Low-Resolution Face Recognition in the Wild via Selective Knowledge Distillation

Shiming Ge, *Senior Member, IEEE*, Shengwei Zhao, Chenyu Li, and Jia Li, *Senior Member, IEEE*

Abstract—Typically, the deployment of face recognition models in the wild needs to identify low-resolution faces with extremely low computational cost. To address this problem, a feasible solution is compressing a complex face model to achieve higher speed and lower memory at the cost of minimal performance drop. Inspired by that, this paper proposes a learning approach to recognize low-resolution faces via selective knowledge distillation. In this approach, a two-stream convolutional neural network (CNN) is first initialized to recognize high-resolution faces and resolution-degraded faces with a teacher stream and a student stream, respectively. The teacher stream is represented by a complex CNN for high-accuracy recognition, and the student stream is represented by a much simpler CNN for low-complexity recognition. To avoid significant performance drop at the student stream, we then selectively distil the most informative facial features from the teacher stream by solving a sparse graph optimization problem, which are then used to regularize the fine-tuning process of the student stream. In this way, the student stream is actually trained by simultaneously handling two tasks with limited computational resources: approximating the most informative facial cues via feature regression, and recovering the missing facial cues via low-resolution face classification. Experimental results show that the student stream performs impressively in recognizing low-resolution faces and costs only 0.15-MB memory and runs at 418 faces per second on CPU and 9433 faces per second on GPU.

Index Terms—Face recognition in the wild, two-stream architecture, knowledge distillation, CNNs.

I. INTRODUCTION

FACE, a fundamental attribute that distinguishes one subject from another, needs to be recognized many times

S. Ge is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China (e-mail: geshiming@iie.ac.cn).

S. Zhao and C. Li are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China.

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China (e-mail: jiali@buaa.edu.cn).

Fig. 1. Low-resolution faces can become recognizable for subjects that are familiar with the corresponding high-resolution faces. Although low-resolution faces may have low quality, blurry textures, poor illumination and diversified occlusions, the knowledge from high-resolution faces can help to extract discriminative features for efficient face recognition.

everyday in modern computer vision and multimedia applications. Among these applications, many well-known face recognition models [1]–[4] need to be re-deployed on mobile phones [5] or even smart cameras [6] to meet the real-world requirements that aim to identify low-resolution faces with extremely low computational cost and memory footprint (*i.e.*, face recognition in the wild [7]). Toward this end, it is necessary to explore a feasible solution that can address a key challenge in face recognition: how to convert an existing complex face model into a more efficient one that still works well on low-resolution faces without remarkable loss of recognition accuracy?

Compared with high-resolution faces, low-resolution faces have their unique visual attributes. As shown in Fig. 1, many details are missing in low-resolution faces. However, they are still recognizable for subjects who are familiar with the corresponding high-resolution faces, implying that the neural systems of human beings may have the capability of recovering missing details of familiar faces. Inspired by this fact, many existing low-resolution face models have been proposed, which can be roughly grouped into two categories: the hallucination category and the embedding category.

Models in the hallucination category propose reconstructing the high-resolution faces before recognition [8]–[10]. For example, Kolouri and Rohde [8] described a transport-based

single frame super-resolution method to automatically construct a nonlinear Lagrangian model of high-resolution facial appearance. After that, the low-resolution facial image was enhanced by finding the model parameters that best fit the given low-resolution data. Jian and Lam [9] observed that the singular values of a face image at different resolutions have approximately linear relationship. Based on this observation, they first applied singular value decomposition for face representation to learn the mapping function between low-resolution and high-resolution face pairs, and then performed both hallucination and recognition of low-resolution faces simultaneously. Similar method proposed by Yang *et al.* [10] used sparse representation to perform joint hallucination and recognition, which can synthesize person-specific versions of low-resolution faces with recognition guarantee. Typically, these approaches exhibit impressive performance in recognizing the reconstructed high-resolution faces, while the super-resolution operation often brings in additional computational cost and leads to low recognition speed.

Different from the hallucination-based models, models in the embedding category directly extract discriminative features from low-resolution faces by using various external face contexts. For example, Biswas *et al.* [11] proposed embedding low-resolution facial images into an Euclidean space such that the distances between them in the transformed space can well approximate the best distances of high-resolution faces. Ren *et al.* [12] proposed coupled kernel embedding to map the facial images with different resolutions onto an infinite subspace. The recognition process was then carried out in the new space by minimizing the dissimilarities captured by their kernel gram matrices in the low-resolution and high-resolution spaces. Generally speaking, the most important process in the embedding-based approaches is transferring the knowledge from high-resolution faces to low-resolution ones. However, a key issue that needs to be carefully addressed in this process is correctly transferring only the desired knowledge rather than all of them from high-resolution domain to low-resolution domain. Such selective knowledge transfer is one of the most important challenges in converting existing face models into more efficient ones that also work well on low-resolution faces.

Inspired by this fact, we propose a selective knowledge distillation approach for low-resolution face recognition in the wild. As shown in Fig. 2, a two-stream CNN is first trained to simultaneously recognize high-resolution faces and their resolution-degraded versions by using two streams. The two streams consist of a teacher stream that operates on high-resolution faces, and a student stream that is much simpler for low-resolution face recognition. To ensure that the student stream has comparable recognition performance with the teacher stream, we then selectively distil only the most informative facial features from the teacher stream by solving a sparse graph optimization problem, which are then used to regularize the fine-tuning process of the student stream. In this way, the student stream is actually trained by simultaneously handling two tasks with limited computational resources: approximating the most informative facial cues via feature regression, and recovering the missing facial cues via low-resolution face classification. Note that the teacher stream can

adopt any architecture of existing deep face models, implying that the proposed approach can convert any existing face model into a much simpler one with higher speed and lower memory at the cost of minimal performance drop. Experimental results on four public datasets show that the student stream performs impressively in recognizing faces at extremely low resolutions. In particular, it uses only 0.15MB memory and runs at about 418 faces per second on a single CPU thread or 9,433 faces per second on GPU.

The main contributions of this paper are summarized as follows. 1) We propose a face model compression method via selective knowledge distillation, which can greatly reduce model size and complexity without remarkable performance drop; 2) We propose graph-based optimization algorithm that can extract the most discriminative facial features from existing face models, which can be used to supervise the training process of low-resolution face models; and 3) We conduct comprehensive experiments to show that the compressed model can achieve an extremely high recognition speed with a comparable accuracy with the state-of-the-art high-resolution face models.

The rest of this paper is organized as follows: Section II reviews related works and Section III presents the selective knowledge distillation approach. Extensive experiments are conducted in Section IV to evaluate the proposed approach, and the paper is concluded in Section V.

II. RELATED WORKS

The approach we proposed in this paper aims to distil knowledge from complex face models for low-resolution face recognition. Therefore, we briefly review related works from three aspects, including the general face recognition models, low-resolution face recognition and knowledge distillation.

A. General Face Recognition Models

Recently, the general face recognition technique has evolved from the classic shallow frameworks [13], [14] to deep ones [1], [3], [4], [15]–[18] with impressive performance improvements. For the deep approaches, a key factor to distinguish them is the loss functions they adopted. For example, DeepFace [1] is an early attempt to ensemble Convolutional Neural Networks (CNNs) by building 3D faces with identification loss. After that, various loss functions have been proposed for training face recognition CNNs, such as triplet loss [3], [4], center loss [15] and range loss [17]. In [18], the tasks of identifying faces and their attributes were simultaneously considered to enhance the recognition performance. For the DeepID series, several small CNNs using different facial patches were first separately trained in [2], and its subsequent works incorporate face verification signals [19] and change the base networks [20] to increase accuracy.

