Appendices for
Probablisitic Knowledge Distillation of Face Ensembles

Jianqing Xu* Shen Li*
Ailin Deng Miao Xiong Jiaying Wu Jiaxiang Wu Shouhong Ding Bryan Hooi
Youtu Lab, Tencent. IDS and SoC, National University of Singapore.
{joejqxu, willjxwu, ericshding}@tencent.com
{shen.li, ailin, miao.xiong, jiayingwu}@u.nus.edu bhooi@comp.nus.edu.sg

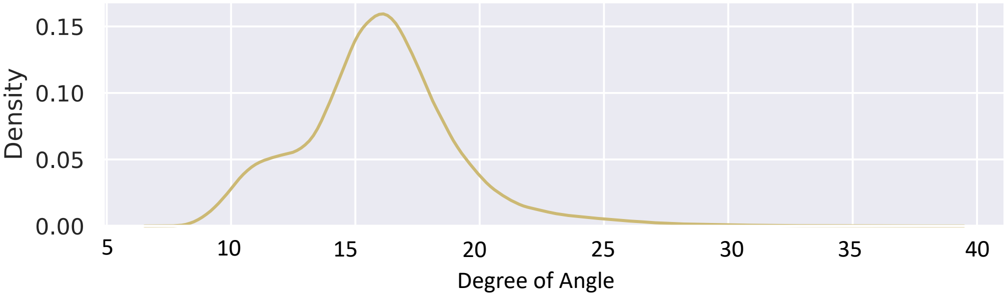

Figure A.1. The distribution of pairwise degree of angle. Most angles are concentrated around ~ 16 degrees, indicating that the angular distances between feature embeddings given by different ensemble members are not trivially close to zero.

A. Discussions on Ensemble Diversity

Ensembles are created by combining several individual models. It is widely accepted [7, 17] that the prediction performance of an ensemble jointly depends of the individual performance and the diversity of its individual members. In particular, diversity has long been recognized as a key factor in ensemble performance [13, 6, 2]. In the case of deep ensembles, diversity is achieved via some implicit techniques such as random initialization [14, 27], tweaking the optimizer [18, 30] or employing different hyperparameter settings [28]. Therefore, in Algorithm 1, we train the ensemble members with different initialization, different optimizers and different hyperparameters including m_1, m_2 and m_3 in Eq. (1). As such, different ensemble members are expected to converge at different local minima, yet without the loss of recognition performance.

To corroborate it, we visualize the distribution of the angular distance between each pair of ensemble members given each face image as shown in Figure A.1. We empirically find that the mode of the empirical distribution (~ 16 degrees instead of zero) indicates features of different ensemble members are not trivially similar. Their diversity achieved by random initialization, optimizer and hyperparameter tweaking gives rise to the empirical improvement of an ensemble over a single model, as shown in Table 1 in the main text.

B. Derivation of Bayesian Ensemble Averaging (BEA)

This section shows how Eq. (5)(6)(7) in the main text are derived. Note that the ensemble posterior for $q(\boldsymbol{z}|\boldsymbol{x})$ to emulate is

$$p(\boldsymbol{z}|f_{\theta_1},...,f_{\theta_n},\boldsymbol{x}) := p(\boldsymbol{z}|f_{\theta_1}(\boldsymbol{x}),...,f_{\theta_n}(\boldsymbol{x})) \tag{A.1}$$

$$= \frac{p(f_{\theta_1}(\boldsymbol{x}),...,f_{\theta_n}(\boldsymbol{x})|\boldsymbol{z}) \cdot p(\boldsymbol{z})}{p(f_{\theta_1}(\boldsymbol{x}),...,f_{\theta_n}(\boldsymbol{x}))} \tag{A.2}$$

*Equal contribution.

Assume that $f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_n}(\boldsymbol{x})$ are conditionally independent given the latent $\boldsymbol{z}$. Then, Eq. (A.2) reads

