HYPERSPHERE FACE UNCERTAINTY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

An emerging line of research has found that *hyperspherical* spaces better match the underlying geometry of facial images, as evidenced by the state-of-the-art facial recognition methods which benefit empirically from hyperspherical 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 (Shi & Jain, 2019). PFE is the first attempt to circumvent this dilemma. However, we theoretically and empirically identify two main failure cases of PFE when it is applied to hyperspherical deterministic embeddings aforementioned. To address these issues, in this paper, we propose a novel framework for face uncertainty learning in hyperspherical space. Mathematically, we extend the *von Mises Fisher* density to its r-radius counterpart and derive an optimization objective in closed form. For feature comparison, we also derive a closed-form mutual likelihood score for latents lying on hypersphere. Extensive experimental results on multiple challenging benchmarks confirm our hypothesis and theory, and showcase the superior performance of our framework against prior probabilistic methods and conventional hyperspherical deterministic embeddings both in risk-controlled recognition tasks and in face verification and identification tasks.

1 INTRODUCTION

Euclidean space is the most commonly used representation space for modelling face images for facial recognition. However, an emerging line of research has found that state-of-the-art face recognition systems empirically benefit from Deep Convolutional Neural Networks (DCNNs) that map a face image from input space into *hyperspherical* 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 (Wang et al., 2017); SphereFace imposed angular discriminative constraints on hypersphere (Liu et al., 2017); CosFace pushed the boundary by adding cosine margin penalty to target logits (Wang et al., 2018b); and ArcFace further improved the discriminative power of face recoginition models by proposing additive angular margin penalty, which is equivalent to the geodesic distance margin on a hypersphere (Deng et al., 2019).

However, while achieving clear successes in face recognition, all these approaches aim at learning *deterministic* mappings from input space to feature space, and thus 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 PFE (Shi & Jain, 2019), 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 face recognition settings.

Probabilistic face representation learning presents a promising avenue for addressing this problem. Far from being a novel idea, probabilistic face modelling has been explored abundantly in the literature (Arandjelovic et al., 2005; Shakhnarovich et al., 2002; Hiremath et al., 2007; Li et al., 2013). Of greatest relevance is PFE (Shi & Jain, 2019), which models latent codes using a multivariate independent Gaussian distribution that is inherently defined in Euclidean space. While improvements have been made, we identify two main failure cases when PFE is applied to hyperspherical embeddings. On one hand, theoretically, the independent Gaussian assumption inevitably fails in the case of hyperspherical embeddings. This is in line with the empirical findings of PFE (Shi & Jain, 2019), "*We also tried implementing ArcFace but it does not converge well in our case. So we did not use it.*" On the other, empirical studies suggest that the framework proposed in PFE leads to unstable training when built on hyperspherical deterministic embeddings, e.g ArcFace and CosFace. This further limits PFE's applicability to the state-of-the-art deterministic embeddings whose ranges are hypersphere. To address these issues, in this paper, we propose a novel framework, Hypersphere Face (HypersFace), for face uncertainty learning in an r-radius hyperspherical space. Unlike PFE defined in Euclidean space, HypersFace captures the most likely feature representation and its local concentration value on hyperspheres. This concentration value can be interpreted as a measure of uncertainty in hyperspherical space, dispensing with the independent Gaussian assumption and over-parameterization of the full covariance matrix. Specifically, as compared to PFE that maximizes the expectation of the mutual likelihood score, our proposed framework minimizes KL divergence between hyperspherical Dirac delta and r-radius vMF, which proves to be superior for face uncertainty learning through extensive experiments.

Our contributions include:

1. We theoretically and empirically identify the downsides of the existing framework for uncertainty learning in the case of hyperspherical embeddings and propose a novel framework for hyperspherical uncertainty learning as a remedy to these issues.
2. By extending the *von Mises Fisher* distribution to its r-radius counterpart, we show our proposed framework admits closed-forms for optimization and for feature comparison.
3. We showcase that our proposed framework outperforms prior probabilistic methods (PFE) and state-of-the-art deterministic embeddings on multiple challenging datasets in risk-controlled face recognition tasks as well as in face verification and identification tasks.