Generally speaking, these deep models have achieved impressive performance in recognizing general faces. As shown in Table I, however, many of such generic models have a large amount of parameters, high dimensional feature representations and complex classification function for inference. The complexity of these models prevent

TABLE I

REPRESENTATIVE DEEP MODELS FOR GENERAL FACE RECOGNITION.
#TRAIN: NUMBER OF TRAINING IMAGES, RES.: INPUT FACE
RESOLUTION, DIM.: OUTPUT FEATURE DIMENSION, #LYR:
NUMBER OF NETWORK LAYERS, #PAR:
NUMBER OF MODEL PARAMETERS

Model	#Train	Res.	Dim.	#Lyr	#Par
DeepFace [1]	4.4M	152×152	4,096	8	120M
DeepID [2]	203K	39×31	9,600	7	17M
DeepID2 [19]	203K	55×47	4,500	7	10M
FaceNet [3]	260M	96×96	128	22	140M
VGGFace [4]	2.6M	224×224	4,096	16	138M
CenterLoss [15]	700K	112×96	1,024	7	6M
CCN-3DMM [16]	500K	100×100	198	101	30M
GTNN [18]	6.0M	128×128	512	10	3M
RangeLoss [17]	1.5M	112×96	1,024	7	6M
VGGFace2 [21]	3.31M	224×224	2,048	50	21.7M

them from being directly deployed in the wild where the computational resource is limited. Although DeepID series take low-resolution faces as the input, the unique attributes of low-resolution faces are not explored. To further enhance the performance of low-resolution face recognition, it is necessary to explore the missing features during the resolution degradation.

B. Low-Resolution Face Recognition

Typically, there are two ways for low-resolution face recognition. The hallucination category aims to reconstruct high-resolution faces before recognition, while the embedding category proposes extracting features directly from low-resolution faces via the embedding schema. In the hallucination category, Kolouri and Rohde [8] constructed a nonlinear Lagrangian model of high-resolution facial appearance and then found the model parameters that best fit the low-resolution faces. Jian and Lam [9] proposed a framework based on singular value decomposition and performed face hallucination and recognition simultaneously. In [10], a joint face hallucination and recognition framework was proposed based on sparse representation. This framework can synthesize person-specific low-resolution faces for recognition. In [22], a system was proposed to recognize faces by using sparse representation with the specific dictionary involving many natural and facial images. Moreover, deep models like [23] and [24] can generate extremely realistic high-resolution images from low-resolution faces. However, the speed of such hallucination or super-resolution based approaches may be a little slow due to the complex high-resolution face reconstruction process, which hinders their direct deployment in real-world scenarios with limited computational resources.

Instead of reconstructing high-resolution faces, a more direct approach is embedding low-resolution faces into various external contexts to recover the missing features during resolution degradation. Inspired by that, some approaches proposed transforming both high-resolution and low-resolution faces into a unified feature space for matching [25]–[31], while in [32] and [33] the multi-scale (multi-resolution) faces were simultaneously analyzed to extract better features. In [34],

the multidimensional scaling was adopted to learn a common transformation matrix to simultaneously transform the facial features of low-resolution and high-resolution training images. Shekhar *et al.* [35] proposed a joint sparse coding technique for robust recognition at low-resolution, while Wang *et al.* [36] attempted to solve very low resolution recognition problem using deep learning methods. In [37], CNNs were adopted with a manifold-based track comparison strategy for low-resolution face recognition in videos.

From these approaches, we find that the core idea of the embedding-based approaches is transferring (or making use of) the knowledge from high-resolution faces. As a result, the performance in low-resolution face recognition is mainly influenced by two key factors: what knowledge to transfer and how to make use of it. In other words, the desired knowledge should be selectively distilled from high-resolution data (or models) and guide the low-resolution face recognition process in a right way. This is also the core idea of this paper.

C. Knowledge Distillation

Instead of mining the knowledge from high-resolution faces, another way to obtain a low-resolution face model (*i.e.*, the student network) is distilling such knowledge directly from pre-trained complex face models (*i.e.*, the teacher network). With the development of much deeper and wider networks, such distillation technique has been adopted in many works [38]–[49] to compress a complex model (or an ensemble) into a simpler model that is much easier to deploy. Among these works, Luo *et al.* [42] utilized the learned knowledge of a large teacher network or the ensemble of some networks as the supervision to train a compact student network for face recognition. In their approach, the most relevant neurons for face recognition were selected at the higher hidden layers for knowledge transfer. Lopez-Paz *et al.* [50] proposed the general distillation framework to combine distillation and learning with privileged information. Su and Maji [51] proposed cross quality distillation to learn models for recognizing low-resolution images, non-localized objects and line-drawings by using labeled high-resolution images, labeled localized objects and color images, respectively. Radosavovic *et al.* [52] proposed data distillation to ensemble predictions from multiple transformations of unlabeled data to automatically generate new training annotations.

To sum up, the core component of knowledge distillation is the trade-off between speed and performance, and such a technique provides an opportunity to convert many complex models into simple models that can be deployed in the wild. Note that in this study we not only try to distil complex face models into simple ones, but also explore the feasibility of using resolution-degraded faces as the input to further speed up the recognition speed while maintaining the recognition accuracy. In this way, the challenges in low-resolution face recognition and knowledge distillation are simultaneously addressed with a single framework.

III. THE APPROACH

Our two-stream knowledge distillation framework consists of a teacher stream and a student stream (see Fig. 2).
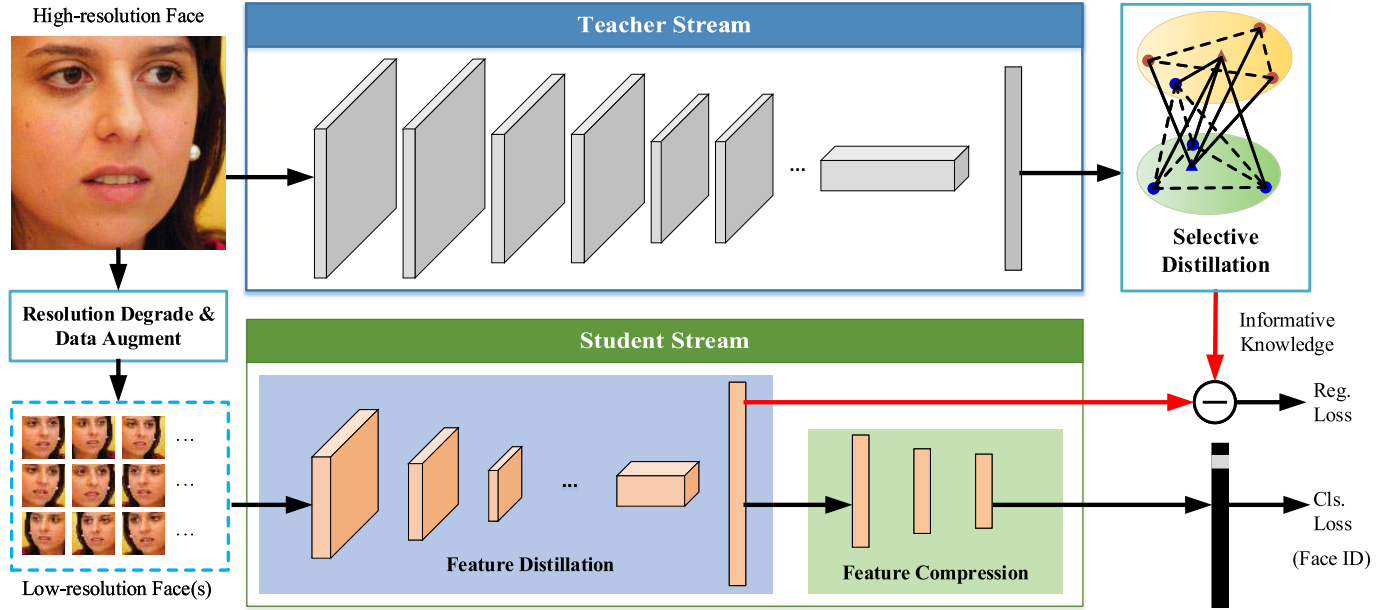
Fig. 2. The framework of our approach. The distillation process consists of three stages. In the first stage, we initialize the teacher and student streams to recognize high-resolution faces and their low-resolution versions, respectively. In the second stage, we selectively distil only the most informative facial features from the teacher stream by solving a sparse graph optimization problem. In the third stage, the selected features are used to regularize the fine-tuning process of the student stream. In this manner, the student stream for recognizing low-resolution faces in the wild is actually trained by simultaneously handling two tasks with limited computational resources: selective feature approximation and low-resolution face identification.

