Resolution Invariant Face Recognition using a Distillation Approach

Syed Safwan Khalid, Muhammad Awais, Zhen-Hua Feng, *Member, IEEE* Chi-Ho Chan,
Ammarah Farooq, Ali Akbari and Josef Kittler *Life Member, IEEE*

***Abstract*—Modern face recognition systems extract face representations using deep neural networks (DNNs) and give excellent identification and verification results, when tested on high resolution (HR) images. However, the performance of such an algorithm degrades significantly for low resolution (LR) images. A straight forward solution could be to train a DNN, using simultaneously, high and low resolution face images. This approach yields a definite improvement at lower resolutions but suffers a performance degradation for high resolution images. To overcome this shortcoming, we propose to train a network using both HR and LR images under the guidance of a fixed network, pretrained on HR face images. The guidance is provided by minimising the KL-divergence between the output Softmax probabilities of the pretrained (*i.e.,* Teacher) and trainable (*i.e.,* Student) network as well as by sharing the Softmax weights between the two networks. The resulting solution is tested on down-sampled images from FaceScrub and MegaFace datasets and shows a consistent performance improvement across various resolutions. We also tested our proposed solution on standard LR benchmarks such as TinyFace and SCFace. Our algorithm consistently outperforms the state-of-the-art methods on these datasets, confirming the effectiveness and merits of the proposed method.**

***Index Terms*—Face Recognition, Resolution Invariance, Low Resolution, Convolutional Neural Networks, Distillation.**

I. Introduction

Face recognition (FR) involves identification/verification of the subject's identity given his/her query face image. It is a task that is routinely applied in real-world applications such as surveillance, passport control, forensic investigations, access control, time-and-attendance systems and many others. Owing to its significance, FR has always been an active topic for research in computer vision [1]–[5]. Recent developments in deep neural networks (DNN) have profoundly increased the accuracy of FR tasks, and the enriched face representation obtained using DNN embeddings[1] has become the primary method for state-of-the-art (SOTA) FR algorithms [6]–[13]. Currently, even for a large scale dataset such as Megaface [14] (with more than 1 million images), the SOTA FR methods give near perfect accuracy [8], [10]. Thus DNN based FR techniques are expected to be fairly accurate even for in-the-wild challenging scenarios with large variations in illumination, pose, background etc. However, these techniques work reliably only if the available face images are of a sufficiently high resolution (HR). For low resolution (LR) images, the accuracy of SOTA FR techniques has been found to degrade significantly [15]–[17].

Since the face images available from standard CCTV cameras are usually LR and we can encounter images of various resolutions in many other FR applications as well, it is important to investigate and develop FR methods that are robust to resolution changes. It is worth noting that in a generic FR scenario, we compare a query image of an arbitrary resolution, against a gallery of images with varying resolutions. Hence the resolution of the true match in the gallery would be generally unknown. This makes a resolution-invariant (RI) FR algorithm a more suitable option as opposed to using two separate solutions, *i.e.,* one for LR-FR and one for HR-FR. Despite the significance of LR/RI FR, it is an under studied topic as compared to the standard HR-FR. A few works in the literature that deal with this challenging task can be divided into two main approaches

1) Algorithms that employ super-resolution techniques to transform LR faces into HR domain. The recognition is then performed in the HR domain [17]–[19].
2) Algorithms that try to minimise the difference between the extracted features of LR and HR face images in some lower dimensional feature space, thus making the solution, to some degree, resolution-invariant [15], [20], [21].

Both these approaches have shown some promising results; however, the latter is more straight-forward and easier to work with. Furthermore, we note from the literature that the former approach does not show any noticeable performance improvement over the latter approach [17], [21], [22]. Therefore, in this paper, we focus on the second approach.

One simple method to minimise the difference between LR and HR facial features in a DNN embedding space is to train a network simultaneously on HR, and corresponding down-sampled LR images [15]. This approach works well for LR-FR, but at the expense of accuracy for HR-FR. Essentially, in order to bring the HR and LR features close to each other, the network learns to discard some information from the HR images. To alleviate this shortcoming, we propose to use a Teacher-Student distillation network [23] when training with HR and LR faces simultaneously. The teacher network is fixed

The authors are with the Centre of Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, UK. S. S. Khalid is also with the COMSATS University, Islamabad, Pakistan. Z.-H. Feng is also with the Department of Computer Science, University of Surrey, Guildford, GU2 7XH, UK.

This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/dstl/MURI project EP/R018456/1.

[1] A DNN embedding or a feature vector refers to the n-dimensional vector in a deep neural network available just before the loss layer.

and pretrained on HR images only. A student network is trained on combined HR and down-sampled LR face images. During training, HR images are fed to the fixed teacher network and its output Softmax probabilities are used as soft-targets for training the student network.

Distillation is traditionally used to distil information from a complex model into a simpler one; however, our approach is different and employs distillation to guide the student network towards a better optimum. We utilise the same network for both student and teacher streams and hence both streams have the same complexity. Furthermore, we reinforce the guidance provided by the teacher network by sharing its Softmax weights with the student network. Consequently, both networks have the same loss-landscape and the DNN embeddings are forced to be close to each other. We test the results of our proposed scheme on two publicly available LR benchmark datasets, *i.e.,* TinyFace [16] and SCFace [24]. While our proposed scheme outperforms state-of-the-art on both these datasets, we realised that these benchmarks contain relatively small number of faces with limited variations. Hence the results on these datasets would not be a true representative of performance in uncontrolled settings. Therefore, we develop two novel LR-FR protocols by artificially down-sampling the publicly available Megaface and Facescrub datasets [14], [25]. While the down-sampled images in the resulting protocols do not contain the blurriness and lack of illumination usually found in native-LR images, still, owing to the large variations in pose and illumination, the proposed protocols are more challenging than the publicly available native-LR datasets. Figure 1, shows some example images from the synthetically down-sampled data as well as some images from the native LR datasets. Our contributions are summarised as follows:

1) We propose a novel mechanism of training a network for LR-FR using combined HR and LR data. We utilise a pretrained network as a teacher/guide and train another student network. The guidance from the teacher network is provided by sharing Softmax weights between the two networks as well as minimising the KL-divergence between the Softmax probabilities of the teacher and the student network.
2) We develop two LR-FR protocols using the publicly available Facescrub and Megaface datasets. The protocols emulate two real-world scenarios of FR under surveillance settings. Our protocols are more challenging than the available benchmarks and are used to show the efficacy as well as limitations of our proposed scheme.
3) We evaluate our proposed scheme on two benchmark datasets, *i.e.,* TinyFace and SCFace. Our proposed scheme is shown to outperform state-of-the-art for both these datasets.

The rest of the paper is organised as follows: in Section II, we give a review of the SOTA and also describe benchmark datasets for LR-FR tasks, including the novel protocols that we have developed. In Section III, we first overview a few baseline methods and then describe the proposed algorithm in detail. In Section IV, we present the experimental results and conclusions are drawn in Section V.