$$
\begin{aligned}
&\frac{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})|\boldsymbol{z}) \cdot p(f_{\theta_n}(\boldsymbol{x})|\boldsymbol{z}) \cdot p(\boldsymbol{z})}{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_n}(\boldsymbol{x}))} && \text{(A.3)}\\
&= \frac{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})) \cdot p(f_{\theta_n}(\boldsymbol{x}))}{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_n}(\boldsymbol{x}))} \cdot \frac{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})|\boldsymbol{z}) \cdot p(f_{\theta_n}(\boldsymbol{x})|\boldsymbol{z}) \cdot p(\boldsymbol{z})}{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})) \cdot p(f_{\theta_n}(\boldsymbol{x}))} && \text{(A.4)}\\
&\propto \frac{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})|\boldsymbol{z}) \cdot p(f_{\theta_n}(\boldsymbol{x})|\boldsymbol{z}) p(\boldsymbol{z})}{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})) \cdot p(f_{\theta_n}(\boldsymbol{x}))} && \text{(A.5)}\\
&= \frac{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x}), \boldsymbol{z}) \cdot p(f_{\theta_n}(\boldsymbol{x}), \boldsymbol{z})}{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})) \cdot p(f_{\theta_n}(\boldsymbol{x})) \cdot p(\boldsymbol{z})} && \text{(A.6)}\\
&= \frac{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x}), \boldsymbol{z})}{p(f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x}))} \cdot \frac{p(f_{\theta_n}(\boldsymbol{x}), \boldsymbol{z})}{p(f_{\theta_n}(\boldsymbol{x})) \cdot p(\boldsymbol{z})} && \text{(A.7)}\\
&= p(\boldsymbol{z}|f_{\theta_1}(\boldsymbol{x}), ..., f_{\theta_{n-1}}(\boldsymbol{x})) \cdot \frac{p(\boldsymbol{z}|f_{\theta_n}(\boldsymbol{x}))}{p(\boldsymbol{z})} && \text{(A.8)}\\
&=: p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_{n-1}}, \boldsymbol{x}) \cdot \frac{p(\boldsymbol{z}|f_{\theta_n}(\boldsymbol{x}))}{p(\boldsymbol{z})} && \text{(A.9)}\\
&\propto p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_{n-1}}, \boldsymbol{x}) \cdot p(\boldsymbol{z}|f_{\theta_n}(\boldsymbol{x})) && \text{(A.10)}
\end{aligned}
$$

Note that Eq. (A.10) is obtained by assuming the prior $p(\boldsymbol{z})$ is a uniform distribution over the hypersphere of radius r, which is a constant (i.e. the inverse of the hypersphere's surface) that is independent from $\boldsymbol{z}$. This assumption is reasonable, because before seeing any samples $f_{\theta_i}(\boldsymbol{x})$ $(i = 1, ..., n)$, the best guess of $\boldsymbol{z}$ is the uniform distribution over a finite support. Also note that our model operates over the hypersphere (a finite support) rather than Euclidean space (an infinite support). Hence, the derivation can proceed. Therefore, we obtain Eq. (5) in the main text. Applying this recursive equation for n times, we have

$$p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_n}, \boldsymbol{x}) \propto \prod_{i=1}^{n} p(\boldsymbol{z}|f_{\theta_i}(\boldsymbol{x})) \tag{A.11}$$

Also note that $p(\boldsymbol{z}|f_{\theta_i}(\boldsymbol{x})) = r\text{-vMF}(\boldsymbol{z}; \boldsymbol{\mu}_i, \kappa_i)$. Then,

$$
\begin{aligned}
\prod_{i=1}^{n} p(\boldsymbol{z}|f_{\theta_i}(\boldsymbol{x})) &= \prod_{i=1}^{n} \frac{\mathcal{C}_d(\kappa_i)}{r^d} \exp\left(\frac{\kappa_i}{r} \boldsymbol{\mu}_i^T \boldsymbol{z}\right) && \text{(A.12)}\\
&\propto \exp\left(\frac{1}{r} \left(\sum_{i=1}^{n} \kappa_i \boldsymbol{\mu}_i\right)^T \boldsymbol{z}\right) && \text{(A.13)}\\
&= \exp\left(\frac{\| \sum_{i=1}^{n} \kappa_i \boldsymbol{\mu}_i \|_2}{r} \left(\frac{\sum_{i=1}^{n} \kappa_i \boldsymbol{\mu}_i}{\| \sum_{i=1}^{n} \kappa_i \boldsymbol{\mu}_i \|_2}\right)^T \boldsymbol{z}\right) && \text{(A.14)}
\end{aligned}
$$

We recognize Eq. (A.14) as an r-vMF distribution (up to a normalization constant) with the mean $\frac{\sum_{i=1}^{n} \kappa_i \boldsymbol{\mu}_i}{\| \sum_{i=1}^{n} \kappa_i \boldsymbol{\mu}_i \|_2}$ and the concentration value $\| \sum_{i=1}^{n} \kappa_i \boldsymbol{\mu}_i \|_2$. This yields Eq. (5)(6)(7) in the main text.

