFER) has become the core component of the next generation of artificial intelligence.

Over the past few years, with the advance of deep learning, FER performance based on deep convolutional neural network (CNN) has gained significant improvement, which is also inseparable from the support of existing facial expression datasets. An informative and usable dataset will help advance FER research. As for now, there are many available datasets that could be used for 2-D or 3-D FER algorithm evaluation, such as extended CohnKanade (CK+) [4], high-resolution thermal facial expression recognition (TFER) datasets [5], Facescape [6] dataset, and Bosphorus [7] dataset. In general, most of the research on FER is to identify a specific emotion category (i.e., anger, disgust, fear, happiness, sadness, and surprise) on given 2-D static images or videos. For instance, Yang *et al.* [8] extracted the information of the expression components through a deexpression learning procedure to identify facial expressions and demonstrated the superior performance on the existing 2-D datasets. Mohan *et al.* [9] proposed a two-branch network for 2-D FER, in which two branches explore the geometric features and holistic features of facial expressions and finally fused to obtain distinctive features describing facial expressions. Although the 2-D data modality-based FER methods have considerable results and can be applied in real life, the stability of the 2-D FER system is greatly affected by illumination or pose variations. In order to alleviate the problem, the 3-D data modality-based FER methods have attracted the attention of many researchers [10]–[12]. Such 3-D FER methods are robust to illumination and head pose variations since they could capture more subtle facial deformations caused by muscle movements. Moreover, some researchers use the time-varying frame sequence to perceive the expression change information on the basis of 3-D data modalities, which is 4-D FER (also called 3-D videos) [13], [14]. For example, Li *et al.* [14] explored the 4-D FER using a dynamic geometrical image and proved the effectiveness of the method on the existing 4-D FER dataset. To meet the requirements of real applications, FER based on the modal fusion of 2-D facial images and 3-D face models has become a promising research direction.

Hence, it is critical to explore the complementarity between different modalities and to combine 2-D facial images with 3-D facial models to obtain good performance. The method of combining 2-D and 3-D image information captured simultaneously is specified as 2-D + 3-D FER. However, most 2-D + 3-D multimodal methods predict the final result by using the feature- or score-level fusion strategies. On one hand, although only different 2-D facial attribute images need to be trained for a single network, the network is designed in parallel for multiple 2-D facial attribute images in the feature

Manuscript received August 22, 2021; revised October 18, 2021; accepted October 26, 2021. Date of publication November 8, 2021; date of current version November 24, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61772169 and in part by the Fundamental Research Funds for the Central Universities under Grant JZ2018YXQN0121. The Associate Editor coordinating the review process was Dr. Jing Lei. *(Corresponding author: Huan Yan.)*

Yu Gu, Huan Yan, and Xiang Zhang are with the School of Computer and Information, Hefei University of Technology, Hefei 230601, China (e-mail: hfut_bruce@hfut.edu.cn; yanhuan@mail.hfut.edu.cn; zhangxiang@mail.hfut.edu.cn).

Zhi Liu is with the Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: liu@ieee.org).

Fuji Ren is with the Department of Information Science and Intelligent Systems, The University of Tokushima, Tokushima 770-8570, Japan (e-mail: ren@is.tokushima-u.ac.jp).

Digital Object Identifier 10.1109/TIM.2021.3125972

1557-9662 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

extraction subset, and such time consumption and memory consumption is large. On the other hand, multiple networks need to be independently trained and finally fused at the result level, which causes large computation time and memory consumption. Some facial expression information is lost when extracting features for each 2-D facial attribute image in the feature extraction subset. Therefore, it is important to find an effective fusion method to extract the distinguishing features related to expressions so that the classifier can correctly classify expressions.

In this article, we explore the multimodal FER method by combining 2-D + 3-D facial expression data. Our goal is to design a new method to generate discriminative affective representations by combining features of different layers in the network. To achieve these, we propose a novel approach, denoted as attention-based multichannel data fusion network (AMDFN), to learn effective face descriptors through an attention-based multichannel data fusion network. First, we map a 3-D face scan into multichannel images and then fuse them in a ResNet-18 backbone to get layered emotion features. Second, we leverage a layer attention model to explore the dependencies between features of different layers to learn discriminative affective cues. In contrast to the previous methods [9], [10], [15]–[18], which used feature- or score-level fusion strategies for expression recognition, our proposed AMDFN framework mitigates the computation time and memory consumption and improve the performance of FER. Moreover, our system also uses a layer attention module (LAM) to model the dependencies between features of different layers in the network.

In summary, the contribution of this work lies in twofold.

1) We propose a novel framework AMDFN, which consists of a multichannel data fusion and LAM. Different from the traditional feature- or score-level fusion methods, AMDFN first fuses the multichannel images into ResNet-18 backbone to get layered emotion features and then leverages LAM to model the dependencies between features of different layers.
2) Extensive experiment results demonstrate that the proposed method is capable of learning distinctive facial features. It achieves good performance on the Facescape subset with 16 expressions and outperforms state-of-the-art methods on the publicly available Bosphorus dataset.

This article is organized as follows. Section II details the related work on FER. Section III introduces the pipeline of the proposed system. Section IV describes the experimental results and discussion. Finally, conclusions are drawn in Section V.

II. Related Works

In this section, we summarize the related work on 2-D FER, 3-D FER, and 2-D + 3-D FER. In addition, Facial ActionCode System (FACS) is also briefly reviewed as the essential guidance of expression arrangement.

A. 2-D FER

The CNN can automatically learn the intrinsic feature representation in the data, which makes it widely used in FER-based vision applications. From the beginning of the 21st century to the present, there are many researchers dedicated to the research of FER [19]–[27]. Fasel [19], [20] proposed a data-driven facial analysis method based on CNN, which not only can extract the features related to a given facial task but also has strong robustness. Hu *et al.* [21] proposed the supervised scoring ensemble (SSE) method to solve the problem of supervising only the output feature layer, resulting in insufficient training of deep CNN models. Kuo *et al.* [25] proposed a frame-based compact FER framework for FER, which uses very few parameters and has very good competitive performance compared to the state-of-the-art methods. The regularization was integrated into the loss function in [26] and optimized using a deep metric learning framework. The comparison of a large number of experiments proved the effectiveness of the proposed method in identifying facial expressions. Recently, in order to describe facial expressions more accurately, Jia *et al.* [28] adopted label distribution learning method for emotion recognition, which can solve how to describe the ambiguity of expressions, and proposed an emotion distribution learning method that exploits label correlations locally. Moreover, Jia *et al.* [28] also used a local low-rank structure to implicitly capture local label associations.

