Facial Expression Guided Diagnosis of Parkinson's Disease via High-Quality Data Augmentation

Wei Huang, Yintao Zhou, Yiu-ming Cheung, *Fellow, IEEE*, Peng Zhang, *Member, IEEE*,
Yufei Zha, *Member, IEEE*, and Meng Pang

***Abstract*—Parkinson's disease (PD) is a neurodegenerative disease which is prevalent among the elder population and severely affects the life quality of patients and their families. Therefore, it is important to conduct an early diagnosis for potential patients with PD, so as to promote prompt treatment and avoid the aggravation of the disease. Recently, the in-vitro PD diagnosis based on facial expressions has received increasing attention because of its distinguishability (i.e., PD patients always possess the characteristics of "masked face") and affordability. However, the performance of the existing facial expression-based PD diagnosis approaches is limited by: 1) the small-scale training data on PD patients' facial expressions, and 2) the weak prediction model. To address these two problems, we propose a new facial expression guided PD diagnosis method based on high-quality training data augmentation and deep neural network prediction. Specifically, the proposed method consists of three stages: Firstly, we synthesize virtual facial expression images with 6 basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) based on multi-domain adversarial learning to approximate the premorbid expressions of PD patients. Secondly, we introduce three facial image quality assessment (FIQA) criteria to measure the quality of these synthesized facial expression images and design a fusion screening strategy that shortlists the high-quality ones to augment the training data. Finally, we train a deep neural network prediction model based on the original and synthesized high-quality facial expression images for PD diagnosis. To show real-world impacts and evaluate the proposed method under different facial expressions, we also create a (currently largest) multiple facial expressions-based PD face dataset in collaboration with a hospital. Extensive experiments are performed to demonstrate the effectiveness of the multi-domain adversarial learning-based facial expression synthesis and the fusion screening strategy, particularly the superior performance of the proposed method for PD diagnosis.**

***Index Terms*—Data augmentation, deep learning, multi-domain adversarial learning, Parkinson's disease diagnosis.**

I. Introduction

PARKINSON'S disease (PD) is a neurodegenerative disease characterized by motor symptoms (e.g., rest tremor, bradykinesia, and hypomimia) and non-motor symptoms like hyposmia, cognitive impairment, and sleeping disorders, just to name a few [1]. Also, it seriously and adversely influences the life quality of PD patients and their families. According to the statistics in [2], PD has already become the second most common type of neurodegenerative diseases only after the Alzheimer's disease (AD). Moreover, the reports in [3] and [4] show that PD affects more than 6 million individuals around the world, which results in a 2.5-times increase in prevalence over the past 30 years. Although it is universally accepted that there is no elixir for PD at the current stage, the early diagnosis and prompt treatment of PD are essential to alleviate the symptoms and avoid the progression of disease [5], [6], [7], [8].

Generally, PD diagnosis can be categorized into two types, i.e., the in-vivo PD diagnosis and the in-vitro PD diagnosis. The former is mainly carried out based on professional imaging diagnostic devices such as positron emission tomography (PET) [9]. Although this in-vivo diagnosis way can achieve a high PD diagnosis accuracy, it still has two drawbacks: 1) *Inconvenience*—The professional in-vivo PD diagnostic devices are always scarce in developing or poverty areas [10]. Also, it is inconvenient for the elderly in these areas to travel a long distance to the hospitals in developed areas for an in-vivo PD diagnosis, especially during the COVID-19 pandemic. 2) *Expensive*—The cost of the in-vivo PD diagnosis using PET is usually expensive and may not be affordable for every family. By contrast, the latter PD diagnosis has become popular recently, as it only collects in-vitro biomarkers, e.g., speech signal [11], [12], [13], [14], [15], [16], [17], gait signal [5], [18], [19], [20], [21], [22], [23] or facial expression [24], [25], [26], [27], [28], from PD patients, which is convenient and affordable.

Manuscript received 9 July 2022; revised 25 September 2022; accepted 19 October 2022. Date of publication 25 October 2022; date of current version 1 November 2023. The work of Wei Huang, Peng Zhang, and Yufei Zha was supported in part by the National Natural Science Foundation of China under Grants 62271239, 61862043, and 61971352, and in part by the Natural Science Foundation of Ningbo under Grants 2021J048 and 2021J049. The work of Yiu-ming Cheung was supported in part by NSFC / Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21, in part by the General Research Fund of RGC under Grant 12201321, and in part by the Hong Kong Baptist University (HKBU) under Grant RC-FNRA-IG/18-19/SCI/03. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo. *(Corresponding author: Meng Pang.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the medical ethics committee of Neurology-NCU and the academic board of SMCS-NCU under Application No. S2021GDKX0936.

Wei Huang and Yintao Zhou are with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China (e-mail: n060101@e.ntu.edu.sg; yintaozhou@email.ncu.edu.cn).

Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong 999077 (e-mail: ymc@comp.hkbu.edu.hk).

Peng Zhang and Yufei Zha are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China (e-mail: zh0036ng@nwpu.edu.cn; yufeizha@nwpu.edu.cn).

Meng Pang is with the School of Mathematics and Computer Sciences and the Institute of Mathematics and Interdisciplinary Sciences, Nanchang University, Nanchang 330031, China (e-mail: pangmeng1992@gmail.com).

Digital Object Identifier 10.1109/TMM.2022.3216961

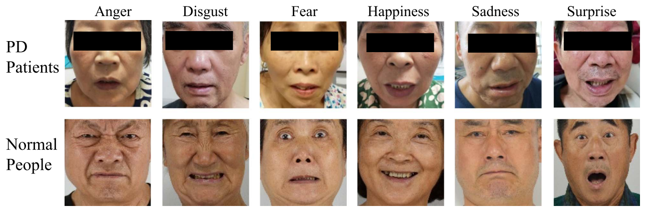

Fig. 1. A comparison between the masked faces of PD patients and the facial expression images of normal persons under six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise). To avoid disclosure of PD patients' identity information, we cut out the eye regions of their facial expression images in the illustration.

Among the above-mentioned three in-vitro biomarkers, facial expression is the most favorable one because it does not require professional sensors like the gait signal. Besides, the speech signal can be severely influenced by the inconsistency of acoustic characteristics among different language speakers [15]. The rationality of using facial expression to diagnose PD lies in the observation that PD patients always possess "masked face" characteristics (or academically called hypomimia[1]) [26], [29]. A comparison between the masked faces of PD patients and the facial expression images of normal persons under six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) is illustrated in Fig. 1. It can be observed that the PD patients suffer from obvious expression disorder and cannot express emotions as naturally as normal persons. The inspiring observation therefore motivates humans to conduct early PD diagnosis of potential PD patients based on their facial expressions.

Despite some success achieved by the existing facial expression-based PD diagnosis techniques [24], [25], [30], [31], these techniques still suffer from three major limitations, with each corresponding to one key challenge. Firstly, existing techniques are designed for handling *simplex* facial expression, e.g., smiling, but cannot generalize to the other expressions, as there is no multiple facial expressions-based PD face dataset available for study nowadays. Secondly, the existing techniques often encounter over-fitting problem during training process due to the small-scale PD training dataset. Lastly, the existing techniques usually predict PD in a two-step procedure, i.e., feature extraction + classification, and are based on the weak conventional handcrafted feature extraction models (e.g., local binary pattern and gray-level co-occurrence matrix).

To address the above-mentioned three challenges, we first create a multiple facial expressions-based PD face dataset in collaboration with the affiliated hospital of Nanchang University for in-vitro PD diagnosis research. This dataset consists of 95 PD patients with each having seven basic types of facial expressions (i.e., neutral, anger, disgust, fear, happiness, sadness, and surprise). Furthermore, we propose an effective in-vitro PD diagnosis approach that is capable of handling multiple facial expressions of potential PD patients, based on a novel training data augmentation strategy and an end-to-end deep neural network PD prediction model. As illustrated in Fig. 2, the proposed method consists of three stages: Firstly, we synthesize multiple identity-preserved facial images of PD patients with 6 basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) by utilizing the star generative adversarial network (StarGAN) [32] based on multi-domain adversarial learning [33], [34], [35], in order to estimate the premorbid facial expressions of these PD patients. Secondly, we introduce three facial image quality assessment (FIQA) criteria to quantitatively measure the quality of the synthesized facial expression images and design a fusion screening strategy to shortlist high-quality ones to augment the training data. Finally, we train an end-to-end deep neural network prediction model based on the original and high-quality synthesized facial expression images for PD diagnosis.

We summarize the contributions of this paper as follows:

- We create a PD facial expression dataset containing multiple facial expressions (i.e., neutral, anger, disgust, fear, happiness, sadness, and surprise) of 95 PD patients, which is currently the largest facial expression dataset for in-vitro PD diagnosis research.
- We introduce three FIQA criteria to measure the quality of the synthesized premorbid facial expression images of PD patients and design a fusion screening strategy to shortlist high-quality ones to augment the training data.
- We propose a new facial expression guided in-vitro PD diagnosis method, which adopts an end-to-end deep neural network prediction model trained on the augmented PD training data for PD diagnosis.
- We conduct extensive experiments to demonstrate the effectiveness of the high-quality PD training data augmentation strategy (i.e., StarGAN-based facial expression synthesis + fusion screening strategy), and the superior performance of the proposed method for PD diagnosis.

The rest of this paper is organized as follows. In Section II, we make an overview of the related works on in-vitro PD diagnosis. In Section III, the technical details of the proposed facial expression guided in-vitro PD diagnosis approach are elaborated. In Section IV, extensive experiments are conducted to demonstrate the effectiveness of the proposed method. In Section V, we present the conclusions and future works.

