MC-GCN: A Multi-Scale Contrastive Graph Convolutional Network for Unconstrained Face Recognition With Image Sets

Xiao Shi, Xiujuan Chai, *Member, IEEE*, Jiake Xie, and Tan Sun

***Abstract*— In this paper, a Multi-scale Contrastive Graph Convolutional Network (MC-GCN) method is proposed for unconstrained face recognition with image sets, which takes a set of media (orderless images and videos) as a face subject instead of single media (an image or video). Due to factors such as illumination, posture, media source, etc., there are huge intra-set variances in a face set, and the importance of different face prototypes varies considerably. How to model the attention mechanism according to the relationship between prototypes or images in a set is the main content of this paper. In this work, we formulate a framework based on graph convolutional network (GCN), which considers face prototypes as nodes to build relations. Specifically, we first present a multi-scale graph module to learn the relationship between prototypes at multiple scales. Moreover, a Contrastive Graph Convolutional (CGC) block is introduced to build attention control model, which focuses on those frames with similar prototypes (contrastive information) between pair of sets instead of simply evaluating the frame quality. The experiments on IJB-A, YouTube Face, and an animal face dataset clearly demonstrate that our proposed MC-GCN outperforms the state-of-the-art methods significantly.**

***Index Terms*— Face recognition, image set, graph convolutional, contrastive information, multi-scale.**

I. Introduction

RECENTLY, the performance of face recognition has been remarkably boosted owing to the advances in deep learning. As many methods have been reported to surpass human performance in single image face recognition [1]–[6], set-based unconstrained face recognition is now gaining more and more attention. A face image set is collected from different segments in videos or images, which suffer a huge intra-set variance due to heterogeneous factors (illumination, posture, media source, etc.). This task is more similar to the real-world biometric scenarios and more challenging compared to single image-based face identification [7].

Many approaches have been proposed to solve this task recently [8]–[14]. An image set-based FR model consists of two important parts: image-level feature extractor and set descriptors aggregation. The former focuses on high-quality face feature description and aims to reduce variations. The latter tries to embed a set of face descriptors to produce a compact representation for template-based face recognition. In this paper, we seek to achieve a more reasonable and explainable method to solve the latter. Average/max pooling is the common aggregation method [8], [9], [15], which handles every frame equally. Since each frame in a set has different importance for identification, many researchers try to perform weighted aggregation based on the quality of the image to reduce variations recently [10]–[14]. However, these methods actually suffer some limitations. First, if there are high-quality elements in both sets for verification, it will be beneficial for verification to perform the high-quality elements more important. As shown in Fig. 1, set A contains high-quality and low-quality images, but set B contains only a low-quality image. In this case, assigning high-quality images with high weights and low-quality images with low weights will destroy the performance of verification. Second, these methods consider the image-level relationship to build attention mechanism, but they ignore the correlation cues between face prototypes. A face prototype represents a subset of similar frame quality and face attributes, which further enriches the semantic representation of the set.

Inspired by this observation, we cast set descriptors aggregation as a node weight prediction problem based on a graph convolutional network (GCN) [16]. Each pair of sets is treated as a graph, and each prototype is treated as a node. Moreover, we propose a novel Multi-scale Contrastive Graph Convolutional Network (MC-GCN) model to overcome the above limitations. This method advocates both high-quality and contrastive images between sets instead of only focusing on quality. MC-GCN introduces a contrastive graph convolutional (CGC) block to leverage both contrastive information and global relationship information. Furthermore, a graph-based multi-scale method was designed to explore richer prototype correlation cues by fusing prototype graphs with multiple levels of fineness.

Manuscript received August 6, 2021; revised March 14, 2022; accepted March 22, 2022. Date of publication April 6, 2022; date of current version April 12, 2022. This work was supported in part by the National Science Foundation of China under Grant 61976219; in part by the Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences under Grant CAAS-ASTIP-2016-AII; and in part by the Fundamental Research Funds for Central Non-Profit Scientific Institution under Grant Y2021LM02, Grant 01020100102036, and Grant JBYW-AII-2021-34. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding author: Xiujuan Chai.)*

Xiao Shi, Xiujuan Chai, and Tan Sun are with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China, and also with the Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China (e-mail: chaixiujuan@caas.cn).

Jiake Xie is with Hangzhou Wangdao Holdings Company Ltd., Hangzhou 310051, China.

Digital Object Identifier 10.1109/TIP.2022.3163851

Fig. 1. An extreme sample in IJB-A. Set A contains high-quality and low-quality images, but set B contains only low-quality image. In this case, it will be better to focus on contrastive images than to simply build quality-based attention model.

The proposed method effectively improves the relationship representation and offers more discriminative and contrastive compact set representations. We evaluate MC-GCN with a general image-level feature extractor VGGFace2 [9] on IJB-A [7], YTF [17] and a novel pig face dataset [18]. MC-GCN achieves a significant improvement in those datasets. The proposed method considerably solves these extreme samples and enhances the weight of those images with similar prototypes (contrastive information) between pair of sets compared to exisiting quality-based aggregation methods [10]–[12]. The main contributions can be summarized as follows:

1) We present a novel GCN-based set descriptors aggregation model to fully exploit the relationship information in face set. By treating face prototypes as nodes in graph, we are able to adaptively incorporate multi-level semantic relationship to model attention mechanism;
2) MC-GCN, a new node weight prediction method is proposed to build the attention control model for set descriptors aggregation. MC-GCN introduces two main modules: contrastive graph convolutional (CGC) block and multi-scale set graph. CGC block advocates high-quality images while further activating the contrastive information between pair of sets. The multi-scale set graph further exploits the rich correlation cues between prototypes at various fineness;
3) MC-GCN achieves state-of-the-art performance on two popular video-based human face recognition datasets. Moreover, we further evaluate MC-GCN on a novel set-based pig face recognition dataset, which achieves far improvement compared previous method and shows its good generalization ability.