The teacher stream can adopt any complex face recognition neural networks that have been previously trained (and the training data may be no longer available). The distillation process aims to learn a simple and compact student stream that imitates the teacher stream for its practical deployment in real-world scenario.

The learning process consists of three stages: 1) the Initialization stage initializes the teacher stream by taking a complex CNN or an ensemble of several CNNs pre-trained on high-resolution face images, and the student stream by classifying low-resolution face images with identity labels; 2) the Selection stage extracts the most informative knowledge from the teacher stream where the "right" knowledge is selected while the "wrong" one is wiped out; and 3) the Fine-tuning stage transfers the selective knowledge from teacher and low-resolution face images to co-supervise the fine-tuning progress of the student stream by jointly performing feature regression and face identity classification. More details of the three stages are described as follows.

A. Definition

For the sake of simplicity, we define the key components of the two-stream CNNs as follows:

Teacher stream. The teacher stream $\phi_t(\mathcal{F}; \mathbb{W}_t)$ is a complex CNN (or an ensemble of several complex CNNs) with the set of parameters $\mathbb{W}_t$ pre-trained for recognizing a high-resolution face $\mathcal{F}$. Here we assume that $\mathbb{W}_t$ absorbs the rich knowledge encoded in massive high-resolution face images from a teacher face set $\mathbb{D}_t$, in which each face is labeled by an integer from the identity set $\mathbb{L}_t$. Generally, the number of training face images $|\mathbb{D}_t|$ is very large and may be invisible

to the student stream (*e.g.*, a CNN model released on the Internet is pre-trained with additional face images from private datasets).

Student stream. The student stream $\phi_s(\tilde{\mathcal{F}}; \mathbb{W}_s)$ is a much simpler CNN for recognizing a low-resolution face $\tilde{\mathcal{F}}$ with parameters $\mathbb{W}_s$. It is learned from the student face set $\mathbb{D}_s = \{(\mathcal{F}_i, \{\tilde{\mathcal{F}}_{ij}\}_{j=1}^N, l_i)\}_{i=1}^{|\mathbb{D}_s|}$, where $|\mathbb{D}_s|$ is the number of high-resolution faces. For each high-resolution face $\mathcal{F}_i$, the student face set also contains its N resolution-degraded versions, and the jth resolution-degraded face is denoted as $\tilde{\mathcal{F}}_{ij}$. Note that both the high-resolution face $\mathcal{F}_i$ and its degraded versions $\{\tilde{\mathcal{F}}_{ij}\}_{j=1}^N$ correspond to the same identity label l_i from the identity set $\mathbb{L}_s$. Here we assume that there are totally C classes of faces in $\mathbb{L}_s$, and the number of high-resolution faces for the cth class is K_c such that $|\mathbb{D}_s| = \sum_{c=1}^{C} K_c$.

B. Initialization of the Two-Stream CNNs

As shown in Fig. 2, our two-stream CNNs simultaneously conduct high-resolution and low-resolution face recognition with a teacher stream $\phi_t(\mathcal{F}; \mathbb{W}_t)$ and a student stream $\phi_s(\tilde{\mathcal{F}}; \mathbb{W}_s)$, respectively. The parameter set $\mathbb{W}_t$ of the teacher stream can be initialized by state-of-the-art face recognition models or their ensemble, such as VGGFace [4] with VGG16 architecture [53], FaceNet [3] with GoogLeNet architecture [54] and VGGFace2 [21] with ResNet50 architecture [55]. As a representative example, we use the architecture of VGGFace [4] in the teacher stream and initialize $\mathbb{W}_t$ with the author-released model. Note that VGGFace is pre-trained on a massive face image dataset $\mathbb{D}_t$, which we assume is no longer available in the knowledge distillation process.
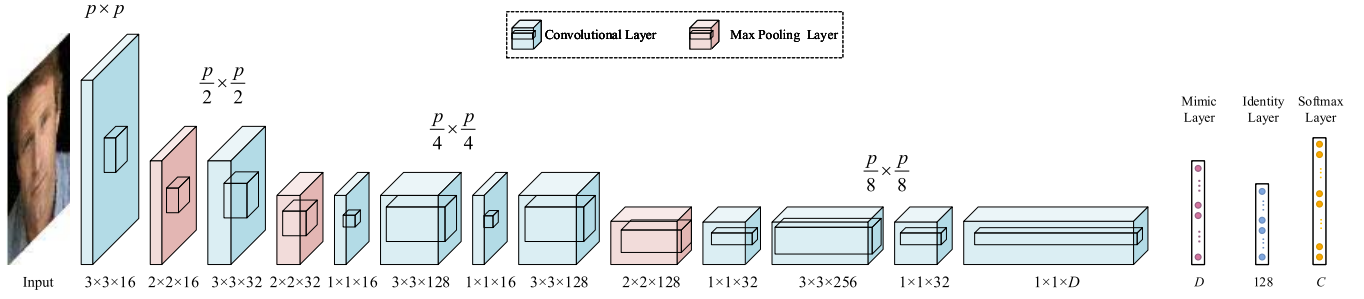
Fig. 3. The structure of the proposed student network. It contains ten convolutional layers, three max pooling layers and three fully-connected layers. It only requires 0.79 million parameters (excluding the last softmax layer), which is much smaller than existing high-resolution face recognition models (*e.g.*, VGGFace).

The student stream $\phi_s(\tilde{\mathcal{F}}; \mathbb{W}_s)$ aims to recognize a low-resolution face $\tilde{\mathcal{F}}$ with a compact network trained on $\mathbb{D}_s$. Therefore, we adopt a lightweight network architecture that is similar to [56]. As shown in Fig. 3, the student stream can take a low-resolution (*e.g.*, 32×32) face as the input with its majority using 3×3 filters and increasing the number of channels after every pooling step. Moreover, the global average pooling is used to make predictions as well as 1×1 filters to reduce the feature dimension between 3×3 convolutions. Note that a $1 \times 1 \times D$ mimic layer is adopted here to receive the knowledge from the teacher stream in the future, where D is the dimension of the learned high-resolution face representation in the teacher stream. Thus, the features from the mimic layer can be used for feature approximation with the teacher network. In addition, since the capacity of the student model is weak, the feature layer that mimics the transferred knowledge should be sufficiently deep, we empirically insert the identity layer between the mimic layer and the softmax layer. The identity layer also plays the role of feature compression. Finally, the architecture has ten convolutional layers, three max pooling layers and three fully-connected layers. As a result, the amount of parameters in $\mathbb{W}_s$ reaches only 0.79M, which is only 0.57% of the teacher parameter set $\mathbb{W}_s$ (*i.e.*, 138M). These parameters are first initialized with xavier and then optimized by minimizing the classification loss over $\mathbb{D}_s$:

$$\mathbb{W}_s^* = \arg\min_{\mathbb{W}_s} \sum_{i=1}^{|\mathbb{D}_s|} \sum_{j=1}^{N} \ell(\phi_s(\tilde{\mathcal{F}}_{ij}; \mathbb{W}_s), l_i), \quad (1)$$

where $\ell(.)$ is a softmax function that measures the classification loss. The minimization problem (1) can be resolved by stochastic gradient descent with standard backpropagation [57].