II. Related Works

Compared to in-vivo PD diagnosis techniques which are mainly based on expensive imaging diagnostic instruments, the in-vitro diagnosis approaches require only the collection of the in-vitro biomarkers from PD patients, which is more convenient and cheaper. Generally, the most commonly used biomarkers for PD are: speech signal [11], [12], [13], [14], [15], [16], [17], gait signal [5], [18], [19], [20], [21], [36] and facial expression [24], [25], [26], [27], [28]. In the following, we make an overview of some representative in-vitro PD diagnosis approaches in the literature.

A. *In-Vitro PD Diagnosis Based on Speech Signal*

Based on the observation that nearly 90% of the PD patients suffer from dysarthria and vocal impairment in the early

[1] This indicates the reduction or loss of spontaneous facial movements and facial expressions in PD patients

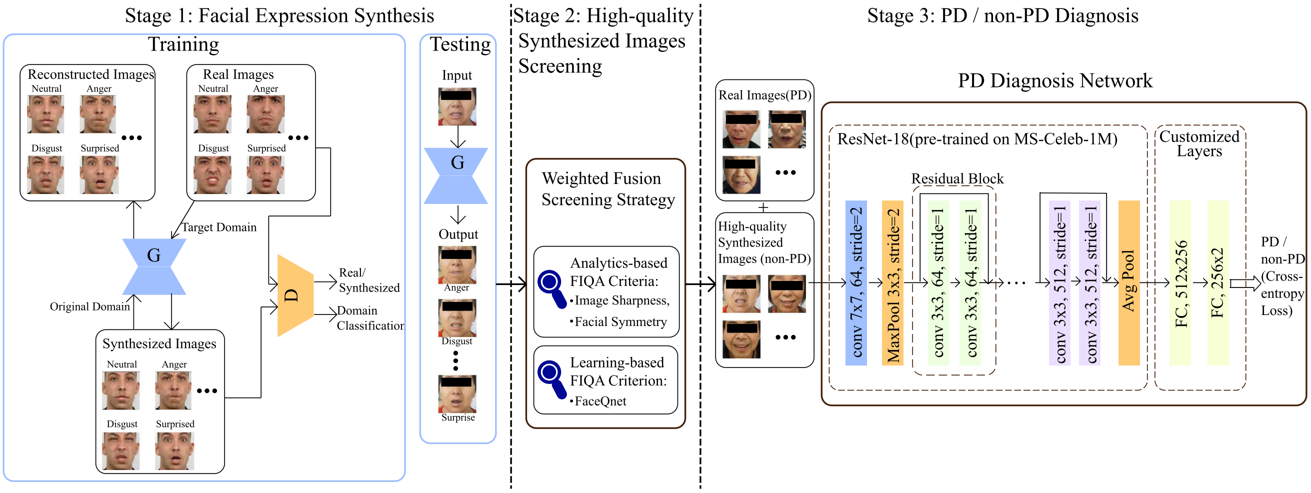

Fig. 2. An overview of the flowchart of the proposed PD diagnosis method based on facial expressions.

stage [11], some attempts [11], [12], [13], [14], [15] have been made to conduct PD diagnosis using the speech signals. In these works, the robust voice features have been extracted and combined with machine learning techniques to conduct the PD diagnosis. In [12], a method which combines minimum-average-maximum-tree with singular value decomposition is proposed for the extraction of voice features. After that, the popular K-nearest neighbor classifier is utilized to classify PD patients and normal persons. In [13], speech features are extracted via principal component analysis (PCA) and fed into the random forest classifier for classification. In [11], feature extraction is realized using the tunable Q-factor wavelet transform, which has higher frequency resolution than the classical transform and thus can capture discriminative information. It is worth noting that, although speech signals are easy to be collected, the language bias would degrade the PD diagnosis performance [15]. Specifically, it is reported in [15] that native Mandarin speakers with PD exhibit significant differences in acoustic features from native English speakers with PD. Therefore, the relatively low robustness and adaptability of the speech signals-based PD diagnosis may not meet the actual diagnostic requirements.

B. *In-Vitro PD Diagnosis Based on Gait Signal*

Motor symptoms such as bradykinesia, rigidity, tremor, and postural instability define the diagnosis of PD [37]. Motor impairment leads to specific gait characteristics in PD, such as shuffling gait, reduced step length, impaired gait initiation, and reduced gait speed [38]. In practice, the statistics of the potential PD patients' gait patterns (e.g., gait speed, swing time, step length) can be collected using biosensors. For example, Barth et al. [18] employed the mobile and lightweight sensors to record the gait signals and used the Fourier transform to perform a frequency based analysis. In [19], various multi-dimensional sensors are inserted into shoe insoles of subjects to collect signals and tensor decomposition is then applied on the multi-sensors data to identify potential PD patients. In [20], temporal features including the stance phase, swing phase, and stride time of the gait signals are employed to distinguish between PD patients and normal persons. In [21], correlation-based features are extracted from gait signals which are collected by a more sophisticated sensor (i.e., the vertical ground reaction force sensor). After that, the classical support vector machine (SVM) algorithm is then employed for PD/non-PD classification based on these features.

Despite that gait signal is a robust biomarker for the in-vitro diagnosis of PD, the wearable sensors are diverse and expensive. Moreover, wearing a lot of sensors may cause inconvenience and discomfort in the elderly.

C. *In-Vitro PD Diagnosis Based on Facial Expression*

The rationality of the facial expression-based PD diagnosis lies in the observation that PD patients always possess the characteristics of "masked face" [26], [29]. In [26], it is found that the distances of facial movements of PD patients are much smaller than those of normal people after quantitatively analyzing the geometric features extracted from expression video. Moreover, it is reported in [39] that PD patients usually have a lower expressivity in exhibiting some facial action units (e.g., brow lowerer, nose wrinkler, upper lip raiser).

Recently, some in-vitro PD diagnosis methods based on facial expressions have been developed. In [28], a PD diagnosis method is proposed by recording specific facial expressions of patients using video. Both geometric feature based on corner angles of landmarks and texture feature based on gray-level co-occurrence matrix (GLCM) are fed into the SVM classifier to conduct classification [28]. In [24], the variances of facial features in selfie photos within a time period are utilized to automatically assess the severity of PD. In [25], Jin et al. collect videos of PD patients with smiling expressions and conduct feature extraction according to the expression amplitude and degree of tremor for PD diagnosis. In [27], Rajnoha et al. use one static face image per subject for training the decision tree classifier, which obtains a poor PD diagnosis accuracy of 67.33%. It is obvious that the performance of the above PD diagnosis methods

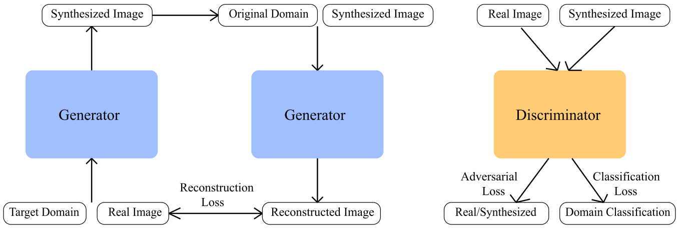

Fig. 3. An illustration of the working mechanism of the generator and discriminator in StarGAN network.

is far from satisfactory. The reasons can be twofold: 1) the training data of the facial expressions of PD patients is small which would easily cause the over-fitting problem; 2) the prediction models based on conventional machine learning methods are weak and cannot extract high semantic features for classification. To address the above issues, we create a facial expression image dataset containing both the neutral and all 6 basic emotional images of 95 PD patients, and it is currently the largest facial expression dataset of PD patients to the best of our knowledge. Furthermore, we propose a new facial expression guided in-vitro PD diagnosis method by further augmenting the training data and adopting the powerful deep neural network as the prediction model.

III. The Proposed Method

The proposed facial expression guided in-vitro PD diagnosis method consists of three stages, as illustrated in Fig. 2. Firstly, we adopt the powerful StarGAN to synthesize the premorbid normal facial expression images of the PD patients. Then, we introduce three FIQA criteria to measure the quality of these synthesized facial expression images and design a fusion screening strategy that shortlists the high-quality ones to augment the training data. Finally, we train a ResNet-based deep neural network based on the mixture of the original and synthesized high-quality face images for PD prediction. The technical details of the three stages are elaborated in Sections III-A, III-B, and III-C, respectively.

A. Facial Expression Synthesis Via StarGAN

Adversarial learning can be utilized to synthesize virtual images based on a two-player game between the generator and discriminator [40]. In this stage, we adopt a multi-domain adversarial learning-based StarGAN [32] to perform multiple facial expression synthesis considering that StarGAN is good at resolving the multi-domain image-to-image translation problem [41], [42], [43]. In image-to-image translation field, the term *domain* is denoted as a set of images sharing the same *attribute* value, and the term *attribute* is denoted as a meaningful feature inherent in an image such as gender, age, or facial expression. The working mechanism of the generator and discriminator in StarGAN is illustrated in Fig. 3. Note that, the generator in StarGAN is composed by a series of convolutional layers and residual blocks, and the discriminator in StarGAN is similar to that of [44] which classifies each local patch of the input image as real or fake.

StarGAN Training: During the training phase in the stage, three well-designed losses, i.e., adversarial loss, classification loss, and reconstruction loss, are combined to optimize the StarGAN model. The details of the above three losses are presented thereinafter.