A Note on Abbreviations: We extensively employ the following abbreviations throughout this paper:

HR/LR:	High resolution/ Low resolution
FR:	Face Recognition
RI:	Resolution Invariant
SR:	Super Resolution
SOTA:	state-of-the-art
DNN:	Deep Neural Network
CNN:	Convolutional Neural Network

II. Related Work and Datasets

A. Face recognition

High Resolution Face Recognition (HR-FR) has seen a tremendous improvement in accuracy in the past few years owing to the advances in DNN architectures and the availability of large scale datasets for training. Modern FR algorithms train a DNN on massive datasets such as CasiaWeb [26] (around 0.5 million images) or MS-Celeb [27] (more than 5.8 million images) using either classification losses (such as Softmax) or metric/contrastive losses (such as triplet loss). Currently, variants of the Normalised Softmax loss such as SphereFace [28], CosFace [7] and ArcFace [8] are providing SOTA results on benchmarks such as LFW [29] and MegaFace [14]. For instance, ArcFace has shown a rank-1 accuracy of 91.75% and 98.35% on MegaFace when trained on CasiaWeb and MS-Celeb, respectively. However, a recent work [17] reported only 40% rank-1 accuracy for ArcFace, when tested on SCFace LR-HR protocol; notwithstanding, that SCFace is far less challenging in terms of variations as compared to MegaFace. This shows that SOTA FR methods are non-robust to resolution changes and perform poorly when tested on LR-FR tasks.

Low Resolution Face Recognition (LR-FR) is an understudied topic as compared to its HR counterpart. Since there are no large scale datasets available for training, most works rely on down-sampled versions of HR face images[2]. As discussed earlier, the approaches to tackle LR-FR can be broadly divided into two categories, *i.e.,* Super-Resolution (SR) based methods and Resolution-Invariant (RI) methods. Some earlier works in SR include [18], [31]; whereas, some early works using the RI approach are [20], [32]. More recently, DNN based methods in both SR [16], [17], [19] and in RI [15], [21] have outperformed these earlier works by a margin and replaced them as a baseline.

SR methods: Super-resolution techniques focus on creating visually appealing outputs, and in the process, these algorithms can loose identity related information. Accordingly, joining a contemporary SR algorithm with a face recognition system does not yield any significant performance improvement for LR-FR tasks. To overcome this deficiency, [19] utilises an SR-identity loss to measure and minimise the identity difference in the super-resolved HR face images. In [16], the construction of a native LR dataset, *i.e., TinyFace* is discussed; also an approach based on joint end-to-end training of an SR sub-network followed by a face recognition network is proposed.

[2] Recently, two native-LR datasets, *i.e.,* TinyFace [16] and SurvFace [30] have been made available publicly, yet the training data in these datasets is still small as compared to CasiaWeb or MS-Celeb.

The authors dub their approach as complement-super resolution and Identity (CSRI) method and report improved accuracy on TinyFace. A recent report [17] discusses a face normalisation technique which is closely related to the super resolution techniques and achieves state-of-the-art performance on SCFace. While SR methods appear to be an intuitive solution for the LR-FR problem, designing an SR method that retains identity-related information is a challenging as well as computationally demanding task. Despite the increase in complexity, SR methods do not appear to have any significant superiority over the RI methods that are discussed below.

RI methods: In [15], a DNN based RI technique was discussed that trained a network on simultaneous HR and LR images and attained much better performance as compared to the classical RI methods. It was further improved in [21] in which additional losses were incorporated, *i.e.,* centre loss and euclidean loss, to boost the performance. These DNN based RI methods gave SOTA performance on SCFace until recently, having been outperformed by the SR method in [17]. However, our proposed method outperforms [17] and sets a new SOTA result for SCFace. A shortcoming of the RI methods, that has been largely ignored in the literature, is a decrease in performance at higher resolutions and more crucially at cross resolutions. A recent unpublished work [33] has also highlighted this issue and is also relevant to our proposed scheme since [33] also proposes to use distillation to develop resolution-invariant DNN based FR; however, the way we employ distillation is significantly different than the one discussed in [33]. Essentially, [33] utilises a scaled euclidean distance for distillation, similar to [34]; whereas, we propose to use KL-divergence coupled with weight-sharing of the Softmax layer. It is worth noting, that out proposed method outperforms [33] on TinyFace.

Comparison of SR and RI methods: The SR methods are intuitively appealing and a number of SR techniques, exhibiting a decent performance, are available in the literature [18], [19], [35], [36]. Since the SR techniques are primarily focused on creating visually appealing images, they are well suited for human-assisted face recognition tasks. However, these techniques cannot be applied directly to automated face recognition and require additional mechanisms to preserve the identity related information in the super-resolved image [16], [17], [19]. The primary advantage of RI techniques is their lack of complexity as compared to the SR methods. Despite this reduced complexity, RI techniques offer similar, if not better, face identification performance, when compared to the SR methods. However, since the RI methods work in the feature domain, they do not offer any advantage in human-assisted face recognition applications.

B. Datasets

1) TinyFace: It is a native LR dataset created by extracting LR faces from the publicly available PIPA [37] and MegaFace [14] datasets. It has 15,975 labelled LR faces corresponding to 5,139 identities. Furthermore, it has 153,428 unlabelled LR faces. The face image height ranges from 6 to 32 pixels with a mean height of 20 pixels. The labelled identities are split into 2,570 IDs for training and 2,569 IDs for testing. The unlabelled faces are used as distractors for the testing protocol. The images for the IDs in the test-set are further split into probe and gallery images. A probe image is compared against all the images in the gallery and also against all the distractors to find the best match. The performance matrices are the Cumulative Characteristic Curve (CMC) and the mean Average Precision (mAP). TinyFace has the advantage of containing native LR images, as opposed to the synthetically down-sampled LR images. However, it is a small dataset with limited variations in pose and illumination. It contains only around 7-8K images for training and most of the images in the test-set are unlabelled distractors. Consequently, we do not use TinyFace for training and only use it to compare our proposed scheme against the state-of-the-art. Another drawback of TinyFace is that all images are unaligned and tightly cropped. The standard face alignment methods require loosely cropped images for proper alignment. To overcome this issue, when evaluating algorithms on TinyFace, we augment our training data with tightly cropped images to make our network robust to distortions that result from aligning tightly cropped faces. Despite its shortcomings, TinyFace is a welcome addition to the scant benchmarks available for LR-FR tasks.

2) SCFace: It is a dataset of face images captured using five video surveillance cameras that were placed at varying distances. Consequently, the dataset contains facial images of varying resolutions and is suitable for testing resolution invariant FR algorithms. There are 130 identities in total with 15 probe images corresponding to each identity at three different distances (*i.e.,* 1.0m, 2.6m and 4.2m) each. For each identity, there is an HR mugshot image available as well. The standard protocol [17], [21] is to use mugshot HR images as gallery and the surveillance camera images as probes. Rank-1 accuracy results are evaluated for each distance separately. SCFace has the advantage of containing images of various qualities and resolutions; however, there is little variation in pose. Also, all images appear to be taken at the same day with the same hairstyles and clothes. Moreover, it is a small-scale dataset and the performance evaluations on SCFace would not correspond to the in-the-wild scenarios. Despite its limited variations, the SOTA HR-FR algorithms such as ArcFace [8] give barely 48% rank-1 accuracy for images taken from 1.0m distance [17]. The algorithms trained for LR-FR perform much better with a state-of-the-art accuracy of around 77% [17].