C. Selective Knowledge Distillation From the Teacher Stream

After the initialization stage, the student stream usually suffers from low recognition accuracy over low-resolution faces since many identity cues are missing during the resolution degradation. As a result, its parameters need to be fine-tuned again under the supervision of the teacher stream to learn how to extract the most discriminative features even when the face resolution is very low. Let $\hat{\phi}_t(\mathcal{F}; \hat{\mathbb{W}}_t)$ and $\hat{\phi}_s(\tilde{\mathcal{F}}; \hat{\mathbb{W}}_s)$ be the sub-networks composed by the first several layers of the teacher stream $\phi_t(\mathcal{F}; \mathbb{W}_t)$ and the student stream

$\phi_s(\tilde{\mathcal{F}}; \mathbb{W}_s)$, respectively. $\hat{\phi}_t(\mathcal{F}; \hat{\mathbb{W}}_t)$ is the feature extraction backend before the softmax layer for extracting the identity features of high-resolution face images, while $\hat{\phi}_s(\tilde{\mathcal{F}}; \hat{\mathbb{W}}_s)$ corresponds to the main feature branch till the mimic layer and is used to extract approximated features to match the teacher. The fine-tuning process of the student stream can be described as mimicking the feature representation $\hat{\phi}_t(\mathcal{F}; \hat{\mathbb{W}}_t)$ with $\hat{\phi}_s(\tilde{\mathcal{F}}; \hat{\mathbb{W}}_s)$ to improve the final recognition ability of $\phi_s(\tilde{\mathcal{F}}; \mathbb{W}_s)$. The problem is: how to conduct the inter-network supervision?

Typically, the teacher stream has very powerful ability to recognize high-resolution faces with the identities available in the teacher face set $\mathbb{D}_t$. However, it can not directly recognize the unfamiliar faces from the student face set $\mathbb{D}_s$ due to the diverse identities from $\mathbb{D}_t$ and $\mathbb{D}_s$. In this case, the knowledge from the teacher stream may contain some errors, which will mislead the fine-tuning process of the student stream. Thus, we selectively distil only the most informative knowledge and reject the wrong one from $\hat{\phi}_t(\mathcal{F}; \hat{\mathbb{W}}_t)$ to improve the feature extraction ability of the sub-network $\hat{\phi}_s(\tilde{\mathcal{F}}; \hat{\mathbb{W}}_s)$ and the face recognition ability of the student stream $\phi_s(\tilde{\mathcal{F}}; \mathbb{W}_s)$.

Toward this end, a feasible solution is finding out the most informative faces from $\mathbb{D}_s$ by using the features given by the teacher stream, and such informative faces can be defined as the ones with small inter-class similarity and large intra-class similarity. Toward this end, we formulate the selective knowledge distillation process as an inference problem on a graph. The nodes represent faces and the edges represent their correlations. As shown in Fig. 4, a densely-connected graph will contain massive edges between all the nodes from C face classes and thus slow down the inference process. In order to conduct the graph-based inference efficiently, we add a centroid node for each face class and then construct a sparse-connected graph $\mathcal{G} = \{\mathbb{V}, \mathbb{E}\}$. In the graph $\mathcal{G}$, the node set $\mathbb{V}$ contains two types of nodes: face nodes $\{\mathcal{F}_i\}_{i=1}^{|\mathbb{D}_t|}$ and centroid nodes $\{\mathcal{U}_c\}_{c=1}^{C}$. The ith face node $\mathcal{F}_i$ and the cth centroid node $\mathcal{U}_c$ is represented by D-dimensional column feature vectors $\mathbf{f}_i$ and $\mathbf{u}_c$, respectively. These two types of feature vectors are extracted from high-resolution face images with the teacher model and can be computed as

$$\mathbf{f}_i = \hat{\phi}_t(\mathcal{F}_i; \hat{\mathbb{W}}_t), \quad \mathbf{u}_c = \frac{\sum_{i=1}^{|\mathbb{D}_s|} \delta(l_i = c)\mathbf{f}_i}{\sum_{i=1}^{|\mathbb{D}_s|} \delta(l_i = c)}, \quad (2)$$
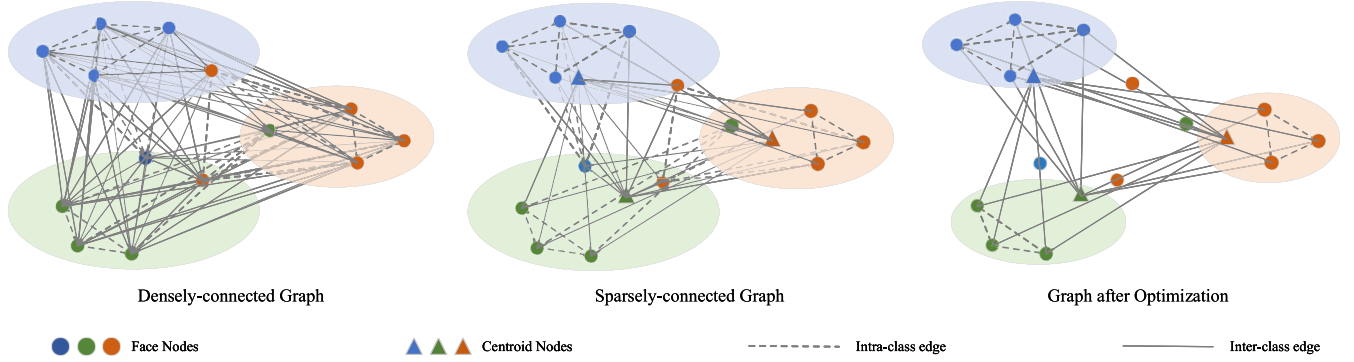
Fig. 4. Selective knowledge distillation with graph optimization. The dense-connected graph (left) has too much edges between feature nodes, on which the optimization problem is difficult to be resolved. In order to address that, a sparse-connected graph (middle) is built via adding a virtual centroid node for each node class. Then, the problem is formulated as a graph labeling problem (right) which can be efficiently solved through energy minimization.

where $\delta(l_i = c)$ is an indicator function which equals 1 if $l_i = c$ and 0 otherwise. We can see that each face node is characterized by the appearance of a specific face, and the centroid node is represented by the average appearance.

With the assistance of centroid nodes, we can construct a sparse graph whose edge set $\mathbb{E}$ consists of densely-connected intra-class edges $\{(\mathcal{F}_i, \mathcal{F}_j)|\forall i, j, l_i = l_j\}$ that link all face nodes within the same class, and sparsely-connected inter-class edges $\{(\mathcal{F}_i, \mathcal{U}_c)|\forall i, c, l_i \neq c\}$ that link face nodes in one class with the rest centroid nodes outside the class. In this way, the sparsely-connected graph $\mathcal{G}$ contains $(|\mathbb{D}_s| + C)$ nodes in total and only $\sum_{c=1}^{C}[K_c(K_c - 1)/2 + K_c(C - 1)]$ edges.

Given the sparse graph, we can select the most informative faces by solving a binary labeling problem:

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{|\mathbb{D}_s|} \left(\alpha_i \sum_{c=1}^{C} \delta(l_i \neq c) \cdot d(\mathbf{f}_i, \mathbf{u}_c) \right)$$
$$+ \lambda \sum_{i=1}^{|\mathbb{D}_s|} \sum_{j=1, i<j}^{|\mathbb{D}_s|} \alpha_i \alpha_j \cdot \delta(l_i = l_j) \cdot d(\mathbf{f}_i, \mathbf{f}_j),$$
$$s.t. \quad \alpha_i \in \{0, 1\}, \forall i = 1, \ldots, |\mathbb{D}_s|, \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_{|\mathbb{D}_s|})$ is a binary vector with $|\mathbb{D}_s|$ components, and its ith component α_i equals 1 if the face $\mathcal{F}_i$ is an informative face and 0 otherwise. Note we use the Cosine distance $d(\cdot)$ to measure the similarity between two feature vectors. We can see that the first term prefers the selection of less informative face nodes that have low similarity with the "average" faces in other classes. λ is a negative weight that balances the two terms so that the second term prefers the face nodes that have high similarity with other faces with the same identity label. In particular, with the non-negative distance measure $d(\cdot)$ and the negative weight λ, the first term tends to select less faces and the second term tends to select more. In practice, we can solve the problem (3) by using the graph cut algorithm [58].