B. 3-D FER

There are mainly two categories of 3-D FER methods, i.e., model-based and feature-based.

Model-based methods typically train 3-D morphable face models and then fit them to each probe face. For instance, Ocegueda *et al.* [29] presented a semiautomatic 3-D FER system based on geometric facial information. First, assemble the 3-D face mesh into the 3-D mesh annotated face model (AFM). Then, calculate the most expressive parts of the face based on specific geometric features and also form an expression map. Mpiperis *et al.* [30] established correspondence among a set of faces and constructed bilinear models that decouple the identity and facial expression factors. The unknown face is fit to bilinear models, and the identity-invariant expression recognition can be realized. Gong *et al.* [31] approximated the 3-D face model to the sum of a basic facial shape component (BFSC) and an expressional shape component (ESC) and then constructed the feature vector using ESC components for 3-D FER. Moreover, Cordea *et al.* [32] proposed a tracking algorithm based on a 3-D model, allowing real-time recovery of the 3-D position, direction, and facial expression of the moving head. The proposed solution can quickly and stably track the extended sequence online, despite the noise and large changes in attitude and expression. However, the model-based methods require to establish dense correspondence when building the model, which causes a large time consumption. Therefore, most of the current 3-D FER methods adopt a feature-based approach.

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

TABLE I
ANALYSIS OF 3-D FACIAL EXPRESSION METHODS ON THE BOSPHORUS DATASET

Method	Category	Data	Feature	Fusion method	Classifier	Accuracy
Li et al. [33]	Model	3D	Normals,LBP	-	MKL	75.83%
Ujir et al. [34]	Model	3D	Surface normals	-	SVM	63.63%
Li et al. [16]	Feature	2D+3D	HOG,SIFT	Feature/Score-level	SVM	79.72%
FERLrTC [10]	Feature	2D+3D	RGB,normals,LBP curvature,depth	Feature-level	SVM	75.93%
DF-CNN [15]	Feature	2D+3D	RGB,normals, curvature,depth	Feature-level	SVM	80.28%
AMDFN	Feature	2D+3D	RGB,normals, curvature,depth	Data-level	Softmax	83.13%

Feature-based methods predominantly extract local descriptors around facial landmarks. Zarbakhsh and Demirel [35] selected the most discriminating measures among some Euclidean distances between 83 keypoints. Berretti *et al.* [36] proposed a fully automatic face recognition method based on facial feature point recognition to filter the feature descriptors of the face depth image, define sampling points starting from the key points of the face, and select the most relevant feature subset. Maalej *et al.* [37] proposed a facial local geometry analysis method combining machine learning technology to perform facial expression classification. Wang *et al.* [38] extracted primitive 3-D facial expression features and then applied the feature distribution to classify the prototypic facial expressions. Zhen *et al.* [12] proposed a muscular movement model (MMM) for 3-D FER, which extracted a set of features within each muscular region and then predicted the expression label in 3-D and 4-D through support vector machine (SVM) and hidden Markov model (HMM).

C. 2-D + 3-D FER

In recent years, due to the strong complementarity of different modalities, it is significantly extensive that the application of FER is using 2-D + 3-D multimodal data [10], [11], [15], [16]. In [10], a 4-D tensor model was constructed for exploring multimodal 2-D + 3-D structural information and correlations, and then, tensor dimensionality reduction technique (i.e., Tucker decomposition) was used to reduce the dimension for classification prediction. Li *et al.* [15] represented a face scan as six types of 2-D facial attribute images, where features were extracted respectively and fused afterward. Gilani and Mian [11] used 2-D + 3-D multimodal data as the network input for face recognition where the three channels correspond to the depth, azimuth, and elevation of the normal vector. Although the feature-based methods have low time consumption, they mostly depend on the recognition ability of local features.

The above 2-D + 3-D FER methods focus on the fusion of feature level and score level. In the former, although only different 2-D facial attribute images need to be trained for a single network, the network is designed in parallel for multiple 2-D facial attribute images in the feature extraction subset, and such time consumption and memory consumption is large. In the latter, multiple networks need to be independently trained and finally fused at the result level, which causes large computation time and memory consumption. Some facial expression information is lost when extracting features for each 2-D facial attribute image in the feature extraction subset. Table I lists some 3-D and 2-D + 3-D FER methods and their key characteristics.

D. FACS

FACS is a system that initially classifies human facial movements through the appearance of the face [39]. It defines the facial action units (AUs) and several facial action descriptions (ADs), which are used to encode any anatomically possible facial expressions. In FACS, there are 86 AUs and ADs (with underlying facial muscles) from different codes (i.e., main codes, head movement codes, eye movement codes, visibility codes, and gross behavior codes). As a matter of fact, AUs and ADs may not all appear on the face for basic expressions. There are many studies on FER with the help of FACS [40]–[43]. Wang *et al.* [40] objectively considered AUs to determine facial expressions and proposed an AU-guided unsupervised domain adaptive FER framework, which alleviates the annotation bias between different domains. Pu *et al.* [41] explored the association between AUs and facial expressions and designed an AU expression association framework to adaptively use AU representations to promote FER. In this article, we only select the facial actions associated with the expression in the main codes on the Facescape dataset to better analyze the 3-D expression recognition.

III. PROPOSED METHOD

A. Overall Framework

The architecture of our proposed method − AMDFN is shown in Fig. 1, and AMDFN mainly contains three modules: data preprossing, multichannel data fusion, and layer attention.

The first module mainly preprocesses the original 3-D face scan and then maps it into eight 2-D facial attribute images. The multichannel data fusion module mainly performs channel fusion of the eight 2-D facial attribute images after mapping, and then, the training and testing sets are fed into the ResNet backbone to extract the features related to the expressions.

For the last part, the interdependencies between different layers in the network are modeled through LAM, which improves the representation ability of features by assigning different attention weights to features of different layers.