3) MegaFace/FaceScrub: MegaFace is a large scale dataset that uses 1 million unlabelled face images as distractors in FR evaluation protocols. The distractors have large variations in pose, illumination and resolution. These distractors are used in combination with the FaceScrub [25] dataset in the MegaFace challenge [14], where the FaceScrub dataset is used as labelled probe set. FaceScrub contains 100k images of 530 identities with large variations across face images of the same identity; however, the MegaFace challenge usually takes a small subset of the FaceScrub dataset containing 80 identities. In [8], it was identified that some images are common in both the distractor and the probe set; some mislabelled images in the probe set were also found. Accordingly, a list of noisy images in the MegaFace/FaceScrub is provided by [38] which we use to clean the MegaFace/FaceScrub dataset.

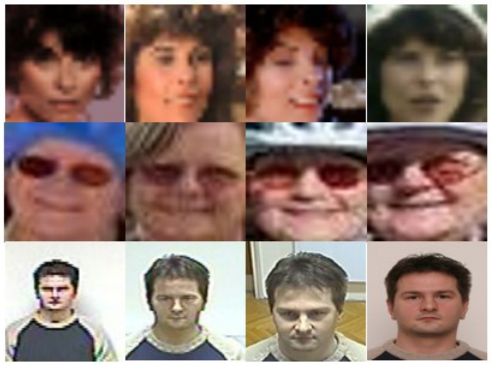


Fig. 1. Example images from various datasets with each row corresponding to a single identity. The top row contains synthetically downsampled images from FaceScrub. The middle row contains images from TinyFace and bottom row is from SCFace. Note the lack of variations in TinyFace and SCFace as opposed to FaceScrub; however, the synthetic LR does not contain the natural blur and lack of illumination present in native LR.

The MegaFace challenge is essentially an HR-FR benchmark and is not suitable for LR-FR evaluations. However, as we noted earlier, the available benchmarks for LR-FR tasks are of small-scale, with limited variations. Therefore, we decided to implement our own protocols by down-sampling the MegaFace and FaceScrub datasets. We implemented two scenarios that represent two important real-world applications:

1) **P1:LR-LR**: In this setting, we down-sampled both the gallery images as well as the probe images to some lower resolution using bi-cubic interpolation of the OpenCV library [39]. We report the results for varying resolutions that correspond to images with face widths ranging from 20 pixels to 112 pixels. This scenario represents the situation when the face of a person-of-interest (POI) is extracted from a surveillance camera footage and it is required to search for that person in some other available footage.
2) **P2:LR-HR**: In this setting, we only down-sample the probe images and the gallery is left untouched. It is a cross-resolution FR task that represents the situation where a list of POIs and their corresponding HR faces are available in a gallery and we need to search for these persons in some surveillance camera footage.

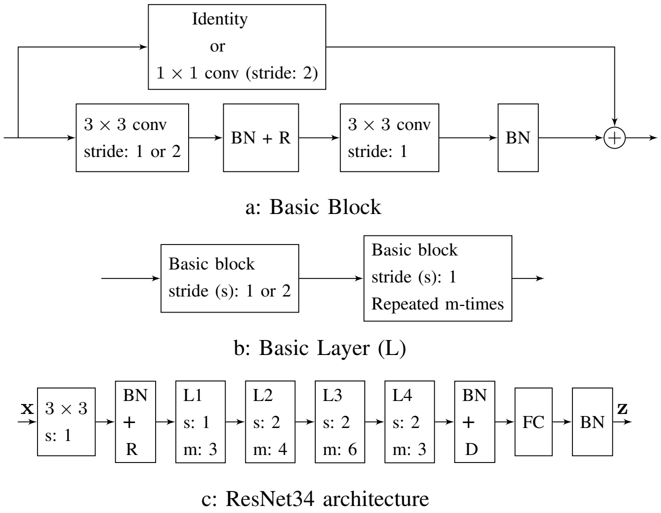
Apart from down-sampling the images, we use the same protocols for identification and verification as has been used in the MegaFace challenge. For the identification scenario, one image of a particular identity from the FaceScrub dataset is added to the list of distractors to make a gallery. All the remaining images of that particular identity are compared against the gallery images to find the best match. This process is repeated for each image of each identity in the probe set. The average rank-1 accuracy across all identities is used as a performance metric. For verification, we create pairs of images as inputs and the algorithms are required to decide whether a given pair belongs to the same identity or not. The true positive rate, the false positive rate and the corresponding Receiver Operating Characteristic (ROC) curve are used as performance metrics. To evaluate the performance, we create all pairs of all the probe images as well as pairs of probe and distractor images. The actual MegaFace challenge employed 1 million distractors; however, since we are working with a large number of different resolutions, it was computationally too demanding for us to use the complete list of distractors. Therefore, we use a subset of 10K distractors in our proposed evaluation protocols, which were selected at random.

III. Proposed Approach

We first discuss some baseline approaches that would motivate the development of our proposed network and then explain in detail the proposed solution.

A. Baseline Approaches

We work with a standard 34 layer ResNet [40] architecture with slight modifications; the details of our network are depicted in Figure 2. Working with a 34 layer network serves two important purposes: firstly, it allows us to have a fair comparison with the works on LR-FR that use a shallow network, *i.e.,* [16], [21]. Secondly, it becomes computationally feasible to perform a range of experiments with various hyper-parameters and input resolutions. We intend to study the effects of employing deeper and more powerful CNNs combined with more training data in a follow-up study.



a: Basic Block

b: Basic Layer (L)

c: ResNet34 architecture

Fig. 2. (a) The basic building block of the ResNet34 architecture used in our work. The residual connection is an identity connection if the first convolution layer in the basic block has stride equal to 1; otherwise, a 1×1 convolution with stride 2 is used to match the dimensions of the residual connection with the output. (b) A basic layer that consists of a first basic block with stride equal to either 1 or 2 and then an m-times repetition of the basic block with stride 1. (c) The overall architecture of the ResNet34 architecture. BN represents batch-norm, R is relu, D is dropout and FC is the fully connected layer. Input $\mathbf{x}$ is a 112×112 image and output $\mathbf{z}$ is a 512-dim feature vector.