II. Related Works

A. Set-Based Face Recognition

Set/video-based face recognition task has been studied for many years. In early approaches, an image set is usually represented as a manifold [19]–[25], the distance is calculated in manifold space. It is difficult for these methods to handle the large variations in the unconstrained FR task.

With the introduction of the IJB-A benchmark by NIST in 2015 [7] and the success of deep learning in image-based face recognition [1]–[4], the problem of unconstrained set-based face recognition attracts more and more attention. This type of method usually consists of two important parts: image-level feature extractor and set descriptors aggregation. Many methods are proposed to learn better image-level features by noticing the noise and variations [3], [4], [13], [26].

Another part is also the task that this article solves, how to better aggregate the image-level features into a discriminative compact representation. Average/max pooling is the common aggregation method [8], [9], [15], which handles each frame equally but ignores the huge inner variances. Literature [10] proposes an attention model to aggregate a set of features by an independent quality assessment module. Zhong *et al.* [11] apply Vector of Local Aggregated Descriptors (VLAD) with ghost clusters to better learn the impact of image quality. The Reinforcement Learning (RL) method learns the inner-set relationship and builds a quality evaluation model [12]. Liu *et al.* [14] use a redundancy-eliminating self-attention to emphasize important samples. Although these works take good account of correlation and image quality, they are hard to work on those sample pairs with a large distributional difference (for example Fig. 1).

In this work, we consider the weighted aggregation in a different way, where we leverage both contrastive information and global relationship information to build attention mechanism. The contrastive information will help the model advocate those images with similar prototypes between set pairs instead of simply upholding the high-quality images.

B. Graph Convolutional Networks

Graph Convolutional Networks (GCNs) [16] is introduced to process non-Euclidean structures. Since the strong capability of modeling complex graphical patterns, GCN has led to considerable performance improvement in computer vision tasks. For example, Kipf and Welling [16] apply the GCNs to semi-supervised classification. Literatures [27], [28] prove the effectiveness of GCNs in Human Pose Regression. Wang *et al.* [29] use the edge convolution method to process point clouds. In recent years, some GCN-based architectures are designed to learn video and large-scale sets [30], [31]. G-TAD [30] detects actions in video with a deep GCN model. Reference [31] adopts GCN to formulate a pipeline similar to the Mask R-CNN [32] for node segmentation.

In this paper, we adopt deep GCN as the basic machinery to capture the contrastive information and correlation cues between pair of sets.

III. Proposed Method

An overview of our method is illustrated in Fig. 2. It takes a pair of face media sets as input and outputs a matching result for the unconstrained set-based face verification. VGGFace2 [9] is adopted as the image-level feature extractor to embed every image into a latent space. We generate prototypes from each set at multi-finenesses and build them into a multi-scale prototype graph. Then the graph pairs for each scale will be fed into the Contrastive GCN consisting of multiple CGC blocks. The CGC blocks encode the correlation and contrastive information by merging and generating two

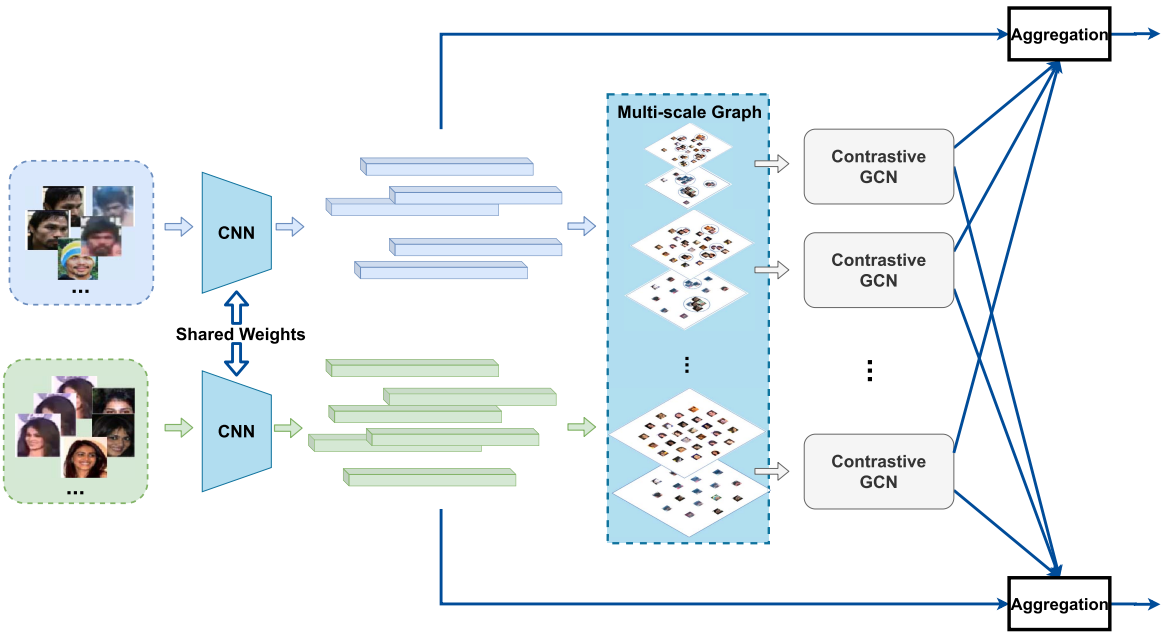


Fig. 2. Overview of Multi-scale Contrastive Graph Convolutional Network (MC-GCN) architecture. Given a pair of face media sets, a common face feature extractor is adopted for image-level feature representations. Each two feature sets are generated to multi-scale graph based on prototype finenesses. Then a contrastive GCN consisting of several CGC blocks to build attention control model that activates the contrastive samples.