B. Data Preprocessing

The data preprocessing procedure can be divided into two stages in order to represent a 3-D face scan using eight attribute images, which are fed into multichannel data fusion

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

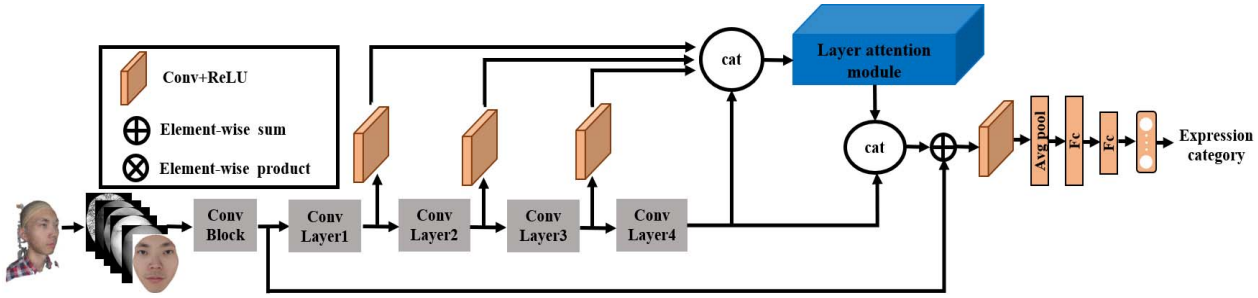

Fig. 1. Framework of the proposed AMDFN, which consists of a multichannel data fusion and LAM. The input of the first module is an image fusion of eight channels, which is mapped from the 3-D face scan, and then, the LAM is used to model the dependencies between features of different layers to improve the representation ability of features.

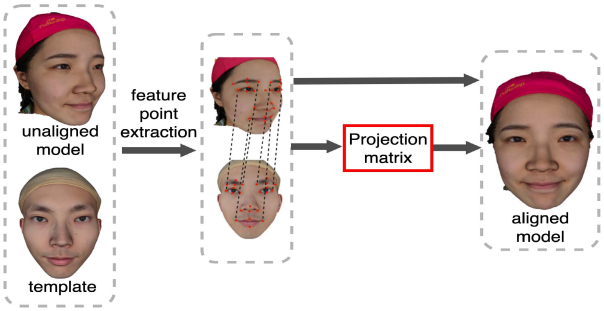

Fig. 2. Illustration of facial pose registration. RGBD data are rendered for the model and template, followed by feature points detection in RGB image. Combined with depth data, 3-D feature points are located. After the calculation of transform matrix between two models with 3-D point set, the pose of the model is matched up with the template.

module. Note that here, we want to explain that data preprocessing is not a contributing point of this article.

1) Face Alignment: In the first stage, to avoid the problem countered by 2-D FER that variations in head pose of samples hinder higher accuracy, face alignment is applied to unify the pose of each 3-D face polygon mesh. For each 3-D mesh, its corresponding RGBD data are rendered, followed by automatically localizing a set of facial feature point on the rendered 2-D texture image, such as nosetip, lip boundary, and eye corners. Then, we project the feature points to their corresponding locations on the 3-D facial mesh exploiting the depth data recorded during the render process. After the process of localizing 3-D feature points, the pose of each 3-D facial mesh is matched up with a template facial mesh as closely as possible by fitting the feature points. Fig. 2 shows the whole face alignment process.

The next step is to find the rotation matrix and transformation vector of the 3-D face mesh to adjust the pose of the 3-D face mesh. As described in [44], we set the facial feature points of the 3-D facial mesh and prepared template to be p' and p, respectively, and their relationship can be expressed as

$$p'_i = Rp_i + T + N_i, \quad i = 1, 2, \ldots, N \tag{1}$$

where R denotes rotation matrix, T denotes the transformation vector, N denotes the number of facial feature points, and N_i denotes the noise vector. Therefore, solving the rotation matrix R and the transformation vector T can be transformed into the following optimization problem:

$$\underset{R,T}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| p'_i - (Rp_i + T) \right\|^2. \tag{2}$$

Then, the least-squares solution of R and T based on the singular value decomposition (SVD) of a 3×3 matrix can be solved [44]. The procedure of 3-D face registration left us with unified models that share the same head pose and easy to process. The geometric center of facial mesh is at the origin and the nosetip is on the Z-axis, while the central axis of the face is parallel to the Y-axis.

2) Get Facial Mesh: Once we obtained the face meshes that share the same pose, nosetip detection and face cropping introduced in [45] is applied in the second stage to acquire the normalized facial mesh. We abandon unnecessary parts such as clothes, shoulder, neck, and hair that have nothing to do with human expression.

3) Mapping to Attribute Image: In the final stage, we take advantage of modeling facial expressions in 3-D by mapping a face scan to eight different 2-D attribute images, named the depth image, curvature image, three normal images, as well as texture image. These attribute images can comprehensively describe geometric and photometric details of a facial mesh, and hence, it is ideal to be fed into multichannel data fusion module to learn different features in different expressions. We first run the render process to generate the RGB texture image (expressed as T_R, T_G, and T_B) and the depth image (D) for every facial mesh. Then, we use the coordinates information of each vertex in the mesh to estimate its normal and curvature value, resulting in three normal images (N_x, N_y, and N_z) along the x-, y-, and z-directions, as well as a normalized curvature image (C). Finally, we can represent a 3-D facial scan with eight 2-D attribute maps: N_x, N_y, N_z, D, C, T_R, T_G, and T_B. The generation process of normal images and curvature image is described as follows.

4) Normal Image: Surface normal is the most basic information for shading a surface. It is the vector that points straight away from the surface at a particular point. Let F be a face in the polygon mesh that is formed by three vertices: V_1, V_2, and V_3, each of which is represented by a 3×1 vector using its coordinate along the x-, y-, and z-directions. Then, we can compute the unit normal vector N_f of F by

$$N_f = \frac{(V_1 - V_2) \times (V_2 - V_3)}{\|(V_1 - V_2) \times (V_2 - V_3)\|}. \tag{3}$$

Given that the normal vector of faces F consists of point V, the unit normal vector N_v at V can be represented as

$$N_v = \frac{\sum_{V \in F} N_f}{\left\| \sum_{V \in F} N_f \right\|}. \tag{4}$$

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

We generate three normal maps: N_x, N_y, and N_z for each facial mesh using three components of a normal vector in the x-, y-, and z-directions. A normal map is one of the most important attributes for a 3-D mesh as it provides a good measure of how bright the surface should be under illumination.