For the network in Figure 2, the input $\mathbf{x}$ is a 112×112 image and the output $\mathbf{z}$ is a 512-dimensional output feature vector. Let $\mathbf{x}_i$ be the ith image of a batch of N training images fed to the network, let $y_i \in \{1, ,2, \cdots, K\}$ be the true class label of $\mathbf{x}_i$, and $\mathbf{z}_i$ be the output feature vector of the network in Figure 2. The feature vector is passed through a Softmax layer to get the output probabilities

$p_k^{(i)} = \exp(\mathbf{w}_k^T \mathbf{z}_i) / \sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{z}_i)$, where $\mathbf{w}_k$ is the weight of the Softmax layer corresponding to the kth class. The loss is computed by evaluating cross-entropy between $p_k^{(i)}$ and the true class labels q_k, *i.e.*, $l_i = -\sum_{k=1}^{K} q_k \log(p_k^{(i)})$, where $q_k = 1$ for $k = y_i$ and zero otherwise. The overall loss for a single batch can be written as

$$L = \frac{1}{N} \sum_{i=1}^{N} l_i = \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{w}_{y_i}^T \mathbf{x}_i)}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}_i)} \qquad (1)$$

The network is trained by minimizing the loss through stochastic gradient descent. We consider the following baseline networks:

1) A network trained on HR images only from the CasiaWeb dataset. Lets call it HR-only network.
2) A network trained on down-sampled images from CasiaWeb with two different resolutions corresponding to images with face widths equal to 20 pixels and 16 pixels, respectively. Note that the images are resized using bicubic interpolation to the required size of 112 × 112 pixels before being fed to the network. We call this network LR-only.
3) A network trained on combined HR and down-sampled LR images. The down-sampling is done to get two different resolutions corresponding to face widths of 20 pixels and 16 pixels. This baseline is similar to the approach discussed in [15]. Lets call it HR+LR network.

It is obvious that the HR-only network would perform well for high resolution images and the LR-only network for low resolution facial images. We would expect the HR+LR network to perform well for both resolutions. However, as shown in detail in the Experiment section (i.e., Section IV-A), the HR+LR network outperforms the LR-only network for low resolution images, but its performance for high resolution images is much lower than the HR-only network. We explain this behaviour as follows: Let $\mathbf{x}_{i,hr}$ and $\mathbf{x}_{i,lr}$ be a set of HR and corresponding LR training examples of a particular identity y_i. Let $\mathbf{z}_{i,hr}$, $\mathbf{z}_{i,lr}$, $p_{k,hr}^{(i)}$ and $p_{k,lr}^{(i)}$ be the corresponding DNN embeddings and Softmax probabilities, respectively. For these inputs, the cross entropy loss will be minimised if both $p_{k,hr}^{(i)} \rightarrow 1$ and $p_{k,lr}^{(i)} \rightarrow 1$ for $k = y_i$. Consequently, from (1), $\mathbf{w}_{y_i}^T \mathbf{z}_{i,hr} >> \mathbf{w}_{k \neq y_i}^T \mathbf{z}_{i,hr}$ and $\mathbf{w}_{y_i}^T \mathbf{z}_{i,lr} >> \mathbf{w}_{k \neq y_i}^T \mathbf{z}_{i,lr}$. Hence, the loss is minimised only when the angle between $\mathbf{z}_{i,hr/lr}$ and $\mathbf{w}_{y_i}$ is minimised. Since both $\mathbf{z}_{i,hr}$ and $\mathbf{z}_{i,lr}$ approach the same weight vector $\mathbf{w}_{y_i}$, they are forced to be close to each other. Essentially, when the same network is trained on both HR and LR images, the Softmax layer would bring the DNN embeddings of a single identity close to each other, for both the LR and HR images. Now as the LR embeddings get close to HR, it results in a performance improvement at LR; however, the network also forces the HR embeddings to be close to LR, thus resulting in a loss of performance at HR. In other words, the network would ignore some information in the HR images so that the HR and LR features can get close to each other. This scenario is graphically represented in Figure 3(a).

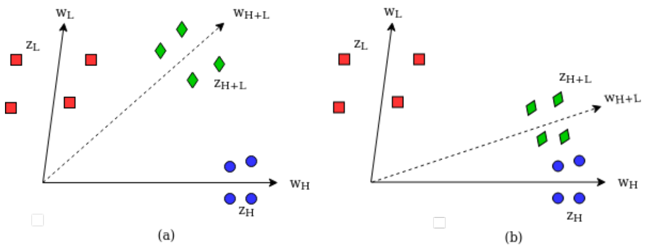


Fig. 3. A graphical representation of the effect of Softmax weight sharing. (w_L, z_L) depict the weights (Softmax) and features of a network trained on LR images, (w_H, z_H) represent the output of a network trained on HR images, and (w_{H+L}, z_{H+L}) depict the output of a network trained on combined LR and HR images. (a) In the absence of weight sharing, the LR weights and features move toward HR, thus increasing the performance at LR; however, the HR weights and features also move towards LR, resulting in a performance loss at HR. (b) Using Softmax weight sharing, we can force the (w_{H+L}, z_{H+L}) to be close to (w_H, z_H) thus resulting in an increase in performance at LR while maintaining the performance at HR better.

B. Proposed Solution

In an ideal solution for cross-resolution face recognition, we would like the HR features to remain similar to those of the HR-only network; whereas, we would like the LR features to get as close as possible, to the HR features. To achieve this end, we propose a strategy depicted in Figure 4.

We take a pretrained network that has been trained on HR images as a teacher network with fixed weights. The teacher network is fed HR images only. A trainable student network is simultaneously fed the same HR images with the corresponding down-sampled LR images, in the same batch. We train the student network by minimising the KL-divergence between the output Softmax probabilities of the teacher network and those of the student network. The resulting loss function for a single input $\mathbf{x}$ can be written as

$$\begin{aligned} l &= -\sum_{k=1}^{K} q_k^{pre} \log \frac{p_k}{q_k^{pre}} \\ &= \underbrace{-\sum_{k=1}^{K} q_k^{pre} \log(p_k)}_{\text{Cross Entropy}} + \underbrace{\sum_{k=1}^{K} q_k^{pre} \log(q_k^{pre})}_{\text{Entropy}}, \end{aligned} \qquad (2)$$

where K is the total number of classes, $p_k \propto \exp(\mathbf{w}_k^T \mathbf{z}_t)$ is the output of trainable Softmax, $q_k^{pre} \propto \exp(\mathbf{w}_{pre,k}^T \mathbf{z}_{pre})$ is the output of the pretrained Softmax and $\mathbf{w}_k, \mathbf{w}_{pre}$ and $\mathbf{z}_t, \mathbf{z}_{pre}$ are the corresponding trainable and pretrained weights and features, respectively. Note that the second term in the loss function, *i.e.,* the entropy of pretrained probabilities is independent of the trainable student network and hence does not play any part in training. So the KL loss function is essentially a cross entropy loss where the true target labels q_k have been replaced with the pretrained probabilities of the teacher network q_k^{pre}. Minimising the KL loss is equivalent to minimising the difference between the pretrained and trainable logits, *i.e.,* $\mathbf{w}_{pre,k}^T \mathbf{z}_{pre}$ and $\mathbf{w}_k^T \mathbf{z}_t$ [23]. However, minimising the difference between the logits does not necessarily make the pretrained and trainable features to be close to each other. Since in FR we are primarily concerned with the feature vectors, we would expect a performance improvement in the trainable system if both $\mathbf{z}_{pre}$ and $\mathbf{z}_t$ are forced to be close to each other. To achieve this end, we share the trainable Softmax layer between the student and the teacher network, as shown in Figure 4. Now for each input image $\mathbf{x}$, we have an HR image fed to the teacher network and an HR/LR image fed to the student network. Hence for each $\mathbf{x}$, the loss will consist of two cross entropy terms, *i.e.,*

