Orthogonalization-Guided Feature Fusion Network for Multimodal 2D+3D Facial Expression Recognition

Shisong Lin, Mengchao Bai, Feng Liu, Linlin Shen, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—As 2D and 3D data present different views of the same face, the features extracted from them can be both complementary and redundant. In this paper, we present a novel and efficient orthogonalization-guided feature fusion network, namely OGF^{2}Net, to fuse the features extracted from 2D and 3D faces for facial expression recognition. While 2D texture maps are fed into a 2D feature extraction pipeline (FE2DNet), the attribute maps generated from 3D data are concatenated as input of the 3D feature extraction pipeline (FE3DNet). The two networks are separately trained at the first stage and frozen in the second stage for late feature fusion, which can well address the unavailability of a large number of 3D+2D face pairs. To reduce the redundancies among features extracted from 2D and 3D streams, we design an orthogonal loss-guided feature fusion network to orthogonalize the features before fusing them. Experimental results show that the proposed method significantly outperforms the state-of-the-art algorithms on both the BU-3DFE and Bosphorus databases. While accuracies as high as 89.05% (P1 protocol) and 89.07% (P2 protocol) are achieved on the BU-3DFE database, an accuracy of 89.28% is achieved on the Bosphorus database. The complexity analysis also suggests that our approach achieves a higher processing speed while simultaneously requiring lower memory costs.**

Manuscript received December 16, 2019; revised April 5, 2020 and May 26, 2020; accepted May 31, 2020. Date of publication June 11, 2020; date of current version May 26, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 61672357, 91959108, and U1713214, in part by the Science and Technology Project of Guangdong Province under Grant 2018A050501014, in part by the Shenzhen Fundamental Research fund under Grants JCYJ20190808163401646 and JCYJ20180305125822769, in part by the Science and Technology Development Fund, Macau SAR (File no. 189/2017/A3), and in part by the Research Committee at the University of Macau under Grants MYRG2016-00123-FST and MYRG2018-00136-FST. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lei Zhang. *(Shisong Lin and Mengchao Bai contributed equally to this work.) (Corresponding author: Linlin Shen.)*

Shisong Lin, Mengchao Bai, and Feng Liu are with Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518172, China (e-mail: linshisong2018@email.szu.edu.cn; baimengchao2017@email.szu.edu.cn; feng.liu@szu.edu.cn).

Linlin Shen is with the Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China, with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518172, China, and also is a Distinguished Visiting Scholar at the Faculty of Science and Technology, University of Macau, Macau 999078, China (e-mail: llshen@szu.edu.cn).

Yicong Zhou is with the Faculty of Science and Technology, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.3001497

***Index Terms*—Multimodal facial expression recognition, feature fusion.**

I. Introduction

FACIAL expression is generated by one or more motions of the muscles beneath the skin of the face. In daily communication, facial expressions are important ways to express the emotional reaction of a person to observers. Generally, basic facial expressions can be divided into six categories: anger, disgust, fear, happiness, sadness and surprise. In the field of computer vision and machine learning, numerous studies have been conducted on facial expression recognition (FER) due to its practical importance in sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems [1].

Existing FER approaches in the literature can be classified into three categories according to their data modalities: 2D FER, 3D FER and multimodal 2D+3D FER. Comprising the majority of FER approaches, 2D FER methods [2]–[9] are based on 2D face images or videos. Despite the significant advances that have been achieved, the 2D FER methods still fail to solve challenges related to variations in illumination and pose [10].

With the increasing development of 3D sensors and scanning technologies, various approaches have been proposed for 3D FER. 3D FER methods can be divided into three categories: model-based methods [11], [12], feature-based methods [13]–[15] and deep learning based methods [16]–[18]. Different from 2D data, 3D scans (captured by infrared devices) are robust to pose, illumination and scale variation, so 3D scans can address more complex scenes. However, infrared images usually fail to capture subtle facial deformations, such as skin wrinkles [19].

Considering the performance limitations of single-mode data and the large complementarity among different modalities, using both 2D and 3D data [10], [20]–[25] for FER is becoming a prospective research topic. Although significant progress has been made in 3D FER and 2D+3D FER, unsolved problems still remain. While some facial expressions are easy to distinguish, others are not. For example, happiness and surprise can be easily identified, but sadness and fear are often confused in [10]. Zhu *et al.* [21] simply design six attention-based pipelines for feature extraction to address this issue. Then, the extracted features are directly fused into a high-level representation for FER. There is no doubt that the features extracted from 2D and 3D attribute

maps are complementary, but the maps are all from the same face and may have similarities as well. The extracted features can be potentially redundant, which may have a negative effect on the FER, so it is necessary to reduce redundancy before feature fusion.

Considering the above issues, in this paper, we propose a novel approach using both 2D and 3D data for FER. We combine all attribute maps generated by a 3D face as the input of the 3D feature extraction pipeline (FE3DNet), and the 2D texture maps are fed into the 2D feature extraction pipeline (FE2DNet). The global weighted pooling (GWP) module is utilized to vectorize the feature maps of the last convolutional layers in FE3DNet and FE2DNet. While 2D and 3D data represent different views of the same face, different backbone networks tailored to various image streams might be more appropriate for extracting different features. As a large number of 2D+3D face pairs are not available in the literature, separately training the two networks and fusing their features at a late stage is a more feasible approach. The capability of 2D and 3D face models can thus be fully explored.

We further design an orthogonal loss to map the two features into an orthogonal space to remove the redundancies among them. The orthogonalized features are then fused to combine different information for further performance improvement.

Our contributions are as follows:

- This paper presents a novel CNN model, namely OGF^{2}Net, for 2D+3D FER. Our model uses two networks, FE2DNet and FE3DNet, to extract facial features from 2D and 3D faces. The two networks are separately trained at the first stage and frozen in the second stage for late feature fusion, which can well address the unavailability of a large number of 3D+2D face pairs.
- As the redundancy between features extracted from 2D and 3D data may have a negative impact on the classification results for FER, we propose a novel orthogonal loss to reduce the correlation of the features learned in the two pipelines.
- The proposed method achieves state-of-the-art performance on both the BU-3DFE and Bosphorus databases.

The rest of paper is organized as follows. Section II briefly introduces the related work on 2D FER, 3D FER and multimodal 2D+3D FER. Section III describes the proposed method, including a detailed description of map generation, the network architecture and the objective function. Experimental results are shown in Sections IV, and V concludes the paper.

