Spherical Confidence Learning for Face Recognition

Shen Li[¶*] Jianqing Xu[‡*] Xiaqing Xu[†] Pengcheng Shen[‡] Shaoxin Li[‡] Bryan Hooi [¶]
[¶] Institute of Data Science, National University of Singapore.
[‡] Youtu Lab, Tencent. [†] Aibee.
shen.li@u.nus.edu joejqxu@tencent.com xqxu@aibee.com quantshen@tencent.com
darwinli@tencent.com bhooi@comp.nus.edu.sg

Abstract

An emerging line of research has found that spherical *spaces better match the underlying geometry of facial images, as evidenced by the state-of-the-art facial recognition methods which benefit empirically from spherical representations. Yet, these approaches rely on deterministic embeddings and hence suffer from the* feature ambiguity dilemma*, whereby ambiguous or noisy images are mapped into poorly learned regions of representation space, leading to inaccuracies. Probabilistic Face Embeddings (PFE) [17] is the first attempt to address this dilemma. However, we theoretically and empirically identify two main failures of PFE when it is applied to spherical deterministic embeddings aforementioned. To address these issues, in this paper, we propose a novel framework for face confidence learning in spherical space. Mathematically, we extend the* von Mises Fisher *density to its r-radius counterpart and derive a new optimization objective in closed form. Theoretically, the proposed probabilistic framework provably allows for better interpretability, leading to principled feature comparison and pooling. Extensive experimental results on multiple challenging benchmarks confirm our hypothesis and theory, and showcase the advantages of our framework over prior probabilistic methods and spherical deterministic embeddings in various face recognition tasks.*

1. Introduction

A plethora of research has demonstrated the advantage of spherical latent space over Euclidean space in modelling certain types of data [4, 14, 22]. Face images are one of these types: as indicated by an emerging line of research, state-of-the-art face recognition methods empirically benefit from Deep Convolutional Neural Networks (DCNNs) that map a face image from input space into *spherical* space. This important idea has been explored in a number of recent works: NormFace pioneered this idea by introducing a normalization operation on both features and weights [18]; SphereFace imposed angular discriminative constraints on hypersphere [12]; CosFace pushed the boundary by adding cosine margin penalty to target logits [19]; and ArcFace further improved the discriminative power of face recognition models by proposing additive angular margin penalty, which is equivalent to minimizing the geodesic distance margin on a hypersphere [3].

However, while achieving clear successes in face recognition, all these approaches aim at learning *deterministic* mappings from input space to feature space, and thus *e pur si muove*, are unable to capture data uncertainty that is ubiquitous in face recognition in the wild. An ambiguous face, for instance, will be mapped into poorly learned regions of the latent space, thus causing a large bias to the facial features of its subject if applied in a deterministic way. First pointed out by Probabilistic Face Embeddings (PFE) [17], this issue was referred to as the *Feature Ambiguity Dilemma*, where ambiguous faces are mapped into a 'dark space' in which the distance metric is distorted, resulting in unwanted effects. Such deterministic mappings act as a bottleneck to further improvement of face recognition performance, especially in unconstrained environments.

Probabilistic face representation learning presents a promising avenue to addressing this problem. Far from being a novel idea, probabilistic face modelling has been explored abundantly in the literature [1, 16, 7, 11]. Of greatest relevance is PFE [17], which assumes that latent codes obey a multivariate independent Gaussian distribution that is inherently defined in Euclidean space. While improvements have been made, we identify two main failures when PFE is applied to spherical embeddings. On one hand, theoretically, the independent Gaussian assumption inevitably fails in the case of spherical embeddings. On the other, further empirical studies suggest that the PFE framework leads to unstable training when instantiated with spherical densities, e.g. r-radius von Mises Fisher as proposed. This further limits PFE's applicability to the state-of-the-art determinis-

*Corresponding authors

tic embeddings whose ranges are strictly sphere.

To address the issues of existing approaches, in this paper, we propose a novel framework, Sphere Confidence Face (SCF), for face confidence learning in an r-radius spherical space. Unlike PFE defined in Euclidean space, SCF captures the most likely feature representation and its local concentration value on spheres. We theoretically show that the concentration value can be interpreted as a measure of confidence, which allows for principled feature comparison and pooling, dispensing with the independent Gaussian assumption and pairwise training undesired. Compared to PFE that maximizes the expectation of the mutual likelihood score, our proposed framework minimizes KL divergence between spherical Dirac delta and r-radius vMF, which proves to be superior for face uncertainty learning through extensive experiments. Code is available at `http://github.com/MathsShen/SCF/`.

2. Background: Dilemmas of PFE

We identify failures of PFE from a theoretical perspective before delving into our proposed framework. Due to space limitations, we refer readers to [17] for details. Recall that the optimization objective of PFE is to minimize the expectation of negative mutual likelihood score,

$$\begin{aligned} &\min \mathbb{E}[-s(\mathbf{x}^i, \mathbf{x}^j)] \\ &= -\mathbb{E}\left[\log \iint_{\mathbb{R}^d \times \mathbb{R}^d} p(\mathbf{z}^i|\mathbf{x}^i) p(\mathbf{z}^j|\mathbf{x}^j) \delta(\mathbf{z}^i - \mathbf{z}^j) d\mathbf{z}^i d\mathbf{z}^j\right] \end{aligned} \tag{1}$$

where $\{\mathbf{x}^i, \mathbf{x}^j\}$ is a genuine pair. For prediction, PFE assumes that a distributional estimate $\mathbf{z}$ of the appearance of a person's face $\mathbf{x}$ follows a multivariate independent Gaussian distribution $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Lambda}_\mathbf{x})$, where $\boldsymbol{\Lambda}_\mathbf{x}$ is a diagonal matrix, assigning uncertainty to each dimension independently. This implies that given $\mathbf{x}$, each latent dimension $\mathbf{z}_l$ $(l = 1, ..., d)$ is independent of one another. However, this independence assumption fails when PFE is applied to the state-of-the-art deterministic embeddings whose ranges are a sphere of radius r (i.e. $\mathbf{z}_1^2 + ... + \mathbf{z}_d^2 = r^2$ in d-dimensional Euclidean space). One might argue that a full covariance matrix can be learned instead to obviate this issue. However, this inevitably leads to inefficiency and difficulty in fitting many more parameters (e.g. at least $d(d+1)/2$ in d-dimensional space) while preserving the positive semidefiniteness of the covariance matrix.

