Appendices for Spherical Confidence Learning for Face Recognition

A. Proof of Theorem 1 and Corollary 1

Proof. 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 $|\arg z| < \pi/2$,

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

we have $\mathcal{I}_{d/2-1}(\kappa) \sim e^\kappa/\sqrt{2\pi\kappa}$ as $\kappa \to \infty$. Then, the r-radius vMF posterior can be rewritten as

$$p(\mathbf{z}|\mathbf{x}) = \frac{\kappa^{d/2-1}}{(\sqrt{2\pi}r)^d \mathcal{I}_{d/2-1}(\kappa)} \exp\left(\frac{\kappa}{r}\boldsymbol{\mu}^T\mathbf{z}\right) = \frac{\sqrt{2\pi\kappa}\cdot\kappa^{d/2-1}}{(\sqrt{2\pi}r)^d \exp\left(\kappa\left(1-\frac{1}{r}\boldsymbol{\mu}^T\mathbf{z}\right)\right)} \tag{A.2}$$

Therefore, when $\boldsymbol{\mu} = \mathbf{z}/r$, $p(\mathbf{z}|\mathbf{x}) \to \infty$ as $\kappa \to \infty$; otherwise, $p(\mathbf{z}|\mathbf{x}) \to 0$ as $\kappa \to \infty$.

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

B. The Culprit for PFE-v Training Instability

This section empirically shows that simply changing Gaussian into r-vMF does not solve the issue within the framework of PFE and demonstrates the culprit for why PFE-v suffers from training instability. After the change of density, the resulting PFE objective becomes:

$$\min_{\kappa} \mathbb{E}_{(\mathbf{x}^i,\mathbf{x}^j)\sim\mathcal{P}}[-s(\mathbf{x}^i, \mathbf{x}^j)] \tag{A.3}$$

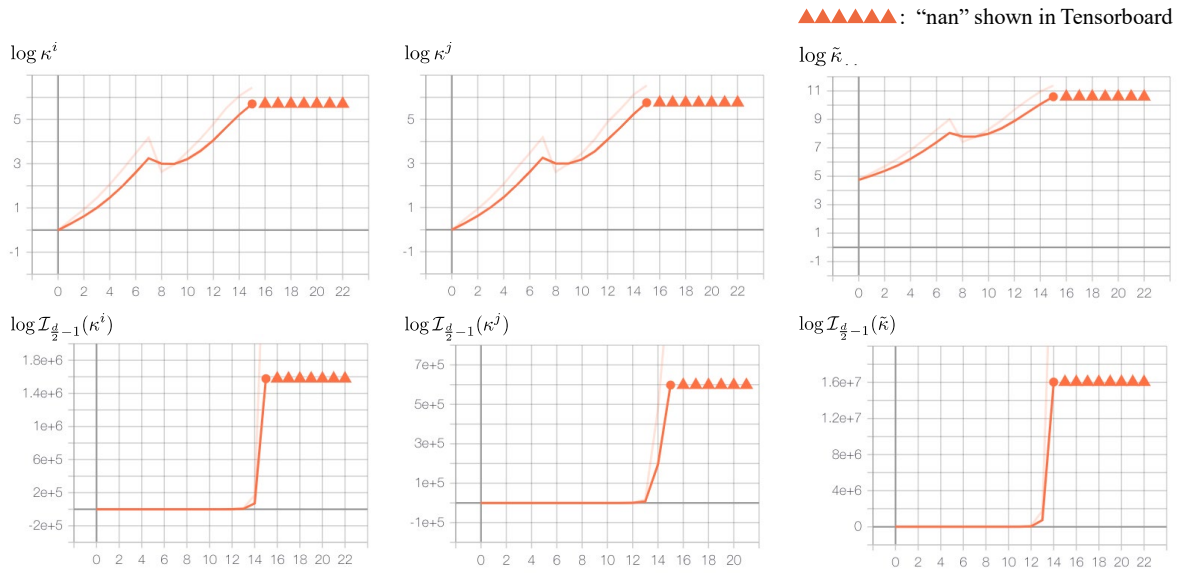

Figure A.1. Training dynamics of all major terms in the PFE-v optimization objective.