5) Curvature Image: Curvature is used to reflect the degree of curvature of geometry. Qualitatively, the more severe the curvature, the greater the curvature of that part. The curvature in the 3-D space includes principal curvatures, Gaussian curvature, and mean curvature. Gaussian curvature and mean curvature are not very indicative of local shape. The two principal curvatures (taken as a pair) are more informative, but one would prefer a single shape indicator rather than a pair of numbers [46]. The principal curvatures are the basic element of curvature, which represents the maximum and minimum normal curvature through a certain point on the surface. The curvature map mentioned in this article is quantified by two principal curvatures, which can be regarded as a second-order differential geometric quantity. The curvature map is quantized by two principal curvatures at the mesh surface [46]. It is formed by the curvature value at each vertex V, which can be defined as

$$\text{ShapeIndex}(V) = \frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{k_1(V) + k_2(V)}{k_1(V) - k_2(V)}\right) \tag{5}$$

where $k_1(V)$ and $k_2(V)$ represent the maximum and minimum curvatures at point V in two principal directions, respectively, ranging from 0 to 1. The shape index value of each vertex is calculated for a 3-D mesh. Then, we generate a curvature image using an interpolation technique.

The principal curvatures at each point can be estimated using the local cubic fitting algorithm [47], where a local coordinate system is created by taking the vertex V as origin and its normal vector N_V as the z-axis. In the plane perpendicular to N_V, the x- and y-axes are randomly generated. Then, a neighborhood point P is transformed into the local coordinate system and fit a cubic surface $z(x, y)$ and its normal vector. The process can be described as

$$\begin{cases} z(x,y) = \dfrac{A}{2}x^2 + Bxy + \dfrac{C}{2}y^2 \\ \qquad\qquad + Dx^3 + Ex^2y + Fxy^2 + Gy^3 \\ z_x = Ax + By + 3Dx^2 + 2Exy + Fy^2 \\ z_y = Bx + Cy + 3Gy^2 + 2Fxy + Ex^2. \end{cases} \tag{6}$$

The equations can be solved by the least-square fitting algorithm and the symmetric matrix can be represented as

$$W = \begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{7}$$

$k_1(V)$ and $k_2(V)$ are the eigenvalues of W.

Different maps for various types of facial expressions of a subject are shown in Fig. 3. These attribute images are selected for their capability of describing details of a 3-D mesh. Moreover, using a 2-D attribute image instead 3-D mesh reduces the considerable computational cost.

C. Multichannel Data Fusion

In the first step, we use the same data mapping method as in [15] to map the original 3-D face scan image into eight 2-D facial attribute images. Therefore, we can represent a 3-D facial scan with eight 2-D attribute maps: texture image, curvature image, depth image, and three normal images along the x-, y-, and z-directions. Such operation has two advantages.

1) The mapping is simple and the basic 2-D attribute images are selected that contain most of the details in the 3-D mesh. Thus, the loss of 3-D facial expression information can be avoided.
2) The learning network can obtain discriminative representations for 3-D FER through mapped information.

We mentioned above that the 2-D + 3-D FER methods focus on the fusion of feature level and score level. In the former, although only different 2-D facial attribute images need to be trained for a single network, the network is designed in parallel for multiple 2-D facial attribute images in the feature extraction subset, and such time consumption and memory consumption is large. In the latter, multiple networks need to be independently trained and finally fused at the result level, which causes large computation time and memory consumption. Some facial expression information is lost when extracting features for each 2-D facial attribute image in the feature extraction subset.

Therefore, the multichannel data are constructed after the mapping operation. The shape of the constructed data is $C \times H \times W$, in which $H \times W$ corresponds to the size of the 2-D facial attribute image and C represents the fusion image channels (i.e., $C = 8$). In order to facilitate the input of the model, we set the spatial dimension of each 2-D facial attribute image to 112×112. Giving a 3-D face scan input I_{fs}, we get eight channels input I_{ai} through data mapping, a convolution block is used to extract the shallow feature F_0 of the I_{fs} input, and the calculation is as follows:

$$F_0 = C_B(M(I_{\text{fs}}), \theta_0) \tag{8}$$

where θ_s denotes the network parameter in the convolution block, C_B denotes the convolution block operation, M denotes the data mapping operation, and $M(I_{\text{fs}}) \in \mathbb{R}^{C\times H\times W}$.

Then, we use the backbone of the ResNet-18 [48] to extract the intermediate features F_i of the I_{fs} input. To make the feature dimension of each intermediate layer the same so that the LAM can capture the dependency of different intermediate features, we perform the dimensionality reduction operation in the first $N-1$ layers to maintain the same dimensionality as the output of the Nth layer. F_i is calculated as follows:

$$F_i = \begin{cases} C_{L_i}(R_i(F_{i-1}, \theta_{r_i})), & i = 1, 2, \ldots, N-1 \\ C_{L_i}(F_N), & i = N \end{cases} \tag{9}$$

where F_i denotes the features extracted at different layers of the network, C_{L_i} denotes the ith convolution layer operation in the ResNet-18, R_i denotes the dimensionality reduction operation of the first $N-1$ layers (e.g., pooling and convolution), and θ_{r_i} is the parameter corresponding to R_i.

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

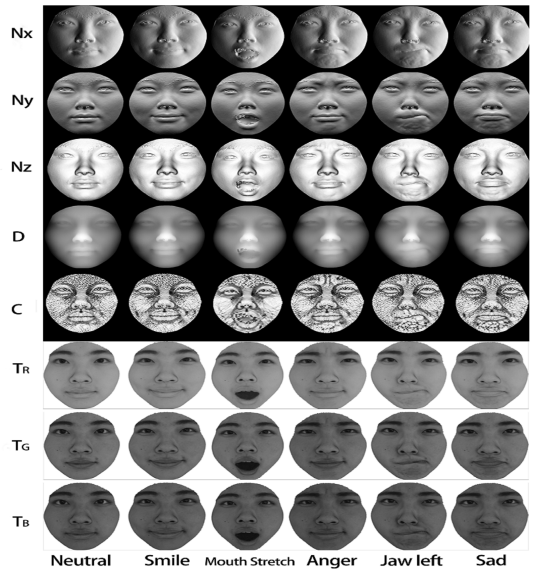

Fig. 3. Demonstration of the eight types of 2-D attribute images generated from six facial expression mesh (subject 078), we illustrate the mapping result with six facial expressions and AU movements (i.e., neutral, smile, mouth stretch, anger, jaw left, and sadness). Display from top to bottom: three normal images (along the x-, y-, and z-directions), the depth image, curvature image, and texture image.