Second, the minimization of negative mutual likelihood score is problematic for spherical embeddings. As shown in Figure 1 (c) and (d), we find that simply changing Gaussian into a spherical density (r-radius vMF) does *not* resolve the issue. Detailed discussions are presented in Section 3.3. Instead of modeling uncertainty, we do the opposite, proposing a new framework suitable for spherical confidence learning which circumvents these two dilemmas.

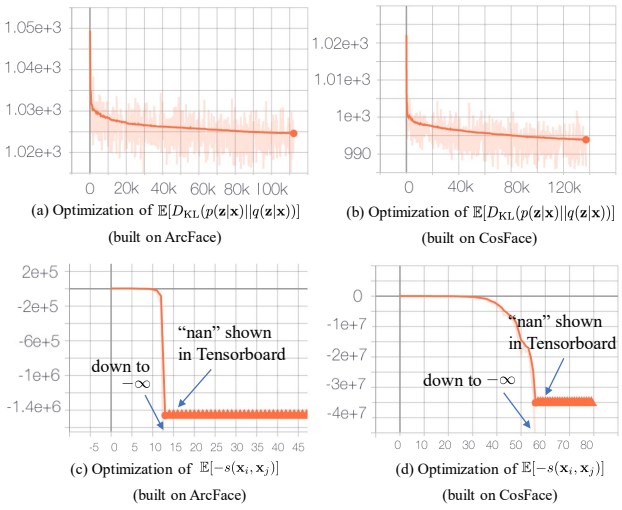

Figure 1. Empirical comparisons of training dynamics between the optimization objectives of two frameworks. Our proposed framework (a)(b) gives rise to a stable training process whereas that of PFE (c) (built on ArcFace) suffers from instability when it is instantiated with r-radius vMF; so does PFE (d) (built on CosFace). The culprit for such instability is discussed in Appendix B. Here, $s(\cdot, \cdot)$ denotes mutual likelihood score, of which the explicit form is given by Equation (13).

3. Proposed Method

3.1. r-Radius *von Mises Fisher* Distribution

Recent advancement on face recognition (e.g. ArcFace and CosFace) suggests that spherical space is better-suited for facial feature representation than is Euclidean space. We adopt this idea and further extend it to probabilistic confidence modelling. Specifically, given a face image $\mathbf{x}$ from input space $\mathcal{X}$, the conditional latent distribution is modelled as a *von Mises-Fisher* (vMF) distribution [4] defined on a d-dimensional unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$,

$$p(\mathbf{z}'|\mathbf{x}) = \mathcal{C}_d(\kappa_\mathbf{x}) \exp\left(\kappa_\mathbf{x} \boldsymbol{\mu}_\mathbf{x}^T \mathbf{z}'\right), \tag{2}$$

$$\mathcal{C}_d(\kappa_\mathbf{x}) = \frac{\kappa_\mathbf{x}^{d/2-1}}{(2\pi)^{d/2} \mathcal{I}_{d/2-1}(\kappa_\mathbf{x})}, \tag{3}$$

where $\mathbf{z}', \boldsymbol{\mu}_\mathbf{x} \in \mathbb{S}^{d-1}$, $\kappa_\mathbf{x} \geq 0$ (subscripts indicate statistical dependencies on $\mathbf{x}$) and $\mathcal{I}_\alpha$ denotes the modified Bessel function of the first kind at order α:

$$\mathcal{I}_\alpha(x) = \sum_{m=0}^{\infty} \frac{1}{m! \Gamma(m + \alpha + 1)} \left(\frac{x}{2}\right)^{2m+\alpha}. \tag{4}$$

The parameters $\boldsymbol{\mu}_\mathbf{x}$ and $\kappa_\mathbf{x}$ are called the mean direction and concentration parameter, respectively. The greater the

value of $\kappa_{\mathbf{x}}$, the higher the concentration around the mean direction $\boldsymbol{\mu}_{\mathbf{x}}$. The distribution is unimodal for $\kappa_{\mathbf{x}} > 0$, and degenerates to uniform on the sphere for $\kappa_{\mathbf{x}} = 0$.

We further extend it to r-radius vMF that is defined over the support of an r-radius sphere $r\mathbb{S}^{d-1}$. Formally, for any $\mathbf{z} \in r\mathbb{S}^{d-1}$, there exists a one-to-one correspondence between $\mathbf{z}'$ and $\mathbf{z}$ such that $\mathbf{z} = r\mathbf{z}'$. Then, the r-radius vMF density (denoted as r-vMF($\boldsymbol{\mu}_{\mathbf{x}}, \kappa_{\mathbf{x}}$)) can be obtained by applying the change-of-variable formula:

$$p(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}'|\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{z}'}{\partial \mathbf{z}} \right) \right| = \frac{\mathcal{C}_d(\kappa_{\mathbf{x}})}{r^d} \exp\left(\frac{\kappa_{\mathbf{x}}}{r} \boldsymbol{\mu}_{\mathbf{x}}^T \mathbf{z} \right) \tag{5}$$

3.2. Sphere Confidence Face (SCF)

State-of-the-art deterministic embeddings, such as ArcFace and CosFace, that are defined in spherical spaces, are essentially Dirac delta $p(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - f(\mathbf{x}))$, where $f : \mathcal{X} \mapsto r\mathbb{S}^{d-1}$ is a deterministic mapping. Here we formally extend Dirac delta into spherical space:

Definition 1 (Spherical Dirac delta)**.** A probability density $p(\mathbf{z})$ on the support of an r-radius sphere $r\mathbb{S}^{d-1}$ is spherical Dirac delta $\delta(\mathbf{z} - \mathbf{z}_0)$ (for some fixed $\mathbf{z}_0 \in r\mathbb{S}^{d-1}$) if and only if the following three conditions hold:

$$\delta(\mathbf{z} - \mathbf{z}_0) = \begin{cases} 0 & \mathbf{z} \neq \mathbf{z}_0 \\ \infty & \mathbf{z} = \mathbf{z}_0 \end{cases}; \quad \int_{r\mathbb{S}^{d-1}} \delta(\mathbf{z} - \mathbf{z}_0) d\mathbf{z} = 1;$$

$$\int_{r\mathbb{S}^{d-1}} \delta(\mathbf{z} - \mathbf{z}_0)\phi(\mathbf{z}) d\mathbf{z} = \phi(\mathbf{z}_0).$$

To address the dilemma encountered in the existing framework, we propose a new training objective by leveraging this extended definition.