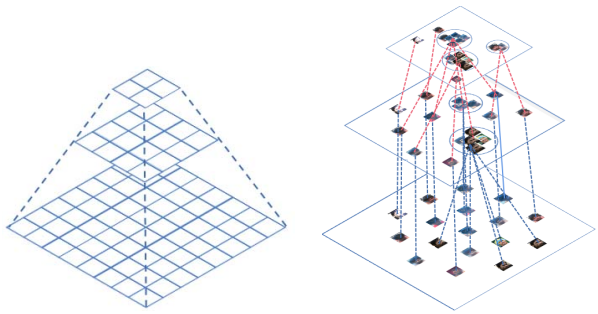
semantic graphs. Finally, the regression outputs of all nodes will be re-fused to build attention mechanism.

A. Multi-Scale Prototype Graph

Given a pair of face sets (A,B), we extract the feature for each image with a trained CNN, forming two sets of features $\mathcal{D}_A = \{f_i\}_{i=1}^{N}$ and $\mathcal{D}_B = \{f_j\}_{j=1}^{M}$, where f_i and f_j are d-dimensional vectors. To construct the basic affinity graph, we regard each image as a vertex, and use cosine similarity to find K nearest neighbors for each sample. By connecting between neighbors whith a edge threshold e_τ, two affinity graphs $\mathcal{G}_A = (v_A, \varepsilon_A)$ and $\mathcal{G}_B = (v_B, \varepsilon_B)$ are obtained.

Instead of processing the correlation at the image-level, we hope to build a prototype-based relationship model to fully consider the variance factors. Each prototype is representative of the subject face under a certain condition in terms of pose, illumination and media modality, which is a sub-set of a face set. Previous methods based on prototypes or clusters usually have a certain fineness [11], [13]. In this paper, multiple fine-grained prototypes are considered simultaneously to describe richer facial semantic relations. So, we propose a multi-scale face prototypes graph based on the basic image affinity graph.

Multi-scale is a popular method to expand the diversity of data scale in computer vision tasks, which describes richer semantic information representation compared to particular resolution [33]. We transfer the method of multi-scale from image to graph. As shown in Fig. 3, the low-level adjacent prototypes are aggregated to a higher-level prototype to achieve prototype graph downsampling. Specifically, We generate the prototype by detecting a sub-graph containing various vertices that are closely connected to each other. To achieve different finenesses prototypes, we gradually enhance the connectivity



(a) Ms for image (b) Ms for prototypes graph

Fig. 3. Schematic diagram of Multi-scale for image and face graph. In b, the high-level face prototype is adaptively aggregated from low-level prototypes with similar properties.

of the subgraphs by reducing the threshold e_τ. Specifically, with the decrease of edge threshold e_τ, more vertices will be linked and the connected subgraphs become larger. Alg. 1 shows the detailed procedure to produce a Multi-scale prototype graph in $\mathcal{G}_A$ and $\mathcal{G}_B$. We consider the basic image affine graph as the bottom graph, where each prototype vertice contains only one image. At the higher level, more intensive prototypes are used to build a deeper semantic relationship.

The pair of basic affinity graphs $\mathcal{G}_A$ and $\mathcal{G}_B$ are generated to multiple pairs of prototype graphs $\{(\mathcal{P}_A^n, \mathcal{P}_B^n)|n \in 1, 2, \ldots, N\}$ with the different semantic relationship, where N is the number of scales.

B. Contrastive Graph Convolutional

In this paper, we formulate a novel Contrastive Graph Convolutional (CGC) block, which encodes prototypes using both their global and contrastive relationship by building

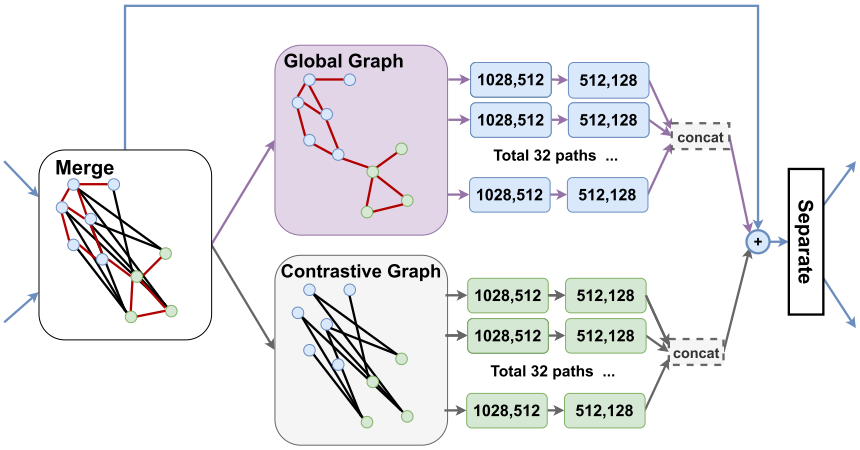


Fig. 4. The architecture of CGC block. Two types of semantic features are extracted separately from the Global and Contrastive graphs, which are constructed from the input features. The feature extractor consists of 32 paths to increase the diversity of transformations. Finally, the two semantic features and input are summed and transported to the next stage.