2 Proposed Method

2.1 Dilemma Encountered in PFE

We first identify failure cases of PFE from a theoretical perspective before delving into our proposed framework. Due to space limitations, we refer readers to Shi & Jain (2019) for details. Recall that PFE gives a distributional estimate $\mathbf{z}$ of the appearance of a person's face $\mathbf{x}$ using a multivariate independent Gaussian distribution $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$. This implies that given $\mathbf{x}$, each latent dimension $\mathbf{z}_i$ 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 hypersphere (i.e $\mathbf{z}_1^2 + ... + \mathbf{z}_d^2 = 1$ in d-dimensional Euclidean space). One might argue that a full covariance matrix can be learned instead to address 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. In contrast, we propose a new framework suitable for hyperspherical uncertainty learning that elegantly resolves this issue.

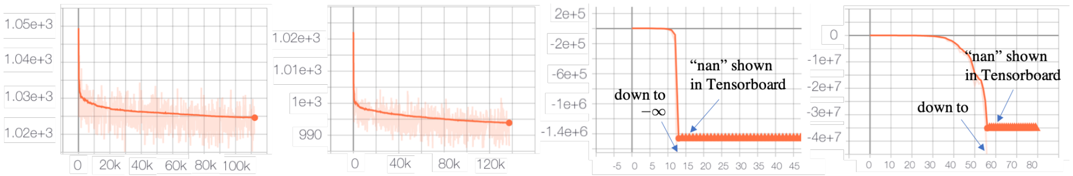
2.2 r-Radius *von Mises Fisher* Distribution

The remarkable recognition performance of state-of-the-art methods (e.g ArcFace and CosFace) indicates that hyperspherical space is better-suited for facial feature representation than Euclidean space. We adopt this idea and further extend it to probabilistic 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 (Fisher et al., 1993) defined on a d-dimensional unit hypersphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$,

$$p(\mathbf{z}'|\mathbf{x}) = \mathcal{C}_d(\kappa_{\mathbf{x}}) \exp\left(\kappa_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^T\mathbf{z}'\right) \quad \text{with} \quad \mathcal{C}_d(\kappa_{\mathbf{x}}) = \frac{\kappa_{\mathbf{x}}^{d/2-1}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa_{\mathbf{x}})} \tag{1}$$

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:

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



(a) Optimization of $\mathbb{E}[D_{\text{KL}}(p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x}))]$ (built on ArcFace)

(b) Optimization of $\mathbb{E}[D_{\text{KL}}(p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x}))]$ (built on CosFace)

(c) Optimization of $\mathbb{E}[-s(\mathbf{x}_i, \mathbf{x}_j)]$ (built on ArcFace)

(d) Optimization of $\mathbb{E}[-s(\mathbf{x}_i, \mathbf{x}_j)]$ (built on CosFace)

Figure 1: Empirical comparison 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). Implementation details can be found in Section 3.2. Here, $s(\cdot, \cdot)$ denotes mutual likelihood score, of which the explicit form is given by Eqn (7).

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 degenerate to uniform on the hypersphere for $\kappa_{\mathbf{x}} = 0$.

We further extend it to r-radius vMF that is defined over the support of the r-radius hypersphere $r\mathbb{S}^{d-1}$. 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

$$p(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}'|\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{z}'}{\partial \mathbf{z}} \right) \right| = \frac{\mathcal{C}_d(\kappa_{\mathbf{x}})}{r^d} \exp\left(\frac{\kappa_{\mathbf{x}}}{r} \boldsymbol{\mu}_{\mathbf{x}}^T \mathbf{z} \right) \tag{3}$$

2.3 Hypersphere Face (HypersFace)

State-of-the-art deterministic embeddings, such as ArcFace and CosFace, that are defined in hyperspherical 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 hyperspherical space:

Definition 1 (Hyperspherical Dirac delta)**.** A probability density $p(\mathbf{z})$ is hyperspherical Dirac delta $\delta(\mathbf{z} - \mathbf{z}_0)$ (where $\mathbf{z}, \mathbf{z}_0 \in r\mathbb{S}^{d-1}$) if and only if it is subject to the following three conditions:

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

To address the dilemma encountered in the existing framework, we propose an alternative as a remedy by leveraging these extended definitions.