$$l = -\sum_{k=1}^{K} q_k^{pre} \log(p_k^t) - \sum_{k=1}^{K} q_k^{pre} \log(p_k^{pre}) + C, \quad (3)$$

where $p_k^t \propto \exp\left(\mathbf{w}_k^T \mathbf{z}_t\right)$, $p_k^{pre} \propto \exp\left(\mathbf{w}_k^T \mathbf{z}_{pre}\right)$ and C represents the entropy terms that are independent of trainable weights. Now let $\gamma_k^t = \mathbf{w}_k^T \mathbf{z}_t$ and $\gamma_k^{pre} = \mathbf{w}_k^T \mathbf{z}_{pre}$ be the logits corresponding to p_k^t and p_k^{pre}, then we can write

$$\frac{\partial l}{\partial \gamma_k^t} = (p_k^t - q_k^{pre}), \qquad \frac{\partial l}{\partial \gamma_k^{pre}} = (p_k^{pre} - q_k^{pre}). \quad (4)$$

Hence the loss will be minimised when both p_k^t and p_k^{pre} approach the same value, *i.e.,* q_k^{pre}. Since $\mathbf{w}_k$ is common in p_k^t and p_k^{pre}, minimising the loss will necessarily make $\mathbf{z}_t$ and $\mathbf{z}_{pre}$ similar to each other. Note that since $\mathbf{z}_{pre}$ is fixed, during training, $\mathbf{z}_t$ will approach $\mathbf{z}_{pre}$ and $\mathbf{w}_k$ will approach $\mathbf{w}_{pre,k}$.This mechanism is graphically depicted in Figure 3(b). Also, note that sharing the Softmax layer achieves the desired effect of bringing the pretrained and trainable features close to each other without requiring any changes in the nature of the loss function. To summarise, we achieve distillation of information from the Teacher network to the Student network using two simultaneous mechanisms, *i.e.,* a KL-loss between the output Softmax probabilities of the Teacher and Student network and a shared Softmax layer between the two networks. The significance of these mechanisms is described below:

1) As the inputs and weights of the Teacher and the Student networks are different, the pretrained features $\mathbf{z}_{pre}$ and trainable features $\mathbf{z}_t$ should lie in entirely different subspaces. However, by sharing the weights of the Softmax layer, we make the outputs of the two networks to share the same subspace and force $\mathbf{z}_{pre}$ and $\mathbf{z}_t$ of the same identity to be close to each other. Note that in the previous works [21], [33], it has been suggested to use a pair-wise metric loss to minimise the distance between the HR and LR features corresponding to the same image. However, a shared Softmax layer will not only minimise the distance between the HR and LR features of the same image, but it will also minimise the distance between all the HR and LR features belonging to the same identity. Hence, it is a more efficient mechanism for achieving cross-resolution face recognition than the SOTA solutions [21], [33]
2) The KL-loss minimises the difference between the Softmax probabilities of the Teacher and the Student network. In the absence of this loss, the networks would be trained using hard-target probabilities (i.e., true class labels). However, there is valuable information in the soft-targets, i.e., in the similarity structure of the Softmax probabilities, available from the Teacher network. By minimizing the KL-loss, we expect the student network to have similar generalisation performance as that of the Teacher network.

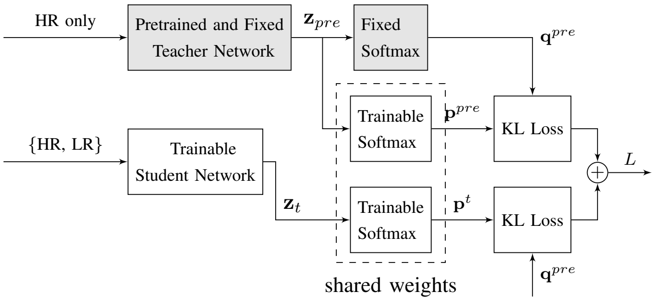


Fig. 4. Block diagram of the training strategy utilised in the proposed approach. The shaded blocks denote fixed weights that have been pretrained on HR-only images.

A note on Face alignment: For training as well as for testing, it is a standard practice in DNN based FR systems to perform a geometric face alignment before passing the image through the network. This alignment is carried out by first extracting facial landmarks and then performing an affine transformation based on the coordinates of five facial landmarks, *i.e.,* eye centres, nose tip and mouth corners. There are many existing facial landmark detection algorithms, such as cascaded shape regression [41], [42] and CNN based methods [43]–[46]. However, standard facial landmark detection approaches such as MTCNN [43] do not work well for LR images, *i.e.,* they fail to detect a large number of LR faces during training as well as testing. One possible solution is first to align the images in the HR domain and then down-sample to create the synthetic LR images. However, this approach would lead to optimistic results on synthetic LR [47] and it cannot be applied to native LR anyhow. Consequently, for low resolution facial landmark detection we train a CNN using Wing loss [48] on the WiderFace [49] dataset. To be more specific, we used a simple CNN-6 model described in [48] which performed well for LR face images.

IV. Experiments

We compare the proposed scheme against the baseline methods discussed in Section III-A as well against a number of state-of-the-art techniques. To give a fair comparison with the state-of-the-art, we use CasiaWeb dataset (*i.e.,* small protocol with < 0.5 million images) to train our network. All networks are trained using PyTorch [50] with a batch size of 64. The learning rate is set to 0.1 and is divided by 10 after 30 epochs and then again after 45 epochs. The training is finished after 65 epochs. The momentum is set to 0.9 and weight decay is set to $1e^{-4}$. Once the training is complete, the Softmax layer is discarded and for each test image, a 512-dimensional feature vector is obtained by passing the image through the network. Recognition is performed by evaluating cosine similarities between the feature vectors for the various test images.