D. Layer Attention Module

The feature map of each layer in the network is regarded as a response to a specific class, and the responses from different layers are related to each other. By exploiting the interdependencies between features of different layers, we could emphasize interdependent feature maps from different layers and improve the feature representation of specific semantics. Therefore, we build an LAM to explicitly model interdependencies between features of different layers.

The structure of LAM is shown in Fig. 4. Inspired by [49], we use a lightweight gating mechanism to model the dependence between different layers in the network to improve the representation ability of extracted features. First, we directly calculate the LAM input features $F_{\text{cat}} \in \mathbb{R}^{N\times C\times H\times W}$ from concatenating the features extracted at different layers $F_i \in \mathbb{R}^{C\times H\times W}$, $i = 1, 2, \ldots, N$. Specifically, we reshape F_{cat} to $\mathbb{R}^{N\times \text{HWC}}$, and then, we use a global average pooling operation to explore the dependency of different intermediate features. Formally, we use $G \in \mathbb{R}^N$ to represent the generated feature vector, and the kth element in G can be expressed as

$$G_k = \text{GAP}(F_{\text{cat}}) = \frac{1}{\text{HWC}} \sum_{i=1}^{\text{HWC}} F_{\text{cat}_k}(i). \tag{10}$$

To make good use of the information integrated by global pooling to better fully capture feature layerwise dependencies, we use a simple gating mechanism with a sigmoid function that can give different attention weights to the features extracted by each layer. Moreover, we perform a feature layerwise multiplication to get the output of the LAM

$$E = F_{\text{cat}} \cdot \sigma(\varphi(F_c(G))) \tag{11}$$

where E denotes the output of the LAM, σ denotes the sigmoid function, φ denotes the rectified linear unit (ReLU) function, and F_c denotes the fully connected (FC) operation.

After obtaining features from both the LAM and the last layer, we integrate these features and shallow features by elementwise summation, which can better stabilize the training process of the deep network. At last, a convolution layer is followed to generate the final prediction map. We let $\mathcal{F}$ represent the whole network prediction function, $\mathbf{w}$ represents the network parameters, y_i represents the label corresponding to the ith input data I_i, and K represents the number of categories, so the AMDFN optimization problem is expressed as

$$\underset{\mathbf{w}}{\text{argmin}} \sum_{r=1}^{R} \sum_{k=1}^{K} \Pi_{[k=y_i]} \log(\mathcal{F}(I_i; \mathbf{w})) \tag{12}$$

where R denotes the total number of samples in the training set, $\log(\cdot)$ denotes the logarithm function, and $\Pi_{[k=y_i]}$ outputs 1 when $k = y_i$ and 0 otherwise. We summarize the training of AMDFN in Algorithm 1, where the whole network parameter solution can be achieved in an end-to-end manner through a standard backpropagation algorithm [50].

Algorithm 1 Training of AMDFN

Input: I_{fs} /* 3D face scans */
Output: f /* Optimized model */
for *epoch* $\leftarrow$ 1 *to epochs* **do**
  Compute shallow feature F_0 and intermediate features F_i; // Eq.(6)(7)
  Compute LAM output E; // Eq.(8)
  Integrate shallow feature F_0 and LAM output E; // Eq.(9)
  Solve optimization problems based on back-propagation algorithm [50]; // Eq.(10)

E. Analysis

Attention can be viewed, broadly, as a tool to bias the allocation of available processing resources toward the most informative components of an input signal [51], [52]. Similarly, our LAM uses this mechanism to design a new learning framework that can automatically model the importance of different network layers to learn discriminative affective cues for effective emotion recognition. Some research work adaptively recalibrates the characteristic response of spatial position or channel direction by explicitly modeling the interdependence between spatial positions or channels [49], [53], [54]

$$\widetilde{H}_{i,c} = H_{i,c} \times A_{i,c} \tag{13}$$

where $H_{i,c}$ and $\widetilde{H}_{i,c}$ represent the input feature and output feature of the attention module for spatial position i or channel c, respectively. In words, the output feature of the attention module is the calibration of the spatial position or channel of the input feature. Take the dependency relationship between modeling channels as an example [49]. Let $e_c \in R^{H\times W\times C}$ be the intermediate feature extracted by CNN, where H and W are, respectively, the height and width of the feature, and C is the number of channels of the feature. Then, an explicit learning module (usually composed of a convolutional layer, an FC layer, and an activation function) is used to adaptively

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

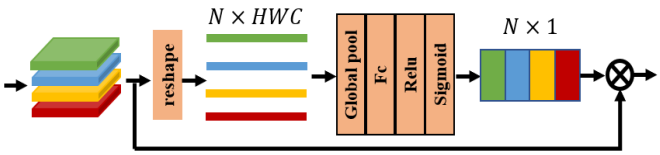

Fig. 4. Network architecture of the LAM.

learn the weights of different channels, which is expressed as w_c, which means the importance of different channels. Finally, the weights are multiplied by the intermediate features to perform feature recalibration

$$\widetilde{E}_c = M(e_c, w_c) = w_c \cdot e_c \tag{14}$$

where $\widetilde{E} = [\widetilde{e_1}, \widetilde{e_2}, \ldots, \widetilde{e_c}]$ and $M(e_c, w_c)$ refers to channel-wise multiplication between the feature map $e_c \in R^{H \times W}$ and the attention weight w_c. The network structure design ideas in this article mainly follow the basic principles of the attention mechanism but focus on the feature dependence of different layers of the backbone network when learning attention weights. In order to explicitly embed the learning module without increasing the burden on the network, we use only global average pooling, FC, and activation functions as components to model the dependencies between different layers.

F. Implementation Details

We use ResNet18-variant [55] as the backbone, which is pretrained on the MS-Celeb-1M [56] dataset. To enable feature extraction for multichannel fusion data, we change the input channel of the first convolutional layer to 8 and the output feature of the last FC layer to 16 or 6 (i.e., Facescape subset or Bosphorus dataset). In this article, there are a total of four layers of output cascaded (i.e., $N = 4$). To make the output feature dimensions of each intermediate layer the same, the input channels of the convolutional layer we added in the first three layers are 64, 128, and 256, the output channels are all 512, and the kernel size is 1×1, with a stride of 1 and a padding of 1. Set the pooling operation for the first three layers, and finally, all the intermediate layers' output is $7 \times 7 \times 512$. In addition, we add a pooling layer and a convolutional layer after the shallow features to convert the output dimension as $7 \times 7 \times 512$. We train the networks using an SGD optimizer, with the learning rate set to 0.01. All the models are trained on a single NVIDIA RTX 2080 Ti using Pytorch for 70 epochs with a batch size of 20 for the Facescape subset and Bosphorus dataset.