Algorithm 1 Multi-Scale Prototype Graph Generation

Require: Affinity graphs $\mathcal{G}$; Initial edge threshold e_τ; Threshold step Δ; Number of scales N;
Ensure: Prototype Graph sets $\mathcal{P}$;
1: $\mathcal{P} = \emptyset$, $i = 0$;
2: $\mathcal{P} = \mathcal{P} \cup \{\mathcal{G}\}$;
3: **while** $n < N - 1$ **do**
4: $\mathcal{G}' =$ PRUNEEDGE($\mathcal{G}$,e_τ);
5: $\mathcal{C} =$ FINDCONNECTEDCOMPONENTS($\mathcal{G}'$);
6: $\mathcal{S} = \{f_c | c \in \mathcal{C}\}$, where f_c is the average feature of the vertices in c.;
7: $\mathcal{P} = \mathcal{P} \cup \{\mathcal{S}\}$;
8: $e_\tau = e_\tau - \Delta$;
9: $n = n + 1$;
10: **end while**
11: **return** $\mathcal{P}$;

two semantic graphs. The global relationship will exploit the variance among prototypes and build a strong quality evaluation model. The contrastive information forces model to focus on those images with similar prototypes (contrastive information) between pair of sets instead of simply evaluating the quality. The CGC block architecture is illustrated in Fig. 4.

As methioned in section III-A, multiple pairs of prototype graphs are generated to describe different semantic relationship information. For each pair of prototypes graphs nodes v_A^n and v_B^n, we merge them to two new graphs — Global Graph $\mathcal{G}_o = (v, \varepsilon_o)$ and Contrastive Graph $\mathcal{G}_c = (v, \varepsilon_c)$, where $v = v_A^n \cup v_B^n$. In this section, we first introduce the global graph and contrastive graph, and then design a deep GCN structures to embed and aggreagte two types of features.

1) Global Graph: In order to describe the global quality correlation between prototypes, we rebuild affinity graph on v. Specifically, we use cosine similarity to find K nearest neighbors for each prototype

$$s_{ij} = \frac{v_i \cdot v_j}{|v_i| \times |v_j|} \tag{1}$$

By connecting between neighbors, we get correlation matrix A_o, where elements $a_{ij} = s_{ij}$ if the v_i and v_j are connected, or zero otherwise. Then a threshold e_τ^o is used to maintain the high connectivity within each prototypes

$$a_{ij} = \begin{cases} s_{ij}, & \text{if } s_{ij} \geq e_\tau^o \quad and \quad (v_i, v_j) \in \varepsilon_o \\ 0, & \text{else} \end{cases} \tag{2}$$

where the ε_o is the global edge builded by the K nearest neighbors.

The global graph provides rich quality correlation cues to exploit the quality impact. The correlation information in the global graph depends on visual features, which only consider the importance of prototypes for the set representation. Therefore, the global graph concerns the amount of information contained in the face images, which explores the quality correlation cues through graph convolution. In the global graph, the quality of the image is positively correlated with its amount of information. Meanwhile, the global graph not only builds more strong quality assessment but also provides some contrastive features due to the connectivity between the sub-graphs with different identities.

2) Contrastive Graph: To further advocate those contrastive prototypes between pairs, we design a contrastive graph, which defined form the pairwise node connected between two graphs. We define the contrastive edge ε_c and the correlation matrix value a_{ij} as follow:

$$\begin{aligned} \varepsilon_c &= \{(v_i, v_j) | v_i \in v_A^n; v_j \in v_B^n\} \\ s_{ij} &= \frac{v_i \cdot v_j}{|v_i| \times |v_j|} \\ a_{ij} &= \begin{cases} s_{ij}, & \text{if } s_{ij} \geq e_\tau^c \quad and \quad (v_i, v_j) \in \varepsilon_c \\ 0, & \text{else} \end{cases} \end{aligned} \tag{3}$$

where, s_{ij} is the cosine distance between v_i and v_j. e_τ^c is a threshold used to remove those connections between nodes with less similarity to highlight those more contrastive prototypes. Similarly, the weight of edge $a_{ij} = s_{ij}$ if the v_i and v_j are connected, or zero otherwise. Through the contrastive graph, the prototypes with the similar pose, face attributes

and clarity of images between sets will be linked to provide contrastive correlation.

3) Graph Convolution: A split-transform-merge strategy GCN method is introduced to increase the diversity of transformations and fuse two types of correlation cues. The essential idea is to update the node representations by propagating information between different prototypes.

We use $X = [x_1, x_2, \ldots, x_G] \in \mathbb{R}^{d \times G}$ to represent the features for all the nodes in both two graphs. For a pair of input sets of length M_1 and M_2, $G = M_1 + M_2$. Given the correlation matrix A, a simple graph convolution is defined as

$$H(X, A, W) = h(\hat{A}XW) \tag{4}$$

where $W \in \mathbb{R}^{d \times d'}$ is a transormation matrix to update the node features from d-dimensional to d'-dimensional. The $\hat{A} \in \mathbb{R}^{G \times G}$ is the normalized version of correlation matrix A, and $h(\cdot)$ denotes a non-linear opeartion followed.

We follow [9] to formulate a dense connection GCN block. As shown in Fig. 4, 32 independent GCN blocks are introduced to increase the diversity of features, where each GCN block contains CGC contains a 1×1 convolution and a graph convolution H to achieve more refined representations. Then, all features are concatenated. We denote the operation as F.

In the CGC block, two graphs with different semantics need to be learned separately and aggregated. Formally, A_o and A_c denote the adjacent matrix of the global graph and contrastive graph, respectively. The aggregation can be described as

$$\begin{aligned} CGC(X, A_o, A_c, W) = PReLU(&\mathcal{F}(X, A_o, W_o) \\ &+ \mathcal{F}(X, A_c, W_c) + X) \end{aligned} \tag{5}$$

where $W = \{W_o, W_c\}$ are trainable weights, PReLU is used to activate the features.