After solving (3), we can select a limited number of informative faces with high intra-class similarity and low inter-class similarity. In this process, the outliers, which are likely to be the errors made by the teacher stream, are discarded from the perspective of feature clustering. In this way, many helpful

knowledge can be accurately distilled and the influences of noisy knowledge introduced by teacher network can be greatly alleviated, which well refines the feature supervision for training the student network. The amount of outliers discarded can be controlled by λ (the influence of λ will be discussed in experiments).

D. Teacher-Supervised Student Stream Fine-Tuning

With the selected informative faces and their features extracted by the teacher stream, the fine-tuning of the student stream will jointly address two issues: 1) approximating the features of informative faces given by the teacher stream via feature regression, and 2) recovering the missing facial cues from low-resolution faces. Thus, we can fine-tune the student stream by solving the minimization problem

$$\min_{\mathbb{W}_s, \hat{\mathbb{W}}_s} \mathcal{L}_{cls}(\mathbb{W}_s; \mathbb{D}_s) + \mathcal{L}_{reg}(\hat{\mathbb{W}}_s; \mathbb{D}_s; \boldsymbol{\alpha}), \tag{4}$$

where the influences of the classification loss and the regression loss are combined together with equal importance to form a multi-task learning problem. The first term $\mathcal{L}_{cls}(\mathbb{W}_s; \mathbb{D}_s)$ is the classification loss of the student stream over all low-resolution faces. Similar to (1), it is defined as

$$\mathcal{L}_{cls}(\mathbb{W}_s; \mathbb{D}_s) = \sum_{i=1}^{|\mathbb{D}_s|} \sum_{j=1}^{N} \ell(\phi_s(\tilde{\mathcal{F}}_{ij}; \mathbb{W}_s), l_i), \tag{5}$$

The term $\mathcal{L}_{reg}(\hat{\mathbb{W}}_s; \mathbb{D}_s; \boldsymbol{\alpha})$ in (4) is the feature regression loss of the sub network $\hat{\phi}_s(\tilde{\mathcal{F}}; \hat{\mathbb{W}}_s)$ formed by feature extraction backend of the student stream. It can be defined as

$$\mathcal{L}_{reg}(\hat{\mathbb{W}}_s; \mathbb{D}_s; \boldsymbol{\alpha}) = \sum_{i=1}^{|\mathbb{D}_s|} \alpha_i \sum_{j=1}^{N} \|\hat{\phi}_s(\tilde{\mathcal{F}}_{ij}; \hat{\mathbb{W}}_s) - \mathbf{f}_i\|^2. \tag{6}$$

By incorporating these two terms into (4), we can solve the classification and regression tasks via the stochastic gradient descent algorithm with standard back-propagation [57]. In this way, the student stream can be fine-tuned under the supervision of the teacher stream in the form of feature regression, leading to improved low-resolution face recognition ability with a limited computational cost.

IV. EXPERIMENTS

In this section, we first introduce the experiment setting and then conduct four experiments to verify the proposed approach. The first experiment is conducted to analyze and discuss the influence of selective knowledge distillation, and the second experiment compares the performance of teacher and student networks in a face verification task. In the third and the fourth experiments, we further compare the student stream with state-of-the-art low-resolution face models in face recognition task and face retrieval task, respectively. Finally, we conduct the efficiency analysis of the learned student models.

A. Experiment Setting

We conduct experiments on four well-known face datasets: UMDFaces [59], LFW (Labeled Faces in the Wild) [60], UCCS (UnConstrained College Students) [61] and SCface (Surveillance Cameras face) [62], which are used to verify the proposed approach from the perspective of selective knowledge distillation, face verification, face identification and face retrieval, respectively. Details of the three datasets (and experimental settings) are listed as follows. We implement all the models with TensorFlow [63] on NVIDIA GPU K80 and single core Intel CPU 2.6G.

The UMDFaces dataset [59] contains $367,888$ images with annotations from $8,419$ subjects, which is obtained by crawling public images on the Internet. In the experiments, we use this dataset to train all the student models and verify the selective distillation operation. For each training face, we first perform face alignment by using the algorithm in [64] to localize facial landmarks. Faces are then cropped and normalized into 224×224 high-resolution images, which are used as the input of the teacher network. Similarly, to form the low-resolution faces for training each student network, we perform random perturbation 16 times on the localized facial landmarks, and then crop and normalize the face regions into face images with size $p \times p$ where the resolution value $p \in \{16, 32, 64, 96\}$.

On the UMDFaces dataset, all the 224×224 face images are first fed into VGGFace, the selected teacher stream, to extract 4096D feature vectors. By solving the graph optimization problem in Eq. (3), informative features are selected and represented with an indicate vector α. After that, with the selected informative features, face identity labels and the student input faces, the student network is trained by using standard BP algorithm. In the training, we set the batch size as 256. Batch normalization layer is introduced to accelerate the network training and prevent over-fitting.

On the LFW dataset [60], we evaluate all student models in the task of face verification. In the experiment, $6,000$ pairs of face images, including $3,000$ positive and $3,000$ negative pairs, are adopted in the evaluation. The performance is reported as the Area under ROC curve (AUC). In the experiment, feature vectors in the hidden layers (mimic layer and identity layer) are first extracted and normalized from a pair of face images. The similarity between them is calculated for verification by using simple threshold. Unlike [19] that
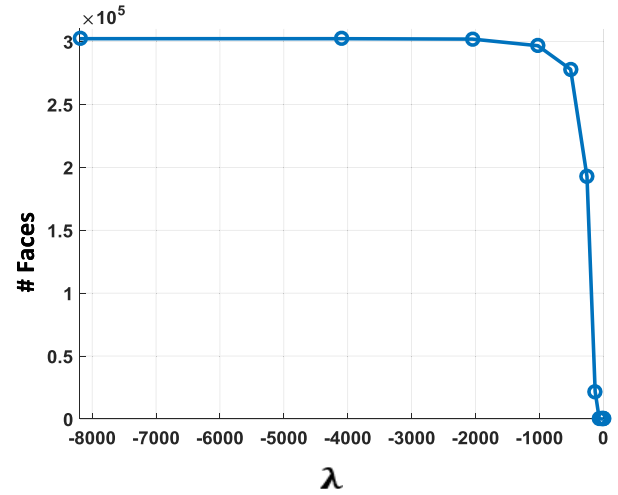

Fig. 5. The influence of the non-positive parameter λ to the number of selected informative faces.

trained Joint Bayesian [65] for face verification, the similarity is used throughout the experiments to directly show the benefit from better supervision utilized to train students.

On the UCCS dataset [61], we compare the student models with state-of-the-arts in the face recognition task. Faces from $1,732$ labeled identities subjects are adopted, where blurry, occluded and badly illuminated images are generally common. Note that the identities in training and testing are exclusive. This dataset is suitable to benchmark more challenging unconstrained face recognition in surveillance conditions.

On the SCface dataset [62], we compare the student models with state-of-the-arts in face retrieval task. The dataset contains 130 subjects, each having one high-resolution frontal face image and multiple low-resolution images, captured from three distances (4.2m, 2.6m and 1.0m, respectively) using different quality surveillance cameras. In the experiment, 50 subjects are randomly selected for training and the rest 80 subjects for testing. Among the testing images, for each subject, one high-resolution face image is used for constructing retrieval dataset and 15 low-resolution face images are used for retrieving. As a result, each low-resolution face image from total $80 \times 15 = 1,200$ is matched with 80 high-resolution face images. The rank-1 recognition accuracies on three subsets with different distances and total set are reported, respectively.

B. Selective Knowledge Distillation

To study selective knowledge distillation, first we would like to explore the influence of different settings of parameter λ on the parse graph optimization algorithm. Therefore, we gradually increase λ from -8192 to 0 with integer power of 2, and then investigate the decreasing tendency of the number of selected informative faces. In Fig. 5, we show the influence of parameter λ on the number of selected informative faces.

When the negative constant λ is very small, the number of the informative faces decreases very slowly. After the λ increases to around -1024, the number of the informative faces starts to decrease sharply. It continues to decrease and remains 0 after λ becomes larger than -32.