C. Proof of Lemma 1

Lemma 1. *For any $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ and $\kappa > 0$, the differential entropy of r-vMF$(\boldsymbol{\mu}, \kappa)$ has an analytic form: $-\kappa - (\frac{d}{2} - 1) \log \kappa + \log \mathcal{I}_{d/2-1}(\kappa) + \frac{d}{2} \log 2\pi$. And it is a monotonically decreasing function of κ in $(0, +\infty)$.*

Proof. According to the definition of differential entropy, the differential entropy of $p(\boldsymbol{z}) = r\text{-vMF}(\boldsymbol{z}; \boldsymbol{\mu}, \kappa)$ can be written as

$$
\begin{aligned}
\mathcal{H}[p] &= -\int_{r\mathbb{S}^{d-1}} p(\boldsymbol{z}) \log p(\boldsymbol{z}) d\boldsymbol{z}\\
&= -\int_{r\mathbb{S}^{d-1}} \left[\frac{\kappa}{r} \boldsymbol{\mu}^T \boldsymbol{z} + \log C_d(\kappa) - d \log r\right] p(\boldsymbol{z}) d\boldsymbol{z}\\
&= -\frac{\kappa}{r} \boldsymbol{\mu}^T \mathbb{E}[\boldsymbol{z}] - \log C_d(\kappa) + d \log r\\
&= -\kappa - \left(\frac{d}{2} - 1\right) \log \kappa + \log \mathcal{I}_{d/2-1}(\kappa) + \frac{d}{2} \log 2\pi
\end{aligned} \tag{A.15}
$$

Note that $\mathcal{H}[p]$ is not dependent on the mean direction $\boldsymbol{\mu}$ but is a univariate function of κ, i.e. $\mathcal{H}[p](\kappa)$ for $\kappa > 0$. Taking derivative wrt κ yields

$$\frac{\partial \mathcal{H}[p](\kappa)}{\partial \kappa} = \left(-1 - \frac{d-2}{2\kappa}\right) + \frac{1}{\mathcal{I}_{d/2-1}(\kappa)} \frac{\partial \mathcal{I}_{d/2-1}(\kappa)}{\partial \kappa} \tag{A.16}$$

For the second term, we exploit the recurrence between the derivative and the Bessel function itself. Specifically, for any $\nu > 0$, $\kappa > 0$, we have the recurrence

$$\frac{\partial \mathcal{I}_\nu(\kappa)}{\partial \kappa} = \frac{\nu}{\kappa}\mathcal{I}_\nu(\kappa) + \mathcal{I}_{\nu+1}(\kappa) \tag{A.17}$$

By virtue of the recurrence, Eq. (A.16) becomes

$$\frac{\partial \mathcal{H}[p](\kappa)}{\partial \kappa} = \frac{\mathcal{I}_{d/2}(\kappa)}{\mathcal{I}_{d/2-1}(\kappa)} - 1 \tag{A.18}$$

Soni [24] showed that $\mathcal{I}_\nu(\kappa) > \mathcal{I}_{\nu+1}(\kappa)$ for all $\nu > -\frac{1}{2}$ and $\kappa > 0$. Taking $\nu = d/2 - 1$ completes the proof. □

D. Proof of Proposition 3.2

Proposition 3.2 *For any $\boldsymbol{x} \in \mathcal{X}$, suppose the first-order moment of the conditional $p(\boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})|\mathcal{D}_\mathcal{X}, \boldsymbol{x})$ exists and that the conditional $p(\varphi^*|\phi^*, \mathcal{D}_\mathcal{X})$ is a point mass, i.e., $p(\varphi^*|\phi^*, \mathcal{D}_\mathcal{X}) = \delta(\varphi^* - \varphi_0^*)$ for some φ_0^*. Then, in the limit of infinite ensemble members, the aleatoric uncertainty of $\boldsymbol{x}$, $\mathcal{A}(\boldsymbol{x}) := \mathbb{E}_{p(\phi^*, \varphi^*|\mathcal{D}_\mathcal{X})}\left[\mathcal{H}\left[p(\boldsymbol{z}|\boldsymbol{x}, \phi^*, \varphi^*)\right]\right]$, is a monotonically decreasing function of the confidence measure $\bar{\kappa}_{\boldsymbol{x}}^{(\infty)}$, where*

$$\bar{\kappa}_{\boldsymbol{x}}^{(\infty)} := \lim_{n\to\infty} \frac{\|\kappa_1\boldsymbol{\mu}_1 + ... + \kappa_n\boldsymbol{\mu}_n\|_2}{n} \propto \kappa_{\varphi_0^*}(\boldsymbol{x}) \tag{A.19}$$

