CCFace: Classification Consistency for Low-Resolution Face Recognition

Anonymous IJCB 2023 submission

Abstract

In recent years, deep face recognition methods have demonstrated impressive results on in-the-wild datasets. However, these methods have shown a significant decline in performance when applied to real-world low-resolution benchmarks like TinyFace or SCFace. To address this challenge, we propose a novel classification consistency knowledge distillation approach that transfers the learned classifier from a high-resolution model to a low-resolution network. This approach helps in finding discriminative representations for low-resolution instances. To further improve the performance, we designed a knowledge distillation loss using the adaptive angular penalty inspired by the success of the popular angular margin loss function. The adaptive penalty reduces overfitting on low-resolution samples and alleviates the convergence issue of the model integrated with data augmentation. Additionally, we utilize an asymmetric cross-resolution learning approach based on the state-of-the-art semi-supervised representation learning paradigm to improve discriminability on low-resolution instances and prevent them from forming a cluster. Our proposed method outperforms state-of-the-art approaches on low-resolution benchmarks, with a three percent improvement on TinyFace while maintaining performance on high-resolution benchmarks.

1. Introduction

One of the key factors of the recent advances in Face Recognition (FR) is the introduction of large-scale datasets [6, 25, 26, 27]. Publicly available training benchmarks, such as CASIA-WebFace [54], VGGFace [3], MS1MV2/3 [12, 6], and WebFace4M [60], are rich in width (thousands of identities) and depth (number of images per identities) [9]. However, large-scale training datasets consist of web-crawled images and mostly contain high-resolution instances [36]. As a result, there is a notable difference between training and real-world testing statistics [7]. In particular, the images captured by security cameras exhibit a lower image resolution than the samples used for training, as seen in Fig. 2 [4, 29, 61]. This disparity leads to a huge performance gap in FR between High-Resolution (HR) and Low-Resolution (LR) since LR data are underrepresented during training [39]. Recently, studies have focused on developing better training objectives to solve this generalization problem [38, 39].

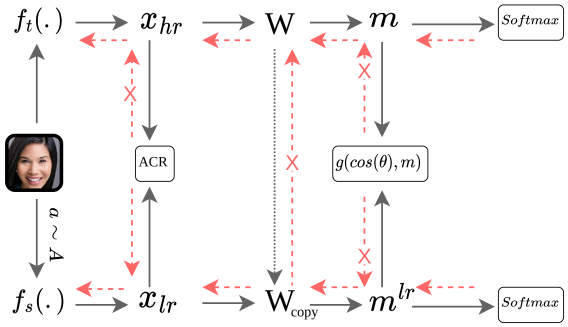

Figure 1. Our method aims to improve FR performance on LR input while maintaining the discriminability of the original HR embeddings. To achieve this, we propose to share the class proxies between student and teacher networks while asymmetrically pushing the LR feature embeddings to have higher mutual information with their HR counterparts.

Whether a set of samples or a proxy represent an identity, FR training criteria can be categorized as proxy-less or proxy-based methods [9]. The former is based on pairwise similarities, such as contrastive or triplet loss learning [38, 40, 51, 43]. In the latter, a prototype (proxy) represents a person's identity and the network tries to learn a classification task (weights of the classifier represent the identities' proxy). In large-scale datasets, challenges concerning computationally expensive sample-mining of the proxy-less loss functions have led the current state-of-the-art (SOTA) FR training loss functions toward proxy-based approaches, such as L-Softmax [27], SphereFace [26], CosFace [48], and ArcFace [6]. Despite the remarkable improvement in numerous benchmarks, such as CFP-FP, LFW, CPLFW, and CALFW, significant performance degradation occurs when a FR is trained on these datasets and then applied to LR images [42, 39].

Two primary approaches have been explored to combat

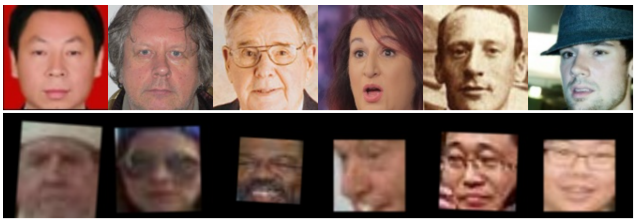

Figure 2. **Top**: Samples from the training datasets are mainly high-quality images. **Bottom**: TinyFace contains real-world low-quality instances. Comparing the top with the bottom, the gap between training and low-resolution testing benchmarks is apparent.

this: 1) construction-based and 2) projection-based methods. Construction-based methods involve enhancing the visual quality of the LR input before recognition, i.e., Face Super Resolution (FSR). This way, the FR process is separated into two tasks: identity-preserving FSR and Super Resolved Face Recognition (SRFR). Among face generation modules, special attention has been placed on Generative Adversarial Networks (GANs) [46, 18]. Despite the remarkable outputs concerning image quality and human perception, GANs add high-frequency components to the synthesized images, which negatively affects the recognition process [50]. Furthermore, since multiple HR faces exist for each LR image, FSR is an ill-posed problem [16]. Also, face images suffer from several other covariates (nuisance) factors, such as head pose, illumination, and expression. These factors result in a large gap between feature embeddings of HR and SR faces in the identity metric space, which significantly deteriorates the final FR performance [50].