II. Related Work

A. 2D FER

As the major FER method, 2D FER approaches can be divided into two main categories based on the type of data: static image and dynamic sequence. Static-based methods [2]–[5] only extract the features containing spatial information from a single image. For example, given a single image, Burkert *et al.* [2] propose a convolutional neural network (CNN) for 2D FER. They achieve 99.6% accuracy and 98.63% accuracy on the CK+ [26] and MMI [27] datasets. Meng *et al.* [3] believe that the subject's identity attributes, i.e., age, gender and personal characteristics, are nonlinearly coupled with facial expressions. They propose an identity-aware CNN network to capture both expression-related and identity-related features for facial expression recognition. Yang *et al.* [5] propose a De-expression Residue Learning (DeRL) technique to extract representations of the expressive components. Using a generative network trained by cGAN [28], the neutral face image is generated from the input face with different expressions. The expressive features deposited in the intermediate layers of the generative model are extracted and concatenated for FER. Different from static-based methods, sequence-based methods [6]–[9] not only focus on the spatial features of a single image, but also consider the temporal features of a continuous sequence. For example, Jung *et al.* [6] use two different deep networks to extract the high-level features. The first deep network extracts temporal appearance features from image sequences, while the other network extracts geometry features from temporal facial landmark points. These two features are combined to boost the performance of FER. They achieve 97.25% accuracy on the CK+ database. Zhang *et al.* [8] propose a part-based hierarchical bidirectional recurrent neural network (PHRNN) to extract the temporal features of facial landmarks from a continuous sequence. In addition, to complement the static appearance information, they also introduce a multi-signal convolutional neural network (MSCNN) to extract the spatial features. Finally, the temporal and spatial features extracted by two networks are fused for FER. However, since 2D data mainly contain appearance information, it is difficult to deal with diverse real scenes by using only 2D data.

B. 3D FER

Traditional 3D FER approaches are mainly model-based [11], [12] and feature-based [13]–[15] methods. Model-based methods use training data to generate a generic expression deformable model, which can be adapted to fit new facial scans. The parameters generated by the models are finally used as expression features for prediction. For example, Mpiperis *et al.* [11] propose a bilinear elastically deformable model to build correspondence among a set of facial meshes. Given a new facial scan, the trained model can estimate the expression and identity parameters, which are used for expression prediction. Gong *et al.* [12] suggest that the shape of an expressional 3D face can be composed of two parts, a basic facial shape component (BFSC) and an expressional shape component (ESC). The BFSC represents the facial structure and neutral-style shape, and the ESC can be used to extract facial expression features. Feature-based approaches extract local expression features and then combine these features for 3D FER. For instance, Berretti *et al.* [13] compute and select the SIFT features on facial landmarks of depth images and feed them into a support vector machine (SVM) for 3D FER. Maalej *et al.* [14] utilize a Riemannian framework to compute the distance of the geodesic path between corresponding patches. Multi-boosting and an SVM classifier are used to recognize the six facial expressions.

However, model-based methods need to build dense correspondence between face scans, and feature-based approaches require extracting the hand-crafted features, which greatly limit

the performance and generalization capabilities of these methods. In recent years, deep learning based methods have been widely applied to 3D FER due to their good performance. For example, Li *et al.* [16] use a pre-trained CNN to extract deep features from attribute maps. Then, expression prediction is implemented by training linear SVMs over the deep features and fusing these SVM scores. The average recognition rate they achieve is 84.87% on BU-3DFE [29]. Yang *et al.* [17] concatenate depth maps, curvature maps and landmark mask maps to create a 3-channel image to train a CNN for FER. The average recognition rate they obtain is 75.9% on BU-4DFE [30]. Chen *et al.* [18] propose a fast and light manifold CNN model (FLM-CNN) that adopts a human vision-inspired pooling structure and achieves better performance in efficiency and feature extraction. They report an average recognition rate of 86.67% on BU-3DFE. However, using only 3D data makes it difficult to focus on the appearance details of facial expression images, resulting in limited performance.

C. 2D+3D FER

Recently, some works [10], [20]–[25] have utilized 2D and 3D multimodal data to further improve the performance of FER. For instance, Li *et al.* [10] generate six 2D facial attribute maps from a 3D face model. These attribute maps are jointly fed into a feature extraction subnet, generating hundreds of multichannel feature maps. All of these feature maps are then fed into a feature fusion subnet, resulting in a high-level deep feature. Then, a linear SVM classifier is used for prediction. They achieve 86.86% accuracy on BU-3DFE. Rather than generating 2D facial attribute maps, Jan *et al.* [22] crop four facial parts (i.e., the eyebrows, eyes, mouth and nose) from texture and depth images according to the located facial landmarks. Then, deep fusion CNN features are learned from these facial parts and input into a nonlinear SVM. Their recognition rate reaches 88.54%. Zhu *et al.* [21] propose an attention-based CNN to focus on emotional salient regions. They also design a dimensional distribution(DD) loss to learn more discriminative features. They report an average accuracy of 88.35% on BU-3DFE. FER based on 2D+3D multimodal data is becoming a promising research topic due to the assumption that there exist large complementarities among different modalities [10].

D. Loss-Guided Feature Transform

Nguyen and Bai [31] try to learn a transformation matrix to minimize the angle θ between feature vectors extracted from two faces of the same subject, i.e., maximize $\cos\theta$. In [32], island loss is proposed by Cai *et al.* to push apart the centers of different clusters to improve the recognition performance. In this work, the value of (1+ $\cos\theta$) is minimized, i.e., the angle between the features of cluster centers is pushed towards π. In [33], the absolute value of $\cos\theta$ is minimized to make person-specific shape features independent of the local relationship features learned from the other stream. The loss is applied to training the two networks for both streams.

The aim of our loss, $\cos^2\theta$, is to map the two features extracted from different streams to an orthogonal space such that the angle θ between them is close to $\frac{\pi}{2}$. Our loss function is differentiable and more easily converges. To address the unavailability of a large number of 3D+2D face pairs, the loss is only applied to supervise the training of the fusion subnetwork. The two networks of 2D and 3D streams are separately trained in the first stage and frozen in the training of the fusion network.

III. The Proposed Method

A. Overview

Fig. 1 illustrates the framework of the proposed OGF^{2}Net for 3D+2D FER. Given a 3D face mesh and its corresponding 2D texture image I_t, we first generate three well-aligned facial attribute maps, i.e., the depth map (I_d), azimuth map (I_a) and elevation map (I_e), to represent the 3D face scan. The three maps are then combined into a three-channel map (I_{dae}). I_t and I_{dae} are fed into two different pipelines, namely, FE2DNet and FE3DNet, to extract two 512-dimensional vectors V_1 and V_2, which are vectorized using the proposed GWP module from the feature maps of the last convolutional layers. Then, we design an orthogonalization-guided (OG) module supervised by the orthogonal (Orth) loss to generate two orthogonal features F_1 and F_2. Finally, we feed the feature fusion subnet with F_1 and F_2 and adopt a classifier with an FC layer to compute the probability of each expression.