As common practice, deep face recognition models map the hyperspherical feature space $r\mathbb{S}^{d-1}$ to a label space $\mathbb{L}$ using a linear mapping that can be represented as a matrix $W \in \mathbb{R}^{n \times d}$, 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 obtained from any given pretrained model by extracting the cth row of W. Our key observation is that, by virtue of these classifier weights, a conventional deterministic embedding as hyperspherical Dirac delta can act as a desired latent prior over the hypersphere, to which regularization can be performed. To this end, we propose to minimize the KL divergence between the hypersphere Dirac delta and the model distribution $q(\mathbf{z}|\mathbf{x})$.

Specifically, our optimization objective is to minimize $\mathbb{E}_{\mathbf{x}}[D_{\text{KL}}(p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x}))]$, where $p(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - \mathbf{w}_{\mathbf{x} \in c})$ and $q(\mathbf{z}|\mathbf{x})$ 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

$$\min_q \mathbb{E}_{\mathbf{x}} \left[D_{\text{KL}}(p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x})) \right] = \mathbb{E}_{\mathbf{x}} \left[-\left(\int_{r\mathbb{S}^{d-1}} p(\mathbf{z}|\mathbf{x}) \log q(\mathbf{z}|\mathbf{x}) d\mathbf{z} \right) - \mathbb{H}_{p(\mathbf{z}|\mathbf{x})}(\mathbf{z}) \right] \tag{4}$$

Note that minimizing Equation (4) with regard to q is equivalent to minimizing the cross-entropy between p and q 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

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}(\mathbf{x}),\kappa(\mathbf{x})) &= -\int_{r\mathbb{S}^{d-1}} \delta(\mathbf{z}-\mathbf{w}_{\mathbf{x}\in c})\left[\frac{\kappa(\mathbf{x})}{r}\mu(\mathbf{x})^T\mathbf{z} + \log C_d(\kappa(\mathbf{x})) - d\log r\right]d\mathbf{z} \\ &= -\frac{\kappa(\mathbf{x})}{r}\mu(\mathbf{x})^T\mathbf{w}_{\mathbf{x}\in c} - \left(\frac{d}{2}-1\right)\log\kappa(\mathbf{x}) + \log(\mathcal{I}_{d/2-1}(\kappa(\mathbf{x}))) + \frac{d}{2}\log 2\pi r^2\end{aligned} \tag{5}$$

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 hypersphere 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 δ encourages the latents that are closer to their corresponding classifier weights to have larger concentration values; and vice versa.

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

Corollary 1. $D_{\text{KL}}(\delta(\mathbf{z}-r\boldsymbol{\mu}_{\mathbf{x}})||r\text{-vMF}(\boldsymbol{\mu}_{\mathbf{x}},\kappa_{\mathbf{x}}))\to 0$ 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 $|\arg z| < \pi/2$,

$$\mathcal{I}_\alpha(z) \sim \frac{e^z}{\sqrt{2\pi z}}\left(1+\sum_{N=1}^{\infty}(-1)^N\frac{\prod_{n=1}^{N}\left(4\alpha^2-(2n-1)^2\right)}{N!(8z)^N}\right) \tag{6}$$

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

Remark 2. Empirical studies further suggest that our proposed framework for hyperspherical face uncertainty 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 hyperspherical embeddings, the optimization objective proposed in PFE (mutual likelihood score maximization) for uncertainty learning in hyperspherical space is empirically difficult to attain, suffering from training instability ('nan' loss value), whereas our proposed objective (5) gives rise to a stable training process. This, again, corroborates the empirical findings of PFE (Shi & Jain, 2019), "*We also tried implementing ArcFace but it does not converge well in our case. So we did not use it.*"

We argue that this stems from two reasons. First, the optimization objective proposed in PFE has to be carried out in a pairwise manner. Selecting pairs in the early training stage requires a tricky and heuristic strategy; otherwise, training tends to become unstable. Second, our proposed objective (5) 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 hyperspherical Dirac delta acts as a desired prior to which the resultant latent distribution can be regularised. Intuitively, the objective proposed in PFE can be understood as an alternative to maximizing the likelihood $q(\mathbf{z}|\mathbf{x})$: if the latent distributions of all possible genuine pairs have a large overlap, then the latent target $\mathbf{z}$ should have a large likelihood $q(\mathbf{z}|\mathbf{x})$ for any corresponding $\mathbf{x}$ (Shi & Jain, 2019). However, maximizing the likelihood $q(\mathbf{z}|\mathbf{x})$ without regularization to $p(\mathbf{z}|\mathbf{x}) := \delta(\mathbf{z}-\mathbf{w}_{\mathbf{x}\in c})$ 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 dispenses with pairwise training and relaxes the independent Gaussian assumption while better capturing uncertainty in hyperspherical space.