As a common practice, deep face recognition classifiers map the spherical feature space $r\mathbb{S}^{d-1}$ to a label space $\mathbb{L}$ via a linear mapping parametrized by a matrix $\mathbf{W} \in \mathbb{R}^{n \times d}$, where n is the number of face identities. Let $\mathbf{w}_{\mathbf{x}\in c}$ denote the classifier weight given a face image $\mathbf{x}$ belonging to class c, which can be readily obtained from any given *pretrained* model by extracting the cth row of $\mathbf{W}$. Our key observation is that, by virtue of these classifier weights, a conventional deterministic embedding as spherical Dirac delta can act as a desired latent prior over the sphere, to which regularization can be performed. To this end, we propose minimizing the KL divergence between the spherical Dirac delta and the model distribution $p(\mathbf{z}|\mathbf{x})$.

Specifically, the optimization objective is to minimize $\mathbb{E}_{\mathbf{x}}[D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))]$, where $q(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - \mathbf{w}_{\mathbf{x}\in c})$ and $p(\mathbf{z}|\mathbf{x})$ is modelled as r-radius vMF parameterized by $\mu(\mathbf{x})$ and $\kappa(\mathbf{x})$ ($||\mu(\mathbf{x})||_2 = 1$ and $\kappa(\mathbf{x}) > 0$; here dependencies on $\mathbf{x}$ are shown in functional forms in place of subscripts). Then, we expand the objective as

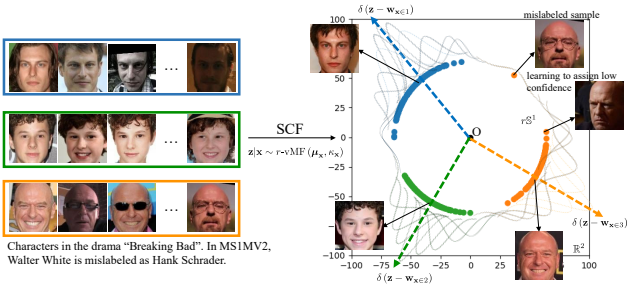

Figure 2. A 2D toy example of training SCF. SCF learns a mapping from the input space $\mathcal{X}$ to an r-radius spherical space, $r\mathbb{S}^1 \subset \mathbb{R}^2$. The latent code of each image is assumed to obey a conditional distribution, i.e. $\mathbf{z}|\mathbf{x} \sim r\text{-vMF}(\boldsymbol{\mu}_{\mathbf{x}}, \kappa_{\mathbf{x}})$, where $\boldsymbol{\mu}_{\mathbf{x}}$ and $\kappa_{\mathbf{x}}$ are parameterized by neural networks. Each identity has a class template $\mathbf{w}_{\mathbf{x}\in c}$ that induces a spherical Dirac delta, for $c = 1, 2, 3$. Optimization proceeds by minimizing $D_{\mathrm{KL}}\big(\delta\left(\mathbf{z} - \mathbf{w}_{\mathbf{x}\in c}\right) \| r\text{-vMF}\left(\boldsymbol{\mu}_{\mathbf{x}}, \kappa_{\mathbf{x}}\right)\big)$. Experiments are carried out using a subset of MS1MV2 containing three identities. We find that there are *mislabeled samples* for the third identity which hamper training otherwise. SCF learns to assign low confidence to such samples in an adaptive manner.

$$\min_p \mathbb{E}_{\mathbf{x}}\left[D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))\right]$$
$$= \mathbb{E}_{\mathbf{x}}\left[-\left(\int_{r\mathbb{S}^{d-1}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{z}|\mathbf{x}) d\mathbf{z} \right) - \mathbb{H}_{q(\mathbf{z}|\mathbf{x})}(\mathbf{z}) \right] \tag{6}$$

Note that minimizing Equation (6) with regard to p is equivalent to minimizing the cross-entropy between q and p with regard to $\boldsymbol{\mu}$ and κ conditional on $\mathbf{x}$. Therefore, it is sufficient to minimize $\mathbb{E}_{\mathbf{x}}[\mathcal{L}(\boldsymbol{\mu}(\mathbf{x}), \kappa(\mathbf{x}))]$ over all $\boldsymbol{\mu}$ and κ, where

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}(\mathbf{x}), \kappa(\mathbf{x})) &= -\int_{r\mathbb{S}^{d-1}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\ &= -\frac{\kappa(\mathbf{x})}{r} \mu(\mathbf{x})^T \mathbf{w}_{\mathbf{x}\in c} - \left(\frac{d}{2} - 1 \right) \log \kappa(\mathbf{x}) \\ &\quad + \log(\mathcal{I}_{d/2-1}(\kappa(\mathbf{x}))) + \frac{d}{2} \log 2\pi r^2. \end{aligned} \tag{7}$$

Figure 2 showcases a 2D toy example of training SCF. Detailed explanations can be found in the figure caption.

3.3. Theoretical Perspective

Remark 1. Unlike PFE which maximizes the expectation of mutual likelihood score of genuine pairs, our proposed framework, by virtue of classifier weights, minimizes the KL divergence between spherical Dirac delta and r-radius vMF. This is a reasonable choice and can be theoretically justified by Theorem 1 and Corollary 1. Intuitively, regularization to the spherical Dirac delta δ encourages the latents

that are closer to their corresponding classifier weights to have larger concentration values (thus higher confidence); and vice versa (see Theorem 2).

Theorem 1. An r-radius vMF density r-vMF$(\boldsymbol{\mu}, \kappa)$ tends to a spherical Dirac delta $\delta(\mathbf{z} - r\boldsymbol{\mu})$, as $\kappa \to \infty$.

Corollary 1. $D_{\mathrm{KL}}(\delta(\mathbf{z} - r\boldsymbol{\mu}_{\mathbf{x}})||r\text{-vMF}(\boldsymbol{\mu}_{\mathbf{x}}, \kappa_{\mathbf{x}})) \to 0$ as $\kappa_{\mathbf{x}} \to \infty$.