Projection-based methods aim to create a shared embedding that can accommodate HR and LR images. To this end, synthetic LR data can be used to increase the resolution diversity of the dataset [42, 20]. However, due to the fixed-angular margin in conventional FR methods, they suffer from convergence problems and cannot fit well with data augmentations such as down-sampling or random cropping [58]. To address this issue, methods have been proposed to adaptively tune the margin based on the difficulty of the samples [57, 23, 32]. MagFace proposes to use the feature norm as the image-quality measure and tunes the margin. The adaptive margin has resolved the convergence problem to some extent. However, the performance still severely deteriorates when dealing with LR images [14]. For instance, typically the face verification accuracy on LFW is above 99%. However, the performance on Tinyface is around 59%.

Another line of work is to use Knowledge Distillation (KD) to obtain resolution agnostic face representation [44]. The main idea is to transfer the prior knowledge from HR images to train a model on the LR instances [42]. A teacher network trained on HR images guides the LR model toward capturing discriminative features from LR instances. Since FR is an open-set problem, forcing the LR model to share the embedding space with the HR model is essential. A straightforward solution is to directly minimize the Euclidian distance between LR and HR model representations, which aligns the embedding spaces; we call it Feature Distillation (FD). Previous methods mainly focused on FD because the embedding of the HR model includes more information than the LR model. In practice, it is shown that more than a FD is needed to align the models' representations [42, 39]. Enforcing a sample-level restriction on cross-resolution FR in a rigorous manner can be sub-optimal, as it eliminates the effect of negative instances and cause the network to prioritize factors such as the pose, glasses, or other facial attributes.

In this paper, we propose the Classification Consistency Face (CCFace) recognition paradigm, which takes advantage of KD and unsupervised representation learning to enforce a consistency between logits and feature embeddings of HR, and LR pairs, see Fig. 3. CCFace shares the same proxy between the HR and LR images and uses the output score of the HR images as a measure of sample hardness to tune the margin penalty of LR samples. In this manner, the training objective is relaxed for the LR inputs to alleviate the overfitting and convergence problem. As the HR embedding is more discriminative, we use two asymmetric methods to maintain the model's performance on HR images: 1) updating the proxies only from HR loss, 2) applying Asymmetric Cross-Resolution (ACR) learning between the HR and LR images with the detached features of HR inputs. Contributions of this work can be summarized as follows:

- We introduce a framework to maintain the consistency between HR and LR face recognition by sharing the proxies between different resolutions and asymmetrically updating the proxies via gradient from HR prediction.
- We introduce sample-mining in tuning the angular margin for LR samples to relax the training for hard samples and reduce overfitting.
- We asymmetrically apply contrastive learning to align the representation of HR and LR images without harming the discriminative power of the model on the original HR images.

2. Related Works

2.1. General Face Recognition

FR is among the oldest and most surfed problems in computer vision and has matured over the years from uti-

lizing handcrafted features and local descriptors toward using deep learning-based models [38]. Existing deep network architectures, such as CNNs and ViT variants [13, 8], are dominant feature extractors in this area [6, 25]. The critical issue is how to train the deep model using a large-scale dataset and prevent overfitting [39]. Deep FR training schemes are either proxy-less or proxy-based. The former utilizes a tuple of similar and dissimilar images. It encourages the network to map the faces with the same identity to close representations and faces with distinct identities to distant representations [38]. The fact that these loss functions directly supervise sample-wise similarities is aligned with the final objective of the FR system. However, the sampling process required for these methods becomes challenging in large-scale datasets, which leads to convergence problems [21]. Therefore, most of the research has been dedicated to classification-based loss functions, such as L-Softmax [27], SphereFace [26], CosFace [48], and ArcFace [6].

2.2. Low-Resolution Face Recognition

LR facial images lack the details of their HR counterpart, such as eyes and skin texture. Therefore, learning discriminative representations from LR samples are much harder, and applying general FR learning methods results in unsatisfactory performance [42]. Two main approaches have been surfed for LR face recognition: 1) construction-based and 2) projection-based methods [56, 49, 2, 42, 10, 30, 59].

Construction-based methods are based on super-resolving the LR input prior to recognition. Apart from the ill-posed nature of the FSR problem, the main scheme of FSR is not optimized for discrimination, resulting in the unsatisfactory performance of the subsequent FR task [38]. To alleviate this, Zhang et al. [56] use the face verification loss in the identity metric space to guide the reconstruction model toward preserving identity information. Several studies focused on dividing the FSR into different sub-tasks. Wang et al. in [49] dedicated one network to reconstructing global details and the other to enhancing local details. Cao et al. [2] used Reinforcement Learning (RL) to localize regions and a local network to attend to the specified regions. Current FSR methods produce visually appealing results; however, their integration with the FR model degrades the overall performance [42]. Also, these methods are computationally intensive [42].

The projection-based approaches try to map both HR and LR images to a unified embedding [10, 30, 59]. To this end, various works focused on distilling well-constructed features from HR teacher to LR student module [59]. Zhu et al. [59] utilized the general KD approach of using soft-logits prediction of the teacher module to guide the student module for the task of LR classification. Numerous studies applied intermediate representation distillation from the HR network to improve LR performance [34, 30, 10]. However, the consistency between HR and LR representations was not imposed, deteriorating the performance on cross-resolution scenarios. Among the facial parts, key parts are essential in FR, such as eyes and ears [22]. Kumar et al. in [22] force the model to generate key points using an auxiliary layer which guides the network toward focusing more on key facial characteristics. These methods mainly focused on improving the results of the LR images and did not maintain the performance of the original HR imagery. Also, guiding the LR network using the proxies of the HR network has not yet been explored for LR face recognition. Therefore, we investigate a prediction consistency knowledge distillation approach to improve the performance on LR images and preserve the model performance on HR images.

