Beyond Vanilla Convolution: Random Pixel Difference Convolution for Face Perception

WENZHE LIU\(^1\), ZHUO SU\(^2\), AND LI LIU\(^{1,2}\), (Senior Member, IEEE)
\(^1\)College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
\(^2\)Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland
Corresponding author: Li Liu (li.liu@oulu.fi)

This work was supported by the National Natural Science Foundation of China under Grant 61872379.

ABSTRACT

Face perception is an essential and significant problem in pattern recognition, concretely including Face Recognition (FR), Facial Expression Recognition (FER), and Race Categorization (RC). Though handcrafted features perform well on face images, Deep Convolutional Neural Networks (DCNNs) have brought new vitality to this field recently. Vanilla DCNNs are powerful at learning high-level semantic features, but are weak in capturing low-level image characteristic changes in illumination, intensity, and texture regarded as key traits in facial processing and feature extraction, which is alternatively the strength of human-designed feature descriptors. To integrate the best of both worlds, we proposed novel Random Pixel Difference Convolution (RPDC) which is efficient alternatives to vanilla convolutional layers in standard CNNs and can promote to extract discriminative and diverse facial features. By means of searched RPDC of high efficiency, we build S-RaPiDiNet, and achieve promising and extensive experiment results in FR (≈0.5% improvement), FER (over 1% growth), and RC (0.25%-3% increase) than baseline network in vanilla convolution, showing strong generalization of RPDC.

INDEX TERMS

Random pixel difference convolution, race categorization, face recognition, emotion recognition.

I. INTRODUCTION

As is well-known, face conveys rich key non-verbal information, and these essential attributes derived from human faces play a vital role in cognitive inference. For social communication, a human must manage to handle faces by means of various methods, including recognizing familiar faces and extracting features, such as race and mental state, on basis of face images. Not only human-to-human interaction, information on the face is also crucial on human-computer interfaces, access control, and surveillance. Due to the unstoppable globalization, artificial intelligence, and frequent personnel transfer in different industries, face perception can affect the daily life of each individual increasingly. In the recent wave of intelligent machines miniaturization and popularization, a growing number of projects have been explored supported by government agencies and research institutions for social and scientific purposes.

Most traditional face processing methods are usually shallow [1], thus are unable to represent high-level semantic information. This leads to limitations since facial features are usually semantically meaningful. Recently, like many other computer tasks, facial feature extraction has been advanced significantly [2] since Deep Convolutional Neural Networks (DCNNs) reentered the field [3]. DCNNs have a strong ability to automatically learn representations of natural images with multiple levels of abstraction [4]. In particular, DCNNs are good at learning deeper semantic features for joint face recognition. However, they are weak in capturing fine-grained image patterns.

Intrinsically, the vanilla convolution naively treats local cues as a weighted sum, which would definitely ease the detailed difference information in local. Therefore, the network built by vanilla convolution encourages similar responses of facial regions, which is contrary to the fact that facial feature is a composition of micropatterns or highly interdependent local pixels, leading to failure in the extraction of discriminative facial features, as shown in Fig. 1. Concretely, baseline built by vanilla convolution lags behind...
In this paper, inspired by the varied pattern representation brought by random pixel difference (as shown in Fig. 1), we intend to promote face feature discrimination and propose a simple and efficient solution: Random Pixel Difference Convolution (RPDC), which is efficient and powerful alternatives to vanilla convolution in DCNNs. Because of the lack of sufficient research on race categorization, we also produce a benchmark result on a large-scale ethnicity estimation, which involves more than twice the number of images and twice the number of race labels in the existing method [6] utilizing the largest face image set yet, by means of residual unit [7] and A-Softmax [2]. When substituting vanilla convolution in the above-mentioned baseline and classic networks by RPDC and developing its gradient variety, the advanced structure reflects considerable inroads into performance promotion in various vision tasks.

Our main contribution are the following:

- First, we propose a new convolution, Random Pixel Difference Convolution, facilitating the extraction of local difference and gradient information, and keep the semantic features extraction capacity of vanilla convolution at the same time.
- Second, the proposed RPDC is universal, and can directly substitute vanilla convolution in existing DCNNs to build S-RaPiDiNet without extra parameters and computational cost. S-RaPiDiNet can successfully achieve performance gains on various vision tasks, such as face recognition, facial expression recognition, and race categorization;
- Third, we fabricate an ethnicity estimation baseline in the newly-released largest-scale race dataset with more than 40,000 face images.

The remainder of this paper is organized as follows. Section II looks back at relevant works in facial tasks and delivers a brief outline of Local Binary Pattern and its extensions. Section III reports our Random Pixel Difference Convolution and a joint baseline network for face perception, leading to the comprehensive experimental evaluations reported in Section IV. Section V abstracts the promising research direction of pixel difference convolution and race categorization. Section VI reports the conclusions, lately.

II. RELATED WORK

A. FACIAL TASKS

Current research efforts in the face perception community focus on the development of algorithms with higher accuracy and robustness in face detection, recognition, and categorization with regard to gender, race, emotion, and age. Face recognition, facial expression recognition, and race categorization are all representative facial tasks in computer vision. Face recognition is the leading biometric technology for identity certification, and facial expression recognition evidently benefits human-computer interactions. Since face recognition and facial expression recognition are sufficiently and adequately studied, and the research process of face processing is quite alike, we mainly focus on the introduction of race categorization in this section.

Race categorization refers to the task of categorizing the individual in a face image in one of several predefined ethnicity classes with a shared origin, customs, and conventions. Admittedly, ethnicity or race is a fundamental morphological trait of human beings, which remains still throughout one’s lifetime. Race is a sensitive and significant component in facial vision and possesses various applications including biometrics-based individual identification [8], health care, population diversity, minority treatment, cross-culture marketing, video and image retrieval, and database indexing.