B. Attribute Map Generation

Each 3D face mesh and its corresponding 2D texture image are used to generate a three-channel image I_{dae} and a gray image I_t, respectively. As the first channel of I_{dae}, the depth image is generated by fitting a surface from the 3D point cloud in the form of $z(x, y)$ using the gridfit algorithm [34]. The surface normal of each point cloud is computed in spherical coordinates (θ, ϕ), where θ and ϕ are the azimuth and elevation angles of the normal vector, respectively. Using a similar (x, y) grid as the depth image, surfaces of the form $\theta(x, y)$ and $\phi(x, y)$ are fitted to the azimuth and elevation angles to generate the second and third channels of I_{dae} as in [35]. Since color information has little effect on the expression category, we convert the color image to a gray image as the input. Fig. 2 illustrates the three attribute maps (from the second row to the fourth row), gray texture images (the first row) and depth-azimuth-elevation maps (the last row).

C. Network Architecture

1) Feature Extraction: There are two pipelines in the feature extraction process, i.e., FE2DNet and FE3DNet.

FE2DNet. The architecture of FE2DNet is a variant of ResNet [36] as demonstrated in Fig. 3(a) FE2DNet is composed of 23 convolutional layers and a GWP module. All convolutional layers use 3 $\times$ 3 filters and follow two design rules: if the size of the output feature map is the same as that of the input feature map, then the convolutional layers have the same number of filters; if the size of the feature map is halved, then the number of filters in the convolutional layer is doubled except for the filters in the first convolutional layer. We implement downsampling by

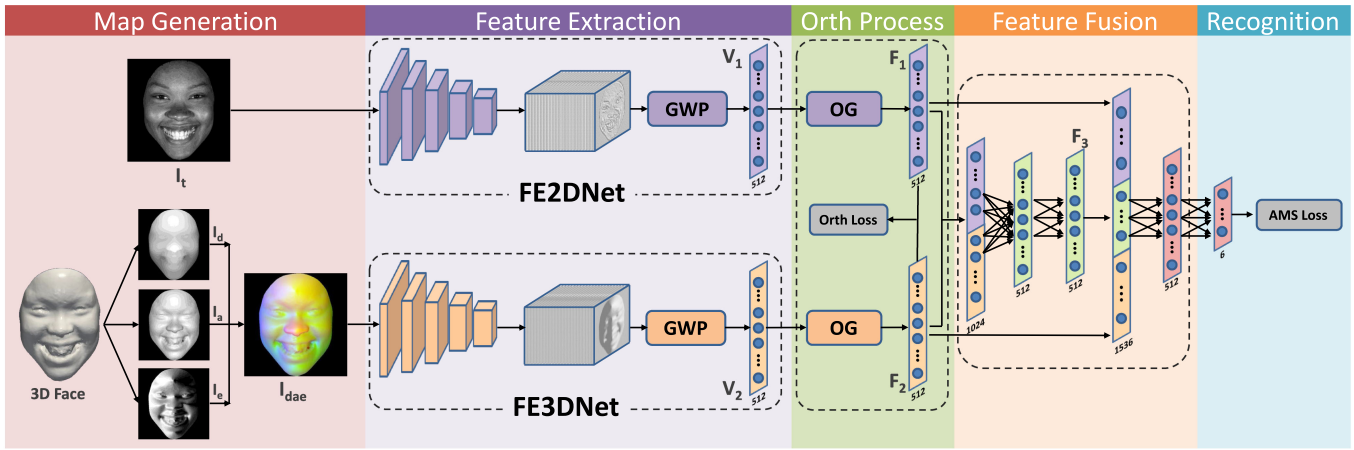


Fig. 1. Overview of OGF^{2}Net. Each 3D face mesh is represented as a three-channel map (I_{dae}). I_{dae} and 2D texture image I_t are fed into the feature extraction pipelines (FE3DNet and FE2DNet, respectively) to generate two 512-dimensional vectors. These two vectors are fed into the OG module, resulting in two orthogonal features, which are then input into a feature fusion subnet. Finally, we adopt a classifier with two FC layers to compute the probability of each expression.

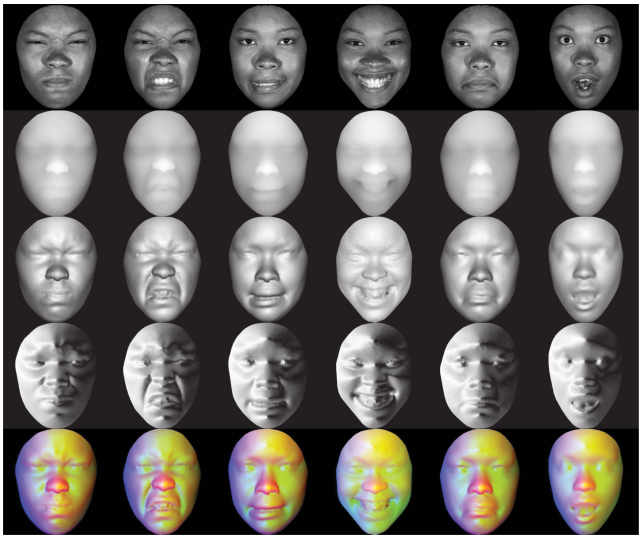
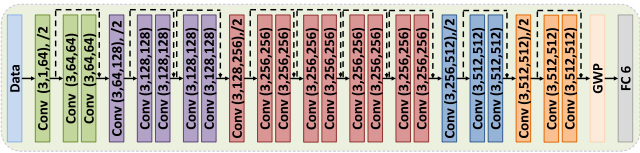
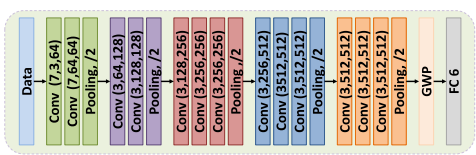


Fig. 2. Examples of gray texture images, three attribute maps and combined RGB images for different expressions, i.e., anger, disgust, fear, happiness, sadness, and surprise. From top to bottom are gray texture maps I_t, depth maps I_d, azimuth maps I_a, elevation maps I_e, and combined maps I_{dae}.



(a) FE2DNet.



(b) FE3DNet.

Fig. 3. Architecture of the proposed FE2DNet and FE3DNet.

convolutional layers with a stride of 2, and each convolutional layer is followed by a batch normalization [37] and PReLu [38] layers. The GWP module is described in detail in Section III-C2.

FE3DNet. As shown in Fig. 3(b), the skeleton architecture of FE3DNet is a modified version of FR3DNet [35], whose last three fully connected layers are replaced with a GWP module. FE3DNet consists of 13 convolutional layers, 5 max pooling layers and a GWP module. Note that each convolutional layer is followed by an ReLu layer just as in [35]. The kernel size of the first two convolutional layers is 7×7, while that of the other convolutional layers is 3×3. All convolutional layers do not halve the size of feature maps. In addition, FE3DNet performs downsampling with the max pooling layer.

2) Global Weighted Pooling Module: Generally, many CNN architectures use a global average pooling (GAP) layer to vectorize the feature maps of the last convolutional layer. GAP is computed as:

$$y^k = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} x_{ij}^k}{MN} \tag{1}$$

where x^k and y^k are the components of the k_{th} feature map and the k_{th} feature vector, respectively. M and N are the length and width of the feature map. GAP achieves good performance in general object recognition. However, different from object recognition, the input face images of FER are well aligned such that corresponding areas contain fixed facial components. In high-level feature maps, each pixel represents a specific area of the input image and contains fixed semantic information. If we use GAP in FER, then it would obviously ignore some semantic information.