1) Adversarial Loss: To stabilize the training process and meanwhile promote the quality of synthesized images, the adversarial loss [45] is introduced into the StarGAN model as follows:

$$\begin{aligned} L_{adv} = \mathbb{E}_x\left[D_{src}(x)\right] - \mathbb{E}_{x,c}\left[D_{src}\left(G(x,c)\right)\right] \\ -\lambda_{gp}\mathbb{E}_{\hat{x}}\left[\left(\left\|\bigtriangledown_{\hat{x}}D_{src}\left(\hat{x}\right)\right\|_2 - 1\right)^2\right], \end{aligned} \tag{1}$$

where D denotes the discriminator of StarGAN and G is the generator; x and c represent the real image and the target domain label, respectively; $\hat{x}$ is sampled uniformly within a pair of real and synthesized images; $D_{src}(x)$ is referred to as the probability distribution over source images; and λ_{gp} represents the hyper-parameter of the gradient penalty.

2) Classification Loss: In addition to determining whether an image is real or synthesized, the discriminator of the StarGAN also aims to predict the domain label of the image, analogous to ACGAN model [46]. The classification loss of the StarGAN is presented as follows:

$$L_{cls} = \mathbb{E}_{x^*,c^*}\left[-logD_{cls}\left(c^* \mid x^*\right)\right], \tag{2}$$

where $D_{cls}(c^* \mid x^*)$ corresponds to the probability distribution over domain labels computed by D; x^* represents either the real image or the synthesized image; and c^* denotes the corresponding domain label.

3) Reconstruction Loss: In order to guarantee that generated images preserve the identity of their input images while changing the domain-related part of the inputs, StarGAN applies a cycle consistency loss [47] to the generator G, defined as

$$L_{rec} = \mathbb{E}_{x,c,c'}\left[\left\|x - G\left(G(x,c),c'\right)\right\|_1\right], \tag{3}$$

where c' denotes the original domain label and $\|\cdot\|_1$ represents the conventional L1-norm.

Based on (1)-(3), the total loss of the StarGAN model is summarized as follows:

$$L_{stargan} = L_{adv} + \lambda_{cls}L_{cls} + \lambda_{rec}L_{rec}, \tag{4}$$

where λ_{cls} and λ_{rec} are two hyper-parameters weighting the classification and reconstruction loss, respectively.

StarGAN Testing: In the testing phase, the neutral facial expression image of one PD patient is fed into the generator to synthesize multiple identity-preserved facial expression images depicting 6 basic emotions including anger, disgust, fear, happiness, sadness, and surprise. This synthesis procedure is formulated in (5) as follows:

$$I_{PD}^{expr} = G\left(I_{PD}^{neutral}, c_{expr}\right), \tag{5}$$

where G is the trained generator, $I_{PD}^{neutral}$ indicates the neutral facial expression image of the PD patient, c_{expr} denotes the target expression label, and I_{PD}^{expr} denotes the corresponding synthesized facial image with the target expression.

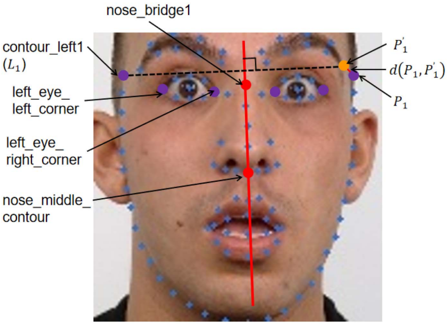

Fig. 4. An illustration of the FIQA on facial symmetry. The example face image is from the public RaFD [48] dataset.

B. FIQA Criteria and Fusion Screening Strategy

After obtaining the synthesized facial expression images of the PD patients, we aim to judge which images are of high quality and then screen them out to facilitate the next stage of training. To this end, we introduce three FIQA criteria, i.e., facial symmetry, image sharpness, and FaceQnet, to measure the image quality, and develop a simple but effective fusion screening strategy by weighting the above three criteria. Specifically, the details of the three FIQA criteria and the fusion screening strategy are described in the following.

1) Facial Symmetry: The face misalignment and tilted posture would easily cause performance degradation in facial expression recognition [49]. In the case, we thus adopt the facial symmetry criterion to measure the quality of the synthesized image. Fig. 4 gives an intuitive illustration of how facial symmetry is measured. First, the coordinates of facial landmarks are acquired by facial landmark detection. The line connecting the landmarks "$nose\_bridge1$" and "$nose\_middle\_contour$" is taken as the midline of the face, dividing it into the left region and the right region. For each key point selected from the left region (e.g., L_1 in Fig. 4), its corresponding symmetric point is projected into the right region (see P_1' in Fig. 4). Subsequently, the distance between the landmarks in the right region (i.e., P_1 in Fig. 4) and those symmetric points P_1' is used to measure the *facial symmetry*, which is calculated as

$$QS_{Symm}(I) = 1 - Norm\left(\sum_{i=1}^{N} d\left(P_i, P_i'\right)\right), \tag{6}$$

where I denotes the input query image, P_i denotes the i^{th} landmark chosen in the right region and P_i' denotes the symmetric point of its corresponding landmark from the left region. $d(\cdot)$ represents the distance between two points. N corresponds to the number of landmark pairs in the whole face region for quality assessment. $Norm(\cdot)$ indicates the normalization operation that keeps the score in the range of [0, 1], which is calculated as $Norm(S) = \frac{S-min(S)}{max(S)-min(S)}$. Note that the larger the value of QS_{Symm} is, the better facial symmetry the query image has.

2) Image Sharpness: Blur occurs ubiquitously in both real images and synthesized ones, adversely influencing the performance of face recognition. Hence, sharpness is one of the most crucial factors of image quality. In this study, the conventional Brenner function [50] is adopted to assess the sharpness of generated facial expression images due to its merits of low computational cost and high efficiency. Subsequently, we measure the *image sharpness* by using the Brenner algorithm to accumulate the squares of difference of horizontally neighboring pixels as follows:

$$QS_{Sharp}(I) = Norm\left(\sum_{x}\sum_{y}\left[I\left(x+2,y\right) - I(x,y)\right]^2\right), \tag{7}$$

where $I(x, y)$ represents the grayscale value of point (x, y) of an M-by-N image. The quality scores are also normalized to [0, 1]. The larger the value of QS_{Sharp} is, the higher sharpness (or less blur) the query image possesses.

3) FaceQnet: FaceQnet [51] is a popular learning-based network because of its unbiased quality label and elaborately-designed architecture. The overall architecture of the FaceQnet is illustrated in Fig. 5. In this case, we adopt the FaceQnet to measure the quality of a query synthesized image through the following two steps:

Groundtruth quality score generation: In the first, we generate the scores for the groundtruth images for reference. For each subject, the image with the highest International Civil Aviation Organisation (ICAO) compliance score obtained by the BioLab framework [52] is selected as the gallery image (i.e. assumed high quality), and others are probe images. Then, 128-dimensional feature vectors are extracted from these probe images and their corresponding gallery images by the popular FaceNet model [53]. Subsequently, the groundtruth quality scores can be obtained by computing the similarity between the gallery image and other samples of the same identity, which is presented as follows:

$$QS_{FaceQnet}(I) = Norm\left(D\left(V_I, V_G\right)\right), \tag{8}$$

where $QS(I)$ represents the quality score of a query image I, V_I and V_G denote the feature vectors of this image and the gallery image of the same identity, respectively. $D(x, y)$ denotes the Euclidean distance between the vectors x and y. For $QS_{FaceQnet}$, the value close to 1 represents high quality while the one close to 0 represents low quality.

Query quality score prediction: Subsequently, we train a convolutional neural network (CNN)-based model to predict the quality score of a query synthesized image. The pre-trained ResNet-50 in [54] is adopted as the backbone and the classification layer of it is replaced with two fully connected (FC) layers. Specifically, the former FC layer reduces the dimension of the feature vector from 2048 to 32, followed by a rectified linear unit (i.e. ReLU) activation function, and the latter acts as an output layer of size 1. We fine tune the parameters of the two FC layers based on the groundtruth data while freezing the parameters in the previous layers. The final architecture of FaceQnet is shown in Fig. 5. With the trained end-to-end FaceQnet model, we could acquire a score between 0 and 1 of an input query image for quality assessment.