As the key demographic attribute of a face, it attracts increasing attention currently because of the algorithmic bias issue in face recognition [9]. Concretely, Phillips et al. [10] point out that the models released by researchers in east Asia come out inferior results when identifying Caucasian and East Asians. Klare et al. [11] put forward that the face recognition system designed by Cognitec, a trusted and authoritative company, identifies African Americans with an accuracy of 5-10% lower than that of identifying White. Accordingly, careful research in the race can benefit bias mitigation and even elimination without a doubt.

Despite its noteworthy significance, few tentative attempts have been made to perform ethnic identification, and the ones usually used relatively quantity-restricted data sets with small sizes and were limited by handcrafted face features. Naturally, another primary drawback is no ground comparison. Regardless of the need to integrate several databases together to perform multi-race classification [12], [13], many face datasets are usually ill-balanced and labeled by non-uniform ethnicity labels. Therefore, there is no convincing method, and no conclusions can be drawn about the capacity of different algorithms.

Anyway, race perception is rather a new and challenging topic in cognitive inference. The major difficulty lies in the relatively large intra-class distance caused by factors including occlusion, gender, aging, miscegenation, and expression. Prior works can be roughly generalized to handcrafted feature-based methods and learning-based methods on account of feature generation. Representative human-designed features include Local Binary Patterns (LBP) features, Histogram of Oriented Gradient (HoG) features, Scale Invariant Feature Transform (SIFT) features, Weber Local Descriptor (WLD) features, Gabor wavelets, etc. Conversely, learning-based models are mainly DCNN-based in the new deep learning wave, but with less concern and contribution in race perception. On the other hand, the ethnicity categorization task is closely related to both overall physical appearance features (e.g., the distance between facial features [14], skin tones) and regional discriminative components (e.g., eye socket and nose). Thus, the research focus of race categorization can be generalized into a holistic method and region-based key components approach on basis of the main analysis component.
W. Liu et al.: Beyond Vanilla Convolution: RPDC for Face Perception

**FIGURE 1.** From left to right: (a) the input images; (b) Features obtained by vanilla convolution; (c) Features obtained by Central Difference Convolution (CDC) [5]; (d) Features obtained by various random pixel difference convolution. Clearly, the feature group extracted by random pixel difference convolutions on the far left highlights gradient information and various patterns, and generates more discriminative features than vanilla convolution and CDC [5].

1) REGION-BASED KEY COMPONENTS APPROACH
Local feature-based methods mainly utilize features extracted in local regions on a face. The primary reasons for local feature-based taking priority over holistic-based methods in face recognition community are robustness of key components, and low demand of data and computing resources. Typical crucial positions on faces include iris texture [15], periocular region [16], nose and nostril, sharing similar characteristics including discrimination in races, robustness to illumination and pose variations [17], and stability regardless of aging or any other later modification. Though key components naturally carry both informative and essential features, the accurate localization of various facial features required by these methods might not be easy from the practical point of view, potentially leading to unsatisfactory when only taking them into account. Consequently, it is reasonable that multimodality approaches shall surpass either mode alone [18].

2) HOLISTICAL METHOD
These overall methods process the input image in a comprehensive way. For instance, Gutta et al. [19] fed extracted features to a hybrid structure, where radial basis functions and inductive detection trees were allocated in sequence, and achieved 94% as the accuracy on the set of 3,006 images. Hosoi et al. [20] generally utilized Gabor wavelets to extract features and employed a followed Support Vector Machines (SVM) as the classifier. On the test set of 1,991 face images tagged by three ethnic labels: African, Asian, and European, they declared a precision of nearly 94%. Likewise, Lu et al. [21] employed two SVMs in a binary classification (i.e., Asian and non-Asian) of 918 face images in total. One SVM worked on the normal intensity image as in previous papers, another SVM acted on the 3D range image captured by the facial scan. The final ethnicity inference came out by combining both results, and its reported accuracy draw near to 98%.

From the above discussion, though significant progress and satisfactory performances have been made in face perception, it can be noted that the test images in race categorization for evaluation of traditional methods are quite insufficient and the natural images are usually cropped to a small size, e.g., 36 × 36 [21] and 60 × 48 [22]. When these well-controlled assumptions do not hold, that is to say, massive face images are in real-world scenarios at meaningful scales, instead of attentively-designed face datasets full of clean and frontal face images, the performance of such descriptors and classifiers cannot be guaranteed. Accordingly, DCNN is likely to be a superior approach, which is more robust to image size and background interference.

It is worth noting that quite a few latest research [23], [24] turns to sub-divided ethnicity fields. For example, distinguish Koreans, Japanese, and Chinese in East Asia or among minorities whose features are not widely separated but closely related. In light of this, race categorization is facing greater challenges and more discriminative feature is imperative for efficient classification.

B. LOCAL BINARY PATTERN AND RELATED CONVOLUTION
The LBP operator [25], [26] is one of the best behaving and the most widely studied local texture descriptors. It describes the relations between a pixel and its surrounding pixels in the neighborhood. A typical LBP pattern is to threshold each pixel in the neighborhood, normally a set of circularly symmetric pixels, with the intensity value of the central pixel, and then treat the binary numerical sequence read clockwise with a predefined starting position as the characterized binary pattern. As a promotion, Li et al. [27], [28] proposed angular difference local binary pattern on a circular grid, where the binary label of each pixel is determined by the adjacent pixel in its clockwise direction. Then, the histogram of these binary labels can be applied as a texture descriptor.
Another brilliant extension of LBP is binary robust independent elementary features (BRIEF) [29], which is powerful in image matching tasks. As shown in Fig. 2 (a), it randomly selects pairwise graylevel pixels on images in a uniform or Gaussian distribution manner, to produce 128, 256, or 512-bit binary descriptors. This descriptor of random pairing inspires our work.