C. Attention Control

For each scale, we gain the node feature sets $\mathcal{F}^n = \{f'_i\}_{i=1}^K$ at n-th scale, where K is the number of prototypes. Each node feature represents embeded correlation for a prototype. The features for all prototypes are passed to regression and softmax operators to generate positive weights $\{q_k\}$ with $\sum_k q_k = 1$. These operations can be described by following equations

$$q_k = \frac{exp(Wf'_k + b)}{\sum exp(Wf'_i + b)} \tag{6}$$

where W and b are the trainable parameters in regression module.

In this way, the weight q_k for prototype p_k is calculated. As metioned above, every prototype p_k is composed of various face samples. We redefine the weights from prototype-level to image-level for multi-scale intergation

$$a_g = \frac{q_k}{t_k}, \quad f_g \in p_k \tag{7}$$

where $\mathcal{A} = \{a_g\}_{g=1}^G$ is the weights for all image-level features $\mathcal{D} = \{f_g\}_{g=1}^G$, the $G = M_1 + M_2$ are numbers of images in the pair of image sets, and t_k is number of elements in p_k.

Generally, the multi-scale attention weights can be represented as: $\mathcal{A}^* = \{a_1^*, a_2^*, \ldots, a_G^*\} \in \mathbb{R}^G$. For each a_g^* is the mean value of a_g in all scales, so that the aggregated feature representation becomes

$$\begin{aligned} R_A &= \sum_i a_i^* f_i, \quad f_i \in D_A \\ R_B &= \sum_j a_j^* f_j, \quad f_j \in D_B \end{aligned} \tag{8}$$

Moreover, we obtain the prototype-based aggregation to each scale separately

$$\begin{aligned} f^*(p_k) &= \frac{\sum f_i}{t_k}, \quad f_i \in p_k \\ r_A^n &= \sum_i a_i f^*(p_i), \quad p_i \in P_A^n \\ r_B^n &= \sum_j a_j f^*(p_j), \quad p_j \in P_B^n \end{aligned} \tag{9}$$

The D_A and D_B are the image-level feature set. The P_A^n and P_B^n are the prototype sets on n-th scale.

The contrastive loss is used to maximize the similarity of the same identity while minimizing the similarity of different identities

$$\begin{aligned} d &= 1 - \frac{r_i \cdot r_j}{|r_i| \times |r_j|} \\ \mathcal{L}_c(r_i, r_j) &= \frac{1}{2n} \sum_{n=1}^{n} yd^2 + (1 - y)max(margin - d, 0)^2 \end{aligned} \tag{10}$$

where y is the verification label, $margin$ is a threshold for contrastive loss.

Overall, the objective function of our N-scales MC-GCN can be formulated as follows

$$\mathcal{L} = \mathcal{L}_c(R_A, R_B) + \sum_n^N \mathcal{L}_c(r_A^n, r_B^n) \tag{11}$$

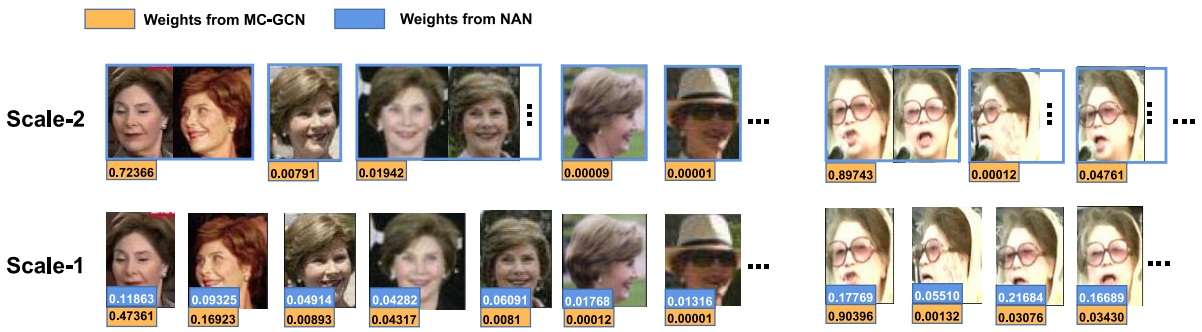
IV. Experiments

In this section, we evaluate MC-GCN for unconstrained set-based face recognition on three datasets, which are IJB-A [7], YTF [17] and a novel pig face dataset. To demonstrate the effectiveness of our method, we show some typical examples of the weighting results comparison to NAN [10] in Fig. 5.