2.3. Knowledge Distillation

Knowledge distillation (KD), first proposed in [15], is a form of model compression that transfers knowledge from a robust teacher model into a small student model. Since its introduction, numerous distillation techniques have been developed [37, 45, 42]. FitNet [37] directly reduces the Euclidian distance between the student and teacher network. Combining representation learning with KD, CRD [45] utilizes contrastive objective to increase the mutual information between the teacher and student model. Despite the remarkable improvement in conventional KD, these methods are mainly designed for closed-set classification and are incompatible with FR. For instance, in [15] Kullback-Leibler (KL) divergence between soft logits of teacher and student network is used to distill knowledge. In an open-set problem such as FR, obtaining the discriminative feature is most important. Furthermore, consistency between features obtained from the teacher and student model is essential because cross-resolution embedding of a specific identity must be close to each other.

3. Method

This section begins with a preliminary proxy-based FR loss function. Then, we explain our proposed classification consistency framework and how we prevent overfitting and alleviate the convergence problem. Then, we analyze the embedding of the FR module when dealing with LR instances and show that LR samples form a cluster well separated from other classes. Then, we introduce our asymmetric cross-resolution framework based on state-of-the-art representation learning methods to restrain the LR samples from forming a cluster.

3.1. Preliminary

The majority of FR models consist of feature extractor $\mathcal{F} : \mathcal{I} \rightarrow \mathcal{X}$ mapping input faces $\mathcal{I}$ from image space to an embedding space $\mathcal{X}$. At the top of the feature extractor,

there is a classifier $\mathcal{W} : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}$ to predict the input identity from the embedding. Using gradient descent, both the feature extractor and classifier will be trained end-to-end. To this end, the widely used Softmax cross-entropy is applied on the predicted labels [1, 19]:

$$L_i = -log\frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{C} e^{W_j^T x_i}}, \quad (1)$$

where $x_i \in \mathbb{R}^d$ is the d-dimensional representation of i-th input. $W \in \mathbb{R}^{d \times C}$ represents the learnable matrix where each column W_j is the proxy of j-th class. When both the x_i and W_j are mapped to a unit hypersphere, then the dot product $W_j x_i$ reflects the cosine of the angle between representation and proxy:

$$L_i = -log\frac{e^{s\ cos(\theta_{y_i})}}{\sum_{j=1}^{C} e^{s\ cos(\theta_j)}}. \quad (2)$$

The introduction of angular penalty to the classification framework has been shown to be effective in increasing inter-class separability and intra-class compactness. SphereFace [26] argues that large-margin classification better aligns with open-set FR and introduces multiplicative angular margin for learning more discriminative features (period of the $cos(\theta_{y_i})$). CosFace proposes using a vertical shift of the function $cos(\theta_{y_i})$, which leads to more powerful feature discrimination and improved stability. ArcFace suggests using additive angular margin (phase shift of $cos(\theta_{y_i})$), which has more clear geodesic interpretation and also improves the performance. In Eq. 3, m_s, m_c, and m_a show the angular penalty introduced by SphereFace, CosFace, and ArcFace, respectively.

$$L_i = -log\frac{e^{s\ cos(m_s\theta_{y_i}+m_a)-m_c}}{e^{s\ cos(m_s\theta_{y_i}+m_a)-m_c} + \sum_{\substack{j=1 \\ j\neq y_i}}^{C} e^{s\ cos(\theta_j)}}. \quad (3)$$

3.2. Classification Consistency

Here unlike conventional KD approaches, the teacher and student networks have the same architecture. Instead, the teacher network only sees the original images, and the student network the down-sampled instances. Given a training set D, the LR samples are obtained from down-sampling with varying interpolations (s): $D_{LR} : \{(I_i^{LR}, y_i) : I_i^{LR} = I_i \downarrow_s)\}_{i=0}^{N}, s \in S$. N is the total number of samples. As presented in Fig. 4, $f_t(.)$ and $f_s(.)$, map the HR and LR faces to a d-dimensional embedding space; $x_i^{HR} = f_t(I_i)$ and $x_i^{LR} = f_s(I_i^{LR})$.

As discussed in section 3.1, in the current SOTA FR objective functions, the similarity between features is guided through the cosine angle between features and softmax proxies. Therefore, the inter-class discrepancy increases if

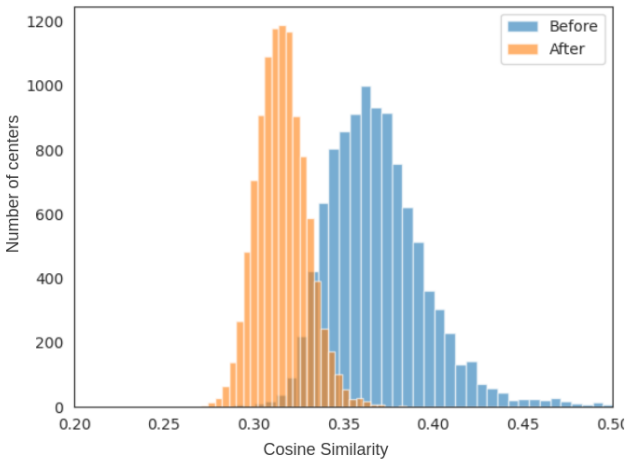

Figure 3. Maximum similarity between classifier proxies, before and after applying CCFace to a model trained by LR instances.