LBP operator has been acknowledged to be strikingly discriminative [25]. Its basic strengths, namely, invariance to monotonic changes in intensity level and high computational efficiency, facilitate it being powerful for challenging image cognition tasks and shed some light on solving extensive computer vision challenges. Although it was originally designed to encode the pixel-wise information in the textured images, LBP has been widely used and has broad applicability at current. Recently, LBP-based methods are becoming more popular in face-related tasks, such as face recognition [30], face detection [31], expression recognition [1], and race categorization [6], [32], as aforementioned handcrafted feature-based methods. By virtue of its low computational burden, Yang et al. [6] used LBP features for demographic research, covering ethnicity categorization. It principally employs the Chi-square distance metric on obtained features to produce a powerful classifier, achieving an error rate of 3.01% on a binary dataset with 11,680 Asian and 1,016 non-Asian faces. To this end, Zhang et al. [32] advanced LBP to multi-scale and multi-ratio version, and integrated the features extracted from 2D intensity face images and corresponding 3D range face images for classification, working out an error rate of 0.42% on the two leading ethnicity groups (i.e., Asia and White) from the FRGC v2 database.

Introducing the traditional texture descriptor into a learning-based method is a new trend in the computer vision community. Recently, on account of LBP and DCNN’s excellent performance on extracting object-level semantic feature and low-level local patterns respectively, Yu et al. [5] hybrid the best of two worlds to introduce CDC, which transfers the threshold mechanism in plain LBP into a different mode, i.e., subtract each pixel in the local region by the central intensity value, as stated in (2). In the meantime, it remains the learning-based framework of DCNNs to derive high-level information. Consequently, CDCN built by CDC realized faithful results on both intra-dataset and cross-dataset tests in face anti-spoofing tasks.

III. THE PROPOSED METHODS

A. RANDOM PIXEL DIFFERENCE CONVOLUTION

As facial cues sharing both holistic attributes and regional subtle variations, it is better to combine the best of features extracted by convolution and patterns detected by LBP. As shown in Fig. 3 and Fig. 2 (b), the random pixel difference convolution not only introduce gradient based on vanilla convolution, but also break the limit of fixed local pattern in the latest new-type convolution, e.g., CDC [5].

Taking a $3 \times 3$ convolution as an example, the vanilla convolution kernel propagates local patches in a weighted
summation manner simply, which would arguably smooth the local detailed cues.

\[ y_5 = \sum_{n \in [1,9]} w_n \cdot x_n \quad (1) \]

where \( w_n \) and \( x_n \) enumerate the values of weights in the convolution kernel and input pixels respectively, and \( y_5 \) stands for the value of current pixel location on the output feature map, namely the location of local patch’s central pixel.

With reference to CDC [5], the central pixel value will subtract from each pixel in the window before being aggregated in the same fashion of vanilla convolution. CDC can directly substitute vanilla convolution in a network without further adjustment. Considering to combine both semantic and gradient information which are crucial for vision tasks, its equation can be generalized as

\[ y_5 = \theta \cdot \sum_{n \in [1,9]} w_n \cdot (x_n - y_5) + (1 - \theta) \sum_{n \in [1,9]} w_n \cdot x_n \quad (2) \]

where \( w_n \) and \( x_n \) share the same depictions as in Eq. (1), and \( y_5 \) denotes the central value in the input local patch. Hyperparameter \( \theta \in [0, 1] \) balances the contribution between intensity and gradient information, and we treat it as a learnable value. The higher value of \( \theta \) represents that gradient features are more decisive.

In terms of Random Pixel Difference Convolution (RPDC), as demonstrated in Fig. 3(c) and Fig. 2 (b), for every pixel in the regional reception field, it can select any other pixel in the window to group a pair to subtract, in order to detect various difference information of various orientations in natural images. The selection of pixels to calculate the difference in an RPDC can be termed as difference pairing. This pairwise pixel difference follows the idea in the BRIEF descriptor [29], but not sampling in uniform or Gaussian distribution. Instead, RPDC adopts differences for every pixel in the window, for the reason that the kernel size in convolution is far less than the height or width in an image, thus, it is better to retain pixel information as much as possible. It is worth noting that, CDC (as shown in Fig. 3(b)) gives up an inward and centripetal gradient information, i.e., \( x_5 - x_i \in [1,9] \), it seems improper when calculating the output value of this exact position. In comparison, RPDC makes up for the lack of certain gradient information in a flexible way. Correspondingly, RPDC can also be in lieu of vanilla convolution in networks without bells and whistles. Hence, this difference pair setting remains in this convolution layer, channel, or other granularities. Further, S-RaPiDiNet can be equipped with searched RPDCs with different pair selections, instead of adopting an RPDC with fixed difference pairing from the very beginning to the end, as in [5]. In order to distinguish S-RaPiDiNets with different pairing settings in each convolution layer, we will mark the serial number of Difference Pairing Configurations (DPC) when needed. To clarify, RPDC can be formulated as below,

\[ y_5 = \theta \cdot \sum_{n \in [1,9]} w_n \cdot (x_n - x_{n,i} \in [1,9]) + (1 - \theta) \sum_{n \in [1,9]} w_n \cdot x_n \quad (3) \]

where \( w_n, x_n, y_5, \) and \( \theta \) share the same connotation as (1) and (2), \( x_{n,i} \in [1,9] \) represent the pixel value \( x_i \) randomly selected in the sampled window to pair with \( x_n \) for difference convolution.

To this end, compared with vanilla convolution, RPDC can pay adequate attention to rich local gradient information, while maintaining the capacity to extract high-level semantic features. In the group of modified convolutions, compared with LBConv [33], the proposed RPDC is learnable and data-driven, in contrast, LBConv utilizes static filters in the operation; on comparison with CDC, RPDC can expand the patterns of difference in small-scale regions exponentially, while remain decisive centripetal gradient information, promoting to detect various gradient information in images, which is accordance with the nature of subtle variation in face images.