where $\mathcal{P}$ is the distribution of all genuine pairs and $s(\cdot,\cdot)$ is calculated by integration over spheres:

$$\begin{aligned}
s(\mathbf{x}^i,\mathbf{x}^j) &= \log p(\mathbf{z}^i=\mathbf{z}^j) \\
&= \log \iint_{r\mathbb{S}^{d-1}\times r\mathbb{S}^{d-1}} p(\mathbf{z}^i|\mathbf{x}^i)p(\mathbf{z}^j|\mathbf{x}^j)\delta(\mathbf{z}^i-\mathbf{z}^j)d\mathbf{z}^i d\mathbf{z}^j \\
&= \log \frac{\mathcal{C}_d(\kappa^i)\mathcal{C}_d(\kappa^j)}{r^{2d}} \int_{r\mathbb{S}^{d-1}} \exp\left(\frac{1}{r}(\kappa^i\boldsymbol{\mu}^i+\kappa^j\boldsymbol{\mu}^j)^T\mathbf{z}\right)d\mathbf{z} \\
&= \log \frac{\mathcal{C}_d(\kappa^i)\mathcal{C}_d(\kappa^j)}{r^d\mathcal{C}_d(\tilde{\kappa})} \underbrace{\int_{r\mathbb{S}^{d-1}} \frac{\mathcal{C}_d(\tilde{\kappa})}{r^d}\exp\left(\frac{\tilde{\kappa}}{r}\tilde{\boldsymbol{\mu}}^T\mathbf{z}\right)d\mathbf{z}}_{=1} \\
&= \log\mathcal{C}_d(\kappa^i)+\log\mathcal{C}_d(\kappa^j)-\log\mathcal{C}_d(\tilde{\kappa})-d\log r \\
&= \left(\frac{d}{2}-1\right)(\log\kappa^i+\log\kappa^j-\log\tilde{\kappa})+\left(\log\mathcal{I}_{\frac{d}{2}-1}(\tilde{\kappa})-\log\mathcal{I}_{\frac{d}{2}-1}(\kappa^i)-\log\mathcal{I}_{\frac{d}{2}-1}(\kappa^j)\right)-\frac{d}{2}\log 2\pi
\end{aligned} \tag{A.4}$$

Here, $\tilde{\kappa}=||\mathbf{p}||_2$, $\mathbf{p}=(\kappa^i\boldsymbol{\mu}^i+\kappa^j\boldsymbol{\mu}^j)$, $\tilde{\boldsymbol{\mu}}=\mathbf{p}/||\mathbf{p}||_2$. As discussed in Remark 4, this PFE objective (A.3) has to be computed in a pairwise fashion. At an early training stage when the learnable parameters of the confidence module are almost randomized, training tends to be unstable: for genuine pairs, the angle between $\boldsymbol{\mu}^i$ and $\boldsymbol{\mu}^j$ are typically smaller than $\pi/2$; as shown in Figure A.1, with fairly large κ^i and κ^j, $\tilde{\kappa}=||\kappa^i\boldsymbol{\mu}^i+\kappa^j\boldsymbol{\mu}^j||_2$ tends to be extremely large (according to the parallelogram principle), resulting in overflow and thus 'nan' values. Unlike the PFE objective, the proposed SCF objective (A.5) does not involve such an exploding term as $\tilde{\kappa}=||\kappa^i\boldsymbol{\mu}^i+\kappa^j\boldsymbol{\mu}^j||_2$, readily leading to a stable training process without bells and whistles.

$$-\frac{\kappa(\mathbf{x})}{r}\mu(\mathbf{x})^T\mathbf{w}_{\mathbf{x}\in c}-\left(\frac{d}{2}-1\right)\log\kappa(\mathbf{x})+\log(\mathcal{I}_{d/2-1}(\kappa(\mathbf{x})))+\frac{d}{2}\log 2\pi r^2 \tag{A.5}$$

C. Estimation Errors Incurred by PFE-G or SCF-G

This section gives a theoretical analysis of uncertainty estimation errors by PFE-G and SCF-G in risk-controlled scenarios. In a risk-controlled recognition task, all facial images are sorted in an ascending order according to their confidence before filtering out and verification. Unlike our proposed framework, PFE-G or SCF-G predicts uncertainty σ_l for each dimension ($l=1,...,d$), and then compute the average of them as an estimate of the overall uncertainty. Then, confidence scores can be obtained by taking the inverse of such estimates. Here we show in Proposition 1 that this treatment is equivalent to finding one single positive scalar σ that minimizes the L^2-Wasserstein distance between an independent Gaussian and an isotropic Gaussian.

Proposition 1. *The overall uncertainty* $\sigma^*=\arg\min_\sigma \mathcal{W}_2(\mathcal{N}(\boldsymbol{\mu},\mathrm{diag}(\sigma_1,...,\sigma_d)),\mathcal{N}(\boldsymbol{\mu},\sigma\mathbf{I}))$, *of which the solution is given by* $\sigma^*=(\sigma_1+...+\sigma_d)/d$.