2.4 Feature Fusion and Inference

We adopt mutual likelihood score proposed in (Shi & Jain, 2019) to compare feature similarity. The mutual likelihood score of two faces, $\mathbf{x}_i$ 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 hyperspherical latents:

$$s(\mathbf{x}_i,\mathbf{x}_j) = \log\mathcal{C}_d(\kappa_i) + \log\mathcal{C}_d(\kappa_j) - \log\mathcal{C}_d(\tilde{\kappa}) - d\log r \tag{7}$$

Table 1: Comparison results on LFW, CFP-FP, CALFW and CPLFW. The results of base embeddings (ArcFace and CosFace) are successfully replicated as reported in the ArcFace officially released repository: https://github.com/deepinsight/insightface/wiki/Model-Zoo.

Models	Training Set	Backbone	LFW	CFP-FP	AgeDB	CALFW	CPLFW
ArcFace	MS1MV2	ResNet24	99.651	96.772	97.121	94.083	88.450
+ PFE-G	MS1MV2	ResNet24	N/A	N/A	N/A	N/A	N/A
+ PFE-v	MS1MV2	ResNet24	N/A	N/A	N/A	N/A	N/A
+ HypersFace	MS1MV2	ResNet24	**99.768**	**97.210**	**98.200**	**94.755**	**89.492**
CosFace	MS1MV2	ResNet24	99.639	96.924	97.093	94.118	88.054
+ PFE-G	MS1MV2	ResNet24	99.690	97.072	97.412	94.399	88.210
+ PFE-v	MS1MV2	ResNet24	N/A	N/A	N/A	N/A	N/A
+ HypersFace	MS1MV2	ResNet24	**99.728**	**97.395**	**98.105**	**94.716**	**88.833**
ArcFace	MS1MV2	ResNet100	99.770	98.270	98.280	96.083	92.700
+ PFE-G	MS1MV2	ResNet100	N/A	N/A	N/A	N/A	N/A
+ PFE-v	MS1MV2	ResNet100	N/A	N/A	N/A	N/A	N/A
+ HypersFace	MS1MV2	ResNet100	**99.783**	**98.286**	**98.350**	**96.133**	**93.333**
CosFace	MS1MV2	ResNet100	99.730	97.440	98.521	94.960	89.149
+ PFE-G	MS1MV2	ResNet100	99.731	97.439	98.530	94.962	89.150
+ PFE-v	MS1MV2	ResNet100	N/A	N/A	N/A	N/A	N/A
+ HypersFace	MS1MV2	ResNet100	**99.740**	**97.452**	**98.540**	**95.015**	**89.154**

where $\tilde{\kappa} = ||\mathbf{p}||_2$, $\mathbf{p} = (\kappa_i \boldsymbol{\mu}_i + \kappa_j \boldsymbol{\mu}_j)$, $\tilde{\boldsymbol{\mu}} = \mathbf{p}/||\mathbf{p}||_2$. In the cases where one subject has multiple face images (observations), it can be shown that a compact hyperspherical distributional representation of the particular subject can be built by iteratively merging learned statistics. Specifically, iterative updating formulae after observing $(n+1)$ observations of a given subject are given by:

$$\tilde{\kappa}_{n+1} = ||\kappa_{n+1}\boldsymbol{\mu}_{n+1} + \tilde{\kappa}_n \tilde{\boldsymbol{\mu}}_n||_2, \quad \tilde{\boldsymbol{\mu}}_{n+1} = (\kappa_{n+1}\boldsymbol{\mu}_{n+1} + \tilde{\kappa}_n \tilde{\boldsymbol{\mu}}_n)/\tilde{\kappa}_{n+1}. \tag{8}$$

Detailed derivations and analyses can be found in Appendix C and D.