Some may raise the question that RPDC-alike latest convolution methods may cause multiple copies of the feature maps, as the two terms in (2) and (3), and further double FLOPs and memory usage in both training and inference. Actually, we efficiently implement the operators to convert RPDCs to a common vanilla convolution layer by calculating the differences of the kernel weights initially. Accordingly, in comparison with vanilla convolution, the efficiency remains. Details of this tweak can be seen in (4).

\[ y_5 = \theta \cdot \left[ w_1 \cdot (x_1 - x_{1,i}) + \cdots + w_9 \cdot (x_9 - x_{9,i}) \right] + (1 - \theta) \cdot \left[ w_1 \cdot x_1 + \cdots + w_9 \cdot x_9 \right] = (w_1 \cdot x_1 + \cdots + w_9 \cdot x_9) + \theta \cdot \left[ -w_1 \cdot x_{1,i} - w_2 \cdot x_{2,i} - \cdots - w_9 \cdot x_{9,i} \right] = \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2 + \cdots + \hat{w}_9 \cdot x_9 = \sum_{i \in [1,9]} \hat{w}_i \cdot x_i \quad (4) \]

B. JOINT FACE BASELINE

The major motivation of baseline making lies on RC task, for the reason that the used BUPT-Xface+RFW is a recently-released large-scale face dataset, and existing methods rarely apply DCNN on this vision task, let alone unified experiment setting, hypotheses, scoring methods, ethnicity categories, and train-test set partition. Therefore, it is unpractical to report a direct comparison among algorithms, and the construction of a convolutional network as the baseline can be the best option at first. To verify the performance and efficiency of the newly designed convolution, we not only applied the classical Resnet18 [7] and advanced Inception-Resnet-V1 [34], but also learn characteristics from other face tasks [2] and construct a network named Sphere32 (as shown in Fig. 4), as the race classification baseline for the BUPT-Xface+RFW dataset. Sphere32 is enlightened by SphereFace [2], with two obvious strengths: in-network residual module and elaborately-designed loss function.

The widely-used residual module utilizes shortcut connections to link shallow and deep features, and relieves the risk of gradient exploding or vanishing in a relatively large neural network, which facilitates us to deepen the network by means
of stacking multiple residual modules once downsampling by convolution whose stride equals to 2.

Empirically, face-related computer vision tasks are marked by extremely small inter-class appearance variations and relatively large intra-class distance due to aging, makeup, and expression. Alternatively, the maximum intra-class distance of optimal face features should be smaller than the minimum inter-class distance. Thus, the motivation of SphereFace comes out naturally: it is necessary to enforce higher compactness for intra-class samples and discrepancy for inter-class samples as much as possible. Taking a multi-class classification task (C classes) as the example, where $W = [w_1, \ldots, w_{c-1}, w_c, w_{c+1}, \ldots, w_C]$. The decision surface of class $c$ and other class $i$ for traditional Softmax loss is $(W_c - W_i) x + b_c - b_i = 0$, where the margin between decision boundaries is zero, thus, the extracted features from different classes are presumably spread pretty close. Alternatively, Angular Softmax Loss [2] normalizes the network on LWF.

IV. EXPERIMENTS
To verify, we explore the performance of RPDC for various tasks covering race categorization (RC), facial expression recognition (FER), and face recognition (FR). In our experiment, the kernel size of RPDC is fixed to $3 \times 3$ on basis of the intrinsic configuration in the baseline model, that is to say, $1 \times 1$ convolution and convolution layers with other shapes remain in a vanilla manner. Moreover, the exact pixel pairing rules for convolution layers in the network are randomly sampled as candidate DPCs, then we grid search to tune them as hyperparameters, and we only illustrate the performance of the network in certain sampled difference pairing configurations in the result section.

A. DATABASES
To illustrate the superiority of the proposed method, we deliberately select widely recognized datasets for each task.

1) FACE RECOGNITION

a: TRAINING DATASET
Every model in the FR task is trained from scratch on CASIA-Webface, a broadly acknowledged dataset for face verification and face identification tasks in general. It contains 494,414 annotated face images of 10,575 unique individuals. In addition, we cropped and aligned photographs into a unified size $112 \times 96$ before network training.

b: TESTING DATASET
Labeled faces in Wild (LFW) [35] is a public dataset of face images for pair matching tasks. The production group collects 13,233 face images in total from 5,749 different celebrities on the Internet.

Cross-Age LFW (CALFW) [36] advances LFW by the selection of 3,000 positive face pairs varying in age, and negative pairs sharing the same gender and ethnicity to magnify intra-class variance in age. It sets a new challenge in the face recognition field. Generally, it will witness a slide of more than 10% in accuracy compared with that on LFW.

Cross-Pose (CPLFW) [37] follows the idea of CALFW to enlarge intra-class variance by varying an attribute in positive pairs and reduce inter-class variance by keeping the same race and gender in negative pairs, but the attribute switches from age to pose. Similarly, it will see a considerable decrease of around 20% in comparison with the performance of the network on LFW.
\textbf{2) FACIAL EXPRESSION RECOGNITION}

Facial Expression Recognition 2013 (FER-2013) \cite{39} contains 28,709 training images, 3,589 public test images, 3589 private test images. All face images are cropped to $48 \times 48$ and annotated by 7 different emotions, including angry, fear, happy, sad, surprise, neutral, and disgust.

Real-world Affective Faces Database (RAF-DB) \cite{40} is a large-scale facial expression database, containing 29,672 face photographs of great diversity collected from the Internet. The basic set used in our experiment is single-labeled by 7 emotions, including angry, disgusted, fearful, happy, sad, surprised, and neutral.

\textbf{3) RACE CATEGORIZATION}

All the images for training and testing are typically aligned and cropped at first by MTCNN \cite{41} into the size of $112 \times 112$.