IV. Experimental Results and Discussion

A. Dataset Description

To validate the effectiveness of AMDFN, we select 251 subjects from the Facescape dataset [6], which has a higher resolution compared with the other 3-D datasets. Among the subjects, we only select facial actions associated with expressions in the main code. Therefore, a total of 16 expressions from 40 facial scans can be analyzed for 3-D expression recognition, as shown in Fig. 5. Moreover, we perform the process of 3-D face scan mapping into 2-D facial attribute images as proposed in [15]. Fig. 3 shows eight types of 2-D facial attribute images with six facial expressions of subject 078 in the Facescape subdataset. To ensure that the experiment is independently trained and tested during the evaluation process (i.e., one sample of the subjects cannot belong to the training and testing sets at the same time), 200 subjects, each with 16 samples (i.e., 3 basic expressions, neutral, and 12 AUs), are randomly selected as the training set from Facescape subset. In other words, a total of 3200 3-D face scans are used to train the network. Then, we use the remaining 51 subjects (i.e., 816 3-D face scans) to test AMDFN.

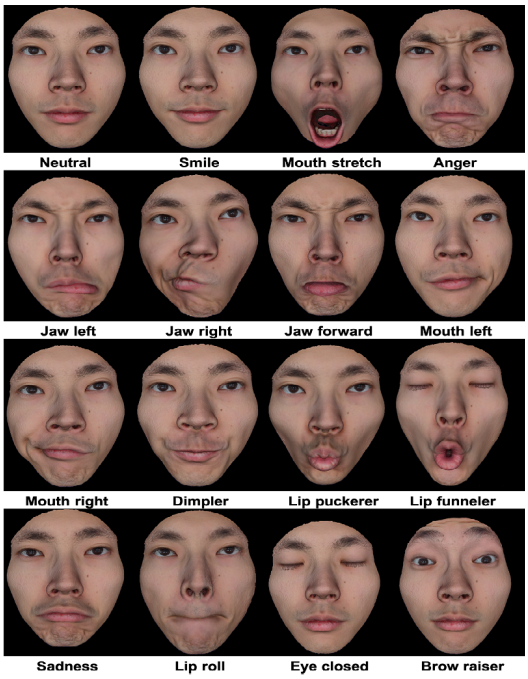

Fig. 5. Illustration of the preprocessed facial mesh in Facescape subdataset (subject 001), with 16 expressions (i.e., neutral, smile, mouth stretch, anger, jaw left, jaw right, jaw forward, mouth left, mouth right, dimpler, lip puckerer, lip funneler, sadness, lip roll, eye closed, and brow raiser).

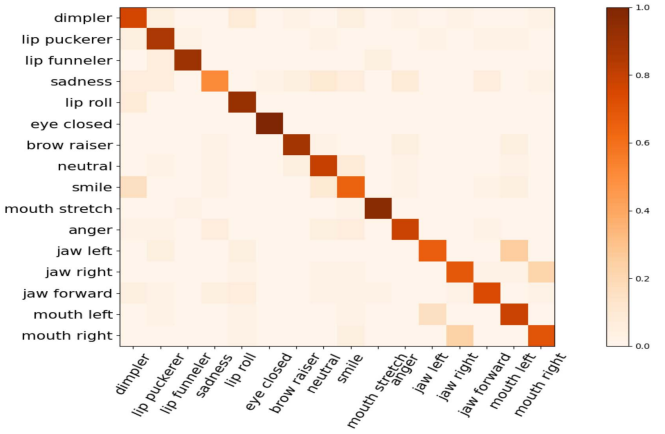

Fig. 6. Confusion matrix on the AMDFN model. The leftmost column shows the ground truth and the row in the bottom shows the type of expression which the sample is classified.

B. AMDFN Evaluation

We first evaluate the AMDFN performance, which recognizes 78.92% accuracy of 16 typical expressions. The detailed

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

TABLE II

COMPARISON OF NETWORK COMPLEXITY OF DATA-LEVEL FUSION, FEATURE-LEVEL FUSION, AND SCORE-LEVEL FUSION

Schemes	Parameters	Running time (per sample)
Score-level fusion	68,795,616	7.28ms
Feature-level fusion	68,795,536	7.16ms
Data-level fusion	29,748,708	1.33ms

performance of AMDFN is provided in the confusion matrix in Fig. 6, in which the leftmost column shows the ground truth label, and the bottom row shows the predicted label. A closer look at the figure reveals that, among the 16 typical expressions, there are six expressions (i.e., *lip puckerer*, *lip funneler*, *lip roll*, *eye closed*, *brow raiser*, and *mouth stretch*) with higher accuracy over 85%; especially, *eye closed* expression recognition accuracy has reached 100%. It keeps consistent with the fact that similarly located AUs can lead to a high misclassification rate, and the expressive AUs tend to get good recognition accuracy. We can also explain this fact from the definition of facial muscles. It can be found from the figure that the *lip roll* has a part identified as *dimpler*. Because *lip roll* expression mainly occurs in orbicular muscle of mouth, this causes the classifier to easily misclassify it to the expression associated with the orbicular muscle of mouth. In addition, we notice that the accuracy of the recognition results is low in the three basic expressions (i.e., *smile*, *anger*, and *sadness*). This is because the basic expression has a variety of AU components, which will be confused with a single AU. For instance, there are five main AU components (i.e., AU1, AU4, AU7, AU15, and AU17) of *sadness*, which causes easy confusion with *jaw forword*. From this fact, it demonstrates the effectiveness of 3-D FER from data-level fusion.

In order to prove that the AMDFN framework reduces computing time and memory consumption, we calculated the network complexity of data-level fusion, feature-level fusion, and score-level fusion on the Facescape subset, as shown in Table II. We use the same backbone network (i.e., ResNet18) to be able to make a fair comparison. In feature-level fusion, we use ResNet18 to extract the 64-D features for each 2-D facial attribute and then cascade the features to perform expression classification. In the score-level fusion, since the final expression classification result is a vote of the expression classification results of different 2-D facial attributes, the parameter amount is a multiple relationship of the expression recognition parameter amount for a single 2-D attribute. From the table, we can see that in the case of the same backbone network, compared to feature-level fusion and score-level fusion, the amount of parameters required for data-level fusion of AMDFN is the lowest, and the average test time per sample is also the lowest.