A. Experiments on MegaFace/FaceSrcub

We compare the performance of our proposed scheme against the baseline approaches on protocols P1:LR-LR and P2:LR-HR described in Section II-B3; the results are depicted in Figure 5 and Figure 6, respectively. We note from Figure 5 that the network trained on only HR images gives an accuracy of around 90% for images with corresponding face width of 100 pixels, the accuracy remains almost unchanged as face width decreases to 60 pixels; however, the accuracy starts to decrease rapidly for lower resolutions and drops to 25% for face widths equal to 20 pixels. On the other hand, the network trained on only LR images performs much better at lower resolutions with an accuracy of around 46% for face width equal to 20 pixels. However, it gives poor performance for higher resolutions, with only 68% rank-1 accuracy for HR images. The network trained on combined HR and LR images has an accuracy of around 78% for HR images, which is much better than the LR-only network; however, it is much lower than the network trained on HR-only images. Interestingly, the HR+LR network outperforms the LR-only network on both low and high resolutions. The proposed scheme gives a rank-1 accuracy of around 51% for face width equal to 20 pixels, which is better than all the baseline approaches. At high resolution, for face width equal to 100 pixels, the accuracy of the proposed scheme is around 83% which is greater than the baseline HR+LR approach. Note that the proposed scheme gives a performance improvement over the baseline HR+LR approach for all resolutions. However, at high resolution, the accuracy is still lower than the HR-only approach. Hence by distilling the information from a pretrained network, we have been able to achieve performance improvement over the baseline HR+LR approach; however, we still have compromised some of our HR performance to gain improvements at other resolutions. Similar results are depicted in Figure 6, with the difference that for face width equal to 20 pixel, the rank-1 accuracy of the proposed scheme for P2:LR-HR is 64%, which is greater than the corresponding accuracy for P1:LR-LR scenario. Here, we would like to emphasise, that in most practical settings, the face images captured by surveillance cameras would almost always be of varying resolutions and hence FR in surveillance scenarios is a cross-resolution problem. Accordingly, the performance improvements shown by the proposed scheme for the cross-resolution protocol P2:LR-HR are of practical importance. We show the results of the verification experiments in Figure 7 and 8 for P1:LR-LR and P2:LR-HR, respectively. The results have been evaluated for downsampled images corresponding to a face width of 25 pixels. Note that the proposed scheme outperforms the baseline approaches in both scenarios. For the LR-LR scenario, the proposed scheme has a true positive rate (TPR) of 50% at a false positive rate (FPR) of 10^{-4}; whereas, both the HR+LR and the LR-only baselines have a TPR of around 45%. The HR-only network gives a very low accuracy of around 30%. For the LR-HR scenario, the proposed scheme has a TPR of 60% at an FPR of 10^{-4}. While the corresponding results for the HR+LR, LR-only and HR-only networks are 54%, 48% and 52%, respectively. These results confirm the superior performance of the proposed method against similar baseline methods, for both face identification and verification.

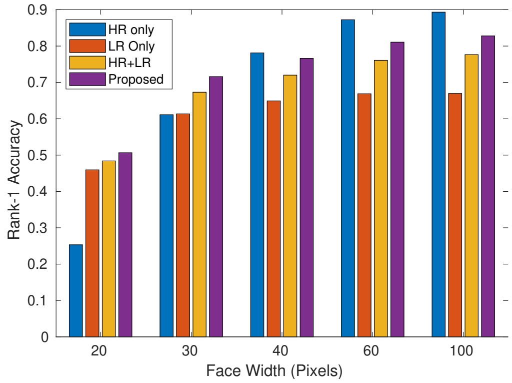


Fig. 5. Comparison of Rank-1 accuracy of the proposed approach against the baseline approaches on protocol P1:LR-LR.

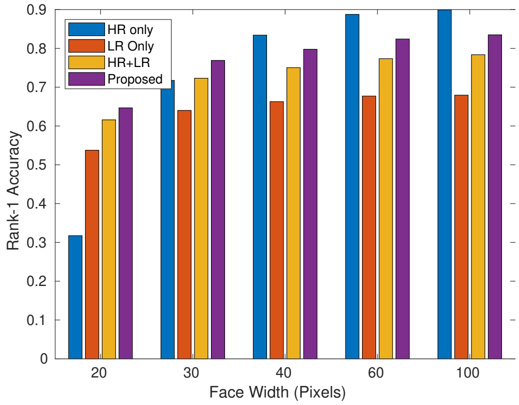


Fig. 6. Comparison of Rank-1 accuracy of the proposed approach against the baseline approaches on protocol P2:LR-HR.

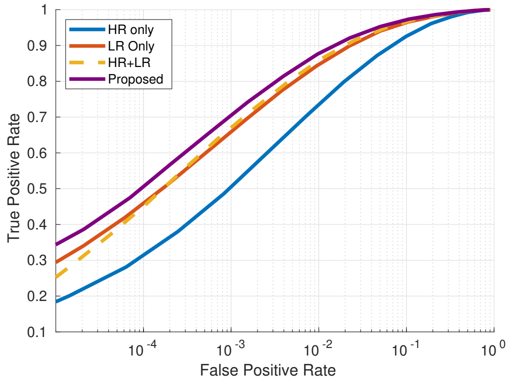


Fig. 7. Verification performance of the proposed approach against the baseline approaches on protocol P1:LR-LR for face width equal to 25 pixels.

Ablation study: As described earlier, our proposed scheme relies on two mechanisms acting simultaneously, i.e., KL-loss and shared Softmax. To evaluate the contribution of each mechanism separately, we do an ablation study in which we compare the performance of the following networks

1) A Student network trained with KL-loss only and no shared Softmax.

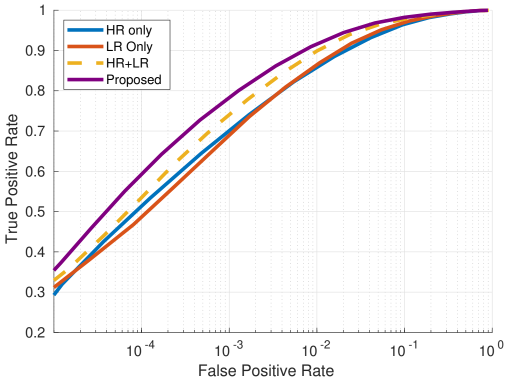


Fig. 8. Verification performance of the proposed approach against the baseline approaches on protocol P2:LR-HR for query face width equal to 25 pixels.

2) A Student network trained with shared Softmax only and no KL-loss, *i.e.,* the loss function is cross-entropy with true target labels.
3) A Student network with both KL-loss and shared Softmax, *i.e.,* the proposed scheme.

We evaluate the performance of these networks on P1:LR-LR and P2:LR-HR and the results are plotted in Figure 9 and Figure 10, respectively. For comparison, we also plot the results of the baseline HR+LR approach. We note from Figures 9 and 10 that both KL-loss and shared Softmax give incremental performance improvement over the baseline HR+LR approach. The proposed scheme, which combines both shared Softmax and KL-loss gives a better performance improvement as opposed to the individual improvements offered by these mechanism acting separately. For instance, for P1:LR-LR, at LR with face width equal to 20 pixels, the baseline HR+LR has a rank-1 accuracy of 48.40%. The distillation approach based on KL-loss-only improves the performance to 49.24%; whereas, distillation using shared- Softmax-only gives an accuracy of 49.22%. The proposed scheme on the other hand has an accuracy of 50.65%. Hence, the combined effect of shared Softmax and KL-loss offers a better distillation of information from the pretrained network. Similarly, for HR images with face width equal to 100 pixels, the accuracy of baseline is 77.64%, that of KL-loss-only is 79.38%, shared Softmax gives 80.62% and the proposed scheme has 82.80% rank-1 accuracy. Note that in the proposed scheme, we are using "soft-targets" from the pretrained network; whereas, in the shared-Softmax-only network, we use "hard-target", *i.e.,* true target labels. Since the proposed scheme is outperforming the shared-Softmax-only network, it reinforces the hypothesis suggested in [23], that there is valuable information in the similarity structure of the probabilities of the Teacher network.