BUPT-Xface and RFW (Racial Faces in wild) \cite{42} is currently the biggest race dataset in the world, to the best of our knowledge. A brief illustration of the main difficulties is shown in Fig. 5, which is adversarial for the immature race classification task even more. Generally, it has 4 labels, i.e., African, Asian, Caucasian, and Indian. As the training dataset, BUPT-Xface contains 4 subsets with different race distributions. On basis of the experimental properties, to balance the number of images among different categories, we randomly select 320,000 face images from Globalface and another 320,000 from Balancedface in BUPT-Xface, and the face images in evenly distributed to each ethnicity. As the result, we got 640,000 face images in total for training, which is still a large database. In terms of the test database, we adopted RFW, which provides roughly 3,000 subjects along with 10,000 face images for each ethnicity.

Color FERET dataset is another widely used \cite{19}, \cite{22} dataset for race classification. Before FERET was released, a great number of papers declared excellent recognition results (normally $>95\%$ accuracy) on databases on a limited scale (<50 individuals in general). It contains 11,333 facial images of 994 individuals, tagged with race labels including Asian, Asian-Middle-Eastern, White, Black-or-African-American, and others. To illustrate the superiority of the proposed method, we utilized 7805 images of 725 subjects as the training set, and the remaining 269 individuals' 3528 images for testing.

\textbf{B. FACE RECOGNITION}

\textbf{a: EXPERIMENT DETAILS}

Instead of substitute all the vanilla convolution with the proposed RPDC, we combine the RPDC into Sphere32 in a partially substitutive way: one only set the first convolution layer in the residual unit as RPDC (named 1PS), whereas the other allocate RPDC on the second in the residual unit (named 2PS). During training, Sphere20, Sphere32 (baseline), CDCN, and S-RaPiDiNet with different convolution settings are trained using 20 epochs with stochastic gradient descent (SGD) optimizer. The learning rate is set as 0.1 at the beginning and is divided by 10 at epoch 10, 15, and 18. The batch size is 256.

\textbf{b: RESULTS ANALYSIS}

As shown in Table 1, we witness a slight increase of S-RaPiDiNet in most cases. The accuracy growth can be extremely difficult when the performance is already high, thus, the modest promotion is meaningful. In detail, S-RaPiDiNet outperforms the baseline model with vanilla convolution with a gap of about 0.14\% on LFW; On CALFW, S-RaPiDiNet surpasses sphere20 by 0.72\% and exceeds baseline sphere32 with vanilla convolution by 0.54\%; On YTF, it outperforms sphere20 by a margin of 0.52\% and achieves 0.16\% precision improvement over baseline.

Besides, we notice that S-RaPiDiNet’s performance is on par with vanilla convolution on CPLFW, while CDCN already lagged behind. We conjecture that the pose variance shows in a broader scale manner on the face image. The improvement of RPDC compared with vanilla convolution is local feature extraction capacity in the small-scale regions, therefore, no obvious increment appears after substituting vanilla convolution with RPDC on CPLFW.

\textbf{C. FACIAL EXPRESSION RECOGNITION}

\textbf{Experiment Details:} We select VGG16 \cite{53} with batch-norm \cite{54} and Resnet18 \cite{7} as the baseline. For a comprehensive comparison, we not only train models from scratch on
facial expression datasets, but also transfer pretrained models to train for expression recognition. For vanilla convolution, it is feasible to obtain pretrained model weights of classic networks, but it can be prohibitively time-consuming for CDCN and S-RaPiDiNet to train on ImageNet [55]. Therefore, in the pre-training stage, we randomly select 100 categories out of 1,000 in the original ImageNet to form a small ImageNet and to train preferable feature extraction capacity rapidly. In detail, we choose 100 labels in completely random. The reason why we do not choose the existing Tiny ImageNet, Imagenet, downsamled ImageNet dataset or other public ImageNet-based small-scale datasets is that they are not sufficient for feature extraction training in face perception. For detail, Imagenet is a subset of only 10 classes from ImageNet; the resolution of images in downsampled ImageNet is resized to 8 × 8, 16 × 16, 32 × 32 or 64 × 64, which is adversarial for face perception in real-world scenario and wild setting; while Tiny ImageNet only has 500, 50 and 50 images for training, validation and test respectively.

In the small ImageNet pre-training stage, we train models with a maximum of 90 epochs while the learning rate (lr) is divided by 10 every 30 epochs. The initial lr, weight decay (wd), and batch size are 0.01, 1e-4, and 256, respectively.

Results Analysis: In the experiment on RFA-DB, we also include the evaluation indicator, true positive rate (TPR, also known as sensitivity), and true negative rate (TNR, also known as specificity) in this multi-class classification task. TPR is defined as the fraction of all images in a certain emotion that were correctly identified as that emotion, and TNR tells the fraction of all images in other emotions that were correctly predicted to not be in that emotion.

\[
\text{TPR} = \frac{TP}{TP + FN}, \quad \text{TNR} = \frac{TN}{TN + FP}
\]

where TP stands for the number of images that are correctly identified as a certain emotion, TN represents the number of images that are correctly detected as other emotions, FN is the number of images in that emotion, but marked as other emotions, FP denotes the number of images that are falsely detected as that emotion. The recorded TPR and TNR performance in Table 3 is averaged on each emotion without bias on the weights.

In the unpretrained setting, S-RaPiDiNet outperforms baseline and CDCN in different networks, different datasets, and different assessment metrics. The accuracy promotion of RPDC on Resnet18 is at least 1%. In terms of VGG16_bn, the accuracy growth of RPDC reaches around 0.5% on FER2013. The improvement in accuracy turns more evident when the S-RaPiDiNet is only pre-trained on small ImageNet.

Given the above, the improved margin of S-RaPiDiNet in unpretrained setting is more obvious than that in a pretrained situation. Far as we are concerned, RPDC can efficiently make up the scarce semantic information extraction capacity of a plain model by its intrinsic pixel difference. Accordingly, RPDC can particularly be a striking method when there are no pretrained weights and needing to train from scratch on specific tasks.