(a) Face nodes before discarding   **(b) Noisy face nodes are marked in gray**   **(c) Face nodes after discarding**

(d) The corresponding discarded face images (marked in the same color with face nodes)
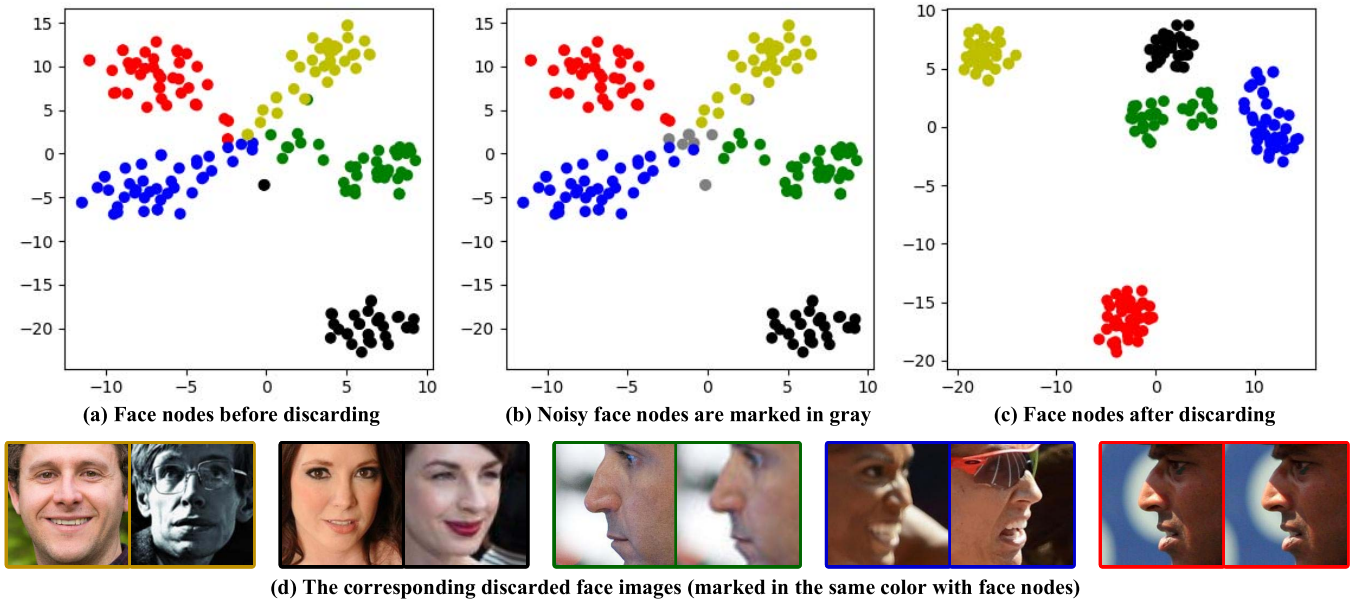
Fig. 6. An example of visualizing selective knowledge distillation on discarding noisy faces with t-SNE [66]. The noisy face nodes (b) which are far away from their own class or closer to other classes are discarded, which leads to a more compact visualization when performing t-SNE again (c). The corresponding discarded face images (d) usually have side postures, heavy occlusions, inconsistent illumination and blurry appearances, which makes them difficult to be identified by the teacher network.

We further delve into the process of discarding faces during selective knowledge distillation. Fig. 6 gives an example for showing the process of discarding faces in five identity classes when λ increases, where we adopt t-Distributed Stochastic Neighbor Embedding (t-SNE) [66] to visualize the high-dimensional face nodes. Through solving the sparse graph optimization problem in Eq. 4, some noisy face nodes (see the original face nodes in Fig. 6(a) and the noisy nodes in Fig. 6(b) marked in gray) that are usually far away from their own class centroid or closer to other classes will be discarded, leading to a more compact visualization effect (see Fig. 6(c)). In Fig. 6(d), we show the discarded noisy face images, where we can see that the discarded faces are often characterized by side postures, heavy occlusions, inconsistent illuminations or blurry appearances. These challenging images, which may be beyond the recognition capability of the teacher stream, are selected and discarded. This implies that selective knowledge distillation indeed selects the more informative faces while discards the less ones. Ideally, the teacher model should have a powerful ability to handle various face variations and cluster the faces correctly, which means that high intra-class similarity and low inter-class similarity are achieved with the extracted features by the teacher model. However, in some challenging cases such as large pose variations, the teacher model may fail and thus causes low intra-class similarity (as shown in Fig. 6). In this case, the extracted teacher knowledge is considered as "wrong" and thus will be discarded by our method in feature regression task.

C. Low-Resolution Face Verification on LFW

With the selected face images, we train many student models to compress the teacher model with different input resolution and various supervision signals. The supervision signals we study are abbreviated as follows:

1) c: only face class supervision (no distillation).
2) s: selective distillation without face class supervision
3) sc: selective distillation with face class supervision
4) dc: direct distillation with face class supervision

For the sake of simplicity, a student model is represented as S-p-s, where supervision signal $s \in \{c, s, sc, dc\}$. For example, model S-32-sc means the student model uses a input resolution of 32×32 and is trained with both selective distillation and class supervision. Note that S-p-c is the baseline student model that are directly trained with the supervision of face classes. Similarly, the teacher model is represented as T-224 with 224×224 input. All the student models are trained with the same architecture as shown in Fig.3.

The performance of various student models is shown in Fig. 7, from which we can see that the recognition accuracy is decreasing along with the lower face resolution. The student model S-96-sc achieves an accuracy of 95.03% by using both selective knowledge distillation and face class supervision, which is only 2.12% lower than the teacher model T-224 without metric learning. Note that the model parameter in S-96-sc is much less than the teacher model VGGFace (0.79M vs. 138M), and the dimension of face feature vectors has a remarkable compression rate of $32\times$. From these results, we can safely claim that this performance is very competitive particularly for practical deployment on resource-limited devices.

From Fig. 7, we can also find that, without the supervision from the teacher stream, the baseline student model S-32-c has a very low accuracy of 70.23%, implying that the baseline model itself may lack the capability of extracting discriminative features when being directly trained on
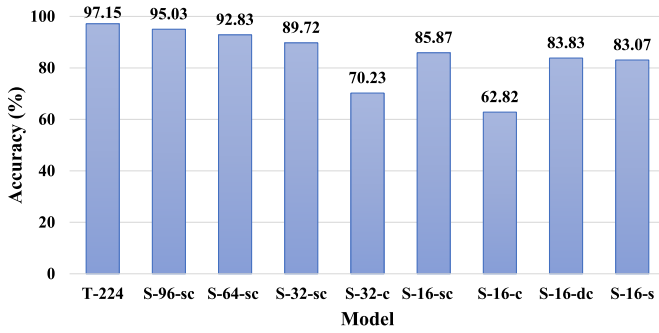
Fig. 7. The recognition accuracy of various teacher and student models on LFW. It shows that the selective knowledge distillation and face class supervision can help to recognizing low-resolution faces.
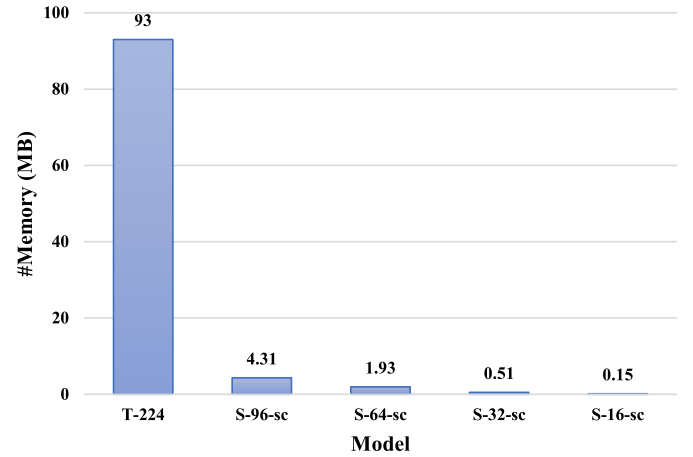

Fig. 8. The inference memory footprint for various models. Compared with the original teacher model, the memory cost of the student model can be greatly reduced.

low-resolution faces. After being trained with joint supervision signals from the teacher stream and face identities, the model S-32-*sc* achieves a sharp improvement of 19.49% in terms of recognition accuracy (*i.e.*, from 70.23% to 89.72%). Similar accuracy improvements can be found between S-16-*sc* and S-16-*c* as well as S-16-*dc* and S-16-*c*, implying that either selective or direct knowledge distillation can effectively transfer the teacher's knowledge into the student network so as to improve the recognition performance remarkably.