Proof. By definition, the L^2-Wasserstein distance between any two Gaussian measures $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ (see Theorem 2.2 in [1]) is given by

$$\mathcal{W}_2(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)),\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2))^2=||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2+\mathrm{tr}\,\boldsymbol{\Sigma}_1+\mathrm{tr}\,\boldsymbol{\Sigma}_2-2\,\mathrm{tr}\sqrt{\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{1/2}} \tag{A.6}$$

and in this case further boils down to

$$\mathcal{W}_2(\mathcal{N}(\boldsymbol{\mu},\mathrm{diag}(\sigma_1,...,\sigma_d)),\mathcal{N}(\boldsymbol{\mu},\sigma\mathbf{I}))^2=||\boldsymbol{\Lambda}^{1/2}-\sigma\mathbf{I}||_F^2 \tag{A.7}$$

where $\boldsymbol{\Lambda}^{1/2}=\mathrm{diag}(\sigma_1,...,\sigma_d)$. Then, simple calculus concludes the proof by showing that the optimal σ^* that minimizes the distance is the arithmetic mean of all dimension-wise uncertainties σ_l, $l=1,...,d$. □

The L^2-Wasserstein distance equals to zero if and only if all predicted dimension-wise uncertainties σ_l's are identical. Otherwise, there always exist uncertainty estimation errors given by PFE-G or SCF-G. In constrast, our proposed framework SCF operates with an r-radius vMF rather than an independent Gaussian, which predicts one single positive scalar κ as confidence in a straightforward yet principled way. This dispenses with dimension-wise uncertainty estimation and post-aggregation that is prone to errors.

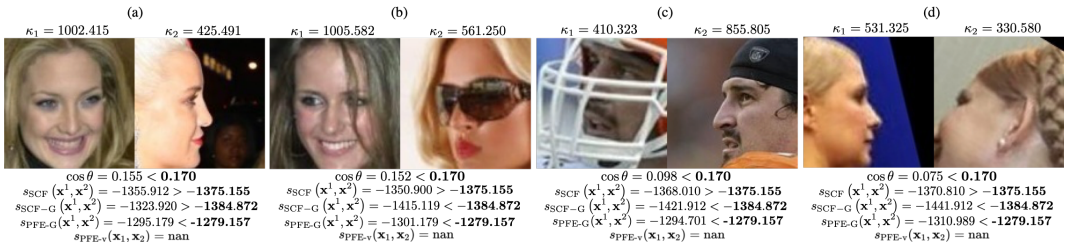

Figure A.2. False negative examples made by PFE or SCF-G while being true positive by SCF, where $\cos\theta$ is the cosine distance of a verification pair $\mathbf{x}^1, \mathbf{x}^2$, $s(\cdot, \cdot)$ is mutual likelihood score and κ^1, κ^2 are the corresponding concentration values. Thresholds are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

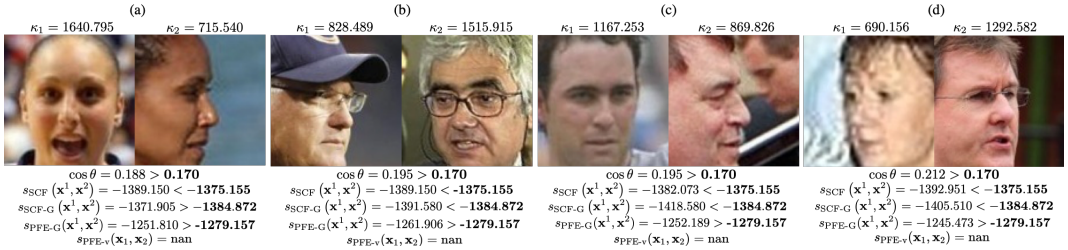

Figure A.3. False positive examples made by PFE or SCF-G while being true positive by SCF, where $\cos\theta$ is the cosine distance of a verification pair $\mathbf{x}^1, \mathbf{x}^2$, $s(\cdot, \cdot)$ is mutual likelihood score and κ^1, κ^2 are the corresponding concentration values. Thresholds are set to -1279.157, -1384.872 and -1375.155 for PFE (accuracy: 89.90), SCF-G (accuracy: 89.83) and SCF (accuracy: 90.80), respectively, on the CPLFW benchmark.

D. Qualitative Analysis

As shown in Figure A.2, PFE or SCF-G fails to make correct predictions given genuine pairs due to the large pose variations (a)(d) and mask or sunglasses wearing (b)(c), whereas SCF is able to assign different concentration values to face images under different conditions. Higher concentration values indicate more confidence (thus less overall uncertainty involved) for our model to make predictions. Specifically, in the cases of large pose variations (a)(d) and severe or partial occlusions (b)(c), SCF adaptively gives lower concentration values to these images, thereby making correct predictions that PFE and SCF-G do not.

Note that in Figure A.2(d), SCF assigns relatively low concentration values to both images but gives the first a slightly higher one than the second, since more facial clues can be seen from the first one. Figure A.3 also demonstrates that PFE or SCF-G is unable to distinguish imposter pairs due to large pose variations (a)(c), partial occlusion (b) and image blur (d), whereas our model, SCF, successfully makes correct predictions by adaptively assigning proper concentration values to corresponding face images also as in Figure A.2.

References

[1] Asuka Takatsu et al. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011. 2