the proxies are well distributed on the hypersphere. Recently, numerous studies have been conducted on uniformly distributing the proxies on the hypersphere [24]. In Fig. 3, we show our studies on the inter-class similarity from the proxies' viewpoint. We illustrate the maximum inter-class cosine value in two scenarios: 1) model trained on the original MS1MV2 dataset, 2) model trained on the down-sampled version of MS1MV2. This observation shows that naively training on the LR samples will result in poor discrimination (increase in inter-class similarity). The inter-class similarity is drastically increased in the model trained on LR images. Therefore, features being supervised through these proxies would not achieve preferred discriminability.

To overcome this issue, we propose sharing the teacher network's proxies with the student model: $W^{HR} = W^{LR} = W$. However, the student model does not adjust the proxies and only uses them as a part of its forward propagation. Also, both networks are being trained using conventional large-angular margin loss function:

$$L_i^{HR} = -log\frac{e^{scos(\theta_{y_i}^{HR}+m)}}{e^{s\ cos(m_s\theta_{y_i}+m)} + \sum_{\substack{j=1 \\ j\neq y_i}}^{C} e^{scos(\theta_j^{HR})}}, \quad (4)$$

$$L_i^{LR} = -log\frac{e^{scos(\theta_{y_i}^{LR}+m^{LR})}}{e^{s\ cos(\theta_{y_i}+m^{LR})} + \sum_{\substack{j=1 \\ j\neq y_i}}^{C} e^{scos(\theta_j^{LR})}}, \quad (5)$$

where m and m^{LR} are the angular margins applied to HR and LR samples, respectively. In this way, the feature representation of LR images is implicitly forced to move toward HR representation in high-dimensional embedding space. In the following section, we explain how we utilize the HR

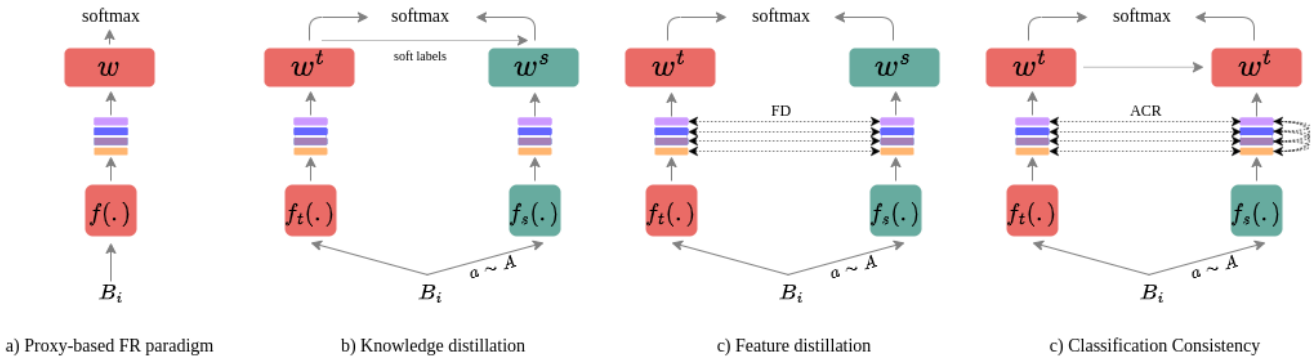

Figure 4. a) The general framework of proxy-based face recognition training has convergence issues when there are many hard samples (augmented data). b) The original knowledge-distillation paradigm in which the predicted probability of the teacher model is used as the target for the student model. c) Feature distillation derived from Knowledge-distillation. The similarity between teacher and student model features is being directly supervised, only via positive pairs. d) Classification consistency framework in which the teacher model knowledge is transferred to the student model through the classifier proxies and the feature similarity supervision. Also, ACR pushes the negative pairs of low-resolution instances to prevent forming a cluster.

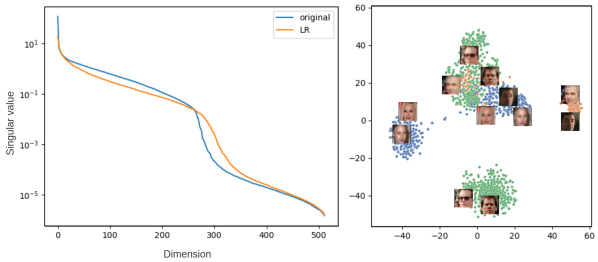

Figure 5. **left**: Singular value spectrum of embedding spaces. The representations were obtained from trained R50 on 1,000 randomly selected samples of the training data and their down-sampled version. Each feature vector has 512 dimensions. **right**: A visualization of hypersphere embeddings of the training dataset (every color represents an identity) generated by t-SNE [27]. Perturbs images, i.e., low-resolution augmentation, causes embeddings to migrate to form a common cluster.

model prediction to tune the angular margin of the LR samples. Also, we further elucidate why simultaneously training two models helps the final performance of the student model.