As shown in Fig. 4, we utilize a global weighted pooling (GWP) module to replace GAP and flatten layer. For each feature map, GWP is implemented by employing a weight map whose size is the same as that of the feature map. The output feature vector is calculated by the dot product of the feature map and

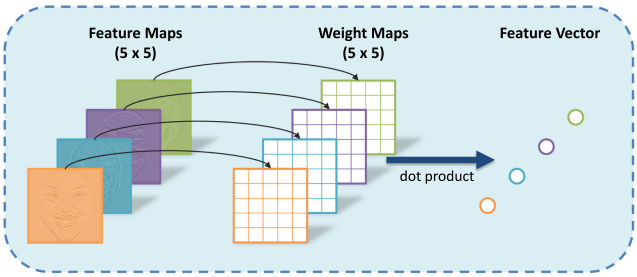


Fig. 4. Implementation of the global weighted pooling module. Each component of the feature vector is computed by the dot product between a feature map and its corresponding weight map.

weight map. GWP is formulated as:

$$y^k = \sum_{i=1}^{M} \sum_{j=1}^{N} x_{ij}^k w_{ij}^k \tag{2}$$

where w^k is the component of the k_{th} weight map. As shown in Eq. (2), GWP sets learnable weights for each pixel of the feature map, while GAP only provides average weights. After training with a large amount of faces, it pays more attention to the specific spatial areas. As shown in Fig. 1, two 512-dimensional vectors, V_1 and V_2, can be extracted by FE2DNet and FE3DNet integrated with GWP.

3) Orthogonal Loss-Guided Feature Orthogonalization: As V_1 and V_2 are extracted from different modalities of the same face, they certainly contain some complementary information about the face. However, there might be redundancy among them as well. To show the complementarity and redundancy of features extracted from different modalities, we visualize the expression-sensitive regions that are important to recognition on heat maps using the Grad-CAM method [39]. The visualization is calculated for features V_1 and V_2. In Fig. 5, the heat maps in the odd and even rows are extracted from FE2DNet and FE3DNet, respectively. Each column displays different facial expressions, i.e., anger, disgust, fear, happiness, sadness and surprise. From these heat maps, we observe some interesting phenomena. First, the mouth and eye are the most important regions for classifying all six expressions. Second, the sensitive regions of the first row are completely different from those of the second row, which represents the complementary characteristics between features extracted from FE2DNet and FE3DNet. However, some sensitive regions (such as the third and fourth rows of the sixth columns) are similar, which shows potential redundancy among them.

Thus, to alleviate the redundancy between features extracted from 2D and 3D data, we propose an orthogonalization-guided (OG) module to force the output of our model to be as orthogonal as possible. As shown in Fig. 6, our OG module separately takes V_1 and V_2 as the input, transforms them with an FC layer and outputs two orthogonal features $F_1 = [\alpha_1, \alpha_2, \ldots, \alpha_{512}]^T$ and $F_2 = [\beta_1, \beta_2, \ldots, \beta_{512}]^T$. The proposed OG module is supervised by the orthogonal loss to ensure the independence between

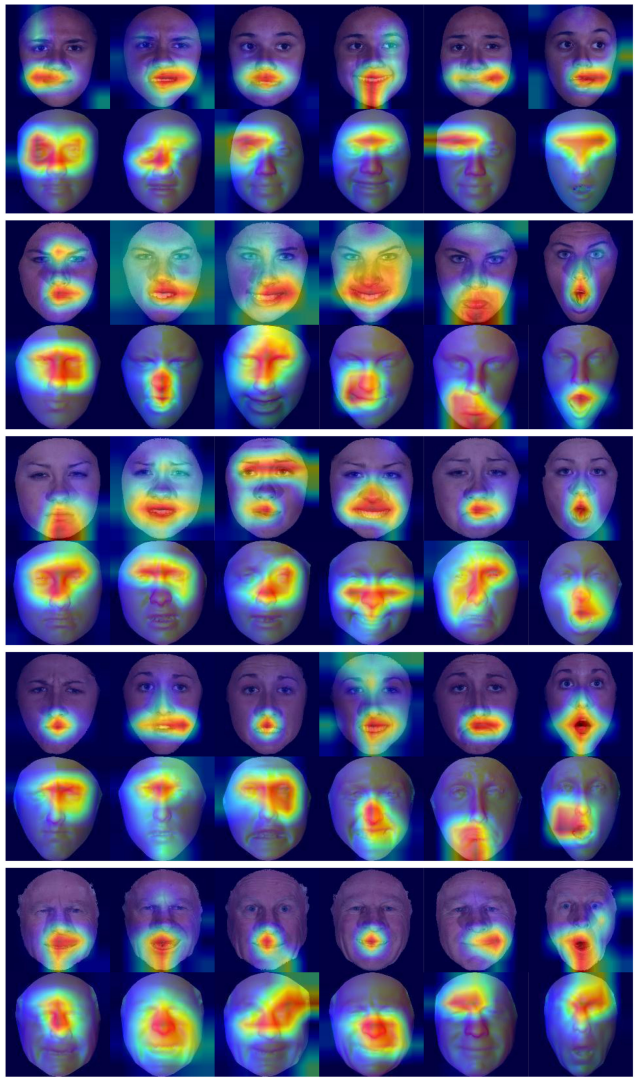


Fig. 5. Visualization of the heat maps from the two feature extraction pipelines (FE2DNet and FE3Dnet). The heat maps in the 1st, 3 rd, 5th, 7th, and 9th rows are extracted from FE2DNet. The heat maps in the 2nd, 4th, 6th, 8th and 10th rows are extracted from FE3DNet.

F_1 and F_2:

$$L_{Orth} = (\cos\theta)^2 = \left(\frac{F_1 \cdot F_2}{\|F_1\| \, \|F_2\|}\right)^2 \tag{3}$$

where θ is the angle between F_1 and F_2. We square $\cos(\theta)$ to ensure that L_{Orth} can be differentiated. In addition, only when $\theta = 90$ does $L_{Orth} = 0$ hold; otherwise, L_{Orth} is greater than 0.