4) Fusion Screening Strategy: Based on the above-mentioned three FIQA criteria, we develop a simple but effective fusion screening strategy to shortlist the high-quality facial expression

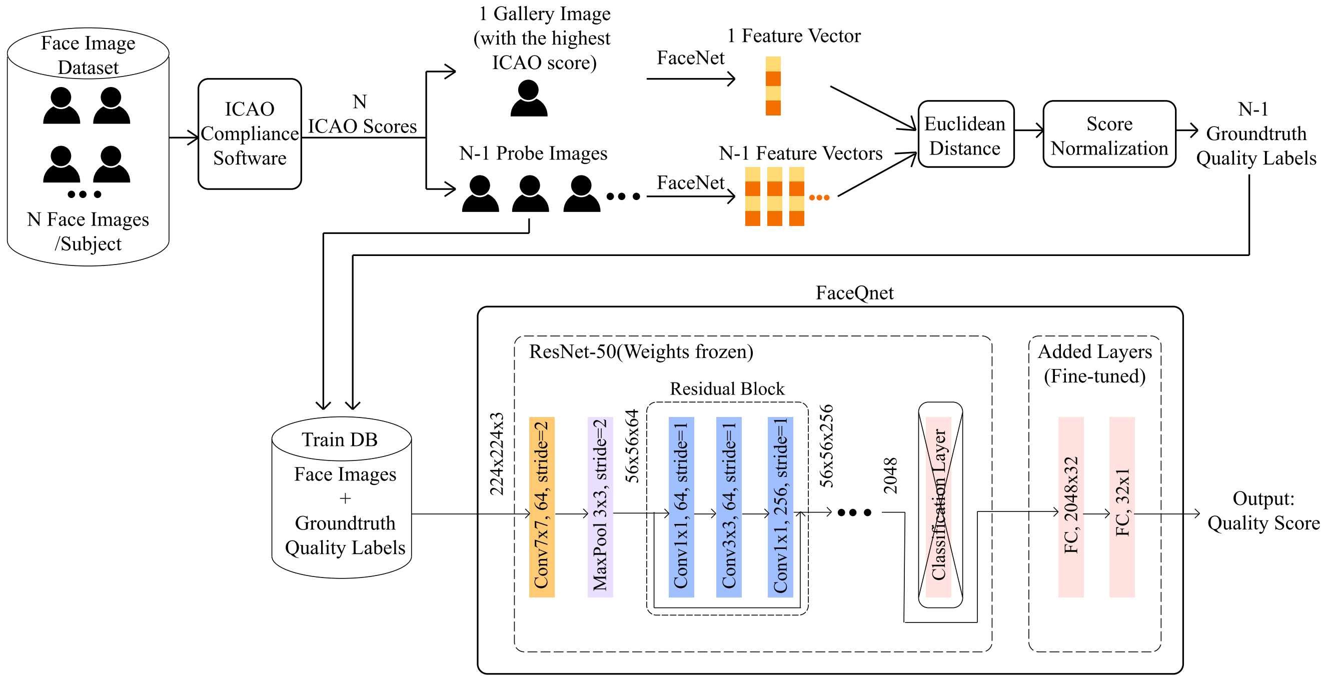

Fig. 5. An illustration of the overall architecture of the FaceQnet.

images by considering both the *analytics-based* FIQA criteria (i.e., facial symmetry and image sharpness) and the *learning-based* FIQA criterion (i.e., FaceQnet). Specifically, we first assign weights to the scores output from the above three FIQA criteria and then sum them up. Formally, this *weighted fusion score* is calculated as follows:

$$QS_{Fusion} = Norm(c_1 QS_{Symm} + c_2 QS_{Sharp} + c_3 QS_{FaceQnet}), \tag{9}$$

where QS_{Symm}, QS_{Sharp} and $QS_{FaceQnet}$ are three different quality scores associated with the facial symmetry, the image sharpness, and the FaceQnet, respectively. c_1, c_2 and c_3 are three weighting coefficients. The value of QS_{Fusion} is also normalized to [0, 1].

After obtaining the weighted fusion score of each synthesized facial expression image, we then sort all of them in the reverse order according to their scores and then shortlist the high-quality ones by defining a reasonable screening threshold.

C. PD Prediction Model Based on ResNet-18

In this stage, we adopt the ResNet-18 network [54] as the backbone of the PD prediction model, whose architecture is illustrated in Fig. 6. Technically, the network consists of 8 residual blocks. Within each block, a convolution operation with a kernel size of 3 is applied, followed by batch normalization (i.e., BN) [55] and ReLU activation. For each hidden layer, the output are pulled back by BN to the standard normal distribution with a mean of 0 and variance of 1. In this way, the input value of the nonlinear transformation function represented by ReLU can be restricted into a sensitive area, and the gradient vanishing problem can be effectively alleviated. Moreover, shortcut connections skipping two layers are additionally constructed in the network to avoid performance degradation brought by more layers. At the end of the network, the classification layer of the original ResNet-18 is substituted with two FC layers: the former is to reduce the dimension of the feature embedding from 512 to 256, and the latter receives the 256-dimensional feature and outputs a 2-dimensional feature vector.

In training of the PD prediction model, the original facial expression images and the high-quality synthesized (premorbid) facial expression images of the training PD patients are fed into the network, while the output is a vector of two elements indicating the probabilities of being PD/non-PD.

IV. Experimental Results

In Section IV-A, we first introduce our created PD facial expression dataset of PD patients and another four public facial expression datasets of normal persons. We then introduce the implementation details and parameter settings of our proposed method in Section IV-B. Subsequently, we conduct the following experiments to demonstrate the effectiveness of the proposed method.

1) In Section IV-C, we evaluate the synthesized facial expression images of PD patients by StarGAN.
2) In Section IV-D, we evaluate the introduced three FIQA criteria, i.e., facial symmetry, image sharpness, and FaceQnet, and the designed fusion screening strategy for shortlisting high-quality synthesized images.
3) In Section IV-E, we evaluate the performance of our proposed method for PD diagnosis and compare with the state-of-the-art counterparts.
4) In Section IV-F, we perform an ablation study to investigate the roles of the facial expression image synthesis by

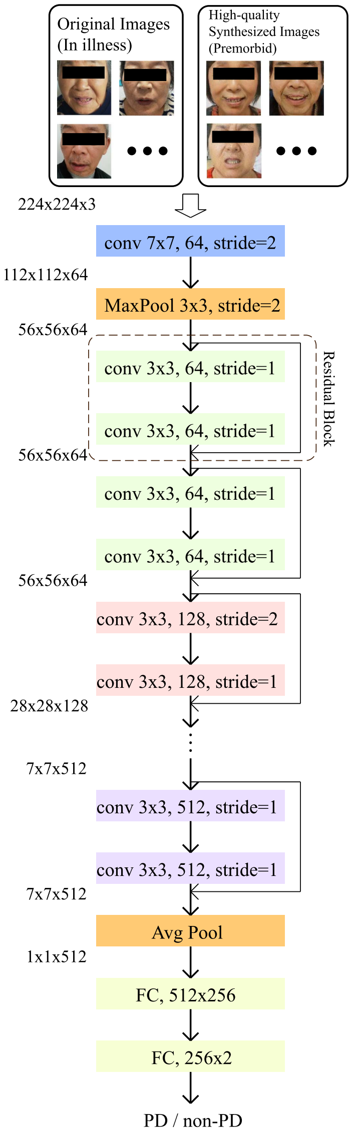

Fig. 6. An illustration of the PD prediction model based on ResNet-18.

StarGAN and the fusion screening strategy on the performance of the proposed method.

A. Datasets Descriptions

In this subsection, we first introduce our created PD facial expression (PDFE) dataset of PD patients, and then describe the other four public facial expression datasets of normal persons, i.e., CK+, RaFD, Oulu-CASIA and Tsinghua-FED.

1) The PDFE dataset: This dataset is collected by our group from an on-going population-based PD study, which has been carried out at the affiliated hospital of the Nanchang University. This dataset contains 95 PD patients (including 55 male and 40 female), in which the average age of these patients is 62.7 with a standard error of ± 9.9. For each patient, there are seven images captured by a CANON EOS 5D Mark III DSLR camera equipped with the EF 24-70 mm f/2.8 L II USM lens, which represent the neutral expression and another 6 types of basic facial expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise). *To the best of our knowledge, this PDFE dataset is currently the largest PD facial expression dataset for PD diagnosis.* It is worth noting that, the data collection and use have obtained the informed consent of all involved PD patients. In order to avoid disclosure of these PD patients' identity information, the eye regions are removed from both the original and synthesized facial expression images of the PD patients in the following illustration figures.

2) The CK+ dataset: The CK+ dataset [56] is an extended version of the original Cohn-Kanade dataset. It contains 593 image sequences from 123 identities. Each image sequence starts from a neutral face and ends with a peak facial expression. In the experiments, the neutral image in the first frame and the images with expressions in the last three frames from 309 image sequences are used.

3) The RaFD dataset: The Radboud Faces Database (RaFD) [48] is a face dataset consisting of 67 identities in which facial expressions, gaze direction, and head orientation vary in a complete factorial design. For each identity, there are 8 expressions (i.e. contemptuous beyond the scope of neutral and 6 basic expressions), 3 gaze directions and 5 different head orientation angles. In the experiments, the neutral images and 6 basic facial expression images in the setting of frontal orientation and straight gaze direction are utilized.

4) The Oulu-CASIA dataset: The Oulu-CASIA dataset [57] consists of 6 basic facial expressions from 80 identities between 23 and 58 years old. All expression images are captured using a visible light (VIS) camera and a near-infrared (NIR) camera under three different illumination conditions, i.e., strong, weak and dark. For each emotion, the facial expression images form a sequence which records the changing process of emotion intensity from plain to peak. In the experiments, we use the subset from VIS domain under strong condition for evaluation, where the first image of each sequence is treated as the neutral image, while the last three images with the highest emotion intensity as the facial expression images.

5) The Tsinghua-FED dataset: The Tsinghua facial expression database (Tsinghua-FED) [58] comprises facial expression images of 110 Chinese young and older adults displaying eight facial emotional expressions (neutral, happiness, anger, disgust, surprise, fear, content, and sadness). In the experiments, the neutral images and 6 basic facial expression images from the 110 identities are used.

For each dataset, a series of data preprocessing steps are carried out as follows: firstly, we detect the facial landmarks and face region via the face detection API of the popular Face++ software;[2] Then, we perform an affine transformation for each image based on two detected facial landmarks (i.e., the center of left eye, the center of right eye); finally, we crop out the face region for each image and then resize it to 128×128 resolution.