3 Experiments

3.1 Datasets

We employ MS1MV2 (Deng et al., 2019) as our training data in order to conduct fair comparison with state-of-the-art deterministic face embeddings including ArcFace (Deng et al., 2019), CosFace (Wang et al., 2018b) and their PFE counterparts (Shi & Jain, 2019). PFE counterparts include PFE with Gaussian (PFE-G) and PFE with vMF (PFE-v). Note that these deterministic embeddings are all in hyperspherical space where independent Gaussian assumption of PFE-G fails and that PFE-v suffers from training issues in various settings. Models are evaluated on seven challenging benchmarks, including LFW (Huang et al., 2008), CFP-FP (Sengupta et al., 2016), AgeDB (Moschoglou et al., 2017), CALFW (Zheng et al., 2017), CPLFW (Zheng & Deng, 2018), MegaFace (Kemelmacher-Shlizerman et al., 2016) and IJB-C (Whitelam et al., 2017).

3.2 Implementation Details

To conduct fair comparison, all experimental settings including data preprocessing, embedding network architectures and related hyperparameters are kept identical. Specifically, data preprossessing is performed by generating normalized face crops (112×112) with five facial points. ResNet100 and ResNet24 (He et al., 2016) are employed as deterministic embedding backbones as in ArcFace(Deng et al., 2019) and CosFace (Wang et al., 2018b). We follow ArcFace and CosFace to 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 under consideration. The concentration module $\kappa(\cdot)$ is parameterized by three-layer perceptrons with the architecture: $\mathrm{FC}(12544) - \mathrm{BN} - \mathrm{ReLU} - \mathrm{FC}(6272) - \mathrm{BN} - \mathrm{ReLU} - \mathrm{FC}(1) - \exp$, where $\mathrm{FC}(d')$ denotes a fully-connected layer with output dimension d', and ReLU and exp denote ReLU and exponent nonlinearity, respectively. HypersFace is trained using an ADAM optimizer with a momentum of 0.9.

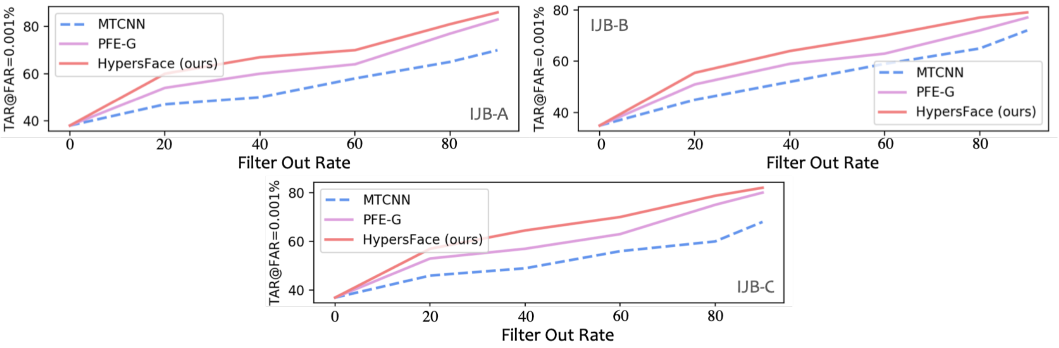


Figure 2: Risk-controlled face recognition on IJB-A, IJB-B and IJB-C.

The learning rate starts at 3×10^{-5} and is dropped by 0.5 every two epochs with the weight decay 0.0005.

3.3 Risk-Controlled Face Recognition

In the 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 uncertainty-aware models, such as PFE and our proposed HypersFace, provide natural solutions for this task. In particular, by performing image-level face verification on IJB datasets, we demonstrate the advantage of HypersFace over PFE in hyperspherical space. Setting aside the original protocols, we take all images from a data set and rank them by confidence scores of uncertainty-aware models (concentration values for HypersFace, the inverse of harmonic mean of variances for PFE-G or the detection score of MTCNN (Wen et al., 2016)). 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, models are trained on MSM1V2 with ResNet24 using AM-Softmax (Wang et al., 2018a). As shown in Figure 2, HypersFace outperforms PFE-G, indicating that our proposed framework is better-suited for face uncertainty learning in hyperspherical space. Note that PFE-v fails in all cases due to the training convergence issue.