Proof Sketch. By leveraging the asymptotic expansion of the modified Bessel function of the first kind (developed by Hermann Hankel): for any complex number z with large $|z|$ and $|\arg z| < \pi/2$,

$$\mathcal{I}_{\alpha}(z) \sim \frac{e^{z}}{\sqrt{2\pi z}} \left(1 + \sum_{N=1}^{\infty} \frac{(-1)^{N}}{N!(8z)^{N}} \prod_{n=1}^{N} \left(4\alpha^{2} - (2n-1)^{2}\right)\right) \tag{8}$$

we have $\mathcal{I}_{d/2-1}(\kappa) \sim e^{\kappa}/\sqrt{2\pi\kappa}$ as $\kappa \to \infty$. Then, these theoretical results (Theorem 1 and Corollary 1) can be readily shown with this fact given. Full proofs can be found in Appendix A.

□

Theorem 2. The quantity $\cos\langle \mu(\mathbf{x}), \mathbf{w}_{\mathbf{x}\in c}\rangle$ is a strictly increasing function of κ^{*} in the interval $(0, +\infty)$, where $\kappa^{*} = \arg\min_{\kappa} \mathcal{L}(\mu, \kappa)$.

Proof. Taking partial derivative of the loss function $\mathcal{L}(\mu, \kappa)$ with regard to κ and setting it to zero yields the equality

$$\frac{\partial \mathcal{L}}{\partial \kappa} := 0 \Longrightarrow \cos\langle \mu(\mathbf{x}), \mathbf{w}_{\mathbf{x}\in c}\rangle = \frac{\mathcal{I}_{d/2}(\kappa^{*})}{\mathcal{I}_{d/2-1}(\kappa^{*})} \tag{9}$$

where $\kappa^{*} = \arg\min_{\kappa} \mathcal{L}(\mu, \kappa)$. Then, for any $u > v \geq 0$ and $\kappa > 0$, define $F_{uv}(\kappa) := \mathcal{I}_{u}(\kappa)/\mathcal{I}_{v}(\kappa)$. According to [9], we obtain the following properties of $F_{uv}(\kappa)$,

$$\lim_{\kappa\to 0} F_{uv}(\kappa) = 0, \quad \lim_{\kappa\to\infty} F_{uv}(\kappa) = 1 \tag{10}$$

Furthermore, $0 < F_{uv}(\kappa) < 1$ and its derivative is always positive in the interval $(0, +\infty)$, i.e. $F'_{uv}(\kappa) > 0$, which concludes the proof. □

Remark 2 (Interpretation of κ). Theorem 2 suggests that the closer $\boldsymbol{\mu}$ gets to $\mathbf{w}_{\mathbf{x}\in c}$ the higher the value of κ^{*}. For models trained with softmax-based loss (ArcFace, CosFace, etc.), the smaller the angle between $\boldsymbol{\mu}$ and its class center $\mathbf{w}_{\mathbf{x}\in c}$, the more confidence we have for prediction. Given one face image alone during the testing phase, predicting its class center for the unknown subject is an ill-posed problem. Our framework bypasses this difficulty by predicting its confidence κ that mathematically measures how close the test face image gets to its unknown class center.

Remark 3 (A mathematical failure of SCF-G). We consider a model variant of SCF, referred to as SCF-G, which operates in Euclidean space, minimizing KL divergence between Euclidean Dirac delta and independent Gaussian $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Lambda}_{\mathbf{x}})$. Here, $\boldsymbol{\Lambda}_{\mathbf{x}}$ is a diagonal covariance matrix. Following the similar derivation of (7), the loss function of SCF-G can be written into (up to a constant, $\log\sqrt{2\pi}$)

$$\mathcal{L}^{\mathrm{G}} = \frac{1}{2}(\boldsymbol{\mu}_{\mathbf{x}} - \mathbf{w}_{\mathbf{x}\in c})^{T}\boldsymbol{\Lambda}_{\mathbf{x}}^{-1}(\boldsymbol{\mu}_{\mathbf{x}} - \mathbf{w}_{\mathbf{x}\in c}) + \frac{1}{2}\log|\det\boldsymbol{\Lambda}_{\mathbf{x}}| \tag{11}$$

A mathematical failure of SCF-G can be seen through finding the optimal diagonal matrix $\boldsymbol{\Lambda}_{\mathbf{x}}^{*}$:

$$\nabla_{\boldsymbol{\Lambda}_{\mathbf{x}}}\mathcal{L}^{\mathrm{G}} := \mathbf{O} \Longrightarrow \boldsymbol{\Lambda}_{\mathbf{x}}^{*} = (\boldsymbol{\mu}_{\mathbf{x}} - \mathbf{w}_{\mathbf{x}\in c})(\boldsymbol{\mu}_{\mathbf{x}} - \mathbf{w}_{\mathbf{x}\in c})^{T} \tag{12}$$

Equality (12) does not hold true in the case of spherical space, as the off-diagonal entries of the matrix on the right hand side can be non-zeros. This mathematical failure exists due to the independence assumption undesired.

Remark 4. Empirical studies further suggest that our proposed framework for spherical face confidence learning exhibits empirical advantages over PFE even when PFE is instantiated with r-radius vMF. As shown in Figure 1, when built on the state-of-the-art spherical embeddings, the optimization objective proposed in PFE (mutual likelihood score maximization) for uncertainty learning in spherical space is empirically difficult to attain, suffering from training instability (the 'nan' loss value), whereas our proposed objective (6) gives rise to a stable training process (see Appendix B for detailed analyses).

We argue that this stems from two reasons. First, the optimization objective of PFE [17] has to be carried out in a pairwise manner. Selecting pairs in the early training stage requires a tricky and heuristic strategy; training tends to become unstable otherwise. Second, our proposed objective (6) resorts to additional information, $\mathbf{w}_{\mathbf{x}\in c}$'s, which can be regarded as class templates for each subject. By leveraging class templates, the spherical Dirac delta acts as a desired prior to which the variational latent distribution can be regularised. Intuitively, the PFE optimization objective can be construed as an alternative to maximizing the likelihood $p(\mathbf{z}|\mathbf{x})$: if the latent distributions of all possible genuine pairs have a large overlap, then the latent target $\mathbf{z}$ should have a large likelihood $p(\mathbf{z}|\mathbf{x})$ for any corresponding $\mathbf{x}$ [17]. However, maximizing the likelihood $p(\mathbf{z}|\mathbf{x})$ without regularization to $q(\mathbf{z}|\mathbf{x}) := \delta(\mathbf{z} - \mathbf{w}_{\mathbf{x}\in c})$ loses the *holistic* control of the latent distribution, inviting unwanted effects. The resultant latent representations tend to bear undesired manifestations confirmed in our empirical studies. Our treatment obviates the need of pairwise training and relaxes the independent Gaussian assumption while better modelling confidence in spherical space.