### D. RACE CATEGORIZATION

1) EXPERIMENTAL DETAILS

Every model is trained from scratch for 20 epochs without any pretraining, except for Sphere32 on the FERET dataset. Due to its relatively data-hungry structure and a small number

| Baseline* | Resnet18 Public | Resnet18 Private | VGG16_bn Public | VGG16_bn Private |
|-----------|----------------|-----------------|----------------|-----------------|
| CDNC*     | 58.96%         | 59.49%          | 60.30%         | 60.49%          |
| S-RaPiDiNet* | 53.16%     | 52.46%          | 48.45%         | 48.87%          |
| S-RaPiDiNet-2PS* | 60.35%      | 60.55%          | 60.99%         | 60.80%          |

| Baseline† | Resnet18 Public | Resnet18 Private | VGG16_bn Public | VGG16_bn Private |
|-----------|----------------|-----------------|----------------|-----------------|
| CDNC†     | 61.80%         | 62.13%          | 64.78%         | 64.28%          |
| S-RaPiDiNet† | 59.01%     | 58.43%          | 62.66%         | 62.52%          |
| S-RaPiDiNet-2PS† | 61.82%      | 62.13%          | 65.31%         | 65.22%          |

TABLE 2. Facial expression recognition accuracy on FER2013 (in percentage). †PS denotes for RPDC only partially substitute the second convolution layer in residual units (Resnet18) or convolution blocks (VGG16_bn). ‡ denotes that the models are trained from scratch on FER2013, † means the models are initially pretrained on the customized small ImageNet derived from ImageNet and then finetuned on FER2013, and ‡ represents the model is pretrained on ImageNet at first. The bold represent the best results.

| Baseline‡ | Resnet18 Public | Resnet18 Private | VGG16_bn Public | VGG16_bn Private |
|-----------|----------------|-----------------|----------------|-----------------|
| CDNC‡     | 68.38%         | 78.88%          | 97.57%         | 97.31%          |
| S-RaPiDiNet‡ | 85.59%     | 76.08%          | 97.20%         | 96.76%          |
| S-RaPiDiNet-2PS‡ | 85.98%      | 76.91%          | 97.30%         | 96.90%          |

| Baseline‡ | Resnet18 Public | Resnet18 Private | VGG16_bn Public | VGG16_bn Private |
|-----------|----------------|-----------------|----------------|-----------------|
| CDNC‡     | 68.38%         | 78.88%          | 97.57%         | 97.31%          |
| S-RaPiDiNet‡ | 85.59%     | 76.08%          | 97.20%         | 96.76%          |
| S-RaPiDiNet-2PS‡ | 85.98%      | 76.91%          | 97.30%         | 96.90%          |

TABLE 3. Facial expression recognition accuracy on RAF (in percentage). †PS denotes for RPDC only partially substitute the second convolution layer in residual units (Resnet18) or convolution blocks (VGG16_bn). ‡ denotes that the models are trained from scratch on RAF-DB, † means the models are initially pretrained on the customized small ImageNet derived from ImageNet and then finetuned on RAF-DB, and ‡ represents the model is pretrained on ImageNet at first. The bold represent the best results, while the second best is underlined.

| Acc | TPR | TNR | Acc | TPR | TNR |
|-----|-----|-----|-----|-----|-----|
| Baseline* | 58.41% | 69.91% | 96.45% | 80.67% | 69.19% | 96.29% |
| CDNC* | 73.44% | 59.20% | 94.70% | 83.38% | 74.88% | 97.02% |
| S-RaPiDiNet* | 80.02% | 70.03% | 96.32% | 84.06% | 73.90% | 97.13% |
| S-RaPiDiNet-2PS* | 81.49% | 69.50% | 96.51% | 84.68% | 76.00% | 97.39% |

| Acc | TPR | TNR | Acc | TPR | TNR |
|-----|-----|-----|-----|-----|-----|
| Baseline‡ | 86.38% | 78.88% | 97.57% | 85.23% | 77.17% | 97.31% |
| CDNC‡ | 85.59% | 76.08% | 97.20% | 87.03% | 76.92% | 97.66% |
| S-RaPiDiNet‡ | 83.60% | 73.10% | 97.09% | 86.83% | 77.99% | 97.52% |
| S-RaPiDiNet-2PS‡ | 85.98% | 76.91% | 97.30% | 86.96% | 78.14% | 97.53% |
of images in FERET, we transfer the weight of Sphere32 (with the same convolution type) pretrained on BUPT-Xface as the initialization scheme. Generally, we use an Adam optimizer [56] with batch size 128. During training, the learning rate is divided by 10 at epoch 10, 15, 18, and weight decay is set to 5e-4 for all. The initial learning rate for resnet18 and IR-V1 is 5e-3, whereas that for Sphere32a is 1e-4.

2) RESULTS ANALYSIS

As for the race categorization task, the performance of S-RaPiDiNet and relative comparison is shown in Table 4, where the best result for each network and experiment dataset is in bold. As mentioned before, since previous works reported in the literature are measured on small sets or their own secret set, it is impossible to draw meaningful comparisons with other methods. Thus, the table only reveals experiments on our own, but still in credibility. It is clear that the proposed S-RaPiDiNet outperforms baseline and CDCN by 0.25%-3% in every experimental setting, reflecting the efficiency and robustness of RPDC not only in different networks but also on different datasets.

E. ABSTRACTION STUDY

As θ is the learnable parameter in both CDCN and S-RaPiDiNet, to elaborate on the effectiveness of RPDC and to find the possibly optimal parameter θ that balances the contribution between intensity and gradient information, we conduct ablation study on FR (Fig. 6), FER (Fig. 7) and RC (Fig. 8) respectively. Once the grid searched pixel pairing configuration is confirmed for each S-RaPiDiNet with θ = 1.0, we sample the parameter θ in S-RaPiDiNet progressively between [0.1, 0.9] in a step of 0.2 in order to finetune and find the best one by grid search. For a fair comparison, we adjust the θ in CDCN the same way as S-RaPiDiNet.