4) Feature Fusion: As shown in Fig. 1, the main components of the feature fusion subnet are FC layers and concatenation operations. The inputs of the feature fusion subnet are two 512-dimensional features (F_1 and F_2) output by the proposed OG module. We concatenate F_1 and F_2, and utilize two FC layers to learn the feature F_3. Then, F_1, F_2 and F_3 are concatenated into a 1536-dimensional feature, which is further mapped by an FC

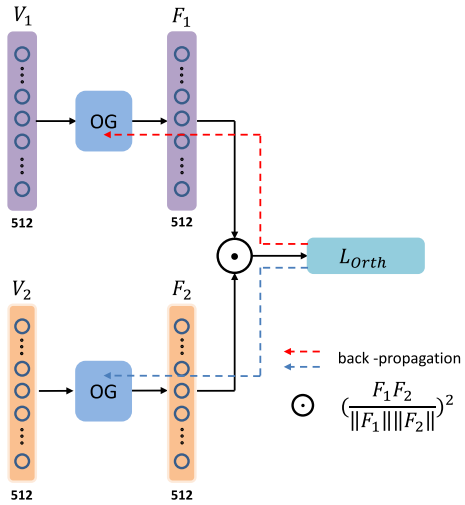


Fig. 6. Details of the orthogonalization process. The weights of the OG module can be updated under the supervision of Orth loss.

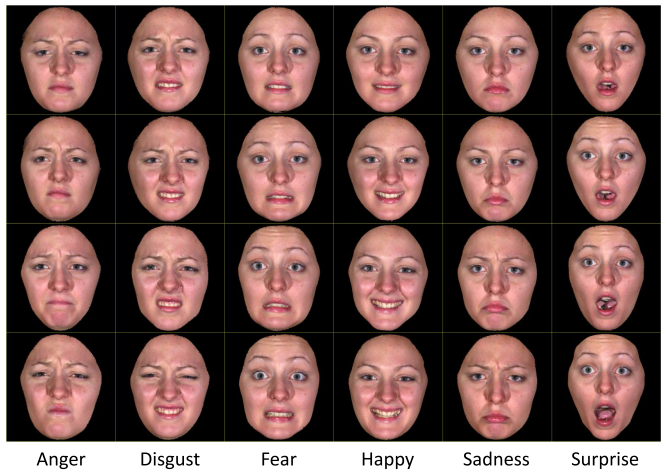


Fig. 7. Samples of 2D texture images of BU-3DFE with different expressions and four levels of expression intensity. From top to bottom are level 1 to level 4.

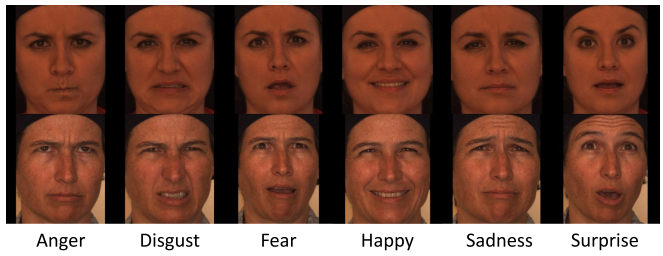


Fig. 8. Samples of 2D texture images of Bosphorus with different expressions. Each expression has only one level of expression intensity.

layer to a 512-dimensional feature. Finally, we adopt a classifier with an FC layer to compute the probability of each expression.

We use additive margin softmax (AMS) loss [40] to update the parameters of the feature fusion subnet and OG module. AMS loss adds an additive angular margin (m) and scaling factor (s) to softmax loss, making the decision more stringent, which is defined as:

$$L_{AMS} = -\frac{1}{N}\sum_i \log\left(\frac{e^{s(W_{y_i}^T f_i - m)}}{e^{s(W_{y_i}^T f_i - m)} + \sum_{j=1, j\neq y_i}^{c} e^{sW_j^T f_i}}\right) \tag{4}$$

Finally, our full objective is defined as:

$$L_{all} = \lambda \times L_{Orth} + L_{AMS} \tag{5}$$

where λ is a hyper-parameter to balance the two terms L_{AMS} and L_{Orth}.

D. Training Strategy

To speed up the convergence of the OGF^{2}Net training, we use a two-step training strategy to train our network in PyTorch. The first step is to train FE2DNet and FE3DNet using I_t and I_{dae}, respectively. The weights of FE2DNet are pre-trained on the publicly available dataset CASIA-WebFace [41]. We initialize the weights of convolutional layers in FE3DNet by the pre-trained model of FR3DNet [35]. During the first step, the hyper-parameters are set as follows:

- Batch size = 32;
- Max epoch = 100;
- The learning rates of the convolutional and FC layers are 0.001 and 0.0001, respectively, and are multiplied by 0.1 at 30 and 60 epochs;
- Adam [42] is used as the learning optimizer.

The second step is to train the feature fusion subnet while the parameters of FE2DNet and FE3DNet learned in the first step are fixed. During the second training step, most of the hyper-parameters have the same value as those in the first step, except for the following:

- Max epoch = 35;
- The learning rate is set as 0.001 and decreased by multiplying it by 0.1 at 16 and 24 epochs;
- $\lambda = 10$.

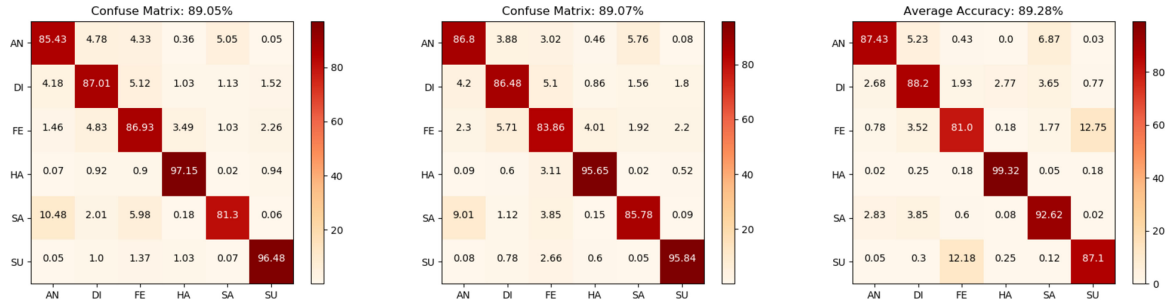
IV. Experimental Results

We compare our results with those of state-of-the-art methods and evaluate the effectiveness of the proposed method on the BU-3DFE [29] and Bosphorus [43] databases. The databases, protocols and results are described in the following sections.

A. Databases and Protocols

1) BU-3DFE Database: The BU-3DFE database contains 2500 scans of 100 subjects (56 females and 44 males), with ages from 18 to 70 years old. Each subject has 25 samples of seven expressions: one sample for neutral and the other 24 samples for six prototypical facial expressions, i.e., happiness, disgust, fear, anger, surprise and sadness. As shown in Fig. 7, each prototypical expression includes four levels of intensity.

2) Bosphorus Database: The Bosphorus database consists of 4666 scans captured from 105 subjects. While the face of 65 subjects present six prototypical expressions, other subjects only



(a) BU-3DFE (P1) (b) BU-3DFE (P2) (c) Bosphorus

Fig. 9. Confusion matrix of OGF^{2}Net for the BU-3DFE and Bosphorus databases. The labels on the vertical and horizontal axis represent the ground truth and predicted expressions, respectively. AN, DI, FE, HA, SA, and SU are abbreviations for anger, disgust, fear, happiness, sadness and surprise, respectively.