3.3. Cross-resolution Angular Margin Adaptivity

Generally, there are three types of samples on the original training dataset: 1) easy, 2) hard, and 3) unrecognizable [39]. Applying augmentation on different samples can produce different types of samples, whether unrecognizable or hard. Angular margin helps the model push the features toward the corresponding positive proxy, i.e., by reducing the value of $cos(\theta)$. Larger margin results in more loss and eventually more gradient to adjust the network weights. For a better illustration of the effect of angular margin on the loss and the magnitude of the gradient that the loss imposes on the network, we conduct a simple experiment. In a 2k classification problem, we changed the value of $cos(\theta_{y_i})$ from zero to one while the negative similarity scores with all of the classes are fixed to 0.1, $cos(\theta_j) = 0.1; j \neq y_i$. Reducing the Eq. 3 to ArcFace:

$$L_i = -log\frac{e^{s\ cos(\theta_{y_i}+m_a)}}{e^{s\ cos(\theta_{y_i}+m_a)} + \sum_{\substack{j=1 \\ j\neq y_i}}^{C} e^{s\ cos(\theta_j)}}, \tag{6}$$

by deriving the gradient of Eq. 6 with regard to x_i:

$$\frac{\partial L_i}{\partial x_i} = (p_{i,y_i} - 1)\frac{\partial cos(\theta_{y_i} - m)}{\partial cos(\theta_{y_i})} w_{y_i} + \sum_{\substack{j=1 \\ j\neq y_i}}^{C} p_{i,j} w_j$$

$$= \sum_{j=1}^{C} w_j (p_{i,j} - \mathbb{1}(y_i = j))(\frac{\partial cos(\theta_j - \mathrm{m}(y_i = j))}{\partial cos(\theta_j)}), \tag{7}$$

$$p_{i,j} = \frac{exp(s\ cos(\theta_j + \mathrm{m}(y_i = j)))}{exp(s\ cos(\theta_{y_i} + m)) + \sum_{\substack{j=1 \\ j\neq y_i}}^{C} e^{s\ cos(\theta_j)}}, \tag{8}$$

where we denote with $\mathrm{m}[E]$ the indicator vector which outputs m if the event E is true and 0 otherwise. Note that the feature and softmax proxies are mapped to the unit hypersphere: $||w_j|| = 1$. Therefore, in Eq. 7 only the term $(p_{i,y_i} - \mathbb{1}(y_i = j))(\frac{\partial cos(\theta_j - \mathrm{m}(y_i = j))}{\partial cos(\theta_j)})$ affects the gradient magnitude. In Fig. 6, we illustrate the gradient magnitude and the loss function value, which shows that the larger angular margin results in more adjustments for the backbone. Therefore, applying the same angular margin on HR and LR samples can result in a convergence problem or overfitting of the student model [20, 32, 39]. We propose dynamically tuning the LR samples' margin value based on the

Table 1. Verification performance comparison on the IJB-B and IJB-C datasets.

Method	IJB-B	IJB-C
	TAR@FAR: 0.0001	
CurricularFace	94.80	96.10
MagFace	94.51	95.97
Mv-Arc-Softmax	93.6	95.2
Ours	94.91	96.29

HR samples' difficulty. One of the well-established sample difficulty measures is the softmax's output score. Here, we utilized the probability output of the teacher network to tune the margin value in the student model:

$$m^{LR} = max(0, m * cos(\theta_{y_i})), \tag{9}$$

where $max(.)$ is for cases when the original sample is either unrecognizable or extremely hard for the teacher model. Therefore, applying margin on the LR version of that sample is not preferred [39]. Since both models are being simultaneously trained, Eq. 9 helps the student model to ignore hard samples at the beginning of the training. Then at the final epochs, the model is able to concentrate more on the hard instances.

3.4. Asymmetric Cross-Resolution Learning

Here, we describe the Asymmetric Cross-Resolution Learning (ACR) learning of CCFace in detail. To begin with, we first discuss the need for the ACR learing. Then we describe our asymmetric loss function, which is based on the state-of-the-art approaches for representation learning.

During the training of angular-based FR algorithms, the network maps the hard instances of each class near the decision boundary [32]. The general idea is that the LR version of images would also be mapped near the class boundary since they are hard samples for the network [7]. However, studies have shown that the LR samples tend to cluster with each other [36, 7]. Fig. 5 shows this counterintuitive phenomenon, which we validate in the following experiment. We experimented with the representation obtained from the original and down-sampled training instances. First, we randomly picked 1000 samples and computed their representations, $X \in R^{512\times1000}$. Then, we compute the covariance matrix, $C \in R^{512\times512}$:

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T, \tag{10}$$

where $N = 1,000$ is the number of samples. Fig. 5 shows the singular value decomposition of C in logarithmic scale and sorted order. As can be seen in this figure, the difference between the order of magnitude of singular values does not imply the dimensional collapse. From the observation