To further verify the importance of knowledge selection, we compare the performance between S-16-*sc* and S-16-*dc*. By carefully selecting informative knowledge, S-16-*sc* achieves an accuracy gain of 2.04% against S-16-*dc* which does not discard noisy faces during training. In addition, the face class supervision signal can also improve the performance, so that the model S-16-*sc* achieves a higher accuracy than S-16-*s*. In summary, our two-stream structure can accurately distil informative knowledge from the teacher stream and recover missing knowledge from the student stream.

D. Low-Resolution Face Identification on UCCS

Since the performance of low-resolution face verification task is promising, we further study low-resolution face identification task on a challenging benchmark, UCCS [61], and compare with the state-of-the-art method proposed in [36], VLRR (very low-resolution recognition). In VLRR, the cropped face regions are normalized into 80×80 as high-resolution faces, which are then down-sampled by a factor 5 for low-resolution images of 16×16. The evaluation is performed on a 180-subject subset by layer-by-layer greedy unsupervised model training. Their model reported the best error rates of 40.97% at top-1 and 22.35% at top-5.

Following the experimental settings of [36], we choose a 180-subject subset of original-resolution images by ranking the subjects according to the number of images. The cropped face regions are then normalized to 16×16 to obtain $4,825$ images. Note that this number is a little smaller than those claimed by VLRR (*i.e.*, 4,500 training images and 935 testing images). After that, to achieve fair comparisons, we randomly separate the images according to a ratio of $4 : 1$ to training and testing sets. Finally, we have $3,918$ images for training and the rest 907 for testing.

On these data, we first train a student model with the input 16×16 directly on the training set of UCCS and then test the performance on its testing set. In this case, the model achieves 58.65% top-1 error rate and 22.71% top-5 error rate, which are worse than VLRR. We suspect that the models, once pre-trained on other datasets, can provide valuable prior knowledge on low-resolution visual recognition problem, as stated in [67]. To verify that, we use the student model S-16-*sc* pre-trained on UMDFaces to fine-tune a new model for face identification on UCCS. First, we fix the parameters before mimicking layer and modify the last softmax layer to 180-way. Then, we train the feature reduction sub-network with the $3,918$ images. The fine-tuned model reaches 32.75% top-1 error rate and 18.3% top-5 error rate, indicating the correct classification of 610 out of 907 testing samples in top-1 results and 741 in top-5, respectively. This implies that our method can achieve better accuracy than VLRR, which may be caused by the fact that the selective supervision from the teacher stream can help the student network learn the discriminative features even when the face resolution is very low.

E. Low-Resolution Face Retrieval on SCface

We further study low-resolution face retrieval task on SCface [62], and compare with the baseline (PCA [62]) and 4 state-of-the-arts, including three embedding-based models (DCA [31], DAlign [68] and LRFRW [70]) and one hallucination-based model (SHSR [69]). Here, LRFRW employs deep learning to perform cross-domain transfer. The results are shown in Table II, where the accuracies on three subsets with different distances and total set are reported respectively. In the experiment, we fine-tune two student models on SCface training set, including S-64-*sc* with the default 128D identity features and S-64-*sc*-1024 with 1024D identity features, respectively.

From the results, we can see several important observations. First, all the models give the accuracies of less than 50% and specially an extremely low total accuracy of 4.73% with the baseline PCA model, showing face retrieval on

Model	Dist 1	Dist 2	Dist 3	Total
PCA [62]	1.82	6.18	6.18	4.73
SHSR [70]	14.70	15.70	19.10	16.50
DCA [31]	12.19	18.44	25.53	18.72
LRFRW [69]	20.40	20.80	31.71	24.30
DAlign [68]	34.37	39.38	_49.37_	41.04
S-64-*sc*	_39.25_	_45.75_	47.75	_44.25_
S-64-*sc*-1024	**43.50**	**48.00**	**53.50**	**48.33**

TABLE III
INFERENCE TIME AND SPEED ON GPU AND CPU

Model	GPU (Nvidia K80) time / #faces per second	CPU (Intel 2.6GHZ) time / #faces per second
T-224	20.4 ms / 49	982.4 ms /1.02
S-96-*sc*	0.61 ms /1,639	67.9 ms /14.7
S-64-*sc*	0.32 ms /3,125	31.92 ms /31.3
S-32-*sc*	0.15 ms /6,667	8.92 ms /112
S-16-*sc*	0.106 ms /9,433	2.39 ms /418

SCface is a very challenging task. Second, as the resolution increases along with the distance getting closer, the recognition accuracy gradually increases, which is as expected, which implies that the resolution is indeed an important effect on recognition performance. Thus, the hallucination-based SHSR super-resolves low-resolution face images for feeding to a pre-trained face recognizor, which improves the total accuracy to 16.50%. Third, the embedding models which transfer knowledge between different domains, such as transfer features from high-resolution to low-resolution faces by discriminant correlation analysis in DCA model and by supervised discriminative cross-resolution learning in LRFRW model, transfer knowledge from near-infrared to visible images by dictionary alignment in DAlign model, achieve the improved total accuracies of 18.72%, 24.30% and 41.04%, respectively. This reveals the impact of the transferred knowledge from other domains. Finally, our two models give better total accuracies than other models, *e.g.*, S-64-*sc* achieves an improved total accuracy of 3.21% over DAlign, implying that the selective knowledge from the teacher stream can facilitate the student network.

F. Efficiency Analysis

Our approach can greatly reduce the amount of model parameters and memory footprint without significant accuracy drop. As shown in Fig. 8, the memory reductions are $22\times$, $48\times$, $182\times$ and $620\times$ for the low-resolution student models with 96×96, 64×64, 32×32 and 16×16, respectively. In particular, for the faces with a very low-resolution of 16×16, the inference memory is only 0.15MB.

In addition, the teacher model, VGGFace (T-224), contains 138 million parameters, while the student network only has about 0.79 million parameters, making a great reduction of

$175\times$ in model complexity at the cost of a very small drop in recognition accuracy. Due to the extreme reductions on the memory cost and the parameter number, the computation complexity can greatly decrease. As shown in Table III, the inference runtime on both high-end GPU and low-end CPU is reduced greatly. With a NVIDIA K80 GPU, the inference time for a face is reduced from 20.4ms with T-224 (VGGFace teacher) to 0.61ms, 0.32ms, 0.15ms and 0.106ms with S-96-sc, S-64-sc, S-32-sc and S-16-sc, respectively. The inference times are also remarkably reduced even in CPU. Our model takes 0.106ms and 2.39ms to recognize a face with a very low-resolution of 16×16 in GPU and CPU respectively, which means 9, 433 faces per second and 418 faces per second.

V. CONCLUSION

At present, the problems of large model parameters and high feature dimension widely exist in face recognition models based on deep learning, which hinders their practical deployment on resource-restricted applications (*e.g.*, on embedded or mobile devices). To address this problem, this paper proposes a knowledge distillation method, adopting original large model as the teacher network and letting the teacher selectively supervise the training of student networks via designing the multi-task loss function combining regression and classification items. We have accomplished combination of high-dimensional deep feature regression and low-resolution facial classification, which achieves the uniform compression of deep network and feature dimension with recognition accuracy rate assured. Experimental results show that the proposed approach can transfer the informative knowledge from the teacher network to student models, leading to compact face recognition models with impressive effectiveness and efficiency.

In our future work, we will tentatively explore the usage of recurrent mechanism that aims to handle the failure cases in the teacher stream. Face attributes such as gender, age and makeup will be incorporated into the multi-task framework to further enhance the performance of the compressed model.