Proof. Note that Eq. (A.19) can be written as

$$\begin{aligned}
\bar{\kappa}_{\boldsymbol{x}}^{(\infty)} &= \|\mathbb{E}[\kappa\boldsymbol{\mu}|\boldsymbol{x}, \mathcal{D}_\mathcal{X}]\|_2 \\
&= \left\|\mathbb{E}_{p(\phi^*, \varphi^*|\mathcal{D}_\mathcal{X})}[\kappa_{\varphi^*}(\boldsymbol{x})\boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})|\boldsymbol{x}]\right\|_2 \\
&= \left\|\iint \delta(\varphi^* - \varphi_0^*)p(\phi^*|\mathcal{D}_\mathcal{X})\kappa_{\varphi^*}(\boldsymbol{x})\boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})d\varphi^* d\phi^*\right\|_2 \\
&= \left\|\int \boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})p(\phi^*|\mathcal{D}_\mathcal{X})\Big(\int \kappa_{\varphi^*}(\boldsymbol{x})\delta(\varphi^* - \varphi_0^*)d\varphi^*\Big)d\phi^*\right\|_2 \\
&= \kappa_{\varphi_0^*}(\boldsymbol{x})\left\|\int \boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})p(\phi^*|\mathcal{D}_\mathcal{X})d\phi^*\right\|_2
\end{aligned} \tag{A.20}$$

Note that for any $\boldsymbol{x} \in \mathcal{X}$, the first-order moment of the conditional $p(\boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})|\mathcal{D}_\mathcal{X}, \boldsymbol{x})$ exists. Then by definition, it can be written as

$$\mathbb{E}[\boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})|\mathcal{D}_\mathcal{X}, \boldsymbol{x}] = \int \boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})p(\phi^*|\mathcal{D}_\mathcal{X})d\phi^* \tag{A.21}$$

Due to its existence, its norm is finite, i.e.,

$$0 < \left\|\int \boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})p(\phi^*|\mathcal{D}_\mathcal{X})d\phi^*\right\|_2 = C < \infty \tag{A.22}$$

Hence, $\kappa_{\varphi_0^*}(\boldsymbol{x}) = \bar{\kappa}_{\boldsymbol{x}}^{(\infty)}/C$. On the other hand, according to Lemma 1, the aleatoric uncertainty $\mathcal{A}(\boldsymbol{x})$ can be written as

$$\begin{aligned}
\mathbb{E}_{p(\phi^*, \varphi^*|\mathcal{D}_\mathcal{X})}\left[\mathcal{H}\left[p(\boldsymbol{z}|\boldsymbol{x}, \phi^*, \varphi^*)\right]\right] &= \mathbb{E}_{p(\phi^*, \varphi^*|\mathcal{D}_\mathcal{X})}\left[-\kappa_{\varphi^*}(\boldsymbol{x}) - (\frac{d}{2} - 1)\log\kappa_{\varphi^*}(\boldsymbol{x}) + \log\mathcal{I}_{d/2-1}(\kappa_{\varphi^*}) + \frac{d}{2}\log 2\pi\right] \\
&= -\kappa_{\varphi_0^*}(\boldsymbol{x}) - (\frac{d}{2} - 1)\log\kappa_{\varphi_0^*}(\boldsymbol{x}) + \log\mathcal{I}_{d/2-1}(\kappa_{\varphi_0^*}(\boldsymbol{x})) + \frac{d}{2}\log 2\pi
\end{aligned} \tag{A.23}$$

Then, using the chain rule, we have

$$\frac{\partial\mathcal{A}}{\partial\bar{\kappa}_{\boldsymbol{x}}^{(\infty)}} = \frac{\partial\mathcal{A}}{\partial\kappa_{\varphi_0^*}} \cdot \frac{\partial\kappa_{\varphi_0^*}}{\partial\bar{\kappa}_{\boldsymbol{x}}^{(\infty)}} = \frac{1}{C} \cdot \frac{\partial\mathcal{A}}{\partial\kappa_{\varphi_0^*}} < 0. \tag{A.24}$$

This concludes the proof. □

Table A.1. Space occupancy (#Params) and computational complexity (MACs) for the setting: ResNet12-KD-5ResNet34. $\kappa_\varphi(\cdot)$ is instantiated by `[CONV(256)-ReLU-BN]`$(\times 3)$`-AvgPool-[FC-ReLU]`$(\times 2)$`-FC-exp`.

Model	[S]: ResNet12	[T]: 5 ResNet34	$\kappa_\varphi(\cdot)$
#Params	40 MB	5×90 MB	2.5 MB
MACs	298.884 M	5×971.072 M	16.003 M

Table A.2. Space occupancy (#Params) and computational complexity (MACs) for the setting: MobileFaceNet-KD-5ResNet34. $\kappa_\varphi(\cdot)$ is instantiated by `[CONV(128)-ReLU-BN]`$(\times 3)$`-AvgPool-[FC-ReLU]`$(\times 2)$`-FC-exp`.