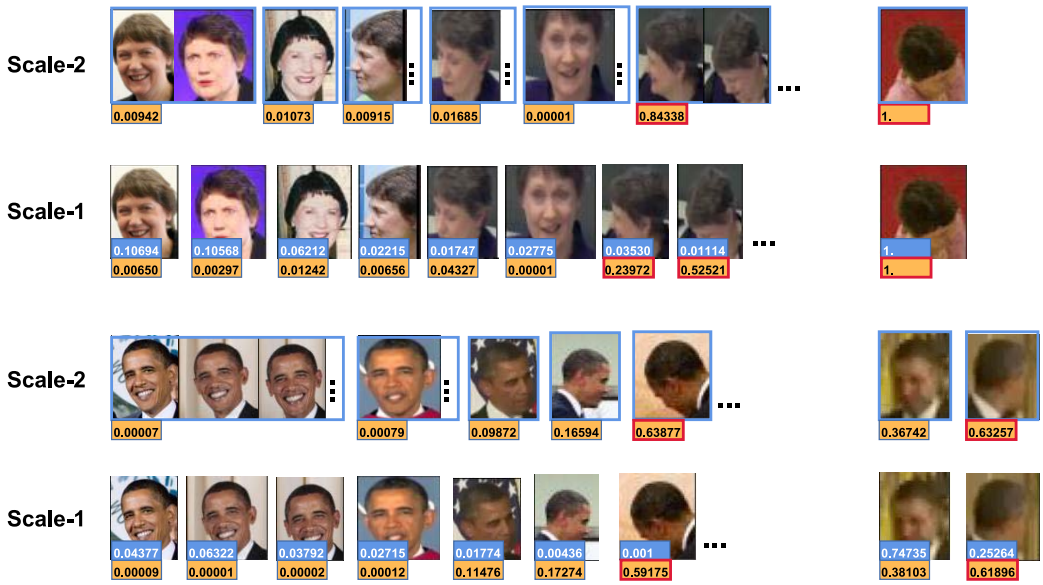
A. Experimental Settings

In our experiments, VGGFace2 [9] is used as the baseline and the image-level feature extractor for MC-GCN. Follow the basic setting in [9], we detect the faces using MTCNN [42] and extend the bounding box by a factor of 0.3. The cropped faces are resized to three smallest dimensions are 224,256,384 and the randomly 224×224 crop is used to process these samples. In the generation of multi-scale prototype graph, the initial edge threshold e_τ is fixed as 0.9 and the threshold steps Δ is set as 0.1. The thresholds e_τ^o and e_τ^c for global and contrastive graphs are set to 0.5 and 0.3.

The Contrastive GCN consists of three CGC blocks and is trained on the target datasets end-to-end. We set the batch size



(a) A typical example in IJB-A.



(b) Extreme examples in IJB-A. MC-GCN focuses on similar (contrastive) prototypes between enroll and verification templates instead of simply evaluating quality.

(c) A typical example in pig face dataset.

Fig. 5. Typical examples showing the weights of the video frames computed by MC-GCN, where the left is the enroll templates and the right is the verification templates.

of all the methods empirically as 128, and the initial learning rate as 1e-3. We train the network for 50 epochs, reducing the learning rate to 0.1 times per 20 epochs. According to the degree of inner-variances within datasets, we build 2-scales MC-GCN for YTF and pig face dataset, 3-scales for IJB-A.

B. Evaluations on IJB-A Dataset

IJB-A [7] is a popular face verification and identification dataset, which contains 5,397 images and 2,042 videos from 500 subjects captured from unconstrained environments with wide variances. In IJB-A, a template is a set of images composed of various video frames and images from different sources. For training and testing, 10 random splits are provided by each protocol, respectively. We evaluate the MC-GCN on face verification and identification by TAR@FAR and TPIR@FPIR.

1) Comparison With Previous Methods: We compare our network with some previous methods. The experimental verification and identification results are shown in Tabel I. It can be seen that the MC-GCN outperforms all the

TABLE I

PERFORMANCE EVALUATION ON THE IJB-A DATASET. THE RESULTS ARE AVERAGED OVER 10 SPLITS ON BOTH VERIFICATION AND IDENTIFICATION PROTOCOL

Methods	Verification			identification		
	TAR@FAR=0.10	TAR@FAR=0.01	TAR@FAR=0.001	TPIR@FPIR=0.10	TPIR@FPIR=0.01	Rank1
GOTS [7]	0.627±0.012	0.406±0.014	0.198±0.008	0.235±0.033	0.047±0.024	0.433±0.021
BCNNs [15]	-	-	-	0.341±0.032	0.143±0.027	0.588±0.020
LSFS [34]	0.895±0.013	0.733±0.034	0.514±0.060	0.613±0.032	0.383±0.063	0.820±0.024
Milti-pose [35]	0.991	0.787	-	0.750	0.520	0.846
DREAM [36]	-	0.944±0.009	0.868±0.015	-	-	0.946±0.011
MPNet [13]	0.980±0.005	0.919±0.013	0.779±0.021	0.831±0.009	0.663±0.042	0.932±0.008
PIFR [14]	0.993±0.003	0.983±0.004	0.955±0.010	0.969±0.009	0.908±0.028	**0.990±0.005**
QAN [37]	0.980±0.006	0.942±0.015	0.893±0.039	-	-	0.955±0.006
NAN [10]	0.978±0.003	0.941±0.008	0.881±0.011	0.917±0.009	0.817±0.041	0.958±0.005
DAC [12]	0.981±0.008	0.954±0.010	0.893±0.010	0.934±0.009	0.855±0.042	0.973±0.011
TDFF [38]	0.988±0.003	0.961±0.007	0.919±0.006	0.941±0.010	0.878±0.035	0.964±0.006
GhostVLAD [11]	0.990±0.002	0.972±0.003	0.935±0.015	0.951±0.005	0.884±0.059	0.977±0.004
VGGFace2 [9]	0.980±0.003	0.950±0.005	0.895±0.019	-	-	-
MC-GCN(1-scale)	**0.994±0.004**	0.986±0.005	0.965±0.010	0.970±0.013	0.911±0.034	0.988±0.006
MC-GCN(2-scales)	0.993±0.004	0.987±0.005	0.967±0.008	0.972±0.011	0.915±0.030	0.988±0.006
MC-GCN(3-scales)	0.993±0.003	**0.990±0.005**	**0.969±0.007**	**0.972±0.009**	**0.917±0.027**	0.987±0.007
MC-GCN(4-scales)	0.982±0.009	0.978±0.014	0.952±0.020	0.964±0.016	0.901±0.045	0.987±0.012

previous state-of-the-art methods. Compared with the baseline VGGFace2, our proposed method achieves an almost 7 percentage points improvement in 1:1 verification while FAR=0.001.

We also report the comparison results with other set descriptor aggregation based methods. The results show that both the verification and identification performance largely improved compared to these methods. The NAN [10], GhostVLAD [11] and DAC [12] are all quality-based methods, which are hard to effectively aggregate the set without high-quality images. In this paper, the multiple prototype scales and prototype-based GCN build a more robust model to describe correlation cues and quality. Besides, contrastive features are considered to deal with these hard samples and improve the discrimination ability of the model. Through these measures, MC-GCN can maintain those features that are more beneficial for verification while opposing low-quality and redundant images.