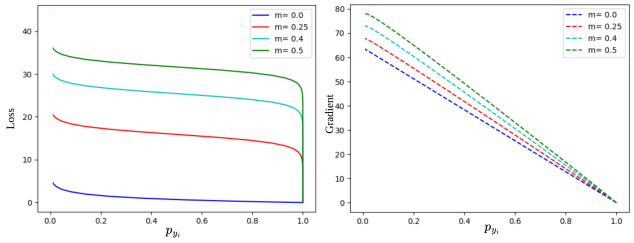

Figure 6. Shows the curves for the loss values (left) and the gradient received to the backbone (right) versus $p_{y_i} = \frac{e^{s(cos(\theta_{y_i}))}}{\sum_{j=1}^{C} e^{s(cos(\theta_j))}}$, when $cos(\theta_{y_i})$ changes from -1 to 1.

in Fig. 5, we can conclude that dimensional collapse is not happening in LR samples, and LR instances are forming a cluster well separated from the other classes. An explicit solution is reducing the pairwise distance between the normalized HR and LR representation of every subject presented in a mini-batch:

$$L_{fd} = \frac{1}{2N}\sum_{i=1}^{N}||\frac{x_i^{LR}}{x_i^{LR}} - \frac{x_i^{HR}}{x_i^{HR}}||^2, \tag{11}$$

where N is the number of samples in a mini-batch. L_{fd} reduces the discrepancy between (HR, LR) pairs; However, rigidly adopting such a sample-level constraint to the cross-resolution FR is sub-optimal, i.e. no negative instances are in Eq. 11. Using L_{fd}, the network may be induced to focus on the frontal pose, the glasses, or other non-identity facial attributes instead of the identity features of the HR-LR image pair. To increase the cross-resolution intra-class compactness, the network should consider different samples of an identity. Furthermore, to prevent LR representations from forming a cluster, the loss function should increase the inter-class discrimination among LR representations. To this end, we proposed ACR to promote discrimination among the LR images and increase the mutual information among the HR and LR counterparts of the same subject:

$$h(x^{LR}, x^{HR}) = \exp(\frac{1}{\tau}\frac{x^{LR}\cdot x^{HR}}{||x^{LR}||\cdot||x^{HR}||}), \tag{12}$$

$$L_{acr,i} = -\log\frac{1}{|P_i|}\sum_{p\in P_i}\frac{h(x_p^{HR}, x_i^{LR})}{\sum_{j\in N_i}h(x_j^{LR}, x_i^{LR})}. \tag{13}$$

In Eq. 13, N_i is a set of all negative LR samples presented in the mini-batch for x_i ($N_i \equiv n \in B : y_n \neq y_i$), P_i is a set of all positive HR samples ($P_i \equiv p \in B : y_p = y_i$) and $|P_i|$ is its cardinality. To maintain the performance on original HR data and since the HR representations have good intra-class compactness and inter-class discrimination, the detached HR features are used in Eq. 13. Therefore, Eq. 13 asymmetrically forces the LR representation to increase

Figure 7. Samples of SCFace. The first column shows the high-quality mugshot instances. For every distance $\{d_1, d_2, d_3\}$, images are taken using seven different cameras. Two images per distance are shown here, demonstrating the effect of changing the camera in the images.

their mutual information with their HR counterparts [38]. Also, it penalizes the similarity between negative LR representations, which prevents them from forming a cluster.

4. Experiments

4.1. Datasets.

We report our experiments based on using MS1MV2 as the training dataset [12, 6]. The MS1MV2 dataset is a refined version of the MS-Celeb-1M dataset with 85k identities and 4 million images. Images were down-sampled using random interpolation to construct the HR and LR pairs. For evaluation, we report our results on LFW, CFP-FP, AgeDB, CPLFW, and CALFW. To validate the proposed method's performance on real-world LR faces, we also report our results on the TinyFace [5] and SCFace [47] datasets. We used an aligned version of these datasets in which samples are resized to 112 by 112 [20]. Also, we report our performance on IJB-B and IJB-C as benchmarks which contain both HR and LR samples.

IJB-B and IJB-C: IJB-B [52] contains around 21.8K images (11.8K faces and 10K non-face images) and 7k videos (55K frames). A total of 1,845 identities are presented in this dataset. Our experimental protocols follow the standard 1:1 verification, which contains 10,270 positive and 8M negative matches. There are 12,115 templates in the protocol, each of which consists of multiple images or frames. Consequently, a template-based matching process is used. Specifically, we average over the instances in a template to obtain the global feature vector for each template. IJB-C [31] is the extended version of IJB-B, including 31.3K images and 117.5K frames from 3,531 identities. The testing protocol of IJB-C is similar to IJB-B.

SCFace: SCface [16] is a challenging cross-resolution FR dataset. It contains HQ mugshot and LQ images captured by surveillance cameras. The images were taken from 130 subjects in an uncontrolled indoor environment using five video surveillance cameras at different distances $d_i \in \{4.2, 2.6, 1.0\}$ (meter); five images at each distance. Also, one frontal mugshot image for each subject is obtained using a digital camera, Fig. 7. In an experiment similar to [16], we employ frontal mugshot images as our gallery and samples taken by surveillance cameras as probes.

Table 2. Identification accuracy on TinyFace

	Architecture	DataSet	Acc@1	Acc@5
AT [55]	ResNet50	CASIA	36.54	50.62
HORKD [10]	ResNet50	CASIA	45.49	54.80
A-SKD [42]	ResNet50	CASIA	47.91	56.55
URL [41]	ResNet100	MS1MV2	63.89	68.67
CurricularFace [17]	ResNet100	MS1MV2	63.68	67.65
CCFace	ResNet100	MS1MV2	**65.71**	**69.25**

4.2. Metrics.

There are two primary approaches to validating an FR paradigm's performance: 1) Recognition and 2) Verification. Recognition is a 1:N task where the network should calculate the similarity score of a given probe image with all the samples in the gallery and determine the identity of the probe image. Face verification is a 1:1 task where the network should determine whether a given pair of images represent the same identity. We report the verification results on the LFW, CFP-FP, AgeDB, CPLFW, IJB-B, IJB-C, and CALFW datasets. The result for identification is reported on the SCFace and TinyFace datasets.