Model	[S]: MobileFaceNet	[T]: 5 ResNet34	$\kappa_\varphi(\cdot)$
#Params	4.8 MB	5×90 MB	0.6 MB
MACs	230.334 M	5×971.072 M	4.004 M

E. The Monte Carlo Method for Estimating Uncertainty

Note that the aleatoric uncertainty of a face image $\boldsymbol{x}$, $\mathcal{A}(\boldsymbol{x}) := \mathbb{E}_{p(\phi^*,\varphi^*|\mathcal{D}_\mathcal{X})}\left[\mathcal{H}\left[p(\boldsymbol{z}|\boldsymbol{x},\phi^*,\varphi^*)\right]\right]$, requires estimating the first order moment of the entropy. We approximate it using the ensemble mean, i.e.,

$$\mathcal{A}(\boldsymbol{x}) \approx \frac{1}{n}\sum_{i=1}^{n}\mathcal{H}\left[p\big(\boldsymbol{z}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big)\right] \tag{A.25}$$

where $\{\phi^*_{(i)},\varphi^*_{(i)}\}$ are the learnable parameters of the ith ensemble member $(i = 1, ..., n)$. To approximate the entropy $\mathcal{H}\left[p\big(\boldsymbol{z}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big)\right]$, we draw m Monte-Carlo samples* from $p\big(\boldsymbol{z}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big)$:

$$\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(m)} \overset{iid}{\sim} p\big(\boldsymbol{z}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big) \tag{A.26}$$

Then, the entropy of the ith ensemble member can be estimated by the following finite sum

$$\mathcal{H}\left[p\big(\boldsymbol{z}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big)\right] \approx -\sum_{j=1}^{m}\log p\big(\boldsymbol{z}^{(j)}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big) \tag{A.27}$$

Therefore,

$$\mathcal{A}(\boldsymbol{x}) \approx -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\log p\big(\boldsymbol{z}^{(j)}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big) \tag{A.28}$$

In experiments, we set $m = 100,000$. Similarly, we can estimate the total uncertainty using the similar technique:

$$\mathcal{T}(\boldsymbol{x}) \approx -\sum_{j=1}^{m}\log\left(\frac{1}{n}\sum_{i=1}^{n}p\big(\boldsymbol{z}^{(j)}|\boldsymbol{x},\phi^*_{(i)},\varphi^*_{(i)}\big)\right) \tag{A.29}$$

F. Compression Rate

This section provides a detailed comparison of compression rate. In the two knowledge distillation settings (ResNet12-KD-5ResNet34 and MobileFaceNet-KD-5ResNet34), we compare the models in term of #Params and MACs. According to Table A.1, the compression rate is $1-(40+2.5)/(5\times 90) = 90.56\%$ in terms of #Params and $1-(298.884+16.003)/(5\times 971.072) = 93.51\%$ in terms of MACs. In the second setting (cf. Table A.2), the compression rate can be even higher: $1-(4.8+0.6)/(5\times 90) = 98.80\%$ in terms of #Params and $1-(230.334+4.004)/(5\times 971.072) = 95.17\%$ in terms of MACs.

G. Ablation Study

Recall that we consider two experimental settings to demonstrate the effectiveness of our proposed framework: (1) ResNet12 [11] is employed as the student to distill knowledge from ResNet34; (2) MobileFaceNet [4] is employed as the student to distill knowledge from ResNet34.

*Note that we are not using the results from Lemma 1 to calculate the entropy. Instead, we use the Monte-Carlo method throughout in order to empirically validate the proposed theory.

For ablation studies in the first setting, we consider the following model variants: Single ResNet12, ResNet12-KD-1ResNet34, ResNet12-KD-1ResNet34+SCF, ResNet12-KD-5ResNet34, ResNet12-KD-5ResNet34+SCF, ResNet12-KD-1SCF(ResNet34). Descriptions of these model variants are given as follows:

- SCF(ResNet12): A state-of-the-art face uncertainty learning method built on ResNet12.
- Single ResNet12: The model is trained using ResNet12 as the backbone and the ArcFace loss as the loss function. No knowledge distillation is employed.
- ResNet12-KD-1ResNet34: ResNet12 is trained to distill knowledge from one ResNet34. No BEA or probabilistic treatment is employed.
- ResNet12-KD-1ResNet34 + SCF: After training ResNet12-KD-1ResNet34, an SCF module is trained based on it.
- ResNet12-KD-5ResNet34: ResNet12 is trained to distill knowledge from five ResNet34's. No BEA or probabilistic treatment is employed.
- ResNet12-KD-5ResNet34 + SCF: After training ResNet12-KD-5ResNet34, an SCF module is trained based on it.
- ResNet12-KD-1SCF(ResNet34): ResNet12 is trained to distill knowledge from one SCF that is built on ResNet34.