Table 1. Averaged results from six seeded models on IJB-B. The evaluation metric is the verification TAR@FAR at 1e-4 and 1e-5, respectively. STDs $\leq$ 0.08@1e-5 and STDs $\leq$0.04@1e-4.

	ResNet34		ResNet100	
TAR@FAR	1e-5	1e-4	1e-5	1e-4
CosFace	84.14	91.31	89.81	94.59
+ PFE-G	84.36	91.42	89.96	94.64
+ PFE-v	N/A	N/A	N/A	N/A
+ SCF-G	84.45	91.40	89.97	94.56
+ **SCF**	**86.55**	**92.16**	**91.02**	**94.95**
ArcFace	83.96	91.43	89.33	94.20
+ PFE-G	83.95	91.58	89.55	94.30
+ PFE-v	N/A	N/A	N/A	N/A
+ SCF-G	84.38	91.61	89.52	94.24
+ **SCF**	**86.46**	**92.21**	**90.68**	**94.74**

Table 2. Averaged results from six seeded models on IJB-C. The evaluation metric is the verification TAR@FAR at 1e-4 and 1e-5, respectively. STDs $\leq$ 0.08@1e-5 and STD $\leq$0.04@1e-4.

	ResNet34		ResNet100	
TAR@FAR	1e-5	1e-4	1e-5	1e-4
CosFace	88.72	93.07	93.86	95.95
+ PFE-G	89.13	93.45	94.09	96.04
+ PFE-v	N/A	N/A	N/A	N/A
+ SCF-G	89.04	93.25	94.15	96.02
+ **SCF**	**90.82**	**93.90**	**94.78**	**96.22**
ArcFace	88.95	93.26	93.15	95.60
+ PFE-G	89.19	93.16	92.95	95.32
+ PFE-v	N/A	N/A	N/A	N/A
+ SCF-G	88.94	93.19	93.85	95.33
+ **SCF**	**90.75**	**94.07**	**94.04**	**96.09**

3.4. Feature Comparison

We adopt mutual likelihood score proposed in [17] to measure feature similarity. The mutual likelihood score of two face images, $\mathbf{x}^i$ and $\mathbf{x}^j$, is defined as $s(\mathbf{x}^i, \mathbf{x}^j) = \log p(\mathbf{z}^i = \mathbf{z}^j)$. We show that a closed-form mutual likelihood score can be obtained for r-radius spherical latents:

$$
\begin{aligned}
& s(\mathbf{x}^i, \mathbf{x}^j) \\
&= \log \iint_{r\mathbb{S}^{d-1} \times r\mathbb{S}^{d-1}} p(\mathbf{z}^i|\mathbf{x}^i) p(\mathbf{z}^j|\mathbf{x}^j) \delta(\mathbf{z}^i - \mathbf{z}^j) d\mathbf{z}^i d\mathbf{z}^j \\
&= \log \frac{\mathcal{C}_d(\kappa^i)\mathcal{C}_d(\kappa^j)}{r^{2d}} \int_{r\mathbb{S}^{d-1}} \exp\left(\frac{1}{r}(\kappa^i \boldsymbol{\mu}^i + \kappa^j \boldsymbol{\mu}^j)^T \mathbf{z}\right) d\mathbf{z} \\
&= \log \frac{\mathcal{C}_d(\kappa^i)\mathcal{C}_d(\kappa^j)}{r^d \mathcal{C}_d(\tilde{\kappa})} \underbrace{\int_{r\mathbb{S}^{d-1}} \frac{\mathcal{C}_d(\tilde{\kappa})}{r^d} \exp\left(\frac{\tilde{\kappa}}{r} \tilde{\boldsymbol{\mu}}^T \mathbf{z}\right) d\mathbf{z}}_{=1} \\
&= \log \mathcal{C}_d(\kappa^i) + \log \mathcal{C}_d(\kappa^j) - \log \mathcal{C}_d(\tilde{\kappa}) - d \log r
\end{aligned}
\tag{13}
$$

where $\tilde{\kappa} = ||\mathbf{p}||_2$, $\mathbf{p} = (\kappa^i \boldsymbol{\mu}^i + \kappa^j \boldsymbol{\mu}^j)$, $\tilde{\boldsymbol{\mu}} = \mathbf{p}/||\mathbf{p}||_2$.

3.5. Feature Pooling with Confidence

In the cases where one subject has multiple face images (observations), it is desirable to obtain one single compact representation from multiple ones before performing face verification using cosine distance.

Given two subjects A and B, each with multiple images $\{\mathbf{x}^{\cdot}_{(m)}\}$ ("$\cdot$" can be either A or B), the proposed model predicts their statistics $\boldsymbol{\mu}^{\cdot}_{(m)}$ and $\boldsymbol{\kappa}^{\cdot}_{(m)}$. Theorem 2 suggests that the proposed framework allows for a natural interpretation of κ^* as a measure of confidence (the inverse of uncertainty). This leads to a principled feature pooling:

$$
\mathbf{z}^A = \frac{\sum_m \kappa^A_{(m)} \boldsymbol{\mu}^A_{(m)}}{\sum_m \kappa^A_{(m)}}, \mathbf{z}^B = \frac{\sum_m \kappa^B_{(m)} \boldsymbol{\mu}^B_{(m)}}{\sum_m \kappa^B_{(m)}} \tag{14}
$$

where $\mathbf{z}^A$ and $\mathbf{z}^B$ are pooled features for A and B, respectively. Then, cosine distance are utilized to measure the similarity, i.e. $\cos\langle \mathbf{z}^A, \mathbf{z}^B \rangle$.

4. Experiments

4.1. Datasets

We employ MS1MV2 [3] as our training data to conduct fair comparison with state-of-the-art deterministic face embeddings including ArcFace [3], CosFace [19] and their PFE counterparts [17]. PFE counterparts include PFE with Gaussian (PFE-G) and PFE with r-vMF (PFE-v). Note that the deterministic embeddings are all in spherical space where independent Gaussian assumption of PFE-G fails and that PFE-v suffers from training issues in various settings. We also implement SCF-G which replaces r-vMF with independent Gaussian to empirically validate the mathematical flaw of SCF-G as analysed in Remark 3. Models are evaluated on eight challenging benchmarks, including LFW [8], CFP-FP [15], AgeDB [13], CALFW [24], CPLFW [23], MegaFace [10], IJB-B and IJB-C [21].