4.3. Implementation details.

We followed the ArcFace setup for preprocessing [6]. All the images are resized to 112×112, aligned to a canonical view, and pixel values are normalized to $[-1, 1]$. To produce synthetic low-resolution samples, we randomly apply different down-sampling interpolations on the I_{HR}, including bilinear, nearest neighbor, and bicubic. Also, the size of the LR image is uniformly sampled from $\{(16 \times 16), (20 \times 20), (32 \times 32)\}$. Together, there would be twenty-seven distinct augmentations. The experiments are conducted with ResNet100 as the backbone [6] unless mentioned. The model is trained for 24 epochs with the Arcface loss. The optimizer is SGD, with the learning rate starting from 0.1 decreased by a factor of 10 at epochs $\{10, 16, 22\}$. The optimizer weight-decay is set to 0.0005, and the momentum is 0.9. During training, the mini-batch size on each GPU is 512, and the model is trained using two Quadro RTX 8000. After training is finished, we disregard the teacher model and only use the student model for evaluation.

4.4. Performance Comparison

In this section, we assess CCFace's performance against the SOTA methods, such as CurricularFace, MagFace, and URL, using TinyFace, SCFace, IJB-B, and IJB-C datasets. Results in Table 1 demonstrate CCFace's competitive performance against other methods. Showing the efficacy of

Table 3. Identification accuracy on SCFace

Method	Distance		
	d1	d2	d3
SKD [11]	43.5	48.0	53.50
CGAN [44]	44.81	49.61	54.30
MDS [33]	60.3	66.0	69.5
DMDS [53]	61.5	67.2	62.9
VGGFace [35]	41.3	75.5	88.8
DCR [28]	73.3	93.5	98.0
CCFace (r50,MS1MV2)	74.8	94.01	99.47

alignment between teacher and student model. Since IJB-B and IJB-C contain both HR and LR samples, this performance shows the ability of the proposed method to generalize across a wide range of resolutions. Table 2 demonstrates the results on TinyFace. According to this result, CCFace can effectively boost the discrimination power over the LR data, which results in over 3 percent improvement in TinyFace data. To evaluate the performance of CCFace for cross-resolution scenarios, Table 3 shows the identification result on the SCFace dataset. CCFace could improve the results of the previous methods by more than one percent. Success across datasets with varying quality levels, from LR to HR, shows the efficacy in alignment between teacher and student embedding space, and maintaining the discriminability on both HR and LR images.

5. Ablation Study

5.1. Impact of Augmentation

One of the major issues with angular-margin-based loss functions, which are dominant in FR, is that their integration with data augmentation is challenging [39]. In order to alleviate this issue, we use the output probability of the teacher network to tune the margin value for augmented samples; see section 3.3 for more detail. Table. 4 shows the student model's performance with different augmentation probabilities. Increasing the augmentation occurrence boosts the FR performance on the LR faces considerably; however, negligible performance degradation is observed on high-quality benchmarks. Table. 4 also shows the performance of the student model without using an adaptive margin. Since the augmentation process increases the chance of unrecognizable instances, the performance gain is inconsistent.

5.2. Components of Proposed Training

Here, we analyze the impact of different elements within our training paradigm on performance. Our experiments utilize R18 as the backbone and CASIA as the training dataset. As shown in Table. 5, training a network solely with a single classifier results in poor performance on the LR samples. However, sharing the classifier between the teacher and student improves performance to some extent. Additionally, incorporating an adaptive margin into the model enhances performance on the LQ dataset. Notably, the application of ACR between the teacher and student significantly improves the cross-quality scenario.

Figure 8. **Top**: Original samples of the training dataset. **Bottom**: The augmented version of training samples. Comparing original and augmented images, the samples' hardness is clearly increased implying that naive training on augmented samples with constant margin results in convergence issues or overfitting.

Table 4. Verification TAR@FAR:0.0001 for the IJB-B and IJB-C datasets with different amounts of augmentation during the training.

Augmentation Prob.	Adaptive Margin	IJB-B	IJB-C
0.0	✓	94.53	95.29
0.1	✓	94.74	95.59
0.2	✓	94.91	96.29
0.2		94.56	95.42

Table 5. Ablation study on the effect of the proposed method on the performance. CR: Cross-resolution matching (every testing pair contain one HR and one LR image), SC: Shared classifier, AM: Adaptive angular margin.