Please refer to Table A.3 for detailed comparison among these models. The other setting is the same as the first except for the change of the student network (from ResNet12 to MobileFaceNet).

We conduct ablation studies to demonstrate the effectiveness of each contributing components of our approach. As shown in the main text, our approach BEA-KD outperforms all its model variants. Specifically, for example, comparison between BEA-KD and ResNet12-KD-5ResNet34 suggests that the proposed Bayesian averaging posterior acts as a good supervisory signal for uncertainty learning. Comparison with "ResNet12-KD-5ResNet34 + SCF" further suggests that training an SCF built on the deterministic KD is inferior to ours: distilling knowledge of feature embeddings is insufficient even when a probabilistic module SCF is built on it for uncertainty learning. More interestingly, comparison with "ResNet12-KD-1SCF(ResNet34)" indicates that distilling uncertainty from multiple teachers is essential to performance improvements. Note that our proposed method can be seen as "ResNet12-KD-5SCF(ResNet34)". Besides, comparison between "Single ResNet12" and "ResNet12-KD-1ResNet34" showcases that KD from a single teacher is successful but with limited improvements when compared with "ResNet12-KD-5ResNet34".

H. State-of-The-Art Comparison

In Section 4.5 of the main text, we showed that our proposed approach outperforms the state-of-the-art KD methods (TAKD [20], DGKD [23], MEAL [21], AE-KD [8], Hydra [25], CA-MKD [29], Eff-KD [9]). Our method consistently outperforms others across all benchmarks. Note that these KD methods are designed for the *closed-set* classification problem (for small C, typically) and operate in the $(C-1)$-simplex space. Here we provide a more detailed analysis of why the performance of these KD methods are inferior to ours. **First**, when applied in large-scale face recognition where C is as large as hundreds of thousands ($C = 617970$ in our case), these closed-set KD methods suffer from *the curse of dimensionality*: the softmax outputs of the teachers are sparse, which come quite close to one-hot encoding, when face images are correctly classified (cf. Figure A.2). Therefore, when the student distills knowledge from such teachers, it makes almost no difference from learning with ground-truth labels (one-hot encoding), thereby resulting in ineffective knowledge distillation. **Second**, as pointed out in prior literature [10, 12, 5], we further confirm that the teacher classifier's softmax outputs exhibit overconfidence: the predictive softmax outputs tend to be much larger at the predicted classes than at the ground-truth labels when face images are misclassified (cf. Figure A.3). Distilling knowledge from such *overconfident* outputs will make the KD process unreliable. **Third**, the feature embeddings given by these models are deterministic, which cannot address the Feature Ambiguity Dilemma [22]. In contrast, our proposed approach BEA-KD operates in the feature space whose dimensionality is typically much smaller than C (the dimensionality of the label space) in million-scale face recognition. Moreover, the proposed approach BEA-KD is designed for open-set recognition KD and can provide the confidence statistic (BEA statistic) that provably correlates to aleartoric uncertainty in a theoretically-grounding manner. These reasons lead to the superiority of our method.

I. Potential Negative Societal Impact

Face recognition has a broad impact on society, and is often deployed in environments of high uncertainty: face images captured in the wild may display low quality, occlusion and poor light conditions. These adverse

Table A.3. Model descriptions. MSE loss denotes the mean squared error between the feature embedding given by the student network and the desired one given by the teacher(s), respectively. The desired feature embedding is computed by Eq. (4) if multiple teachers are used.

Model	BEA	Loss Function	Similarity measure for inference	Description
SCF(ResNet12)	-	Eq. (3)	Eq. (15) and Eq. (16)	A state-of-the-art face uncertainty learning method built on ResNet12.
BEA-KD (Ours)	✓	Eq. (14)	Eq. (15) or Eq. (16)	As described in Section 3.4
Single ResNet12	-	ArcFace loss	Cosine distance	The model is trained using ResNet12 as the backbone and the ArcFace loss as the loss function. No knowledge distillation is employed.
ResNet12-KD-1ResNet34	-	MSE loss	Cosine distance	ResNet12 is trained to distill knowledge from one ResNet34. No BEA or probabilistic treatment is employed.
ResNet12-KD-1ResNet34 + SCF	-	MSE loss followed by Eq. (3)	Eq. (15) and Eq. (16)	After training ResNet12-KD-1ResNet34, an SCF module is trained based on it.
ResNet12-KD-5ResNet34	-	MSE loss	Cosine distance	ResNet12 is trained to distill knowledge from five ResNet34's. No BEA or probabilistic treatment is employed.
ResNet12-KD-5ResNet34 + SCF	-	MSE loss followed by Eq. (3)	Eq. (15) or Eq. (16)	After training ResNet12-KD-5ResNet34, an SCF module is trained based on it.
ResNet12-KD-1SCF(ResNet34)	-	Eq. (14)	Eq. (15) or Eq. (16)	ResNet12 is trained to distill knowledge from one SCF that is built on ResNet34.