B. Experiments on TinyFace and SCFace

In the previous section, we evaluated the performance of the proposed scheme on synthetically down-sampled images from FaceScrub and MegaFace. Now we focus on two native LR datasets, *i.e.,* TinyFace and SCFace. The details of these datasets have already been described in Section II-B. TinyFace has LR images in both gallery and query and hence is an LR-LR benchmark. In contrast, SCFace has query images of varying resolutions and an HR gallery, making it an LR-HR benchmark. Note that TinyFace has been published relatively recently and only a few works have reported results for TinyFace. On the other hand, SCFace is a more established benchmark and many of recent works on LR-FR use SCFace to evaluate their algorithms.

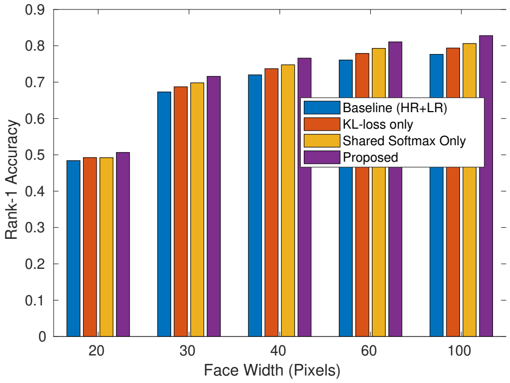


Fig. 9. Ablation study of the effect of KL-loss and shared Softmax in the proposed scheme for P1:LR-LR.

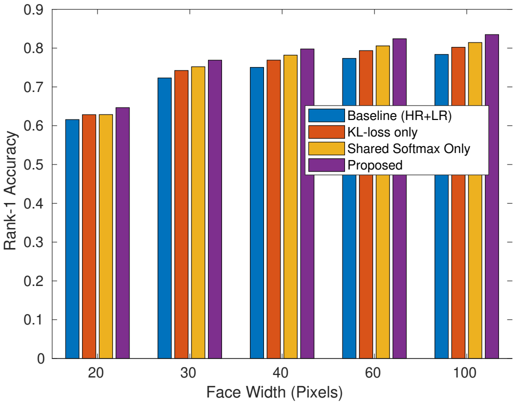


Fig. 10. Ablation study of the effect of KL-loss and shared Softmax in the proposed scheme for P2:HR-LR.

The results of the proposed scheme as well as SOTA algorithms on TinyFace are tabulated in Table I. Until recently, the best accuracy on TinyFace was reported by [16] using the super-resolution based CSRI approach. Recently, the unpublished work in [33] has reported SOTA results on TinyFace with a rank-1 accuracy of 58.6%. Note that the accuracy of our proposed scheme on TinyFace is 70.4% which is significantly better than the SOTA results. The proposed scheme has an mAP score of 63.2 which is approximately 10 points better than the previous state-of-the-art. Hence our proposed establishes a new SOTA accuracy on TinyFace. This also shows that the proposed scheme, despite being trained on synthetically down-sampled LR images only, works effectively for native LR images. Note that the performance on TinyFace is better than the performance /evaluated for the P1:LR-LR protocol (cf. Figure 5), notwithstanding, that P1:LR-LR uses synthetically down-sampled images and does not contain the usual blur and lack of illumination found in native LR datasets. This is attributed to the fact that TinyFace has a limited variation in pose and illumination, as compared to the Facescrub and MegaFace datasets. Hence, despite being a native LR dataset, it is not as challenging as a large scale LR dataset created out of down-sampled images from FaceScrub/MegaFace.

TABLE I
COMPARISON OF THE PROPOSED SCHEME WITH STATE-OF-THE-ART ALGORITHMS ON TINYFACE.

Method	Rank-1	Rank-20	Rank-50	mAP
RPCN	18.6	25.3	27.4	12.9
VGGFace	30.4	40.4	42.7	23.1
CentreFace	32.1	44.5	48.4	24.6
CSRI [16]	44.8	60.4	65.1	36.2
C-T [33]	58.6	73.0	76.3	52.7
Proposed	**70.4**	**82.2**	**85.4**	**63.2**

The performance of the proposed scheme on SCFace and its comparison with various SOTA algorithms is tabulated in Table II. Note that the proposed scheme outperforms both the SR based FAN [17] and RI based DCR [21] for all distances, *i.e.,* d1, d2 and d3 that correspond to different resolutions. Specifically, the difference in performance is significant for distance d1 that corresponds to LR images. Hence, not only the proposed scheme is working better than the SOTA methods for LR images, it is also exhibiting better resolution invariance than any of the other techniques. This also shows that the proposed scheme would work well for images obtained from surveillance cameras in an uncontrolled scenario. Again, we note that the performance of the proposed scheme for SCFace is much better as compared to TinyFace (refer to Table I) and the P1:LR-LR and P2:LR-HR protocols (cf. Figure 5 and 6). This is owing to the fact that SCFace is a small scale dataset with limited variations. Specifically, SCFace contains only frontal images and does not account for performance loss owing to pose variations that play a vital role in limiting the performance of an FR algorithm.

TABLE II
COMPARISON OF THE PROPOSED SCHEME WITH STATE-OF-THE-ART ALGORITHMS ON SCFACE.

Distance	d1	d2	d3	avg.
RICNN [15]	23.0	66.0	74.0	54.3
LDMDS [22]	62.7	70.7	65.5	65.5
LightCNN-FT [21]	49.0	83.8	93.5	75.4
ArcFace (Resnet50) [8], [17]	48.0	92.0	99.3	79.8
ArcFace-FT (Resnet50) [17]	67.3	93.5	98.0	86.3
DCR-FT [21]	73.3	93.5	98.0	88.3
FAN-FT [17]	77.5	95.0	98.3	90.3
Proposed	**88.3**	**98.3**	**98.6**	**95.0**

V. CONCLUSION

In this work, we addressed the problem of resolution invariant face recognition, specifically focusing on low resolution face identification. We proposed a novel strategy that employs a fixed Teacher network, which is pretrained on high resolution images and a trainable Student network, which is trained simultaneously on high and low resolution images. Information in the Teacher network is distilled into the Student network by minimising the KL-divergence between the output Softmax probabilities of the two networks as well as by sharing their Softmax weights. The resulting solution was tested against a number of baseline methods on synthetic LR images from FaceScrub/MegaFace and showed consistent performance improvements. The proposed scheme was also tested on native LR benchmarks, *i.e.,* TinyFace and SCFace and showed considerable performance improvements over the state-of-the-art.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[2] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 2041–2056, 2015.