Res	CR	SC	AM	ACR	LFW	CFP-FP	CPLFW	CALFW	Age-DB
16		✓			91.2	73.3	74.3	75.3	68.7
		✓	✓		92.3	75.9	79.9	78.4	71.5
		✓	✓	✓	95.9	83.6	83.0	82.0	75.4
	✓	✓	✓	✓	92.1	73.9	80.1	74.2	68.8
64		✓			99.7	98.3	93.4	95.9	97.9
		✓	✓		99.7	98.5	93.3	96.01	98.12
		✓	✓	✓	99.8	98.8	93.3	96.13	98.3
	✓	✓	✓	✓	99.7	98.0	94.4	95.6	97.9

6. Conclusion

The proposed CCFace framework strives to maintain its performance on a wide variety of resolutions and generalize well for cross-resolution FR by simply sharing the classifier between the student and teacher network. We established that identities' proxies constructed from HR images are simple yet effective knowledge that can help the model to compensate for the information loss in the LR images and find a discriminative representation for the LR instances. Furthermore, our method prevents LR samples from forming a distinct cluster by applying pushing force directly between LR samples in a contrastive manner.

References

[1] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Morph detection enhanced by structured group sparsity. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 311–320, 2022.

[2] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 690–698, 2017.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[4] Z. Cheng, X. Zhu, and S. Gong. Low-resolution face recognition. In *Asian Conference on Computer Vision*, pages 605–621. Springer, 2018.

[5] Z. Cheng, X. Zhu, and S. Gong. Low-resolution face recognition. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 605–621. Springer, 2019.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[7] S. Deng, Y. Xiong, M. Wang, W. Xia, and S. Soatto. Harnessing unrecognizable faces for improving face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3424–3433, 2023.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] H. Du, H. Shi, Y. Liu, J. Wang, Z. Lei, D. Zeng, and T. Mei. Semi-siamese training for shallow face learning. In *European Conference on Computer Vision*, pages 36–53. Springer, 2020.

[10] S. Ge, K. Zhang, H. Liu, Y. Hua, S. Zhao, X. Jin, and H. Wen. Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10845–10852, 2020.

[11] S. Ge, S. Zhao, C. Li, and J. Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, 2018.

[12] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

[13] 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*, pages 770–778, 2016.

[14] M. He, J. Zhang, S. Shan, and X. Chen. Enhancing face recognition with self-supervised 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4062–4071, 2022.

[15] G. Hinton, O. Vinyals, J. Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[16] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2439–2448, 2017.

[17] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[18] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee. Complete face recovery GAN: Unsupervised joint face rotation and deocclusion from a single-view image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3711–3721, 2022.

[19] H. Kashiani, S. M. Sami, S. Soleymani, and N. M. Nasrabadi. Robust ensemble morph detection with domain generalization. *arXiv preprint arXiv:2209.08130*, 2022.

[20] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022.

[21] Y. Kim, W. Park, and J. Shin. Broadface: Looking at tens of thousands of people at once for face recognition. In *European Conference on Computer Vision*, pages 536–552. Springer, 2020.

[22] A. Kumar and R. Chellappa. S2ld: Semi-supervised landmark detection in low-resolution images and impact on face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 758–759, 2020.

[23] H. Liu, X. Zhu, Z. Lei, and S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019.

[24] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song. Learning towards minimum hyperspherical energy. *Advances in neural information processing systems*, 31, 2018.

[25] W. Liu, Y. Wen, B. Raj, R. Singh, and A. Weller. Sphereface revived: Unifying hyperspherical face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[26] 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*, pages 212–220, 2017.

[27] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

[28] Z. Lu, X. Jiang, and A. Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(4):526–530, 2018.

[29] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips. A meta-analysis of face recognition covariates. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–8. IEEE, 2009.

[30] F. V. Massoli, G. Amato, and F. Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99:103927, 2020.

[31] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[32] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.

[33] S. P. Mudunuri and S. Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):1034–1040, 2015.

[34] W. Park, D. Kim, Y. Lu, and M. Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[35] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

[36] W. Robbins and T. E. Boult. On the effect of atmospheric turbulence in the feature space of deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1618–1626, 2022.

[37] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[38] M. S. E. Saadabadi, S. R. Malakshan, S. Soleymani, M. Mostofa, and N. M. Nasrabadi. Information maximization for extreme pose face recognition. *arXiv preprint arXiv:2209.03456*, 2022.

[39] M. S. E. Saadabadi, S. R. Malakshan, A. Zafari, M. Mostofa, and N. M. Nasrabadi. A quality aware sample-to-sample comparison for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6129–6138, January 2023.

[40] 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*, pages 815–823, 2015.

[41] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6817–6826, 2020.

[42] S. Shin, J. Lee, J. Lee, Y. Yu, and K. Lee. Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition. In *European Conference on Computer Vision*, pages 631–647. Springer, 2022.

[43] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27, 2014.

[44] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi. Attribute-guided coupled gan for cross-resolution face recognition. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2019.

[45] Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[46] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017.

[47] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Inter-session variability modelling and joint factor analysis for face authentication. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2011.

[48] 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*, pages 5265–5274, 2018.

[49] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *International journal of computer vision*, 106(1):9–30, 2014.

[50] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 738–753, 2018.

[51] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[52] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 90–98, 2017.

[53] F. Yang, W. Yang, R. Gao, and Q. Liao. Discriminative multidimensional scaling for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(3):388–392, 2017.

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

[55] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[56] 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)*, pages 183–198, 2018.

[57] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019.

[58] Y. Zhang, S. Herdade, K. Thadani, E. Dodds, J. Culpepper, and Y.-N. Ku. Unifying margin-based softmax losses in face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3548–3557, 2023.

[59] 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)*, pages 3762–3766. IEEE, 2019.

[60] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.

[61] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011.