C. Ablation Studies

The influence of different attribute images and without LAM on FER is also shown in Table III. We can observe that regardless of the input multichannel image or single-channel image, the result of using the LAM is always higher than that of not using the LAM, which proves the importance of the LAM. The LAM improves the representation ability of features by assigning different attention weights to features of different layers, thus improving the final FER performance.

TABLE III

EFFECTIVENESS OF THE PROPOSED METHOD FOR FER

Method	Without LAM	With LAM
$AMDFN-C$	71.81%	73.14%
$AMDFN-D$	73.75%	74.68%
$AMDFN-T$	72.67%	74.16%
$AMDFN-N_x$	74.16%	75.23%
$AMDFN-N_y$	74.54%	75.67%
$AMDFN-N_z$	73.65%	74.45%
$AMDFN$	76.84%	78.92%

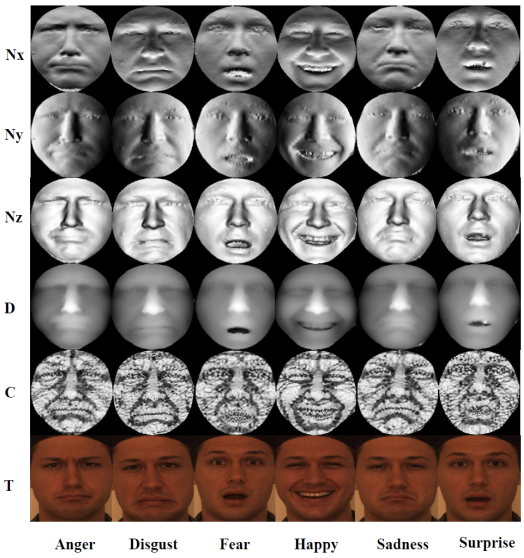

Fig. 7. Display of eight 2-D attribute images mapped from six expressions (i.e., anger, disgust, fear, happy, sadness, and surprise) of the Bosphorus dataset (subject 008). From top to bottom, three normal images (along the x-, y-, and z-directions), the depth image, curvature image, and texture image.

In addition, we found from the results that no matter which channel data is input, the final multichannel fusion result is always the highest. This is because multichannel data fusion combines multiple channel information after mapping 3-D facial scans, avoiding the loss of expression features when using a single-channel number for FER.

D. Comparison With State-of-the-Art Studies

Facescape subset is a newly proposed expression dataset. There are currently no state-of-the-art studies on this dataset. Thus, we compare the proposed method with several state-of-the-art FER methods on the Bosphorus dataset [7], which contains 105 subjects (i.e., 4666 3-D face scans) and 2-D face images with different AUs, facial expressions, poses, and occlusions. We select 60 subjects from the Bosphorus dataset, which performs the six basic expressions with near frontal view. Fig. 7 shows eight types of 2-D facial attribute images with six facial expressions of subject 008 in the Bosphorus subdataset. In this experiment, eight 2-D facial attribute images of these six basic expressions are used and ten-fold cross validation is performed. Table I presents the comparison of performance between the best of the proposed approach and state-of-the-art studies on the Bosphorus dataset. From Table I,

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

we can find that the proposed method achieves 83.13% accuracy, which is significantly better than the others. Note that the multimodal method (2-D + 3-D) [10], [15], [16] performs better than the single modal method [33], [34], which is due to the strong complementarity of multiple data modalities.

V. Conclusion

In this work, we proposed a novel and effective 3-D FER method, which learned discriminative expression representations related to facial expression mainly through AMDFN. In this method, we first merged the 2-D facial attribute images mapped from the 3-D face scan into the ResNet-18 backbone to learn facial expression features, and then, the dependencies between features of different layers are obtained through the layer attention network, which improves the representation ability of features by assigning different attention weights to features of different layers. Experimental results provided reliable evidence that the proposed method can recognize basic expressions and AUs on the Facescape subset and outperforms state-of-the-art methods on the Bosphorus dataset.

In the future, we will use two or more emotionally rich and tightly coupled modalities to solve the emotional recognition problem. Moreover, we will build a large-scale and high-precision 3-D in-wild expression dataset and plan to adopt a domain adaptation method to reduce domain shift.

References

[1] J. J. Gross and R. F. Muñoz, "Emotion regulation and mental health," *Clin. Psychol., Sci. Pract.*, vol. 2, no. 2, pp. 151–164, 1995.

[2] M. Berking and P. Wupperman, "Emotion regulation and mental health: Recent findings, current challenges, and future directions," *Current Opinion Psychiatry*, vol. 25, no. 2, pp. 128–134, 2012.

[3] R. Chai *et al.*, "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 715–724, May 2017.

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[5] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof, "A thermal infrared face database with facial landmarks and emotion labels," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 5, pp. 1389–1401, May 2019.

[6] Y. T. H. Zhu. (2019). *Facescape Database*. [Online]. Available: https://cite.nju.edu.cn/facescape.html

[7] A. Savran *et al.*, "Bosphorus database for 3D face analysis," in *Proc. Eur. Workshop Biometrics Identity Manage.* Berlin, Germany: Springer, 2008, pp. 47–56.

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

[9] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[10] 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 Process.*, vol. 161, pp. 74–88, Aug. 2019.

[11] S. Z. Gilani and A. Mian, "Learning from millions of 3D scans for large-scale 3D face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1896–1905.

[12] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model-based automatic 3D/4D facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1438–1450, Jul. 2016.

[13] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino, and J. Torresen, "A facial expression recognition system using robust face features from depth videos and deep learning," *Comput. Electr. Eng.*, vol. 63, pp. 114–125, Oct. 2017.

[14] W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4D facial expression recognition using dynamic geometrical image network," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 24–30.

[15] 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.

[16] 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.

[17] X. Zhao, D. Huang, E. Dellandrea, and L. Chen, "Automatic 3D facial expression recognition based on a Bayesian belief net and a statistical facial feature model," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3724–3727.

[18] A. Savran, B. Sankur, and M. T. Bilge, "Facial action unit detection: 3D versus 2D modality," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 71–78.

[19] B. Fasel, "Robust face analysis using convolutional neural networks," in *Proc. Object Recognit. Supported User Interact. Service Robots*, Aug. 2002, pp. 40–43.