B. Implementation Details and Parameter Settings

At Stage 1, we train the StarGAN model based on the neutral face images and the other 6 face images with facial expressions

[2] Face++: [Online]. Available: https://www.faceplusplus.com/

from the above-mentioned 4 public datasets including CK+, RaFD, Oulu-CASIA and Tsinghua-FED. The hyper-parameters λ_{cls} and λ_{rec} in (4) are set at 1 and 10 respectively, in accordance with [32]. The StarGAN model is trained within 200000 iterations using Adam [59] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ (β_1 and β_2 are exponential decay rates for the moment estimates in the Adam optimizer). Note that, the initial learning rate is set at 0.0001 for the first 100000 iterations and then linearly decays to 0 over the rest iterations. The weight update is performed once on the generator after 5 times on the discriminator. At Stage 2, the pretrained FaceQnet model [51] is employed to measure the quality of the synthesized facial expression images. The number of landmark pairs for evaluating the facial symmetry is set at 3. Regarding the three hyper-parameters in (9), i.e., c_1, c_2, and c_3, we follow the works in [60], [61], [62], [63] and tune the values of the three hyper-parameters via the *grid search strategy*. Specifically, we perform cross-validation on the training set and conduct a grid search for the best combination of hyper-parameters, by varying c_1, c_2, c_3 from 1 to 5. Empirically, we observe that our method achieves the best performance when c_1, c_2, and c_3 are set at 2, 1, and 3, respectively, and fix the values in all the testing experiments. At Stage 3, we adopt the ResNet-18 pretrained on MS-Celeb-1M [64] as the backbone for feature extraction followed by a fine tune step for PD diagnosis. The proposed model is implemented using PyTorch and the source codes are released[3] All experiments are carried out on a workstation (CPU: Intel Xeon i7-7700 k, 32 G RAM, GPU: Nvidia GTX TITAN V 12 G RAM).

C. Evaluation on Facial Expression Synthesis

In this subsection, we evaluate the synthesized facial expression images by StarGAN in terms of neutral and six emotions of anger, disgust, fear, happiness, sadness, and surprise. Furthermore, the synthesized results of another popular adversarial learning-based CycleGAN [47], which is designed for image-to-image translation, are also reported for reference.

The synthesized facial expression results of StarGAN and CycleGAN are illustrated in Fig. 7. It can be observed that, StarGAN successfully synthesizes realistic-looking identity-preserved facial expression images of the four randomly selected PD patients. Compared to the original facial expression images that always cannot match with the correct emotions, the synthesized facial expression images by StarGAN depict the six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) correctly in most cases, which are believed to approximate the premorbid normal facial expression images of these PD patients. In addition, CycleGAN performs worse than StarGAN in terms of facial expression synthesis. For example, the synthesized "surprise" facial expression image by CycleGAN is vague and distorted. The inspiring results in Fig. 7 verify the effectiveness of StarGAN in multi-domain facial expression synthesis.

D. Evaluation on FIQA Criteria and Fusion Screening Strategy

This subsection evaluates the introduced three FIQA criteria, i.e., facial symmetry, image sharpness, and FaceQnet, and the designed fusion screening strategy. In Fig. 8, we illustrate 6 examples of the synthesized facial expression images annotated by the quality scores using the three FIQA criteria and their weighted fusion (see Eq. (9)). From Fig. 8, it can be seen that: 1) For each single FIQA criterion, as the score increases, the quality of the synthesized image also improves in terms of this criterion (e.g., from asymmetric to symmetric in the facial symmetry criterion or from blurry to clear in the image sharpness criterion); 2) The synthesized facial expression image scored high (or low) in one FIQA criterion may obtain low (or high) score in the other. For example, the synthesized facial expression image in red box achieves a high score of 0.8861 in the facial symmetry criterion but only obtains a low score of 0.1024 in the image sharpness criterion; and the synthesized image in green box obtains a low score of 0.0693 in the facial symmetry criterion but acquires the score of 0.6419 in the FaceQnet criterion. The results indicate that the score based on single FIQA criterion may not correctly reflect the true quality of the synthesized image sometimes; 3) By fusing the three FIQA scores and assigning appropriate weights, the sorting of the synthesized images becomes more reasonable by reference to multiple FIQA criteria. For example, the synthesized asymmetric&blurry image in yellow box obtains a very low score of 0.0405, while the two synthesized symmetric&clear images in purple box are graded high scores of 0.8611 and 0.9347, respectively. Furthermore, we calculate the distribution percentage of all the synthesized facial expression images in Table I, according to the above four scoring patterns. It can be observed from Table I that the distributions of the synthesized facial images under the facial symmetry and FaceQnet scoring patterns are biased. By contrast, the distribution of the synthesized facial images under our weighted fusion scoring pattern is balanced and close to the normal distribution. The promising results in Fig. 8 and Table I justify the rationality and effectiveness of the weighted fusion scoring pattern in the fusion screening strategy.

E. Evaluation on PD Diagnosis Accuracy

After acquiring the high-quality synthesized facial expression images using the fusion screening strategy, we then mix them with the original facial expression images of PD patients to train the PD prediction model. In this subsection, we evaluate the PD diagnosis accuracy of the trained prediction model of our method in the testing set. Specifically, the evaluation protocol and the results are described in the following.

The evaluated PDFE dataset is divided evenly into 5 folds. We select 4 folds (76 subjects) for training while the rest 1 fold (19 subjects) for testing. Furthermore, we select the facial expression images of 47 aged subjects (above 60 years old) from the Tsinghua-FED dataset to supplement the test set. In training, the original facial expression images of the PD patients are labeled with "PD," while these high-quality synthesized facial expression images which estimate the premorbid status of the

[3] [Online]. Available: https://github.com/CherryChou98/FE-Guided-PD-Diagnosis.

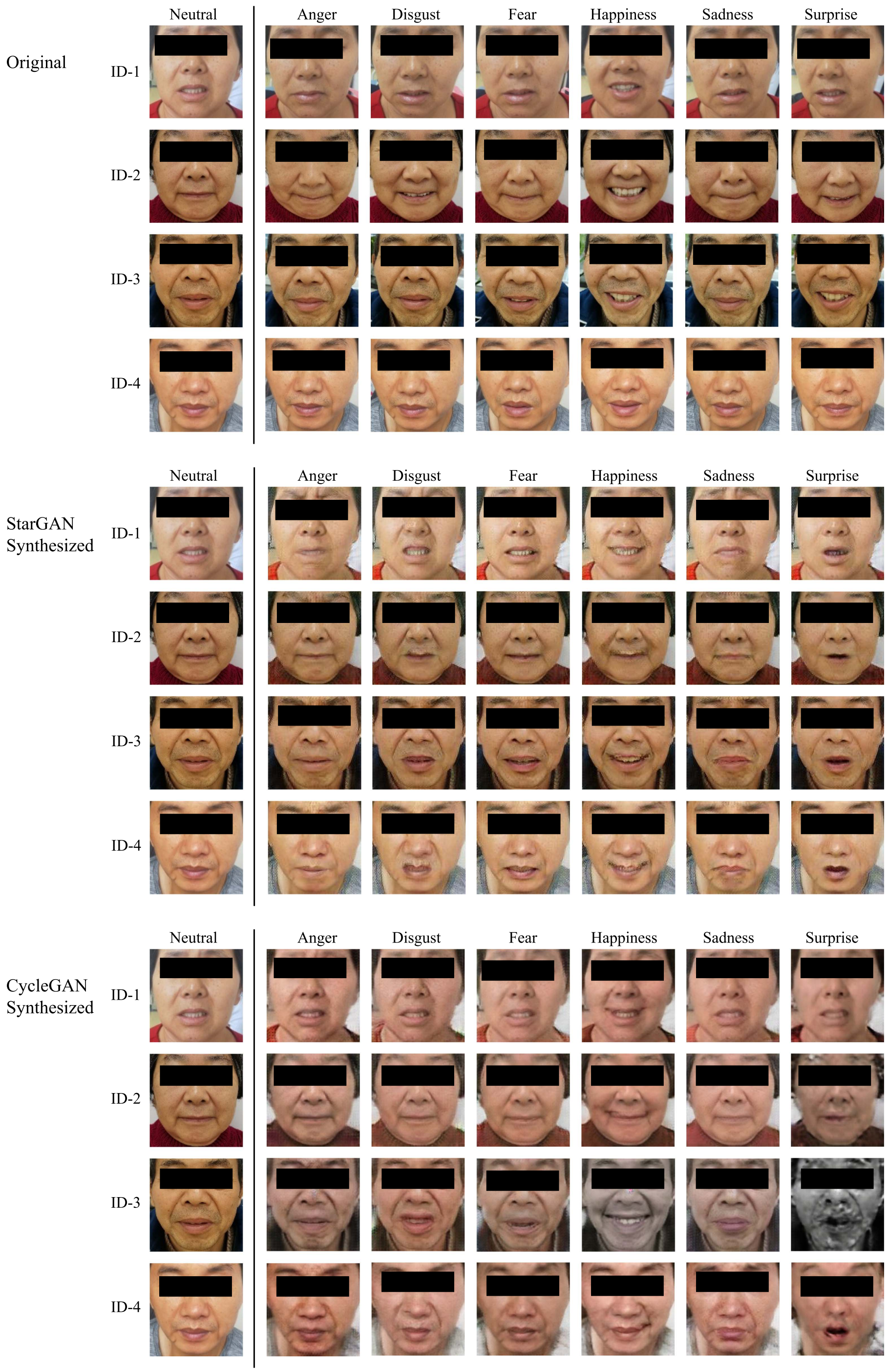

Fig. 7. The synthesized facial expression images by StarGAN and CycleGAN from four randomly selected PD patients. The figures from the top to bottom blocks are the original facial expression images of four PD patients, and the synthesized images by StarGAN and CycleGAN which approximate the premorbid facial expression images of the above four PD patients.