In Fig. 6, 7 and 8, blue lines represent RPDC, red lines stand for CDC, and black lines denote vanilla convolution as the baseline for comparison. Different linetypes (i.e., solid, dashed, and dotted) mean experiments on different datasets or networks. In most cases, RPDC surpasses CDC and vanilla convolution by a clear margin. Moreover, according to the observations, the value of θ for S-RaPiDiNet to reach the peak performance (red flags placed) is usually larger than the optimal value for CDCN. As an illustration, in Fig. 8, S-RaPiDiNet performs best when θ is around 0.9, whereas the optimal value for CDCN is 0.3. In light of this, we can draw a solid conclusion that S-RaPiDiNet extracts more discriminative features than CDCN, because only when CDCN hybrids more information from vanilla convolution (i.e., intensity-level cues), CDCN can realize its best efficiency. Particularly, for FR in Fig. 6, we deem that performance cannot be promoted intensely and fluctuation will certainly appear due to the high performance already. Moreover, when observing Resnet18 in FER (solid lines in Fig. 7) and Sphere32 in RC (dotted lines in Fig. 8), both CDC and RPDC tend to be adversarial for feature extraction, but when θ corrected to be a larger and appropriate value (1.0 and 0.9 respectively), S-RaPiDiNets realize global optimum, indicating the significance of θ, and both vitality and strong dynamic adjustment ability of S-RaPiDiNet, especially when compared with CDCN.

V. PERSPECTIVE

Initially, the granularity of random pixel difference convolution in this paper is layer-wise, it can no doubt go down into sharing richer different convolution patterns among different channels or convolution kernels by means of neural architecture search or advanced computing algorithm. Besides, advanced research on the directions and orientations of pixel difference pairing may further benefit computer vision. Additionally, in spite of the promising performance of RPDC, the selection among vanilla convolution, central difference convolution and random pixel difference convolutions of various pairing modes can also be realized by neural architecture search.
In terms of race categorization, it is a basic biometric technology with numerous potential cognitive applications related to face processing and recognition. There exists decisive cognitive evidence suggesting that humans employ information from various visual cues for face recognition, and one of the main difficulties of face recognition lies in race
bias. As everybody knows, people tend to recognize faces of their own race better and easier than faces of other races. Once a better race perception is achieved, it can provide useful clues for face recognition and enhance performance. In spite of the above, it is still underestimated and relative research is limited. In the first place, the majority of face databases were insufficient in data size for learning-based methods and were collected under very controlled situations, such as frontal or full profile images. Apart from this, it could not be better if any authoritative work can determine unified datasets, image division, and experiment settings, thus, the following researchers can be interpreted from the perspective of previous studies and of the working assumptions. In the meantime, findings and implications in RC can be discussed as much as possible, and thus, future research directions may also be highlighted and more attractive.

VI. CONCLUSION

In this paper, our work looked into a wide range of face perception tasks, which are both canonical and vital in modern society. We introduced a novel random pixel difference convolution, which is expert in extracting various local spatial difference features. Thus, we combined it with both classical networks and specially-designed networks. As illustrated, our method is proved to be efficient and highly robust in different networks, tasks, and datasets, including complex scenes, such as occlusion and background clutter.

What is noteworthy is that there are some difficulties for face perception: (1) Paucity of normative evaluation protocol: Different quantitative indicators and labels in datasets are selected by different researchers in experiments. It raises the barriers for a broader range of potential researchers and increases the difficulty of comparisons of algorithms on equal grounds. (2) Label imbalance in face datasets. The distribution of some attributes (e.g., race, emotion, and age) in datasets is the same as that in reality, due to a lack of artificial control in data collection. Because face perception is well-known for relatively small inter-class distance and large intra-class distance, uneven distribution can easily prevent promising network learning and real-world applications.

For future work, we are confident that RPDC can be capable in other computer vision tasks, including not only other facial tasks (e.g., face anti-spoofing, which is proved to get improved by CDC) but also tasks with various local patterns (e.g., edge detection).