3.4 Comparison with State-of-The-Art

The uncertainty module of HypersFace, $\kappa(\cdot)$, can be plugged into any hyperspherical embedding given by backbones of different depths. To demonstrate the applicability of the proposed framework,

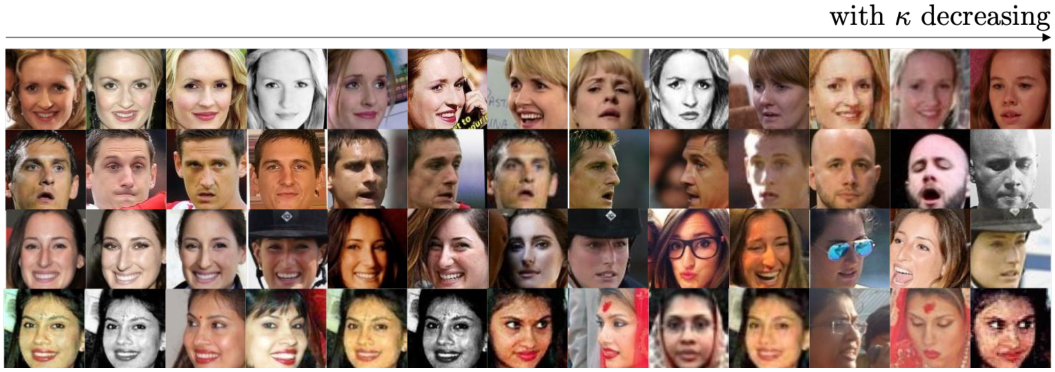


Figure 3: Identity versus concentration value. Each row corresponds to one single identity sorted from left to right with concentration values decreasing.

we first train two backbones, ResNet24 (a shallower backbone) and ResNet100 (a deeper backbone), with a regular classifier on the training set, MS1MV2. Then, uncertainty modules are jointly trained based upon these backbones, respectively.

Table 2: Comparison results on MegaFace. Models are trained on MS1MV2. “Ver.” refers to face verification TAR(@FAR=1e-6). “Id.” denotes rank-1 identification accuracy.

	ResNet24		ResNet100	
Metric	Id.	Ver.	Id.	Ver.
SphereFace	73.12	87.19	75.63	89.40
CosFace	78.84	92.09	80.56	96.56
+ PFE-G	78.91	91.78	80.23	96.51
+ PFE-v	N/A	N/A	N/A	N/A
+ HypersFace	**80.03**	**93.12**	**80.95**	**96.72**
ArcFace	79.12	92.35	81.11	97.00
+ PFE-G	N/A	N/A	N/A	N/A
+ PFE-v	N/A	N/A	N/A	N/A
+ HypersFace	**80.42**	**92.93**	**81.43**	**97.01**

Table 3: Comparison results on IJB-C. Models with different backbones are trained on MS1MV2. The evaluation metric is 1:1 verification TAR@FAR at 1e-4 and 1e-5, respectively.

	ResNet24		ResNet100	
TAR@FAR	1e-4	1e-5	1e-4	1e-5
SphereFace	87.91	85.10	89.32	88.51
CosFace	92.63	89.39	95.34	93.13
+ PFE-G	92.60	89.55	95.42	93.15
+ PFE-v	N/A	N/A	N/A	N/A
+ HypersFace	**93.43**	**90.67**	**95.75**	**93.24**
ArcFace	92.81	89.77	95.65	93.15
+ PFE-G	N/A	N/A	N/A	N/A
+ PFE-v	N/A	N/A	N/A	N/A
+ HypersFace	**93.57**	**90.43**	**95.76**	**93.23**

Results on LFW, CFP-FP, AgeDB, CALFW, CPLFW. As shown in Table 1, our proposed framework yields state-of-the-art performance on LFW, CFP-FP, AgeDB, CALFW and CPLFW when built upon ArcFace and CosFace with various deep backbones, whereas PFE either fails or exhibits inferior improvements in different cases. This showcases the effectiveness and the wide applicability of our proposed framework.