REFERENCES

[1] 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. (CVPR)*, Jun. 2014, pp. 1701–1708.
[2] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1891–1898.
[3] 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.
[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, vol. 1, no. 3, p. 6.
[5] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," Carnegie Mellon School Comput. Sci., Pittsburgh, PA, USA, 2016.
[6] A. Pentland and T. Choudhury, "Face recognition for smart environments," *Computer*, vol. 33, no. 2, pp. 50–55, Feb. 2000.
[7] D. Liu, B. Cheng, Z. Wang, H. Zhang, and T. S. Huang. (2017). "Enhance visual recognition under adverse conditions via deep networks." [Online]. Available: https://arxiv.org/abs/1712.07732
[8] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4876–4884.

[9] M. Jian and K.-M. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1761–1772, Nov. 2015.

[10] M.-C. Yang, C.-P. Wei, Y.-R. Yeh, and Y.-C. F. Wang, "Recognition at a long distance: Very low resolution face recognition and hallucination," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 237–242.

[11] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2019–2030, Oct. 2012.

[12] C.-X. Ren, D.-Q. Dai, and H. Yan, "Coupled kernel embedding for low-resolution face image recognition," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3770–3783, Aug. 2012.

[13] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2004, pp. 469–481.

[14] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2007, pp. 1–8.

[15] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 499–515.

[16] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3D morphable models with a very deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 1493–1502.

[17] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tail," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 5409–5418.

[18] G. Hu *et al.*, "Attribute-enhanced face recognition with neural tensor fusion networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 3764–3773.

[19] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27. 2014, pp. 1988–1996.

[20] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015). "DeepID3: Face recognition with very deep neural networks." [Online]. Available: https://arxiv.org/abs/1502.00873

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

[22] T. Uiboupin, P. Rasti, G. Anbarjafari, and H. Demirel, "Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring," in *Proc. IEEE Conf. Signal Process. Commun. Appl.*, May 2016, pp. 437–440.

[23] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2014, pp. 184–199.

[24] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 105–114.

[25] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.

[26] P. Zhang, X. Ben, W. Jiang, R. Yan, and Y. Zhang, "Coupled marginal discriminant mappings for low-resolution face recognition," *Opt.-Int. J. Light Electron Opt.*, vol. 126, no. 23, pp. 4352–4357, Dec. 2015.

[27] J. Jiang, R. Hu, Z. Wang, and Z. Cai, "CDMMA: Coupled discriminant multi-manifold analysis for matching low-resolution face images," *Signal Process.*, vol. 124, pp. 162–172, 2016.

[28] X. Wang, H. Hu, and J. Gu, "Pose robust low-resolution face recognition via coupled kernel-based enhanced discriminant analysis," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 2, pp. 203–212, Apr. 2016.

[29] X. Xing and K. Wang, "Couple manifold discriminant analysis with bipartite graph embedding for low-resolution face recognition," *Signal Process.*, vol. 125, pp. 329–335, Aug. 2016.

[30] J. Shi and C. Qi, "From local geometry to global structure: Learning latent subspace for low-resolution face image recognition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 554–558, May 2015.

[31] M. Haghighat and M. Abdel-Mottaleb, "Low resolution face recognition in surveillance systems using discriminant correlation analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May/Jun. 2017, pp. 912–917.

[32] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3765–3773.

[33] K.-H. Pong and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognit.*, vol. 47, no. 2, pp. 556–567, Feb. 2014.

[34] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 1034–1040, May 2016.

[35] S. Shekhar, V. M. Patel, and R. Chellappa. (2017). "Synthesis-based robust low resolution face recognition." [Online]. Available: https://arxiv.org/abs/1707.02733

[36] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4792–4800.

[37] C. Herrmann, D. Willersinn, and J. Beyerer, "Low-resolution convolutional neural networks for video face recognition," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 221–227.

[38] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. ACM Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 535–541.

[39] G. Hinton, J. Dean, and O. Vinyals, "Distilling the knowledge in a neural network," in *Proc. Neural Inf. Process. Syst. (NIPS) Workshop*, 2014, pp. 1–9.

[40] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.

[41] T. Chen, I. Goodfellow, and J. Shlens, "Net2Net: Accelerating learning via knowledge transfer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[42] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, "Face model compression by distilling knowledge from neurons," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 3560–3566.

[43] Z. Li and D. Hoiem, "Learning without forgetting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 614–629.

[44] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2016, pp. 1317–1327.

[45] G. Urban *et al.*, "Do deep convolutional nets really need to be deep and convolutional?" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–13.

[46] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 742–751.

[47] Y. Chen, N. Wang, and Z. Zhang, "DarkRank: Accelerating deep metric learning via cross sample similarities transfer," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 2852–2859.

[48] T. Chen, L. Lin, W. Zuo, X. Luo, and L. Zhang, "Learning a wavelet-like auto-encoder to accelerate deep neural networks," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 6722–6729.

[49] G. Zhou, Y. Fan, R. Cui, W. Bian, X. Zhu, and G. Kun, "Rocket launching: A universal and efficient framework for training well-performing light net," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 1–8.

[50] D. Lopezpaz, L. Bottou, B. Scholkopf, and V. Vapnik, "Unifying distillation and privileged information," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–10.

[51] J.-C. Su and S. Maji, "Adapting models to signal degradation using distillation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–14.

[52] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4119–4128.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[54] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[55] 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.

[56] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, p. 7263–7271.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[58] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[59] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "UMDFaces: An annotated face dataset for training deep networks," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 464–473.

[60] E. Learned-Miller, G. B. Huang, A. Roychowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in Face Detection and Facial Image Analysis*. Cham, Switzerland: Springer, 2016, pp. 189–248.

[61] A. Sapkota and T. E. Boult, "Large scale unconstrained open set face database," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep./Oct. 2014, pp. 1–8.

[62] M. Grgic, K. Delac, and S. Grgic, "SCface—Surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, 2011.

[63] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. USENIX Symp. Operating Syst. Design Implement. (OSDI)*, vol. 16. 2016, pp. 265–283.

[64] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1233–1245, Mar. 2016.

[65] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 566–579.

[66] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[67] B. Cheng, D. Liu, Z. Wang, H. Zhang, and T. S. Huang, "Visual recognition in very low-quality settings: Delving into the power of pre-training," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 8065–8066.

[68] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, "Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 886–896, Apr. 2019.

[69] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Majumdar, "Identity aware synthesis for cross resolution face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2018, pp. 479–488.

[70] P. Li, L. Prieto, D. Mery, and P. Flynn. (2018). "Low resolution face recognition in the wild." [Online]. Available: https://arxiv.org/abs/1805.11529

Shengwei Zhao received the B.S. degree from the School of Mathematics and Statistics, Wuhan University, in 2017. He is currently pursuing the master's degree with the Institute of Information Engineering, Chinese Academy of Sciences, and the School of Cyber Security, University of Chinese Academy of Sciences. His major research interests are deep learning and computer vision.

Chenyu Li received the B.S. degree from the School of Electronics and Information Engineering, Tongji University. She is currently pursuing the Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, and the School of Cyber Security, University of Chinese Academy of Sciences. Her research interests are computer vision and deep learning.

Shiming Ge (M'13–SM'15) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China in 2003 and 2008, respectively. He was a Senior Researcher at Shanda Innovations and a Researcher at Samsung Electronics and the Nokia Research Center. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His research mainly focuses on computer vision, deep learning, and AI security, especially high-performance deep models toward scalable applications.

Jia Li (M'12–SM'15) received the B.E. degree from Tsinghua University in 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He served at Nanyang Technological University, Peking University, and Shanda Innovations. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. He has authored or co-authored over 50 technical articles in refereed journals and conferences, such as TPAMI, TIP, IJCV, ICCV, and CVPR. His major research interests are cognitive vision toward evolvable algorithms and models. His research interests include computer vision and multimedia big data, especially the deep-learning-based visual content understanding.