PD patients are labeled with "non-PD". In testing, the total 66 subjects are used for PD diagnosis. We run the experiment 5 times and report the average PD diagnosis accuracy. For clarity, we provide an illustration of the evaluation protocol in Fig. 9.

In Table II, we present the PD diagnosis accuracies and the training time of our proposed method under different screening ratios (i.e., 0%, 10%, 25%, 40%, 55%, 70%, 85%, and 100%) in the fusion screening strategy. In addition, we adopt the CycleGAN using the same fusion screening strategy (denoted as CycleGAN-Fusion) as a baseline for comparison. From Table II, we have the following four key observations: 1) The diagnosis performance of our method is greatly improved from 54.63% to 93.57% as the screening ratio increases from 0% to 10% (i.e., only screening 10% synthesized images). This indicates the importance of the training data augmentation by introducing the high-quality "non-PD" synthesized facial expression images into the original PD dataset. 2) Our proposed method achieves the highest diagnosis accuracy of 95.43% when the screening

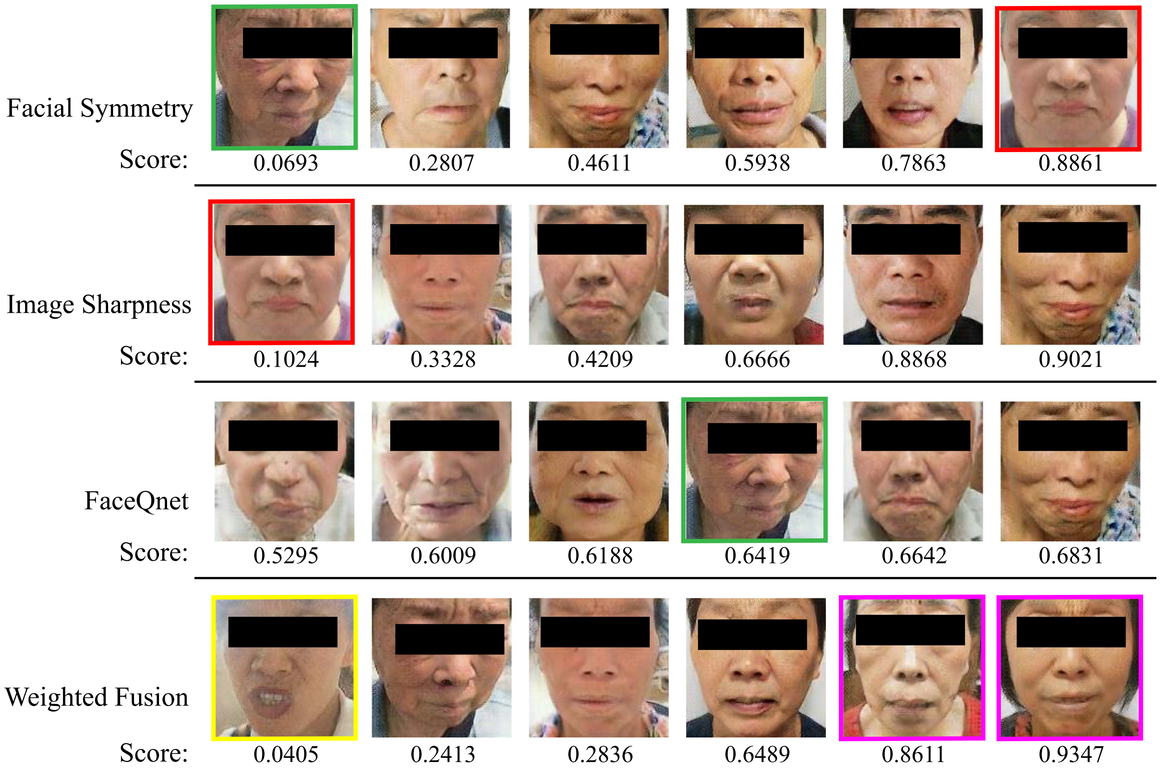

Fig. 8. Illustration of six examples of synthesized facial expression images and the corresponding quality scores using three FIQA criteria (i.e., facial symmetry, image sharpness, and FaceQnet) and their weighted fusion.

TABLE I
THE DISTRIBUTION PERCENTAGE (%) OF ALL THE SYNTHESIZED FACIAL EXPRESSION IMAGES UNDER FOUR SCORING PATTERNS

Interval Scoring pattern	[0,0.1)	[0.1,0.2)	[0.2,0.3)	[0.3,0.4)	[0.4,0.5)	[0.5,0.6)	[0.6,0.7)	[0.7,0.8)	[0.8,0.9)	[0.9,1.0]
Facial symmetry	0.35	0.53	2.06	3.46	2.64	2.88	11.34	30.38	32.08	14.28
Image sharpness	4.04	8.30	12.92	10.99	14.33	19.12	14.33	9.30	4.68	1.99
FaceQnet	0	0	0	0	0	22.57	71.64	5.79	0	0
Weighted fusion	1.23	3.0	6.11	10.58	22.39	21.5	20.09	11.1	3.35	0.65

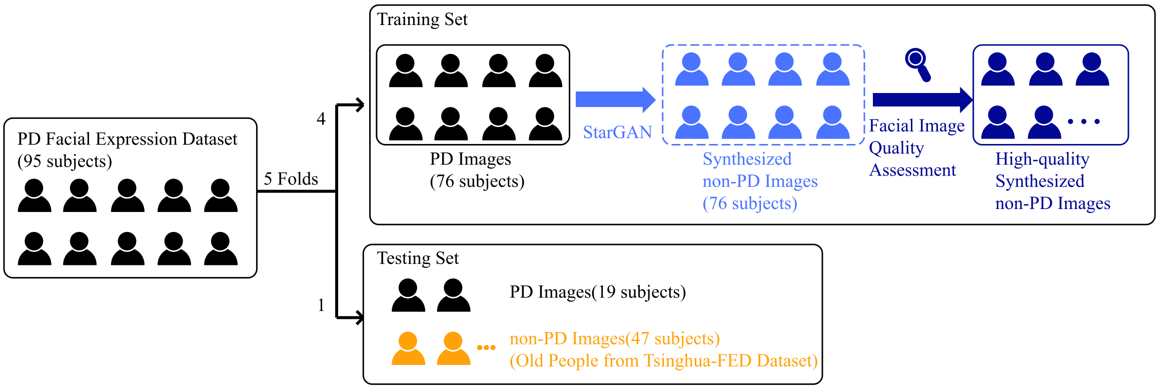

Fig. 9. An illustration of the evaluation protocol for PD diagnosis.

TABLE II
PD DIAGNOSIS ACCURACIES (%) / TRAINING TIME COSTS (S) OF OUR PROPOSED METHOD UNDER DIFFERENT SCREENING RATIOS. 0% (OR 100%) INDICATES THAT NO SYNTHESIZED IMAGES (OR ALL SYNTHESIZED IMAGES) ARE SCREENED OUT

Screening Ratio (%)	CycleGAN-Fusion	Ours
0	54.63 / 280.6	54.63 / 286.6
10	**71.36** / 298.8	93.57 / 303.6
25	64.91 / 329.6	**95.43** / 338.5
40	61.85 / 354.5	94.43 / 360.1
55	57.53 / 383.6	93.28 / 388.8
70	61.91 / 412.4	92.44 / 416.4
85	63.15 / 440.0	92.59 / 442.6
100	58.58 / 475.1	93.43 / 480.6

TABLE III
COMPARISON BETWEEN OUR METHOD AND THE OTHER PD DIAGNOSIS METHODS W.R.T. PD DIAGNOSIS ACCURACY (%)

Diagnosis Method	Accuracy(%)
GLCM+SVM [28]	46.60
DeiT-small [63]	71.93
ConvNeXt-Tiny [64]	73.36
EfficientNetV2 [65]	71.25
FaceQNet [51]	76.03
InceptionResnetV1 [66]	84.07
Ours	**95.43**

TABLE IV
PD DIAGNOSIS ACCURACIES (%) OF TWO VARIANTS OF OUR METHOD BY REMOVING THE FACIAL EXPRESSION SYNTHESIS AND THE FUSION SCREENING STRATEGY, RESPECTIVELY

Methods	Accuracy(%)
Ours w/o facial expression synthesis (setting 1)	54.63
Ours w/o facial expression synthesis (setting 2)	45.37
Ours w/o facial expression synthesis (setting 3)	71.34
Ours w/o fusion screening strategy	93.43
Ours	**95.43**

ratio reaches 25%, and then suffers from performance degradation as the ratio rises. 3) The performance of CycleGAN-Fusion are far inferior to that of our method, again demonstrating the rationality of choosing StarGAN for multi-domain facial expression image synthesis. 4) The training time of the proposed method tends to increase linearly as the screening ratio increases, and consumes only 480.6 seconds (ratio = 100%) even all synthesized facial expression images are screened for training.