Results on MegaFace. There are two testing scenarios in MegaFace including identification and verification. As shown in Table 2, our model achieves the highest rank-1 accuracy and TAR(@FAR=1e-6) accuracy performance among all of the existing methods.

Results on IJB-C. We evaluate models by the 1:1 verification TAR@FAR protocol on IJB-C. As shown in Table 3, HypersFace outperforms PFE as well as state-of-the-art deterministic embeddings at different FARs (1e-4 and 1e-5). This benefits from the theoretical correctness of HypersFace as compared to PFE; for PFE, improper feature fusion accumulates representation error, leading to suboptimal recognition performance.

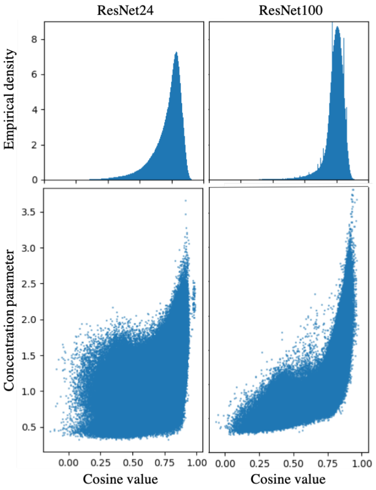


Figure 4: Empirical correlation (bottom) between cosine value $\cos\langle\mu(\mathbf{x}), \mathbf{w}_{\mathbf{x}\in c}\rangle$ and concentration value κ and its marginalized empirical density of cosine value (top) on two backbones.

Empirically, we note that the proposed framework exhibits clearer advantages over PFE with a shallower backbone than with a deeper one. In the next section, we will make a quantitative analysis of this phenomenon.

3.5 Quantitative Analysis

We demonstrate the latent manifold learned by our framework, HypersFace. As illustrated in Figure 4, 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, where $\langle\cdot,\cdot\rangle$ denotes the angle between two vectors. The closer the angular distance between $\mu(\mathbf{x})$ and $\mathbf{w}_{\mathbf{x}\in c}$, the higher the concentration value becomes. This indicates 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. We also visualize what the learned concentration values represent. As shown in Figure 3, higher concentration values correspond to high quality frontal face images whereas lower ones correspond to those with large pose variations, low quality and partial occlusion, or even mislabeled examples. Along with Figure 4, this indicates that concentration values estimated by our uncertainty module indeed captures the data uncertainty in hyperspherical space.

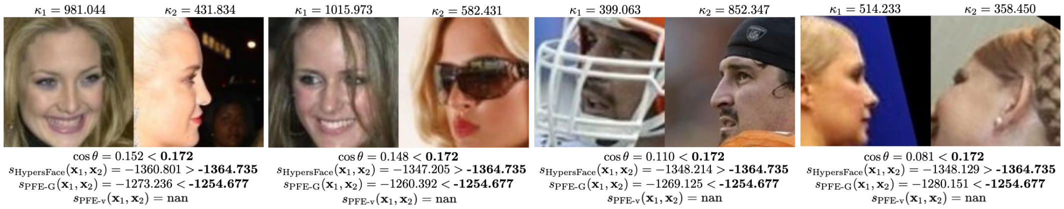

$\kappa_1 = 981.044$
$\kappa_2 = 431.834$
$\cos\theta = 0.152 < \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1360.801 > \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1273.236 < \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$
$\kappa_1 = 1015.973$
$\kappa_2 = 582.431$
$\cos\theta = 0.148 < \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1347.205 > \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1260.392 < \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$
$\kappa_1 = 399.063$
$\kappa_2 = 852.347$
$\cos\theta = 0.110 < \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1348.214 > \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1269.125 < \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$
$\kappa_1 = 514.233$
$\kappa_2 = 358.450$
$\cos\theta = 0.081 < \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1348.129 > \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1280.151 < \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$

Figure 5: False negative examples made by PFE while being true positive by HypersFace, 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 -1254.677 and -1364.735 for PFE (accuracy: 88.210) and HypersFace (accuracy: 88.883), respectively, on the CPLFW benchmark.