[3] X. Song, Z.-H. Feng, G. Hu, and X.-J. Wu, "Half-face dictionary integration for representation-based classification," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 142–152, 2017.

[4] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin, "Gaussian mixture 3d morphable face model," *Pattern Recognition*, vol. 74, pp. 617–628, 2018.

[5] X. Song, Z.-H. Feng, G. Hu, J. Kittler, and X.-J. Wu, "Dictionary integration using 3d morphable face models for pose-invariant collaborative-representation-based classification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2734–2745, 2018.

[6] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[7] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[9] R. Ranjan, A. Bansal, J. Zheng, H. Xu, J. Gleason, B. Lu, A. Nanduri, J.-C. Chen, C. D. Castillo, and R. Chellappa, "A fast and accurate system for face detection, identification, and verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 2, pp. 82–96, 2019.

[10] X. Cheng, J. Lu, B. Yuan, and J. Zhou, "Face segmentor-enhanced deep feature learning for face recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 4, pp. 223–237, 2019.

[11] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Context-aware local binary feature learning for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1139–1153, 2018.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[13] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4269–4282, 2017.

[14] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[15] D. Zeng, H. Chen, and Q. Zhao, "Towards resolution invariant face recognition in uncontrolled scenarios," in *2016 International Conference on Biometrics (ICB)*. IEEE, 2016, pp. 1–8.

[16] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 605–621.

[17] X. Yin, Y. Tai, Y. Huang, and X. Liu, "Fan: Feature adaptation network for surveillance face recognition and normalization," *arXiv preprint arXiv:1911.11680*, 2019.

[18] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[19] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 183–198.

[20] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal processing letters*, vol. 17, no. 1, pp. 20–23, 2009.

[21] Z. Lu, X. Jiang, and A. Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 526–530, 2018.

[22] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 388–392, 2017.

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[24] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[25] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 343–347.

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

[27] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world," *Electronic imaging*, vol. 2016, no. 11, pp. 1–6, 2016.

[28] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[29] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[30] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," *arXiv preprint arXiv:1804.09691*, 2018.

[31] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *IEEE transactions on image processing*, vol. 12, no. 5, pp. 597–606, 2003.

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

[33] F. V. Massoli, G. Amato, and F. Falchi, "Cross-resolution learning for face recognition," *arXiv preprint arXiv:1912.02851*, 2019.

[34] M. Zhu, K. Han, C. Zhang, J. Lin, and Y. Wang, "Low-resolution visual recognition via deep feature distillation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3762–3766.

[35] J. Cai, H. Han, S. Shan, and X. Chen, "Fcsr-gan: Joint face completion and super-resolution via multi-task learning," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019.

[36] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

[37] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, "Beyond frontal faces: Improving person recognition using multiple cues," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4804–4813.

[38] "Insightface," https://github.com/deepinsight/insightface/tree/master/src/megaface, 2018.

[39] Itseez, "Open source computer vision library," https://github.com/itseez/opencv, 2015.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539.

[42] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, "Dynamic Attention-Controlled Cascaded Shape Regression Exploiting Training Data Augmentation and Fuzzy-Set Sample Weighting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2481–2490.

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

[44] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2129–2138.

[45] Z.-H. Feng, J. Kittler, and X.-J. Wu, "Mining Hard Augmented Samples for Robust Facial Landmark Localisation with CNNs," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 450–454, 2019.

[46] Z.-H. Feng, J. Kittler, M. Awais, and X.-J. Wu, "Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks," *International Journal of Computer Vision*, 2019.

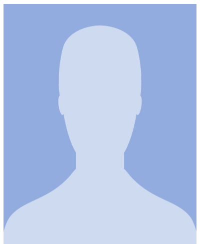
[47] Y. Peng, L. J. Spreeuwers, and R. N. Veldhuis, "Low-resolution face recognition and the importance of proper alignment," *IET biometrics*, vol. 8, no. 4, pp. 267–276, 2019.

[48] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2235–2245.

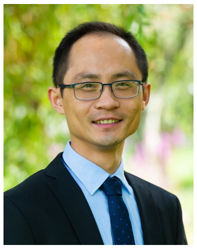
[49] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[50] A. e. Paszke, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Syed Safwan Khalid received the B.Sc. degree in Electrical Engineering from National University of Sciences and Technology, Pakistan in 2006, M.Sc. degree from University of Surrey, UK in 2009 and Ph.D. degree from COMSATS University Islamabad, Pakistan in 2018. He is currently a Research Fellow at the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey, U.K. His research interests include Deep Learning, Machine Learning and Bayesian Signal Processing.



Muhammad Awais received the B.Sc. degree in mathematics and physics from the AJK University in 2001, B.Sc. degree in computer engineering from UET Taxila in 2005, M.Sc in signal processing and machine intelligence and PhD in machine learning from the University of Surrey in 2008 and 2011. He is currently a senior research fellow at the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey. His research interests include image processing, computer vision, pattern recognition, machine learning and deep learning.



Zhen-Hua Feng (S'13-M'16) received the Ph.D. degree from the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, U.K. in 2016. He is currently a Senior Research Fellow at CVSSP, the University of Surrey. His research interests include computer vision, machine learning and pattern recognition.

He has published more than 40 scientific papers in top-ranking conferences and journals, including IJCV, CVPR, ICCV, IEEE TIP, IEEE TIFS, IEEE TCSVT, IEEE TCYB, ACM TOMM, Pattern Recognition, Information Sciences, etc. He has received the 2017 European Biometrics Industry Award from the European Association for Biometrics (EAB) and the 2018 AMDO Best Paper Award for Commercial Application.

Chi-Ho Chan received his Ph.D. degree from the University of Surrey, U.K. in 2008. He is currently a research fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey. From 2002 to 2004, he served as a researcher at ATR International (Japan). His research interests include Image Processing, Pattern Recognition, Biometrics, and Vision-Based Human-Computer Inter- action.

Ammarah Farooq is currently pursuing her doctoral studies at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, U.K. She is working on deep learning for Biometrics applications, specifically focusing on learning embeddings from multi-modal data. Her research interests include deep learning, pattern recognition, natural language processing and artificial intelligence.

Ali Akbari received the PhD degree in Telecommunications from the Sorbonne University, Paris, France in March 2018. Since July 2018, he joined the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK as a research fellow to enrich his experiences in the field of face recognition. He has published two book chapters and several papers in peer-reviewed journals and conference proceedings. He has served as an Associate Editor for the IEEE Open Journal of Circuits and Systems. His research interests include computer vision, deep learning, dictionary learning and image and video coding.

Josef Kittler (M'74-LM'12) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He published the textbook Pattern Recognition: A Statistical Approach and over 700 scientific papers. His publications have been cited more than 66,000 times (Google Scholar).

He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, Pattern Analysis and Applications. He also served as a member of the Editorial Board of IEEE Transactions on Pattern Analysis and Machine Intelligence during 1982-1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982-2005, President of the IAPR during 1994-1996.