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON BU-3DFE

Approach	Feature	Accuracy(%)	
		P1	P2
Yang et al.(2015) [44]	Hand-crafted	84.80	82.73
Zhen et al.(2015) [15]	Hand-crafted	84.50	83.20
Li et al.(2015) [20]	Hand-crafted	86.32	-
Li et al.(2017) [10]	Deep learning	86.86	-
Chen et al.(2018) [18]	Deep learning	86.67	85.96
Wei et al.(2018) [45]	Deep learning	88.03	-
Jan et al.(2018) [22]	Deep learning	88.54	-
Zhu et al.(2019) [21]	Deep learning	88.35	87.06
FE3DNet	Deep learning	85.20	85.13
FE2DNet	Deep learning	86.58	86.69
OGF^{2}Net	Deep learning	**89.05**	**89.07**

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON BOSPHORUS

Approach	Feature	Accuracy(%)
Li et al.(2015) [20]	Hand-crafted	79.72
Yang et al.(2015) [44]	Hand-crafted	77.50
Li et al.(2017) [10]	Deep learning	80.28
FE3DNet	Deep learning	83.55
FE2DNet	Deep learning	85.73
OGF^{2}Net	Deep learning	**89.28**

present some of the six expressions. For each of the 65 subjects, a 3D+2D pair is available for each expression, as shown in Fig. 8.

3) Protocols: For the BU-3DFE database, we follow the two protocols (i.e., P1 and P2) presented in [15], [21], [44]. In P1, we randomly select 60 subjects with the third and fourth expression intensity levels from all 100 subjects. These 60 subjects are fixed in the entire experiment. Then the 10-fold cross-validation method is adopted: the 60 subjects are divided into 10 subsets. Every subset includes 6 subjects and is used as testing data, while the remaining 9 subsets are used as training data. The validation is repeated 100 times to obtain a stable result. In P2, the only difference from P1 is that we randomly select 60 subjects from all 100 subjects for each testing sequence.

As for the Bosphorus database, we follow the same protocol as that of [10]: 60 subjects are randomly selected from 65 subjects and divided into 10 subsets. Every subset including 6 subjects is used as testing data, while the remaining 9 subsets are used as training data. The testing is repeated 100 times and the average accuracy is recorded.

B. Results

1) BU-3DFE Database: Table I lists the performance comparison against other approaches that follow the same protocols (i.e., P1 and P2) for the BU-3DFE database. As we can see, the proposed OGF^{2}Net outperforms both FE2DNet and FE3DNet, which suggests the effectiveness of our feature fusion. For both the P1 and P2 protocols, although the accuracies of FE2DNet and FE3DNet are lower than those of the other algorithms, OGF^{2}Net achieves the highest accuracy of 89.05% and 89.07%, which beats all of the competing methods.

Fig. 9(a) and (b) show the confusion matrix of OGF^{2}Net for the P1 and P2 protocols on the BU-3DFE database, respectively. The diagonal of these two matrices suggests that our method performs remarkably well in recognizing the expressions of happiness and surprise. In particular, the accuracy of our approach in P1 for the fear expression is approximately 7.69% higher than that of Li *et al.* [10], i.e., 79.24%.

To further illustrate the effectiveness of OGF^{2}Net, the first two rows of Fig. 10 visualize the features of the last FC layer in FE2DNet, FE3DNet and OGF^{2}Net using t-SNE [46] for BU-3DFE. The first and second rows are visualized for the P1 and P2 protocols, respectively. Since the feature distribution of each validation set is independent, we only use one validation set that contains 72 samples to illustrate the features. As depicted in the first two rows of Fig. 10, the features of OGF^{2}Net are densely clustered and have distinct boundaries for each expression category, while the features of FE2DNet and FE3DNet are not well discriminative and have ambiguous boundaries. This result suggests that OGF^{2}Net is effective and more discriminative.

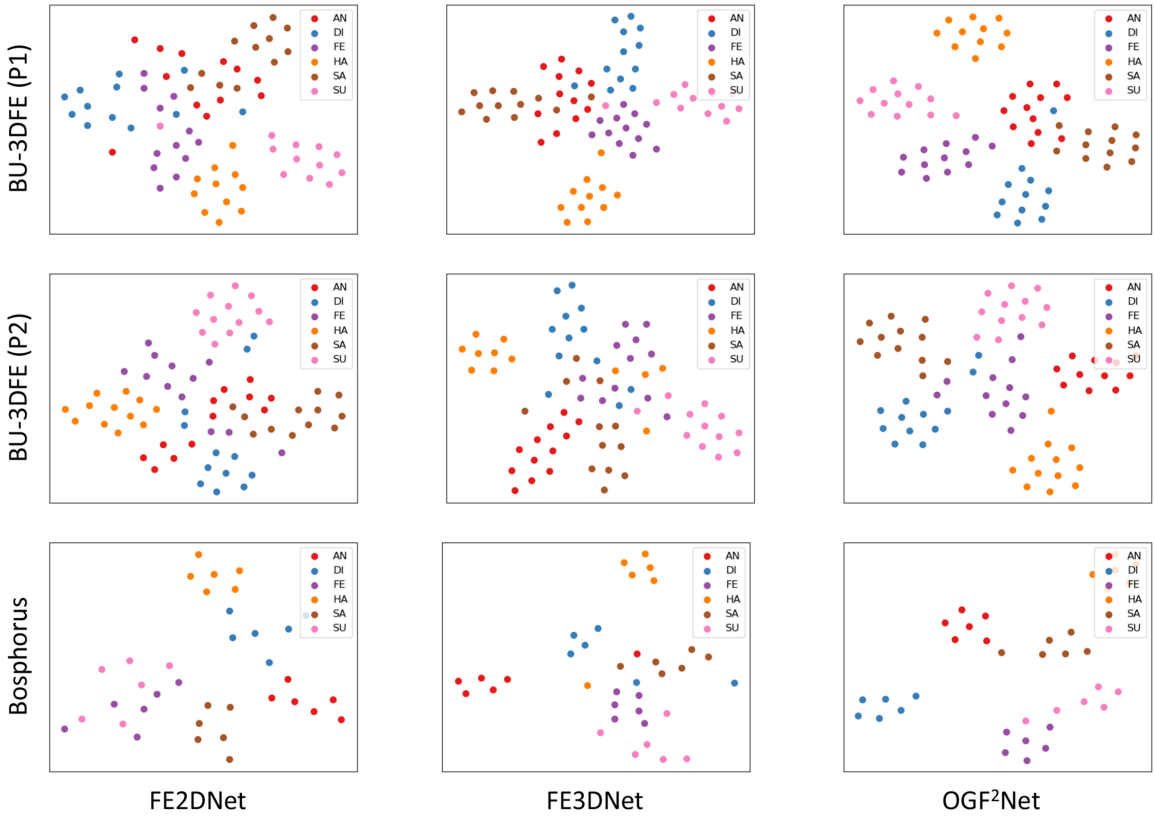


Fig. 10. t-SNE visualization of the features extracted by FE2DNet, FE3DNet, and OGF^{2}Net on the BU-3DFE (first and second rows) and Bosphorus (third row) databases.