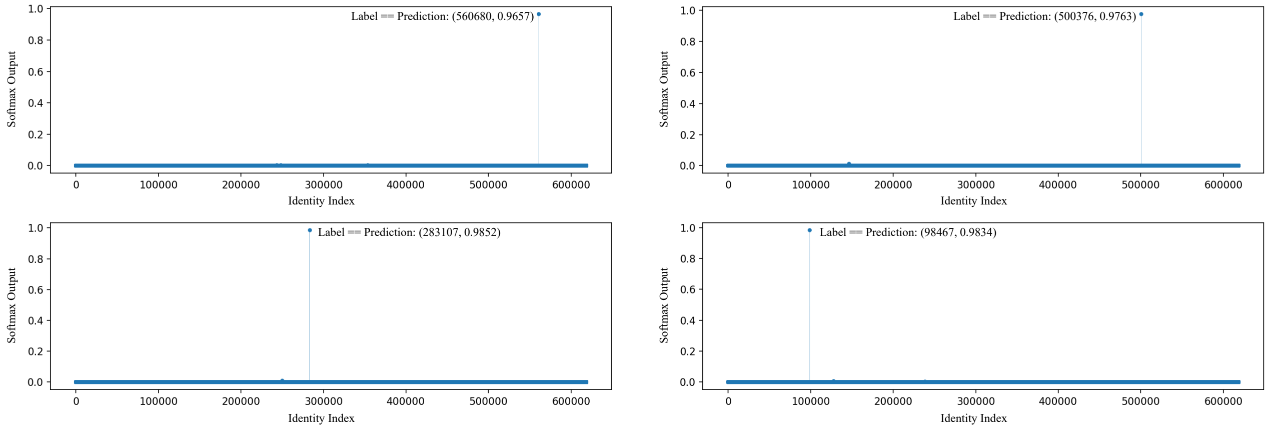

Figure A.2. The softmax outputs of the teacher network (ResNet34) when the input face images are correctly classified (Label == Prediction). Each panel corresponds to the softmax output of one correctly-classified face image (uncurated). We observe that the softmax outputs come quite close to one-hot encoding: the predictive scores get close to 1 (e.g. 0.9657 at 560680) with the rest of the scores dropping down to zeros. The experiment is carried out on WebFace12M which contains 617970 identities in total.

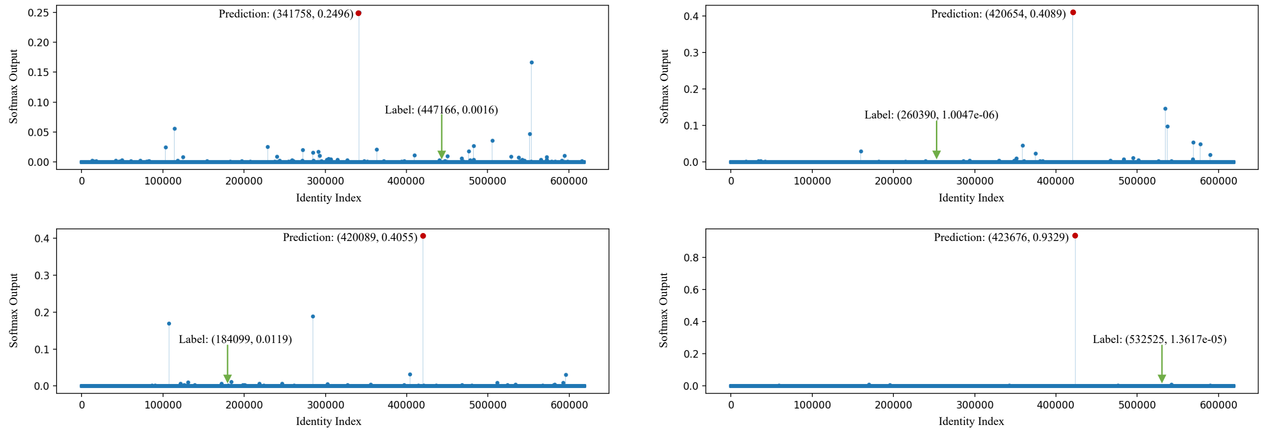

Figure A.3. The softmax outputs of the teacher network (ResNet34) when the input face images are misclassified. Each panel corresponds to the softmax output of one misclassified face image (uncurated). We observe that the softmax outputs display overconfidence: the predictive scores are more than hundreds of times larger than the scores at the groundtruth labels, e.g. 0.2496 at 341758 (marked by red) versus 0.0016 at 447166 (marked by green arrows). The experiment is carried out on WebFace12M which contains 617970 identities in total.

factors cause inaccuracies in face recognition systems. Our research provides a general framework for modelling this uncertainty within the scope of open-set large-scale face recognition.