2) Ablation Study: To demonstrate the effectiveness of multi-scale for prototypes graph, we also show the comparison of different scales in Tabel I. The 1-scale MC-GCN is image-based contrastive GCN, where the nodes of the graph are image-level features and the concept of the prototype is dismantled. Since IJB-A has complex posture, illumination and other attributes, the 3-scales model will perform better. However, when the number of layers rises to a higher layer, variances will occur within the prototype and affect performance.

In order to further prove the effectiveness of contrastive information for recognition, we also perform a further ablation study on CGC. Specifically, we tried Simple GCN, GCN (Knn Graph), GCN (Contrastive graph), and the CGC proposed in this article. We show the ablation study in Table II, and simple GCN (knn graph) achieves a good performance due to excellent relationship modeling ability. GCN (only contrastive graph) pays too much attention to the relationship between sets, and the effect is slightly worse than simple GCN. The CGC introduces the contrastive graph in global relationship modeling and significantly solves hard samples (Fig. 5).

TABLE II

ABLATION STUDY OF CGC (IJB-A)

Methods	Verification TAR	
	FAR=0.01	FAR=0.001
GhostVald	0.972	0.935
GCN for signle set	0.973	0.932
GCN (Knn graph)	0.978	0.943
GCN (Constrasrive graph)	0.968	0.940
CGC	**0.986**	**0.965**

TABLE III

AVERAGE ACCURACY COMPARISON ON YTF

Methods	Accuracy
DeepFace [1]	0.914
CenterLoss [39]	0.949
SphereFace [40]	0.950
ArcFace [4]	0.980
NAN [10]	0.9572
DAN [41]	0.9428
DAC [12]	0.9601
MC-GCN(1-scale)	0.9879
MC-GCN(2-scales)	**0.9889**
MC-GCN(3-scales)	0.9883

3) Effectiveness of MC-GCN: In order to better demonstrate and explain the effectiveness of MC-GCN, we show the weights of some typical examples. For comparison, we show both weights of NAN and MC-GCN on the 1-th scale.

TABLE IV
PERFORMANCE EVALUATION ON THE IJB-C DATASET

Methods	Verification TAR		
	FAR=1e-3	FAR=1e-4	FAR=1e-5
FaceNet [2]	0.665	0.487	0.330
VGGFace2 [9]	0.927	0.862	0.768
ArcFace [4]	0.960	0.921	0.861
MPNet [13]	0.940	0.898	0.827
MC-GCN(1-scale)	0.960	0.927	0.883
MC-GCN(3-scales)	**0.962**	**0.933**	**0.892**

As shown in Fig. 5 (a), when the input pair of sets both contain high-quality pictures, the high-quality images will be given higher weights, while the blurry and other posture samples are almost ignored. In addition, we show the higher-level prototype weights. Due to the richer semantic relationship between prototypes, high-quality and contrastive prototypes will be further highlighted compared to low-level.

As mentioned in Section IV-B.1, MC-GCN is far superior to previous quality-based methods, because it is difficult for quality-based method to deal with extreme sample pairs. For example, Fig. 5 (b) show two extreme template pairs, where the enroll templates consist of different quality images but the verification templates contain only low-quality samples. From the weights of images computed by NAN, it can be seen that the high-quality images are advocated and the low-quality images are ignored in enroll templates. However, the low-quality images in verification templates have a far distance from those high-quality prototypes. Quality-based approaches perform poorly on these extreme pairs. In contrast, our method pays more attention to similar (contrastive) prototypes between enroll and verification templates. Specifically, when the main images in verification templates are poor, MC-GCN will not simply activate high-quality images but supports low-quality images with similar prototypes (as shown in Fig. 5 (b)).

C. Evaluations on YTF Dataset

The YouTube Face (YTF) dataset is a popular video face verification dataset, which contains 3,425 videos of 1,595 different subjects downloaded from YouTube. In experiments, we follow the standard verification protocol as [10], [12].

The Accuracies of the proposed MC-GCN with other state-of-the-arts on YTF are reported in Table III and MC-GCN achieves a 2.88 percentage points improvement compared to DAC [12].

D. Evaluations on IJB-C Dataset

We also verified the performance on the IJB-C to further verify the superiority of the proposed MC-GCN. IJB-C is a more challenging dataset captured from in-the-wild environments to avoid the near frontal bias, which contains 31,334 images and 11,779 videos from 3,531 subjects and provides template-based verification protocol.

We choose the arcface [4] as the image-level feature extractor in this section. As shown in the table IV, we show the comparison result of TAR@FAR. 3-scales MC-GCN improves 0.9% compared to 1-scale MC-GCN under FAR=1e-5. Compared with other methods, MC-GCN improves the baseline by 3.1% under FAR=1e-5, and is significantly better than other set-based methods, which well verified the superiority and excellent generalization of MC-GCN.

TABLE V
VERIFICATION COMPARISON IN TERMS OF TAR AT FAR ON PIG FACE DATASETS

Methods	Verification TAR		
	FAR=0.01	FAR=0.001	FAR=0.0001
VGG [1]	0.903	0.861	0.523
SphereFace [40]	0.923	0.892	0.637
ArcFace [4]	0.925	0.899	0.648
RiseNet [18]	0.936	0.909	0.720
NAN [10]	0.941	0.916	0.760
MC-GCN(1-scale)	0.951	0.931	0.815
MC-GCN(2-scales)	**0.956**	**0.939**	**0.827**

E. Evaluations on Animal Dataset

In order to prove the generalization ability of our method, we conduct experiments on a novel animal (pig) face dataset, which was collected from complex breeding environments with huge variances in pose and illumination. There are 1766 videos from 506 pigs. We create 1000 pairs of pig face templates for train and test, where each face template contains several images randomly selected from video frames. 10-fold cross-validation is considered as the evaluation method.