2) Bosphorus Database: Table II lists the results of the proposed OGF^{2}Net, together with that of FE2DNet, FE3DNet and other approaches in the literature. As shown in Table II, OGF^{2}Net, exceeds the existing methods on Bosphorus, as do FE2DNet and FE3DNet. In particular, the result of OGF^{2}Net reaches a new state-of-the-art performance, with an increase of 9% over the approaches of Li *et al.* [10], i.e., 80.28%. As shown from the confusion matrix of the Bosphorus database in Fig. 9(c), OGF^{2}Net has a remarkable recognition performance for the expression of happiness. Again, we visualize the feature distribution of FE2DNet, FE3DNet and OGF^{2}Net using one of the validation sets that contains 36 samples in the third row of Fig. 10. The features of OGF^{2}Net have a more obvious clustering structure for different expressions than those of FE2DNet and FE3DNet. This emphasizes that OGF^{2}Net is more discriminative in distinguishing different expressions.

C. Ablation Study

In this section, we design three control experiments (i.e., on the pre-trained models for FE2DNet and FE3DNet, the rationality of GWP, and the effectiveness of Orth loss) to validate the effectiveness of our OGF^{2}Net. Note that all these control experiments are implemented on BU-3DFE with the P1 protocol.

1) Results With/Without the Pre-Trained Models: As described in Section III-D, we utilize two pre-trained models to initialize FE2DNet and FE3DNet. In this section, we compare the performances with and without the pre-trained models to validate the importance of the pre-trained models. To maintain a single control variable, we use the GWP module to vectorize the feature maps in this experiment. As shown in Table III, it is obvious that the accuracies using the pre-trained models are higher than those without using the pre-trained models (FE2DNet: 84.86% vs 86.58%, FE3DNet: 81.09% vs 85.20%, OGF^{2}Net: 86.79% vs 89.05%). This significant improvement proves the importance of the pre-trained models for FE2DNet, FE3DNet and OGF^{2}Net.

2) Rationality of the GWP Module: As mentioned in Section III-C2, there are three options for vectorizing the feature maps of the last convolutional layers in FE2DNet and FE3DNet: Flatten, GAP and GWP. We compare the memory requirements and accuracies of them in Table III. The size of parameters of GWP is slightly larger than that of GAP, but much smaller than that of Flatten (FE2DNet: 96.63 MB vs 71.63 MB vs 71.69 MB, FE3DNet: 81.83 MB vs 56.83 MB vs 56.88 MB, and OGF^{2}Net: 186.48 MB vs 136.48 MB vs 136.59 MB). Moreover, GWP achieves better performance than GAP and Flatten (FE2DNet: 85.58% vs 86.16% vs 86.58%, FE3DNet: 85.08% vs 83.01% vs 85.20%, and OGF^{2}Net: 88.54% vs 88.59% vs 89.05%). Thus, we suggest that the designed GWP module is more effective and reasonable than GAP and Flatten for FER.

3) Results of the Early Fusion Strategy: While our work uses the late fusion scheme to fuse the features extracted by FE2DNet and FE3DNet, we test the performance of early fusion in this section. For early fusion, the images I_t and I_{dae} are concatenated into a 4-channel image and fed into a deep

TABLE III
ABLATION STUDY OF DIFFERENT MODELS ON BU-3DFE (P1)

Model	Pre-trained Model	Vectorization Approach			Early Fusion	Concatenation of F_1 and F_2	Orth Loss	Parameters (MB)	Accuracy (%)
		Flatten	GAP	GWP					
FE2DNet	✓	✓						96.63	85.58
	✓		✓					71.63	86.16
				✓				-	84.86
				✓	✓			-	**85.01**
	✓			✓				71.69	**86.58**
FE3DNet	✓	✓						81.83	85.08
	✓		✓					56.83	83.01
				✓				-	81.09
				✓	✓			-	**81.58**
	✓			✓				56.88	**85.20**
OGF^{2}Net	✓	✓				✓	✓	186.48	88.54
	✓		✓			✓	✓	136.48	88.59
				✓		✓	✓	-	86.79
	✓			✓			✓	-	88.33
	✓			✓		✓		-	88.47
	✓			✓		✓	✓	136.59	**89.05**

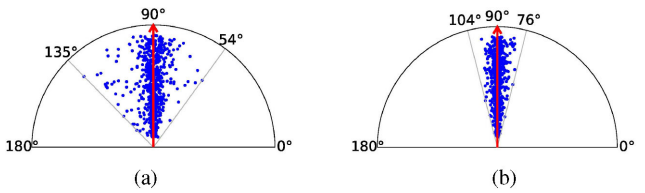


Fig. 11. Comparison of angles between features F_1 and F_2 generated without and with Orth loss. The angle between a blue dot and the x-axis illustrates the angle between facial feature vectors F_1 and F_2, while the radius of the blue dot from the origin is randomly generated.

TABLE IV
COMPARISON OF COMPLEXITY WITH OTHER METHODS

Approach	Parameter (MB)	FPS
Jan et al. [22]	≈ 327	≈33
Zhu et al. [21]	≈ 463	≈10
OGF^{2}Net	≈ 137	≈93

network for classification. As listed in the column of "Early Fusion" in Table III, the performance (85.01%) of FE2DNet is approximately 3% higher than that (81.58%) of FE3DNet. However, both of them are much lower than that (89.05%) of our OGF^2 Net using the late fusion strategy.

4) Effectiveness of Orth Loss: To evaluate the effectiveness of the proposed Orth loss, we compare the performances with and without Orth loss. As we can see in Table III, Orth loss improves the accuracy of OGF^{2}Net from 88.47% to 89.05%, which is a state-of-the-art performance.

We also calculate the offset of the angle between F_1 and F_2 from $\frac{\pi}{2}$ to evaluate the orthogonality of the two features learned with and without the proposed Orth loss, which is defined as:

$$\Delta = \left| \arccos\left(\frac{F_1 \cdot F_2}{\|F_1\| \, \|F_2\|} \right) - \frac{\pi}{2} \right| \tag{6}$$

The smaller the offset is, the better the orthogonality between features F_1 and F_2. For 720 samples in one of the validation sets, the average offsets for features learned with and without Orth loss are 2.8290°and 6.2040°, respectively. It seems that the proposed Orth loss can greatly reduce the offset of the angle between F_1 and F_2, from $\frac{\pi}{2}$ and thus orthogonalize the two features.

Fig. 11 also visualizes the angles between F_1 and F_2 for the 720 samples. Each blue dot in this figure represents the angle between each pair of features F_1 and F_2. While the radius of blue dots are randomly generated, the angles between the blue dots and the 90°axis are determined by the angles between F_1 and F_2. As observed from the figure, the angles for the features learned with Orth loss are much more clustered and located much closer to the 90°axis.