4.2. Implementation Details

To conduct fair comparisons, all experimental settings including data preprocessing, network architectures and hyperparameters are kept identical. In particular, data preprocessing is performed by generating normalized face crops (112×112) with five facial points. ResNet100 and ResNet34 [6] are employed as deterministic embedding backbones. Following ArcFace and CosFace, we set the hypersphere radius r to 64 and choose the angular margin 0.5 for ArcFace and 0.35 for CosFace. The mean direction module $\boldsymbol{\mu}(\cdot)$ is initialized by deterministic embeddings and fixed throughout the training. The concentration module $\kappa(\cdot)$ is parameterized by a regular perceptron: `FC-BN-ReLU-FC-BN-exp`, where `FC` denotes

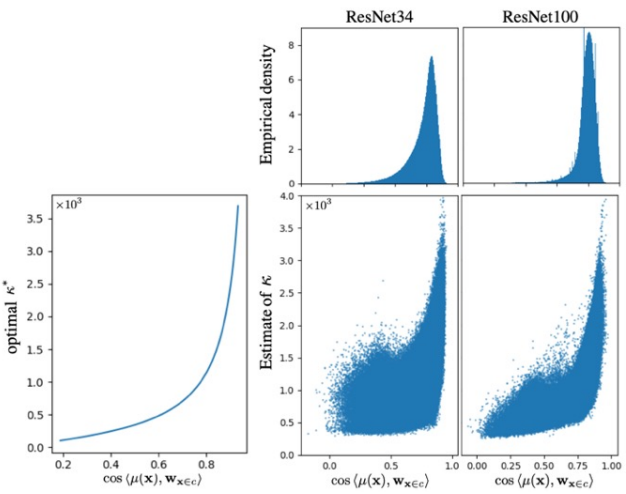

Figure 3. Left: the function plot of the inverse of Equation (9). Bottom right: the empirical correlation between cosine value $\cos\langle\mu(\mathbf{x}), \mathbf{w}_{\mathbf{x}\in c}\rangle$ and concentration value κ. Top right: marginalized empirical densities of cosine value on two backbones.

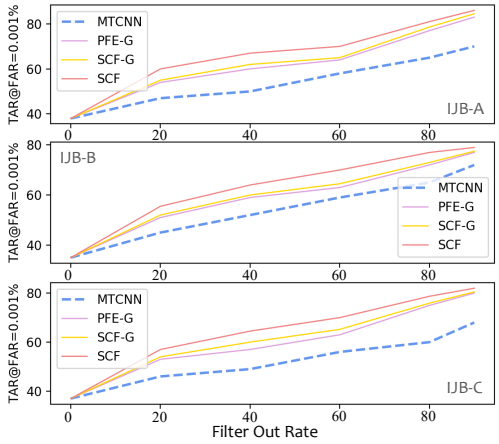

Figure 4. Risk-controlled face recognition on IJB-A, IJB-B and IJB-C.

fully-connected layers, `BN` refers to batch normalization layers, and `exp` denotes exponent nonlinearity that ensures the positivity of concentration values. Note that PFE-G or SCF-G uses the same architecture for dimension-wise uncertainty estimation. Models are trained using an ADAM optimizer with a momentum of 0.9. The batch size is set to 1024. The learning rate starts at 3×10^{-5} and is dropped by 0.5 every two epochs with the weight decay 0.0005. Experiments are performed using 16 Tesla V100 32GB GPUs.

4.3. Risk-Controlled Face Recognition

In real-world scenarios, one may expect a face recognition system to be able to reject input images with low confidence of being faces, as those highly undermine the recognition performance. Such images may exhibit large pose variations, poor image quality and severe or partial occlusion. Conventional deterministic embeddings including ArcFace and CosFace are unable to handle such cases whereas probabilistic models, such as PFE and the proposed SCF, provide natural solutions for this task. In particular, by performing image-level face verification on IJB datasets, we demonstrate the advantage of SCF over PFE and SCF-G in spherical space. Setting aside the original protocols, we take all images from a data set and rank them by confidence scores of probabilistic models (concentration values for SCF; the inverse of the variance mean for PFE-G and SCF-G or the detection score of MTCNN [20]). Then the system is able to filter out a proportion of all images according to the rankings in order to achieve better verification performance. For fairness, all methods employ original deterministic embeddings and cosine similarity for matching. To avoid saturated results, all models are trained on MS1MV2 with ResNet34 using the CosFace loss. As shown in Figure 4, SCF outperforms PFE-G and SCF-G, indicating that our proposed framework is better-suited for face confidence learning in spherical space. Note that PFE-v fails in all cases due to the training convergence issue. We provide a detailed theoretical analysis in Appendix C, showing that uncertainty scores given by PFE-G and SCF-G lead to estimation errors in the case of feature pooling.

Table 3. Averaged comparison results from six seeded models on MegaFace. "Ver." refers to face verification TAR(@FAR=1e-6). "Id." denotes rank-1 identification accuracy. STDs ≤ 0.10.

	ResNet34		ResNet100	
Metric	Id.	Ver.	Id.	Ver.
CosFace	77.52	92.61	80.56	96.56
+ PFE-G	77.35	92.42	80.44	96.49
+ PFE-v	N/A	N/A	N/A	N/A
+ SCF-G	77.64	92.69	80.57	96.61
+ **SCF**	**77.98**	**93.12**	**80.93**	**96.90**
ArcFace	77.48	92.51	81.03	96.98
+ PFE-G	77.24	92.35	80.53	96.43
+ PFE-v	N/A	N/A	N/A	N/A
+ SCF-G	77.52	92.60	81.23	97.11
+ **SCF**	**78.00**	**92.99**	**81.40**	**97.15**

4.4. Comparison with State-of-The-Art

The confidence module of SCF, $\kappa(\cdot)$, can be plugged into any spherical embedding given by backbones of different depths. To demonstrate the applicability of the proposed framework, we train two backbones, ResNet34 (a shallower backbone) and ResNet100 (a deeper backbone), with a regular ArcFace or CosFace classifier using the training set—MS1MV2. Then, confidence modules are further trained on top of these backbones. For the sake of statistical relevance, we train six different seeded models for each model variant in question (SCF, SCF-G, PFE), and report the averaged results with standard deviations (STDs).