Furthermore, we compare our proposed method with a representative facial expression-based PD diagnosis method [28] based on conventional machine learning techniques of GLCM and SVM (GLCM+SVM), and five state-of-the-art deep learning models, i.e., DeiT [65], ConvNeXt [66], EfficientNetV2 [67], FaceQNet [51], and InceptionResnetV1 [68]. Note that, DeiT, ConvNeXt, and EfficientNetV2 are pretrained on the ImageNet dataset [69], FaceQNet is pretrained on the VGGFace2 dataset [70], and InceptionResnetV1 is pretrained on the CASIA-Webface dataset [71]. All of the above five deep learning models are fine tuned using the original facial expression images from the PDFE dataset and the facial expression images of normal persons from three public datasets, i.e., Oulu-CASIA, RaFD and CK+. The testing procedure follows the evaluation protocol mentioned above (refer to Fig. 9). The diagnosis accuracies of our method and the other compared methods are listed in Table III. From Table III, it can be observed that,

- Our method achieves a high PD diagnosis accuracy of 95.43% which is much better than that of the other six compared methods. The superiority of our method attributes to two aspects: 1) the effective training data augmentation scheme (i.e., StarGAN-based facial expression synthesis + fusion screening strategy) provides high-quality synthesized facial expression images of PD patients to facilitate the training process; 2) the deep neural network PD prediction model extracts high-semantic features to boost PD/non-PD classification.
- The conventional machine learning-based GLCM+SVM performs worse than the five deep learning methods, and is far inferior to our method in terms of PD diagnosis accuracy, which demonstrates the good representation learning capability of deep neural networks.
- FaceQNet and InceptionResnetV1 pretrained on professional face datasets could achieves better PD diagnosis performance than ConvNeXt-Tiny, EfficientNetV2, and DeiT-small pretrained on the ImageNet dataset.

F. Ablation Study

In this subsection, we perform an ablation study to explore the roles of the two modules, i.e., facial expression synthesis by StarGAN and the fusion screening strategy, on the PD diagnosis performance of our proposed method. Accordingly, we first construct two variants of our method denoted as "Ours w/o facial expression synthesis" and "Ours w/o fusion screening strategy" by removing the two modules, respectively. For "Ours w/o facial expression synthesis," there exist three different training settings: 1) training our model only using the PDFE dataset of PD patients, 2) training our model only using the public facial expression datasets of normal persons, i.e., Oulu-CASIA, RaFD, and CK+, and 3) training our model using the mixture of PDFE, Oulu-CASIA, RaFD and CK+ datasets. From Table IV, It can be observed that,

- "Ours w/o facial expression synthesis (setting 1)" and "Ours w/o facial expression synthesis (setting 2)" perform poor in PD diagnosis, because their training datasets only include a single category label (either PD or non-PD).
- "Ours w/o facial expression synthesis (setting 3)" obtains a higher PD diagnosis accuracy than that of "Ours w/o facial expression synthesis (setting 1)" and "Ours w/o facial expression synthesis (setting 2)," which indicates that the mixture of the PDFE dataset of PD patient and the three public facial expression datasets of normal persons can facilitate the training of the PD/non-PD classifier in PD diagnosis model.
- The PD diagnosis performance of our method (with facial expression synthesis) is far superior to that of "Ours w/o facial expression synthesis (setting 3)," which indicates that

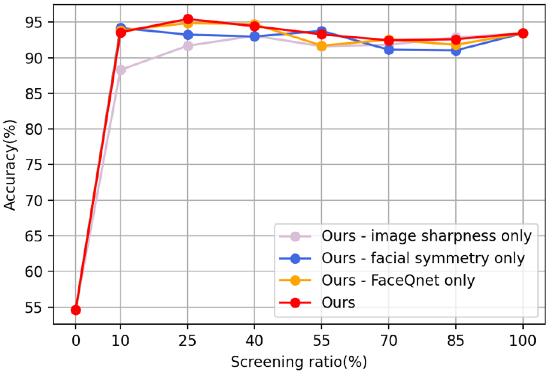

Fig. 10. PD diagnosis accuracies (%) of three variants of our method, which only use the facial symmetry, image sharpness, and FaceQnet for screening high-quality synthesized facial expression images, respectively.

the introduction of the synthesized premorbid facial expression images of PD patients can further promote the training of the diagnosis model.

- The performance of “Ours w/o fusion screening strategy” is not competitive with that of our method (with fusion screening strategy), which shows that the fusion screening strategy also contributes to the PD diagnosis performance by filtering out some low-quality synthesized facial expression images.

Furthermore, we also report the diagnosis accuracies of three variants of our method by only using the facial symmetry, image sharpness, or FaceQnet for screening high-quality synthesized facial expression images, respectively. As shown in Fig. 10, the three variants achieve their best performance under different screening ratios, and all of them perform worse than our method (i.e., with fusion screening strategy) in almost all cases. This again verifies the superiority of the weighted fusion scoring pattern in the fusion screening strategy over the other scoring patterns based on single criterion such as facial symmetry, image sharpness or FaceQnet.

V. Conclusion and Future Works

This paper has proposed a new facial expression guided in-vitro PD diagnosis method. It addresses the two problems (i.e., limited training data and weak prediction model) in traditional in-vitro PD diagnosis methods based on facial expressions by introducing an effective data augmentation scheme (i.e., StarGAN-based facial expression synthesis + fusion screening strategy) and a powerful deep neural network PD prediction model. Empirical studies have demonstrated the superior performance of the proposed method for PD diagnosis. It is worth noting that this paper has created a PDFE dataset containing both the neutral image and 6 basic facial expression images of 95 PD patients, which is currently the largest facial expression dataset of PD patients for in-vitro PD diagnosis and will be released soon in the future.

It is worth mentioning that, PD patients may not have completely lost their ability to express emotions in the early stage of the disease, and they may be able to express certain expressions correctly. Under the circumstances, using a single facial image of a PD patient to perform diagnosis may carry the risk of misjudgment. To enhance robustness, we plan to extend our proposed approach to a comprehensive diagnosis based on the patient’s six basic facial expression image combinations (i.e., anger + disgust+fear + happiness+sadness + surprise). Besides, we will make an effort to combine the other in-vitro diagnosis techniques based on speech or gait signals with our proposed method based on facial expressions, so as to increase the error tolerance in PD diagnosis. We will leave the interesting study as the future research work. Furthermore, we will continue to collect more facial expression images of new PD patients to enlarge our created PDFE dataset.

References

[1] J. Jankovic, “Parkinson’s disease: Clinical features and diagnosis,” *J. Neurol., Neurosurgery Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] W. Dauer and S. Przedborski, “Parkinson’s disease: Mechanisms and models,” *Neuron*, vol. 39, no. 6, pp. 889–909, 2003.

[3] V. L. Feigin et al., “Global, regional, and national burden of neurological disorders, 1990–2016: A systematic analysis for the global burden of disease study 2016,” *Lancet Neurol.*, vol. 18, no. 5, pp. 459–480, 2019.

[4] E. Dorsey, T. Sherer, M. S. Okun, and B. R. Bloem, “The emerging evidence of the Parkinson pandemic,” *J. Parkinson’s Dis.*, vol. 8, no. s1, pp. S3–S8, 2018.

[5] R. Guo, X. Shao, C. Zhang, and X. Qian, “Multi-scale sparse graph convolutional network for the assessment of parkinsonian gait,” *IEEE Trans. Multimedia*, vol. 24, pp. 1583–1594, 2022.

[6] X. Chen, X. Chen, R. K. Ward, and Z. J. Wang, “A joint multimodal group analysis framework for modeling corticomuscular activity,” *IEEE Trans. Multimedia*, vol. 15, pp. 1049–1059, 2013.

[7] F. L. Pagan, “Improving outcomes through early diagnosis of Parkinson’s disease,” *Amer. J. Managed Care*, vol. 18, no. 7, 2012, Art. no. S176.

[8] B. R. Brewer, S. Pradhan, G. Carvell, and A. Delitto, “Application of modified regression techniques to a quantitative assessment for the motor signs of Parkinson’s disease,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 6, pp. 568–575, Dec. 2009.

[9] E. Tolosa, A. Garrido, S. W. Scholz, and W. Poewe, “Challenges in the diagnosis of Parkinson’s disease,” *Lancet Neurol.*, vol. 20, no. 5, pp. 385–397, 2021.

[10] E. Tolosa, G. Wenning, and W. Poewe, “The diagnosis of Parkinson’s disease,” *Lancet Neurol.*, vol. 5, no. 1, pp. 75–86, 2006.

[11] C. O. Sakar et al., “A comparative analysis of speech signal processing algorithms for Parkinson’s disease classification and the use of the tunable q-factor wavelet transform,” *Appl. Soft Comput.*, vol. 74, pp. 255–263, 2019.

[12] T. Tuncer, S. Dogan, and U. R. Acharya, “Automated detection of Parkinson’s disease using minimum average maximum tree and singular value decomposition method with vowels,” *Biocybernetics Biomed. Eng.*, vol. 40, no. 1, pp. 211–220, 2020.

[13] A. K. Shukla, P. Singh, and M. Vardhan, “Medical diagnosis of Parkinson disease driven by multiple preprocessing technique with scarce lee Silverman voice treatment data,” in *Engineering Vibration, Communication and Information Processing*. Singapore: Springer, 2019, pp. 407–421.

[14] C. Quan, K. Ren, and Z. Luo, “A deep learning based method for Parkinson’s disease detection using dynamic features of speech,” *IEEE Access*, vol. 9, pp. 10239–10252, 2021.

[15] S.-C. Hsu et al., “Acoustic and perceptual speech characteristics of native Mandarin speakers with Parkinson’s disease,” *J. Acoustical Soc. Amer.*, vol. 141, no. 3, pp. EL293–EL299, 2017.

[16] C. Laganas et al., “Parkinson’s disease detection based on running speech data from phone calls,” *IEEE Trans. Biomed. Eng.*, vol. 69, no. 5, pp. 1573–1584, May 2022.