REFERENCES

[1] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” Image Vis. Comput., vol. 27, no. 6, pp. 803–816, 2009.
[2] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 212–220.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NIPS, 2012, pp. 1097–1105.
[4] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, May 2015.
[5] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, “Searching central difference convolutional networks for face anti-spoofing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 5295–5305.
[6] Z. Yang and H. Ai, “Demographic classification with local binary patterns,” in Proc. ICB, Berlin, Germany: Springer, 2007, pp. 464–473.
[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[8] J. G. Cazavos, E. Noyes, and A. J. O’Toole, “Learning context and the other-race effect: Strategies for improving face recognition,” Vis. Res., vol. 157, pp. 169–183, Apr. 2019.
[9] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, “Uncovering and mitigating algorithmic bias through learned latent structure,” in Proc. AAAI/ACM Conf. AI, Ethics, Soc., Jan. 2019, pp. 289–295.
[10] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O’Toole, “An other-race effect for face recognition algorithms,” ACM Trans. Appl. Perception, vol. 8, no. 2, pp. 1–11, Jan. 2011.
[11] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, “Face recognition performance: Role of demographic information,” IEEE Trans. Inf. Forensics Security, vol. 7, no. 6, pp. 1789–1801, Dec. 2012.
[12] F. S. Manesh, M. Gharabrami, and Y. P. Tan, “Facial part displacement effect on template-based gender and ethnicity classification,” in Proc. 11th Int. Conf. Control Automat. Robot. Vis., Dec. 2010, pp. 1644–1649.
[13] S. H. Salah, H. Du, and N. Al-Jawad, “Fusing local binary patterns with wavelet features for ethnicity identification,” in Proc. World Acad. Sci., Eng. Technol., vol. 79, 2013, p. 471.
[14] R. Brunelli and T. Poggio, “Face recognition: Features versus templates,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 10, pp. 1042–1052, Oct. 1993.
[15] Y. Bouzouina and L. Hamami, “Multimodal biometric: Iris and face recognition based on feature selection of iris with GA and scores level fusion with SVM,” in Proc. 2nd Int. Conf. Bio-Eng. Smart Technol. (BioSMART), Aug. 2017, pp. 1–7.
[16] J. R. Lyle, P. E. Miller, S. J. Pundlik, and D. L. Woodard, “Soft biometric classification using periodic and regional feature regions,” in Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS), Sep. 2010, pp. 1–7.
[17] B. Zhang, Y. Gao, S. Zhao, and J. Liu, “Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor,” IEEE Trans. Image Process., vol. 19, no. 2, pp. 533–544, Feb. 2010.
[18] S. Fu, H. He, and Z.-G. Hou, “Learning race from face: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 12, pp. 2483–2509, Dec. 2014.
[19] S. Gutta, H. Wechsler, and P. J. Phillips, “Gender and ethnic classification of face images,” in Proc. 3rd FG, 1998, pp. 194–199.
[20] S. Hosoi, E. Takikawa, and M. Kawade, “Ethnicity estimation with facial images,” in Proc. 6th IEEE Int. Conf. Automat. Face Recognit., May 2004, pp. 195–200.
[21] X. Lu, H. Chen, and A. K. Jain, “Multimodal facial gender and ethnicity identification,” in Proc. ICB, Berlin, Germany: Springer, 2006, pp. 554–561.
[22] G. Muhammad, M. Hussain, F. Alenezy, G. Bebis, A. M. Mirza, and H. Aboalsamh, “Race classification from face images using local descriptors,” Int. J. Artif. Intell. Tools, vol. 21, no. 5, Oct. 2012, Art. no. 1250019.
[23] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, “The CAS-PEAL large-scale Chinese face database and baseline evaluations,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 38, no. 1, pp. 149–161, Jan. 2008.
[24] J. Muhammad, W. Wang, C. Wang, K. Zhang, and Z. Sun, “CASIA-faceAfrica: A large-scale African face image database,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 3634–3646, 2021.
[25] T. Ojala, M. Pietikäinen, and T. Maenpää, “Multiresolution grayscale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.
[26] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, “Local binary features for texture classification: Taxonomy and experimental study,” Pattern Recognit., vol. 62, pp. 135–160, Feb. 2017.
[27] L. Liu, L. Zhao, Y. Long, G. Kang, and P. Fieguth, “Extended local binary patterns for texture classification,” Image Vis. Comput., vol. 30, no. 2, pp. 86–99, Feb. 2012.
[28] Z. Su, M. Pietikäinen, and L. Liu, “BIIRD: Learning binary and illumination robust descriptor for face recognition,” in Proc. BMVC, 2019, p. 102.

[29] M. Calonder, V. Lepetit, M. Ozzyal, T. Trzcinski, C. Strecha, and P. Fua, “BRIEF: Computing a local binary descriptor very fast,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, pp. 1281–1298, Jul. 2012.

[30] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face description with local binary patterns: Application to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[31] J.-C. Chen, J. Wang, L.-P. Zhao, and J. He, “Branch-structured detector for fast face detection using asymmetric LBP features,” Signal, Image Video Process., vol. 14, no. 8, pp. 1699–1706, Nov. 2020.

[32] G. Zhang and Y. Wang, “Multimodal 2D and 3D facial ethnicity classification,” in Proc. 5th Int. Conf. Image Graph., Sep. 2009, pp. 928–932.

[33] F. Juefei-Xu, V. N. Boddeti, and M. Savvides, “Local binary convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 19–28.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in Proc. AAAI Conf. Artif. Intell., 2017, vol. 31, no. 1, pp. 1–7.

[35] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[36] T. Zheng, W. Deng, and J. Hu, “Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, p. 1708.8197.

[37] T. Zheng and W. Deng, “Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments,” Beijing Univ. Posts Telecommun., Beijing, China, Tech. Rep. 18-01, Feb. 2018.

[38] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” in Proc. CVPR, Jun. 2011, pp. 526–533.

[39] Challenges in Representation Learning: Facial Expression Recognition Challenge. Accessed: Oct. 11, 2021. [Online]. Available: https://www.kaggle.com/c/challenges-in-representation-learning/facial-expression-recognition-challenge/

[40] S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 356–370, Jan. 2019.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multilayer cascaded convolutional networks,” IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[42] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, “Racial faces in the wild: Reducing racial bias by information maximization adaptation network,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 692–702.

[43] I. K. Beli and C. Guo, “Enhancing face identification using local binary patterns and K-nearest neighbors,” J. Image Anal., vol. 3, no. 3, 2017, pp. 37, Sep. 2017.

[44] O. A. Arigbabu, S. M. S. Ahmad, W. A. W. Adnan, S. Yussof, and S. Mahmood, “Soft biometrics: Gender recognition from unconstrained face images using local feature descriptor,” 2017, arXiv:1702.02537. [Online]. Available: http://arxiv.org/abs/1702.02537.

[45] L. Lenc and P. Král, “Automatic face recognition system based on the SIFT features,” Comput. Electr. Eng., vol. 46, pp. 256–272, Aug. 2015.

[46] G. Du, F. Su, and A. Cai, “Face recognition using SURF features,” Proc. SPIE, vol. 7496, Oct. 2009, Art. no. 749628.

[47] Z.-H. Huang, W.-J. Li, J. Shang, J. Wang, and T. Zhang, “Non-uniform patch based face recognition via 2D-DWT,” Image Vis. Comput., vol. 37, pp. 12–19, May 2015.

[48] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1701–1708.

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

[50] Y. Sun, Deep Learning Face Representation by Joint Identification-Verification. Hong Kong: The Chinese Univ. Hong Kong, 2015.

[51] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in Proc. Brit. Mach. Vis. Conf., vol. 1, no. 3, 2015, p. 6.

[52] X. Wu, R. He, Z. Sun, and T. Tan, “A light CNN for deep face representation with noisy labels,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[53] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556.

[54] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. IML, 2015, pp. 448–456.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[56] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 2017, arXiv:1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980