Table 4. Averaged comparison results from six seeded models on LFW, CFP-FP, AgeDB, CALFW and CPLFW with STDs quoted[1].

Models	Training Set	Backbone	LFW	CFP-FP	AgeDB	CALFW	CPLFW
ArcFace	MS1MV2	ResNet34	99.68	94.04	96.60	95.00	89.30
+ PFE-G	MS1MV2	ResNet34	99.71	94.19	96.70	95.49	89.90
+ PFE-v	MS1MV2	ResNet34	N/A	N/A	N/A	N/A	N/A
+ SCF-G	MS1MV2	ResNet34	99.69	94.25	96.73	95.57	89.92
+ **SCF**	MS1MV2	ResNet34	**99.73**	**94.87**	**97.16**	**95.79**	**90.62**
CosFace	MS1MV2	ResNet34	99.65	94.17	96.50	94.80	89.50
+ PFE-G	MS1MV2	ResNet34	99.71	94.38	96.76	95.33	89.80
+ PFE-v	MS1MV2	ResNet34	N/A	N/A	N/A	N/A	N/A
+ SCF-G	MS1MV2	ResNet34	99.70	94.44	96.75	95.02	89.75
+ **SCF**	MS1MV2	ResNet34	**99.73**	**95.19**	**97.08**	**95.76**	**90.68**
ArcFace	MS1MV2	ResNet100	99.77	98.27	98.28	96.07	92.70
+ PFE-G	MS1MV2	ResNet100	99.78	98.33	98.21	96.08	92.82
+ PFE-v	MS1MV2	ResNet100	N/A	N/A	N/A	N/A	N/A
+ SCF-G	MS1MV2	ResNet100	99.79	98.31	98.23	96.09	93.10
+ **SCF**	MS1MV2	ResNet100	**99.82**	**98.40**	**98.30**	**96.12**	**93.16**
CosFace	MS1MV2	ResNet100	99.78	98.45	98.03	96.03	92.75
+ PFE-G	MS1MV2	ResNet100	99.80	98.56	98.15	96.10	92.82
+ PFE-v	MS1MV2	ResNet100	N/A	N/A	N/A	N/A	N/A
+ SCF-G	MS1MV2	ResNet100	99.79	98.54	98.14	96.11	92.98
+ **SCF**	MS1MV2	ResNet100	99.80	**98.59**	**98.26**	**96.18**	**93.26**

[1] STDs $\leq 0.22\%$ with ResNet34; STDs $\leq 0.10\%$ with ResNet100.

Results on IJB-B and IJB-C. Models are evaluated by using the verification TAR@FAR protocol on IJB-B and IJB-C. In IJB benchmarks, one subject has multiple face images. Feature pooling is carried out for verification. As shown in Table 1 and 2, SCF outperforms PFE and SCF-G by clear margins ($\sim 2\%$ improvement with shallower backbones and $\sim 0.5\%$ improvement with deeper backbones) at different FARs (1e-4 and 1e-5). This benefits from the theoretical correctness of SCF as compared to PFE; for PFE and SCF-G, improper feature fusion accumulates representation error, leading to suboptimal recognition performance.

Results on LFW, CFP-FP, AgeDB, CALFW, CPLFW. As shown in Table 4, our proposed framework yields promising performance on LFW, CFP-FP, AgeDB, CALFW and CPLFW when built upon ArcFace and CosFace with various deep backbones, marginally surpassing the performance of prior probabilistic frameworks in various cases.

Results on MegaFace. Table 3 demonstrates the comparison results on MegaFace in terms of face identification and face verification. SCF achieves competitive performance when implemented with backbones of different depths.

These marginal improvements suggest the limitation of SCF deployed in one-on-one settings (including LFW, CFP-FP, AgeDB, CALFW, CPLFW and MegaFace): SCF may give a biased estimate of κ given one single face image. In set-to-set settings (e.g. IJB), however, SCF performs better thanks to the theoretical guarantee that κ serves as a natural weight for feature pooling, leading to faithful compact representation aggregated from multiple face images. Noticeably, SCF exhibits clearer advantages over other variants with a shallower backbone than with a deeper one (cf. a detailed analysis in the next section).

4.5. Quantitative Analysis

Equation (9) shows the mathematical relation between the optimal κ^* and the cosine value. Note that $F_{d/2,d/2-1}$ is a strictly increasing function of κ^* and thus has its inverse (implicit though). We plot the inverse form shown in Figure 3 (left), i.e. $\kappa^* = F^{-1}_{d/2,d/2-1}(\cos\langle\mu(\mathbf{x}), \mathbf{w}_{\mathbf{x}\in c}\rangle)$. We also demonstrate the latent manifold empirically learned by our framework SCF. As illustrated in Figure 3 (right), there is a strong correlation between the cosine value $\cos\langle\mu(\mathbf{x}), \mathbf{w}_{\mathbf{x}\in c}\rangle$ and the concentration parameter κ The closer the angular distance between $\mu(\mathbf{x})$ and $\mathbf{w}_{\mathbf{x}\in c}$, the higher the concentration value (confidence) becomes. This

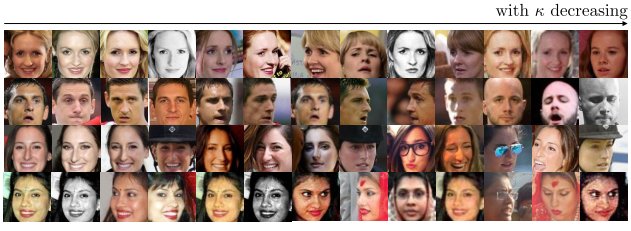

Figure 5. Identity versus concentration value (confidence). Each row corresponds to one single identity sorted from left to right with concentration values decreasing.