[20] B. Fasel, "Head-pose invariant facial expression recognition using convolutional neural networks," in *Proc. 4th IEEE Int. Conf. Multimodal Interface*, Oct. 2002, pp. 529–534.

[21] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 553–560.

[22] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[23] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Trans. Affective Comput.*, vol. 9, no. 3, pp. 343–350, Jul./Sep. 2018.

[24] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[25] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2121–2129.

[26] W. Xie, X. Jia, L. Shen, and M. Yang, "Sparse deep feature learning for facial expression recognition," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106966.

[27] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5683–5692.

[28] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9841–9850.

[29] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "Expressive maps for 3D facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1270–1275.

[30] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-D face and facial expression recognition," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 498–511, Sep. 2008.

[31] 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 (MM)*, Oct. 2009, pp. 569–572.

[32] M. D. Cordea, E. M. Petriu, and D. C. Petriu, "Three-dimensional head tracking and facial expression recovery using an anthropometric muscle-based active appearance model," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 8, pp. 1578–1588, Aug. 2008.

[33] 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 Recognit. (ICPR)*, Nov. 2012, pp. 2577–2580.

[34] 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. 1–10, 2014.

[35] P. Zarbakhsh and H. Demirel, "Fuzzy SVM for 3D facial expression classification using sequential forward feature selection," in *Proc. 9th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Sep. 2017, pp. 131–134.

[36] S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo, "3D facial expression recognition using sift descriptors of automatically detected keypoints," *Vis. Comput.*, vol. 27, no. 11, p. 1021, 2011.

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.

[37] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3D facial expression recognition," *Pattern Recognit.*, vol. 44, no. 8, pp. 1581–1589, 2011.
[38] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1399–1406.
[39] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: The Manual on CD ROM*. Salt Lake City, UT, USA: A Human Face, 2002, pp. 254–277.
[40] K. Wang, Y. Gu, X. Peng, P. Zhang, B. Sun, and H. Li, "AU-guided unsupervised domain adaptive facial expression recognition," 2020, *arXiv:2012.10078*.
[41] T. Pu, T. Chen, Y. Xie, H. Wu, and L. Lin, "AU-expression knowledge constrained representation learning for facial expression recognition," 2020, *arXiv:2012.14587*.
[42] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4678–4683.
[43] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1–9.
[44] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 698–700, Sep. 1987.
[45] A. Mian, M. Bennamoun, and R. Owens, "Automatic 3D face detection, normalization and recognition," in *Proc. 3rd Int. Symp. 3D Data Process., Visualizat., Transmiss. (3DPVT)*, Jun. 2006, pp. 735–742.
[46] J. J. Koenderink and A. J. Van Doorn, "Surface shape and curvature scales," *Image Vis. Comput.*, vol. 10, no. 8, pp. 557–564, 1992.
[47] J. Goldfeather and V. Interrante, "A novel cubic-order algorithm for approximating principal direction vectors," *ACM Trans. Graph.*, vol. 23, no. 1, pp. 45–63, Jan. 2004.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[50] Y. LeCun, D. Touresky, G. Hinton, and T. Sejnowski, "A theoretical framework for back-propagation," in *Proc. Connectionist Models Summer School*, vol. 1. Pittsburgh, PA, USA: Morgan Kaufmann, 1988, pp. 21–28.
[51] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.
[52] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
[53] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[54] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[55] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," 2020, *arXiv:2008.00923*.
[56] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-CELEB-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 87–102.

Yu Gu (Senior Member, IEEE) received the B.E. degree from the Special Classes for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2004, and the D.E. degree from the University of Science and Technology of China in 2010.

In 2006, he was an Intern with Microsoft Research Asia, Beijing, China, for seven months. From 2007 to 2008, he was a Visiting Scholar with the University of Tsukuba, Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor and the Dean Assistant with the School of Computer and Information, Hefei University of Technology, Hefei. His current research interests include pervasive computing and affective computing.

Dr. Gu is a member of the Association for Computing Machinery (ACM). He was a recipient of the IEEE Scalcom2009 Excellent Paper Award and the NLP-KE2017 Best Paper Award.

Huan Yan was born in Guizhou, China, in 1995. He received the B.E. degree from the Hefei University of Technology, Hefei, China, in 2017, where he is currently pursuing the Ph.D. degree.

His research interests include intelligent information processing and wireless sensing and affective computing.

Xiang Zhang was born in Anhui, China, in 1996. He received the B.E. degree from the Hefei University of Technology, Hefei, China, in 2017, where he is currently pursuing the Ph.D. degree.

His research interests include intelligent information processing and wireless sensing and affective computing.

Zhi Liu (Senior Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2009, and the Ph.D. degree in informatics from the National Institute of Informatics, Tokyo, Japan, in 2014.

He was a Junior Researcher (Assistant Professor) at Waseda University, Tokyo, Japan, and a JSPS Research Fellow with the National Institute of Informatics. He is currently an Assistant Professor at Shizuoka University, Shizuoka, Japan. His research interests include video network transmission, vehicular networks, and mobile edge computing.

Dr. Liu is a member of IEICE. He was a recipient of the IEEE Stream-Comm2011 Best Student Paper Award, the 2015 the Institute of Electronics, Information and Communication Engineers (IEICE) Young Researcher Award, and the ICOIN2018 Best Paper Award. He has been serving as the chair for a number of international conferences and workshops. He has been a Guest Editor of journals, including *Wireless Communications and Mobile Computing*, *Sensors*, and IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS.

Fuji Ren (Senior Member, IEEE) received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 1982 and 1985, respectively, and the Ph.D. degree from Hokkaido University, Sapporo, Japan, in 1991.

He is currently a Professor with the Faculty of Engineering, University of Tokushima, Tokushima, Japan. His research interests include information science, artificial intelligence, language understanding and communication, and affective computing.

Dr. Ren is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Chinese Association for Artificial Intelligence (CAAI), the Institute of Electrical Engineers of Japan (IEEJ), the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence (JSAI), and the Asia–Pacific Association for Machine Translation (AAMT). He is a fellow of the Japan Federation of Engineering Societies. He is also the President of the International Advanced Information Institute.

Authorized licensed use limited to: HEFEI UNIVERSITY OF TECHNOLOGY. Downloaded on August 11,2022 at 03:40:32 UTC from IEEE Xplore. Restrictions apply.