Face recognition requires great care and thought into its development and deployment. On one hand, it presents significant risks: face recognition systems can produce biased results, such as higher misclassification rate in subgroups due to disparities in the training data [3]; moreover, poorly controlled usage of face recognition systems may compromise the privacy of individuals. Hence, face recognition systems should be managed within a framework embracing principles of privacy and consent: for example, enforcing privacy protections on all personally identifiable information, and requiring informed consent from individuals before they are included in the database [26].

At the same time, face recognition has seen uses, e.g. in the medical domain, which improve patient safety and quality of care: for example, for earlier diagnosis of medical and genetic conditions [16, 19, 15]. Hence, it is important to carefully weigh the relative costs and benefits of face recognition systems. Moving forward, by contributing toward a stronger framework for uncertainty quantification in state-of-the-art face recognition systems, we aim to strengthen existing efforts to develop algorithmic tools for mitigating bias in face recognition [1]. For example, modelling uncertainty can be used in approaches for identifying under-represented training examples [1]. Improving the accuracy of uncertainty modelling could help to identify subgroups of high uncertainty, where the algorithm may be performing poorly, initiating investigation by human domain experts, or the addition of training data used for improving performance on those subgroups.

References

[1] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019. 8

[2] Gavin Brown, Jeremy L Wyatt, Peter Tino, and Yoshua Bengio. Managing diversity in regression ensembles. *Journal of machine learning research*, 6(9), 2005. 2

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 8

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 5

[5] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *arXiv preprint arXiv:1910.04851*, 2019. 6

[6] Padraig Cunningham and John Carney. Diversity versus quality in classification ensembles based on feature selection. In *European Conference on Machine Learning*, pages 109–116. Springer, 2000. 2

[7] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 2

[8] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *advances in neural information processing systems*, 33:12345–12355, 2020. 6

[9] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701, 2017. 6

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 6

[11] 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. 5

[12] Heinrich Jiang, Been Kim, Melody Y Guan, and Maya Gupta. To trust or not to trust a classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5546–5557, 2018. 6

[13] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003. 2

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[15] Edith R Lederman, Connie Austin, Ingrid Trevino, Mary G Reynolds, Holly Swanson, Bryan Cherry, Jennifer Ragsdale, John Dunn, Susan Meidl, Hui Zhao, et al. Orf virus infection in children: clinical characteristics, transmission, diagnostic methods, and future therapeutics. *The Pediatric infectious disease journal*, 26(8):740–744, 2007. 8

[16] Hartmut S Loos, Dagmar Wieczorek, Rolf P Würtz, Christoph von der Malsburg, and Bernhard Horsthemke. Computer-based recognition of dysmorphic faces. *European Journal of Human Genetics*, 11(8):555–560, 2003. 8

[17] Zhenyu Lu, Xindong Wu, Xingquan Zhu, and Josh Bongard. Ensemble pruning via individual contribution ordering. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 871–880, 2010. 2

[18] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[19] Nicole Martinez-Martin. What are important ethical implications of using facial recognition technology in health care? *AMA journal of ethics*, 21(2):E180, 2019. 8

[20] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020. 6

[21] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4886–4893, 2019. 6

[22] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6902–6911, 2019. 6

[23] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9395–9404, 2021. 6

[24] RP Soni. On an inequality for modified bessel functions. *Journal of Mathematics and Physics*, 44(1-4):406–407, 1965. 4

[25] Linh Tran, Bastiaan S Veeling, Kevin Roth, Jakub Swiatkowski, Joshua V Dillon, Jasper Snoek, Stephan Mandt, Tim Salimans, Sebastian Nowozin, and Rodolphe Jenatton. Hydra: Preserving ensemble diversity for model distillation. *arXiv preprint arXiv:2001.04694*, 2020. 6

[26] American Civil Liberties Union. An ethical framework for facial recognition. https://www.ntia.doc.gov/files/ntia/publications/aclu_an_ethical_framework_for_face_recognition.pdf. 8

[27] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020. 2

[28] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems*, 33:6514–6527, 2020. 2

[29] Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502. IEEE, 2022. 6

[30] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019. 2