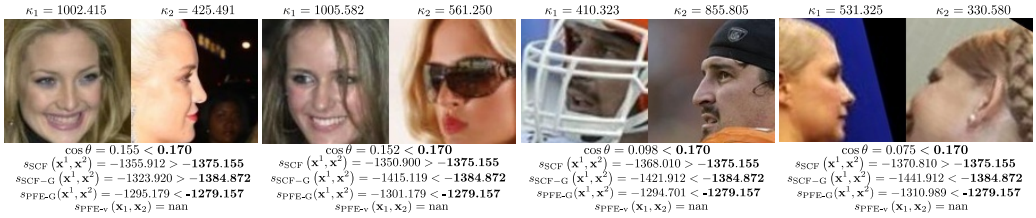

Figure 6. False negative examples made by PFE or SCF-G while being true positive by SCF, where $\cos\theta$ is the cosine distance of a verification pair $\mathbf{x}^1, \mathbf{x}^2$, $s(\cdot, \cdot)$ is mutual likelihood score and κ^1, κ^2 are the corresponding concentration values. Thresholds are set to -1279.157, -1384.872 and -1375.155 for PFE-G (accuracy: 89.90), SCF-G (accuracy: 89.83) and SCF (accuracy: 90.80), respectively, on the CPLFW benchmark.

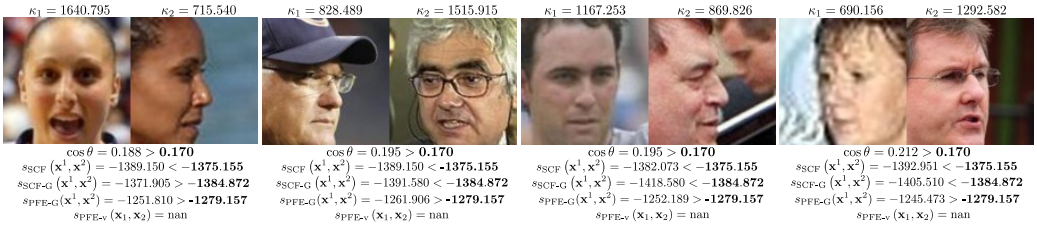

Figure 7. False positive examples made by PFE or SCF-G while being true positive by SCF, where $\cos\theta$ is the cosine distance of a verification pair $\mathbf{x}^1, \mathbf{x}^2$, $s(\cdot, \cdot)$ is mutual likelihood score and κ^1, κ^2 are the corresponding concentration values. Thresholds are set to -1279.157, -1384.872 and -1375.155 for PFE-G (accuracy: 89.90), SCF-G (accuracy: 89.83) and SCF (accuracy: 90.80), respectively, on the CPLFW benchmark.

corroborates Theorem 1, indicating that our model indeed learns the latent distribution that is unimodal vMF for each single class and forms a mixture of vMFs overall, which confirms our hypothesis. Visualization shown in Figure 5 further confirms Theorem 2, suggesting that higher concentration values correspond to more facial clues for recognition with high confidence whereas lower ones correspond to those with large pose variations, low quality and partial occlusion, or even mislabeled examples that might undermine recognition performance.

Noticeably, as shown in Table 1, 2, 3 and 4, the improvements of the proposed confidence learning framework using the shallower backbone ResNet34 are consistently higher than that using the deeper backbone ResNet100. The empirical density of cosine value marginalized from the joint density in Figure 3 also sheds lights on why this is the case: a deeper deterministic backbone itself leads to latent embeddings more concentrated around the mean direction than otherwise. Such deeper deterministic embeddings already exhibit high separability in latent spherical space with *fewer* ambiguous samples lying on the classifier boundaries, which acts as a bottleneck to further improvement. A shallower deterministic backbone, on the other hand, gives rise to spherical embeddings more scattered around the mean direction, whereby the confidence module shows its clearer advantage in assigning proper concentrating values (confidence), thereby making more correct predictions.

4.6. Qualitative Analysis

We conduct qualitative analysis of the advantage of the proposed framework over that of PFE, SCF-G and other model variants. As shown in Figure 6 and Figure 7, CosFace and PFE both fail to make correct predictions due to the large pose variations and low-quality images whereas SCF is able to assign proper concentration values (confidence) to face images under different conditions, thereby making correct predictions. More detailed analyses are relegated to Appendix D.

5. Concluding Remarks

A plethora of research has demonstrated the advantage of spherical latent space in modelling certain types of data [4, 14, 22]. Yet, modelling uncertainty in spherical space remains unexplored. Our work bridges this gap by proposing a probabilistic framework for confidence learning in spherical space. Towards going *beyond* face recognition, e.g. text modeling [5] and link prediction [2], we believe that the presented research sheds light on a promising direction towards confidence learning with general data whose manifold is not trivially Euclidean.

From the theoretical and empirical views, we have identified two main failures of the existing framework for uncertainty learning when it is applied to spherical embeddings. To address these issues, we have proposed a novel framework for spherical face confidence learning, which empirically proves to be superior to prior probabilistic methods on multiple challenging benchmarks. Future work includes theoretical comparison and analyses of these two frameworks in the context of general probabilistic modelling.

Acknowledgments

We thank all anonymous reviewers for constructive suggestions on the manuscript of this paper. We are grateful to Jiaying Wu for attentive proofreading.

References

[1] Ognjen Arandjelovic, Gregory Shakhnarovich, John Fisher, Roberto Cipolla, and Trevor Darrell. Face recognition with image sets using manifold density divergence. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 581–588. IEEE, 2005.

[2] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 856–865. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[4] Nicholas I Fisher, Toby Lewis, and Brian JJ Embleton. *Statistical analysis of spherical data*. Cambridge university press, 1993.

[5] Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] PS Hiremath, Ajit Danti, CJ Prabhakar, K Delac, and M Grgic. Modelling uncertainty in representation of facial features for face recognition. *Face recognition*, 10:183–218, 2007.

[8] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[9] A. L. Jones. An extension of an inequality involving modified bessel functions.

[10] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[11] Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, and Jianchao Yang. Probabilistic elastic matching for pose variant face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3499–3506, 2013.

[12] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[13] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–59, 2017.

[14] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. Spherical topic models. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 903–910, 2010.

[15] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[16] Gregory Shakhnarovich, John W Fisher, and Trevor Darrell. Face recognition from long-term observations. In *European Conference on Computer Vision*, pages 851–865. Springer, 2002.

[17] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6902–6911, 2019.

[18] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

[19] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei 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.

[20] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[21] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen 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.

[22] Richard C Wilson, Edwin R Hancock, Elżbieta Pekalska, and Robert PW Duin. Spherical and hyperbolic embeddings of data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2255–2269, 2014.

[23] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5, 2018.

[24] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.