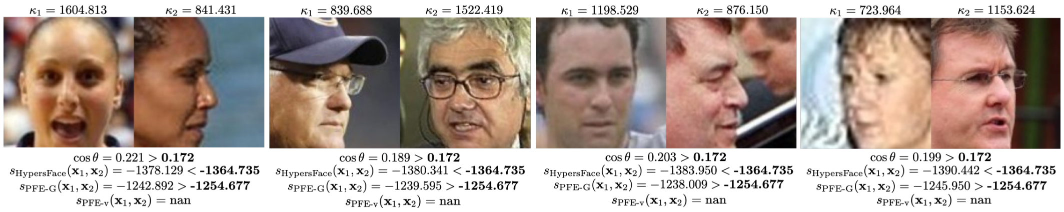

$\kappa_1 = 1604.813$
$\kappa_2 = 841.431$
$\cos\theta = 0.221 > \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1378.129 < \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1242.892 > \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$
$\kappa_1 = 839.688$
$\kappa_2 = 1522.419$
$\cos\theta = 0.189 > \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1380.341 < \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1239.595 > \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$
$\kappa_1 = 1198.529$
$\kappa_2 = 876.150$
$\cos\theta = 0.203 > \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1383.950 < \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1238.009 > \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$
$\kappa_1 = 723.964$
$\kappa_2 = 1153.624$
$\cos\theta = 0.199 > \mathbf{0.172}$
$s_{\text{HypersFace}}(\mathbf{x}_1, \mathbf{x}_2) = -1390.442 < \mathbf{-1364.735}$
$s_{\text{PFE-G}}(\mathbf{x}_1, \mathbf{x}_2) = -1245.950 > \mathbf{-1254.677}$
$s_{\text{PFE-v}}(\mathbf{x}_1, \mathbf{x}_2) = \text{nan}$

Figure 6: False positive examples made by PFE while being true positive by HypersFace, 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 -1254.677 and -1364.735 for PFE (accuracy: 88.210) and HypersFace (accuracy: 88.883), respectively, on the CPLFW benchmark.

Noticeably, as shown in Table 1, 2 and 3, the improvement of the proposed uncertainty learning framework using the shallower backbone ResNet24 is consistently higher than that using the deeper backbone ResNet100. The empirical density of cosine value marginalized from the joint density in Figure 4 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 highly separable latent hyperspherical 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 hyperspherical embeddings more scattered around the mean direction, whereby the uncertainty module shows its clearer advantage in assigning proper concentrating parameters, thereby making correct predictions.

3.6 Qualitative Analysis

We conduct qualitative analysis of the advantage of the proposed framework over that of PFE. As shown in Figure 5 and Figure 6, CosFace and PFE both fail to make correct predictions due to the large pose variations and low-quality images whereas HypersFace is able to assign proper concentration values to face images under different conditions, thereby making correct predictions. More detailed analyses are relegated to Appendix E.

4 Concluding Remarks

A plethora of research has demonstrated the advantage of spherical latent space in modelling certain types of data (Fisher et al., 1993; Reisinger et al., 2010; Wilson et al., 2014). Yet, modelling uncertainty in hyperspherical space remains underexploited. Our work bridges this gap by proposing a general framework for uncertainty learning in hyperspherical space. Towards going *beyond* face recognition, e.g text modeling (Guu et al., 2018) and link prediction (Davidson et al., 2018), we believe that the presented research sheds light on a promising direction towards learning uncertainty of general data whose manifold is not trivially Euclidean.

From the theoretical and empirical views, we have identified two main failure cases of the existing framework for uncertainty learning when it is applied to hyperspherical embeddings. To address these issues, we have proposed a novel framework for hyperspherical face uncertainty 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 uncertainty learning.

REFERENCES

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, pp. 581–588. IEEE, 2005.

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*, pp. 856–865. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

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*, pp. 4690–4699, 2019.

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

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.

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*, pp. 770–778, 2016.

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.

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.

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*, pp. 4873–4882, 2016.

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*, pp. 3499–3506, 2013.

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*, pp. 212–220, 2017.

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*, pp. 51–59, 2017.

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)*, pp. 903–910, 2010.

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)*, pp. 1–9. IEEE, 2016.

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

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

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*, pp. 1041–1049, 2017.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018a.

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*, pp. 5265–5274, 2018b.

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

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*, pp. 90–98, 2017.

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.

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.

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.