5) Concatenation of F_1 and F_2: Though F_3 is the fusion of F_1 and F2, we believe that complementary information might still exist between F3 and F_1/F_2. To test if the concatenation of F_1 and F_2 with F_3 is necessary, we compare the performances of OGF^{2}Net with and without extra concatenation of F_1 and F_2 in Table III. As shown in the table, the addition of F_1 and F_2 further improves the accuracy of F_3 from 88.33% to 89.05%.

D. Complexity Analysis

To further analyze the complexity of our approach, we compare in Table IV the memory cost of the parameters and the processing speed FPS (frames per second) of the proposed method with those of previous methods [21], [22], which are recorded using a workstation equipped with an Intel Xeon CPU (E5-2690 v4, 2.6 GHz) and an NVIDIA Tesla-P100 GPU. While the memory cost of our approach (137 MB) is significantly lower, our approach is much more efficient in terms of processing speed.

V. CONCLUSION

In this paper, we present an efficient 2D+3D facial expression recognition (FER) network based on a novel orthogonalization-guided feature fusion network, OGF^{2}Net, which extracts and

fuses complementary features for 2D+3D FER. The orthogonal (Orth) loss is proposed to reduce the correlation and redundancy among the features learned by FE2DNet and FE3DNet. Furthermore, we use the global weighted pooling (GWP) module to vectorize the feature maps in FE2DNet and FE3DNet to learn more discriminative features and achieve a better result. Experimental results show that our proposed method achieves better performance than other state-of-the-art methods. For BU-3DFE, we achieve 89.05% and 89.07% accuracy for the P1 and P2 protocols, respectively. For Bosphorus, our recognition rate reaches as high as 89.28%.

REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, p. 1, 2020.

[2] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "DeXpression: Deep convolutional neural network for expression recognition," 2015, *arXiv:1509.05371*.

[3] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 558–565.

[4] C. Zhang, P. Wang, K. Chen, and J.-K. Kämäräinen, "Identity-aware convolutional neural networks for facial expression recognition," *J. Syst. Eng. Electron.*, vol. 28, no. 4, pp. 784–792, 2017.

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

[6] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2983–2991.

[7] Y. Ding, Q. Zhao, B. Li, and X. Yuan, "Facial expression recognition from image sequence based on lbp and taylor expansion," *IEEE Access*, vol. 5, pp. 19 409–19 419, 2017.

[8] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[9] S. E. Kahou *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[10] 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.

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

[12] 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.

[13] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected sift features for 3D facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 4125–4128.

[14] 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.

[15] 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.

[16] H. Li, J. Sun, D. Wang, Z. Xu, and L. Chen, "Deep representation of facial geometric and photometric attributes for automatic 3D facial expression recognition," 2015, *arXiv:1511.03015*.

[17] H. Yang and L. Yin, "CNN based 3D facial expression recognition using masking and landmark features," in *Proc. 17th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 556–560.

[18] 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.

[19] Y. Huang, Y. Li, and N. Fan, "Robust symbolic dual-view facial expression recognition with skin wrinkles: Local versus global approach," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 536–543, Oct. 2010.

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

[21] K. Zhu *et al.*, "Discriminative attention-based convolutional neural network for 3D facial expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–8.

[22] A. Jan, H. Ding, H. Meng, L. Chen, and H. Li, "Accurate facial parts localization and deep learning for 3D facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 466–472.

[23] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, Nov. 2013.

[24] S. Wang *et al.*, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.

[25] S. Wang *et al.*, "Analyses of a multimodal spontaneous facial expression database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 34–46, Jan.–Mar. 2013.

[26] P. Lucey *et al.*, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.-Workshops*, 2010, pp. 94–101.

[27] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2005, pp. 317–321.

[28] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[29] 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.*, 2006, pp. 211–216.

[30] L. Yin, X. C. Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, vol. 126, pp. 1–6.

[31] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conf. Comput. Vision*, 2010, pp. 709–720.

[32] J. Cai *et al.*, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 302–309.

[33] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan, "Local relationship learning with person-specific shape regularization for facial action unit detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 11 917–11 926.

[34] J. DErrico, "Surface fitting using gridfit," in *MATLAB Central File Exchange*, 2008.

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

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 770–778.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1026–1034.

[39] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 618–626.

[40] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[41] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7–9, 2015.

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

[44] 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.*, 2015, vol. 1, pp. 1–6.

[45] 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.*, 2018, pp. 31–37.

[46] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Shisong Lin received the B.Sc. degree from the College of Electronic Science and Technology, Shenzhen University, Shenzhen, China, in 2017. Currently he is working toward the M.Sc. degree with the School of Computer Science & Software Engineering, Shenzhen University. His current research interests include 3D face recognition and 3D facial expression recognition.

Mengchao Bai received the B.Sc. degree from the College of Electronic and Information Engineering from Shenzhen University, in 2017. Currently he is working toward the M.Sc. degree majoring pattern recognition with the School of Computer Science & Software Engineering, Shenzhen University. His research interests include facial expression recognition and image translation.

Feng Liu received the B.Sc. and M.Sc. degrees from Xidian University, Xi'an, China and the Ph.D. degree in computer science from the Department of Computing, HongKong Polytechnic University, in 2014. She is currently an Assistant Professor with the School of Computer Science & Software Engineering, Shenzhen University. Her research interests include pattern recognition and image processing, especially focus on their applications to fingerprints. Dr. Liu has published more than 40 papers in academic journals and conferences, and participated in many research projects either as principal investigators or as primary researchers. She is a reviewer for many renowned field journals and conferences and a member of the IEEE.

Linlin Shen received the B.Sc. and M.Eng. degrees from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree from the University of Nottingham, Nottingham, U.K. He is currently a Pengcheng Scholar Distinguished Professor with the School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is also a Honorary professor with the School of Computer Science, University of Nottingham, U.K. He serves as the director of Computer Vision Institute, AI Aided Medical Image Analysis & Diagnosis Research Center and China-U.K. joint research lab for visual information processing. He was a Research Fellow with the University of Nottingham, working on MRI brain image processing. His research interests include deep learning, facial recognition, analysis/synthesis and medical image processing. He is listed as the Most Cited Chinese Researchers by Elsevier. He received the Most Cited Paper Award from the journal of Image and Vision Computing. His cell classification algorithms were the winners of the International Contest on Pattern Recognition Techniques for Indirect Immunofluorescence Images held by ICIP 2013 and ICPR 2016.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees from Tufts University, Medford, MA, USA, all in electrical engineering. He is currently an Associate Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security. He is a Senior Member of the International Society for Optical Engineering (SPIE). He was a recipient of the Third Price of Macau Natural Science Award in 2014. He is also the Co-Chair of the Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and four other journals.