We choose the RiseNet [18] for animal face recognition as the image-level feature extractor and show the verification results of NAN and MC-GCN. It can be seen in Table V that MC-GCN far exceeds the quality-based methods under the same baseline. We verify the effectiveness of the multi-scale prototype on more challenging animal face recognition task. It can be seen that due to the complex scenarios and various poses, the performance of 2-scales MC-GCN is improved significantly compared to 1-scale MC-GCN. Furthermore, we compare the weights between NAN and MC-GCN in Fig. 5 (c). MC-GCN does not blindly activate those high-quality images but gives more weight to more contrasting frames.

V. Conclusion

We introduce a novel Multi-Scale Contrastive Graph Convolutional Network (MC-GCN) to advocate the contrastive samples instead of simply quality evaluation, which has a strong ability of semantic relation modeling due to the multi-scale prototype graph and the CGC block. MC-GCN outperforms state-of-the-arts and solves the problem that previous methods cannot deal with low-quality sets. Moreover, MC-GCN can further improve the performance by combining with a more advanced image-level feature extractor. In the future, we will exploit how to apply the multi-scale graph and CGC to other problems, such as Re-ID, video localization and action recognition, etc.

References

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2015, pp. 41.1–41.12.

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

[3] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[5] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.

[6] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen, "SFace: Sigmoid-constrained hypersphere loss for robust face recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 2587–2598, 2021.

[7] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.

[8] T. Hassner *et al.*, "Pooling faces: Template based face recognition with pooled face images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 59–67.

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[10] J. Yang *et al.*, "Neural aggregation network for video face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4362–4371.

[11] Y. Zhong, R. Arandjelović, and A. Zisserman, "GhostVLAD for set-based face recognition," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer-Verlag, 2018, pp. 35–50.

[12] X. Liu, B. Kumar, C. Yang, Q. Tang, and J. You, "Dependency-aware attention control for unconstrained face recognition with image sets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 548–565.

[13] J. Zhao *et al.*, "Multi-prototype networks for unconstrained set-based face recognition," 2019, *arXiv:1902.04755*.

[14] X. Liu *et al.*, "Permutation-invariant feature restructuring for correlation-aware image set-based recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4986–4996.

[15] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear CNNs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[17] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. CVPR*, Jun. 2011, pp. 529–534.

[18] X. Shi, C. Yang, X. Xia, and X. Chai, "Deep cross-species feature learning for animal face recognition via residual interspecies equivariant network," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer-Verlag, 2020, pp. 667–682.

[19] O. Arandjelović and R. Cipolla, "An information-theoretic approach to face recognition from face motion manifolds," *Image Vis. Comput.*, vol. 24, no. 6, pp. 639–647, Jun. 2006.

[20] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on Grassmann manifold with application to video based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 140–149.

[21] Z. Huang, J. Wu, and L. Van Gool, "Building deep networks on Grassmann manifolds," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[22] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2003, p. I.

[23] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2273–2286, Nov. 2011.

[24] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.

[25] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2496–2503.

[26] Y. Mao, R. Wang, S. Shan, and X. Chen, "COSONet: Compact second-order network for video face recognition," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer-Verlag, 2018, pp. 51–67.

[27] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3D human pose regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3425–3435.

[28] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.

[29] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Nov. 2019.

[30] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10156–10165.

[31] L. Yang, X. Zhan, D. Chen, J. Yan, C. C. Loy, and D. Lin, "Learning to cluster faces on an affinity graph," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2298–2306.

[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[34] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," 2015, *arXiv:1507.07242*.

[35] W. Abdalmageed *et al.*, "Face recognition using deep multi-pose representations," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[36] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5187–5196.

[37] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5790–5799.

[38] L. Xiong *et al.*, "A good practice towards top performance of face recognition: Transferred deep feature fusion," 2017, *arXiv:1704.00438*.

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

[40] 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.

[41] Y. Rao, J. Lin, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3781–3790.

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

Xiao Shi received the B.E. degree in computer science from Shenyang Normal University, China, in 2019. He is currently pursuing the M.S. degree with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. His research interests include deep learning, computer vision, face recognition.

Xiujuan Chai (Member, IEEE) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2007. She has been a Full Professor with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences, since 2019. She has also been selected for the "Agricultural Talents" Project of CAAS. She is currently the Principal Scientist of the Machine Vision and Agricultural Robot Innovation Team, CAAS. She has published over 70 articles and has more than 20 patent applications. Her research interests cover intelligent perception, machine vision, and agricultural robot. She was a recipient of the China's State Natural Science Award in 2015 for her research work.

Jiake Xie received the bachelor's degree from the Zhengzhou University of Light Industry. He is currently the Computer Vision Algorithm Engineer at Hangzhou Wangdao Holdings Company Ltd. He specializes in computer vision related deep learning algorithms with research interests in object detection, semantic segmentation, salient object detection, image matting, video matting.

Tan Sun received the Ph.D. degree in management science. He is currently a Professor, a Doctoral Supervisor, and the Vice President of the Chinese Academy of Agricultural Sciences (CAAS). His current professional positions are the Director of the Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs of China; the Vice Chairperson of the Library Society of China; and the Vice Chairperson of the Association for Science and Technology of CAAS. In recent years, he has published more than 90 academic papers and five monographs. He is the Principal Investigator of more than ten high level projects funded by the National Science and Technology Support Program of the Ministry of Science and Technology and the National Social Science Fund of China. His main research directions include digital information description and knowledge organization, big data mining, smart agriculture.