[17] O. Y. Chén et al., “Building a machine-learning framework to remotely assess Parkinson’s disease using smartphones,” *IEEE Trans. Biomed. Eng.*, vol. 67, no. 12, pp. 3491–3500, Dec. 2020.

[18] J. Barth et al., “Biometric and mobile gait analysis for early diagnosis and therapy monitoring in Parkinson’s disease,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 868–871.

[19] R. N. Noella, D. Gupta, and J. Priyadarshini, "Diagnosis of Parkinson's disease using gait dynamics and images," *Procedia Comput. Sci.*, vol. 165, pp. 428–434, 2019.

[20] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, and V. Venkatraman, "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease," *Future Gener. Comput. Syst.*, vol. 83, pp. 366–373, 2018.

[21] E. Balaji, D. Brindha, V. K. Elumalai, and K. Umesh, "Data-driven gait analysis for diagnosis and severity rating of Parkinson's disease," *Med. Eng. Phys.*, vol. 91, pp. 54–64, 2021.

[22] G. Prateek et al., "Modeling, detecting, and tracking freezing of gait in Parkinson disease using inertial sensors," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 10, pp. 2152–2161, Oct. 2018.

[23] R. Guo, X. Shao, C. Zhang, and X. Qian, "Sparse adaptive graph convolutional network for leg agility assessment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2837–2848, Dec. 2020.

[24] A. Grammatikopoulou, N. Grammalidis, S. Bostantjopoulou, and Z. Katsarou, "Detecting hypomimia symptoms by selfie photo analysis: For early Parkinson disease detection," in *Proc. 12th ACM Int. Conf. PErvasive Technol. Related to Assistive Environments*, 2019, pp. 517–522.

[25] B. Jin et al., "Diagnosing Parkinson disease through facial expression recognition: Video analysis," *J. Med. Internet Res.*, vol. 22, no. 7, 2020, Art. no. e18697.

[26] A. Bandini et al., "Analysis of facial expressions in Parkinson's disease through video-based automatic methods," *J. Neurosci. Methods*, vol. 281, pp. 7–20, 2017.

[27] M. Rajnoha et al., "Towards identification of hypomimia in Parkinson's disease based on face recognition methods," in *Proc. 10th Int. Congr. Ultra Modern Telecommun. Control Syst. Workshops*, 2018, pp. 1–4.

[28] X. Hou, S. Qin, and J. Su, "Visual detection of Parkinson's disease via facial features recognition," in *Proc. Chin. Intell. Automat. Conf.*, 2022, pp. 249–257.

[29] L. Ricciardi et al., "Hypomimia in Parkinson's disease: An axial sign responsive to levodopa," *Eur. J. Neurol.*, vol. 27, no. 12, pp. 2422–2429, 2020.

[30] M. R. Ali et al., "Facial expressions can detect Parkinson's disease: Preliminary evidence from videos collected online," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–4, 2021.

[31] P. Wu et al., "Objectifying facial expressivity assessment of Parkinson's patients: Preliminary study," *Comput. Math. Methods Med.*, vol. 2014, pp. 1–12, 2014.

[32] Y. Choi et al., "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.

[33] W. Huang et al., "Identity-aware facial expression recognition via deep metric learning based on synthesized images," *IEEE Trans. Multimedia*, vol. 24, pp. 3327–3339, 2022.

[34] D. Zhou et al., "Towards multi-domain face synthesis via domain-invariant representations and multi-level feature parts," *IEEE Trans. Multimedia*, vol. 24, pp. 3469–3479, 2022.

[35] Y. Yan, Y. Huang, S. Chen, C. Shen, and H. Wang, "Joint deep learning of facial expression synthesis and recognition," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2792–2807, Nov. 2020.

[36] M. Ullrich et al., "Detection of unsupervised standardized gait tests from real-world inertial sensor data in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2103–2111, 2021.

[37] H. Masur, "Skalen und scores in der neurologie," *Neurol. Sci.*, vol. 21, 2000, Art. no. 254.

[38] H. Stolze, P. Vieregge, and G. Deuschl, "Gangstörungen in der neurologie," *Der Nervenarzt*, vol. 79, no. 4, pp. 485–499, 2008.

[39] G. Simons, M. C. S. Pasqualini, V. Reddy, and J. Wood, "Emotional and nonemotional facial expressions in people with Parkinson's disease," *J. Int. Neuropsychol. Soc.*, vol. 10, no. 4, pp. 521–535, 2004.

[40] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 2672–2680, 2014.

[41] X. Zhang, Y. Zhu, W. Chen, W. Liu, and L. Shen, "Gated switchGAN for multi-domain facial image translation," *IEEE Trans. Multimedia*, vol. 24, pp. 1990–2003, 2022.

[42] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Trans. Multimedia*, vol. 24, pp. 3859–3881, 2022.

[43] L. Chen, L. Wu, Z. Hu, and M. Wang, "Quality-aware unpaired image-to-image translation," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2664–2674, Oct. 2019.

[44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[45] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5769–5779, 2017.

[46] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[48] O. Langner et al., "Presentation and validation of the radboud faces database," *Cogn. Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[49] T. Schlett et al., "Face image quality assessment: A literature survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–49, 2020.

[50] L. Her and X. Yang, "Research of image sharpness assessment algorithm for autofocus," in *Proc. IEEE 4th Int. Conf. Image, Vis. Comput.*, 2019, pp. 93–98.

[51] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "Faceqnet: Quality assessment for face recognition based on deep learning," in *Proc. Int. Conf. Biometrics*, 2019, pp. 1–8.

[52] M. Ferrara, A. Franco, D. Maio, and D. Maltoni, "Face image conformance to ISO/ICAO standards in machine readable travel documents," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 4, pp. 1204–1213, Aug. 2012.

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

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

[55] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[56] 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. Vis. Pattern Recognit. - Workshops*, 2010, pp. 94–101.

[57] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

[58] T. Yang et al., "Tsinghua facial expression database–a database of facial expressions in Chinese young and older women and men: Development and validation," *PLoS One*, vol. 15, no. 4, 2020, Art. no. e0231304.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–41.

[60] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, pp. 6829–6836, 2018.

[61] M. Pang, B. Wang, Y.-M. Cheung, Y. Chen, and B. Wen, "VD-GAN: A unified framework for joint prototype and representation learning from contaminated single sample per person," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 2246–2259, 2021.

[62] M. Pang, B. Wang, S. Huang, Y.-M. Cheung, and B. Wen, "A unified framework for bidirectional prototype learning from contaminated faces across heterogeneous domains," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 1544–1557, 2022.

[63] M. Pang et al., "DisP V: A unified framework for disentangling prototype and variation from single sample per person," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 17, 2022, doi: 10.1109/TNNLS.2021.3103194.

[64] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1 m: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.

[65] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[66] Z. Liu et al., "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.

[67] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[68] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, pp. 4278–4284, 2017.

[69] J. Deng et al., “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[70] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 67–74.

[71] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” Nov. 2014, *arXiv:1411.7923*. [Online]. Available: https://arxiv.org/abs/1411.7923

Wei Huang received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, and the Ph.D. degree from the Nanyang Technological University, Singapore. He was with the University of California San Diego, San Diego, CA, USA, and the Agency for Science Technology and Research, Singapore, as a Postdoctoral Research Fellow. He is currently a Full Professor with the Department of Computer Science and acts as the Dean of the School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China. He has authored or coauthored more than 100 academic journal or conference papers, including the IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON MULTIMEDIA, MICCAI, and *ACM Multimedia*. His main research interests include machine learning, pattern recognition, computer vision, and multimedia. He is a Principal Investigator in studies supported by nearly 20 national or provincial grants, including six NSF-China projects and four NSF key projects in Jiangxi Province, China. He was the recipient of the Jiangxi Provincial Natural Science Award, Best Paper Award of MICCAI-MLMI, most interesting Paper Award of ICME-ASMMC, Best Paper Award of ICITBE, and Best Paper Award of ICCEAI. In 2020, he was designated as the Academic Leader of Jiangxi, China.

Yintao Zhou received the B.Eng. degree in computer science and technology in 2020 from Nanchang University, Nanchang, China, where he is currently working toward the M.Eng. degree under the supervision of Prof. Wei Huang. His research interests mainly include computer vision and machine learning.

Yiu-ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. He is currently a Chair Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning and visual computing, and their applications. Prof. Cheung is also an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, and *Neurocomputing*. He is a fellow of AAAS, IET, and BCS.

Peng Zhang (Member, IEEE) received the B.E. degree from Xi’an Jiaotong University, Xi’an, China, in 2001, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2011. He is currently a Full Professor with the School of Computer Science, Northwestern Polytechnical University, Xi’an, China. He has authored or coauthored more than 80 research papers in journals or conferences, including CVPR, *ACM Multimedia*, *Neurocomputing*, *Pattern Recognition*, *Signal Processing*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON MEDICAL IMAGING. His current research interests include computer vision, pattern recognition, and machine learning. He is the PI for studies supported by three grants from NSFC. He is also the Chief Scientist in Mekitec OY, Oulu, Finland.

Yufei Zha (Member, IEEE) received the Ph.D. degree in information and communication engineering from Airforce Engineer University, Xi’an, China, in 2009. He is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University, Xi’an, China. His research interests include object detection, visual tracking, and machine learning.

Meng Pang received the B.Sc. and M.Sc. degrees in software engineering from Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China, in 2019. From 2020 to 2022, he was a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently a Distinguished Professor with the School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China. His research interests include artificial intelligence security and artificial intelligence medical.