Probabilistic Knowledge Distillation of Face Ensembles

Jianqing Xu* Shen Li*
Ailin Deng[2] Miao Xiong[2] Jiaying Wu[2] Jiaxiang Wu[1] Shouhong Ding[1] Bryan Hooi[2]
[1]Youtu Lab, Tencent. [2]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

Abstract

Mean ensemble (i.e. averaging predictions from multiple models) is a commonly-used technique in machine learning that improves the performance of each individual model. We formalize it as feature alignment for ensemble in open-set face recognition and generalize it into Bayesian Ensemble Averaging (BEA) through the lens of probabilistic modeling. This generalization brings up two practical benefits that existing methods could not provide: (1) the uncertainty of a face image can be evaluated and further decomposed into aleatoric uncertainty and epistemic uncertainty, the latter of which can be used as a measure for out-of-distribution detection of faceness; (2) a BEA statistic provably reflects the aleatoric uncertainty of a face image, acting as a measure for face image quality to improve recognition performance. To inherit the uncertainty estimation capability from BEA without the loss of inference efficiency, we propose BEA-KD, a student model to distill knowledge from BEA. BEA-KD mimics the overall behavior of ensemble members and consistently outperforms SOTA knowledge distillation methods on various challenging benchmarks.

1. Introduction

Knowledge Distillation (KD) is an active research area that has profound benefits for model compression, wherein competitive recognition performance can be achieved by smaller models (student models) via a distillation process from teacher models. As such, smaller models can be deployed into space-constrained environments such as mobile and embedded devices.

There has been abundant literature in KD for face recognition [22, 33, 34]. However, all the existing approaches fall into the "one-teacher-versus-one-student" paradigm. This learning paradigm has several limitations. Firstly, a single teacher can be biased, which further results in biased estimates of face feature embeddings given by a student after knowledge distillation from the biased teacher. Secondly, it only yields point estimates of face feature embeddings, unable to provide uncertainty measure for face recognition in a safety-sensitive scenario.

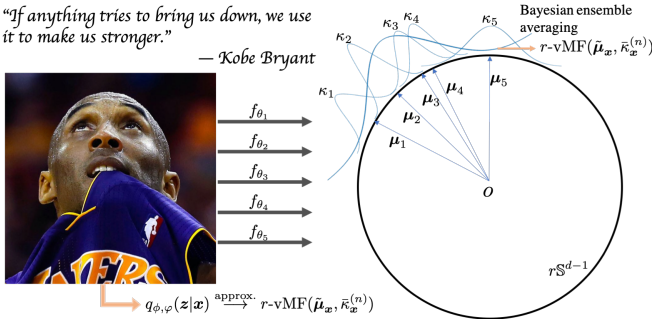

Figure 1. A conceptual illustration of BEA and BEA-KD. Given a face image $\boldsymbol{x}$, we have $n = 5$ probabilistic ensemble members $\{r\text{-vMF}(\boldsymbol{\mu}_i, \kappa_i)\}_{i=1}^{n}$ (marked by light blue). Bayesian ensemble averaging (marked by dark blue) returns a single r-vMF($\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}, \tilde{\kappa}_{\boldsymbol{x}}^{(n)}$) that accounts for the expected positions and confidence by all the ensemble members. To emulate the ensemble's probabilistic behavior, we employ a parametrized distribution $q_{\phi,\varphi}(\boldsymbol{z}|\boldsymbol{x})$ to approximate BEA.

Compared with single-teacher KD, KD from multiple teachers (a.k.a. ensemble KD) is beneficial and has been extensively explored in literature [9, 11, 29, 32, 36, 37]. However, these approaches are designed solely for *closed-set* classification tasks, distilling logits in a fixed simplex space via KL divergence (as the label set remains the same throughout training and test). In contrast, face recognition is inherently an *open-set* problem where classes cannot be known a priori. More specifically, face identities appearing during the inference stage scarcely overlap with those in the training phase. Consequently, without a fixed simplex space for logit distillation, existing approaches cannot be readily applied to face recognition. As will be shown in our empirical studies, existing closed-set KD approaches

*Equal contribution.

exhibit inferior performance in face recognition tasks with million-scale label sets.

How we treat teachers highly affects KD performance. Unlike prior art [20, 26] that takes average of predictions (termed 'mean ensemble' throughout this paper), we treat teachers as draws in a probabilistic manner. We find that this treatment leads to a generalization of mean ensemble, namely Bayesian Ensemble Averaging (BEA), which further brings up two practical benefits that existing methods could not provide: (1) the uncertainty of a face image can be evaluated and further decomposed into aleatoric uncertainty and epistemic uncertainty [6, 7], the latter of which can be used as a measure for out-of-distribution detection of faceness, i.e., distinguishing non-face images from face images; (2) a BEA statistic provably reflects the aleatoric uncertainty of a face image, which acts as a measure for face image quality, thereby improving recognition performance.

In addition, as ensemble methods are known to be computation costly during inference, we expect a more efficient model to inherit the uncertainty estimation capability from BEA. To this end, we propose BEA-KD, a student model that distills from BEA not only its feature embeddings but also the BEA statistic (and thus the aleatoric uncertainty). Consequently, BEA-KD consistently outperforms SOTA KD methods on various benchmarks. Our contribution can be summarized as follows:

(1) We recognize the benefit of multiple teachers (ensemble) for face recognition and present Bayesian Ensemble Averaging (BEA) as a generalization through the lens of probabilistic modelling.

(2) We verify that the proposed BEA can capture aleatoric uncertainty and epistemic uncertainty theoretically and empirically.

(3) We propose BEA-KD, a single smaller (hence efficient) model that inherits the power of uncertainty estimation from BEA yet reduces the high computational cost of BEA inference.

(4) We verify the BEA and BEA-KD's superior performance, respectively, as compared with mean ensemble and SOTA KD methods through extensive experiments.

2. Preliminaries and Background

Notations. Throughout the paper, we let $(\boldsymbol{x}, y) \sim \mathcal{D}$ be a training pair, where $\boldsymbol{x} \in \mathcal{X}$ denotes a face image in the input space $\mathcal{X}$ and $y \in \mathbb{L} := \{1, ..., C\}$ the corresponding label defined in the training label space (here, C is the number of identities seen in the training set). Note that the label space in the test phase is unavailable due to the nature of face recognition. The marginal distribution of $\boldsymbol{x}$ is denoted by $\mathcal{D}_{\mathcal{X}}$. We let $\boldsymbol{z}$ denote a latent embedding in the latent space $\mathcal{Z}$. Due to the hyperspherical treatment, we choose $\mathcal{Z}$ to be the r-radius spherical space, i.e. $\mathcal{Z} := r\mathbb{S}^{d-1}$, where d is the latent dimensionality. We let f_θ denote a deterministic face feature extractor that maps a face image to its spherical embedding, i.e. $f_\theta : \mathcal{X} \mapsto r\mathbb{S}^{d-1}$, where θ is the learnable parameter. Note that as a common practice, a deep face recognition classifier is built upon a feature extractor f_θ in the training phase, which maps the spherical embedding $\boldsymbol{z}$ into the label space $\mathbb{L}$. This mapping is typically a linear transform followed by a nonlinear activation. This linear transform can be parameterized by a row-nomalized matrix $\mathbf{W} \in \mathbb{R}^{C\times d}$. Then, the angular distance between the face feature and the ith identity's class center is $\vartheta_i = \langle \mathbf{W}_i, f_\theta(\boldsymbol{x})\rangle$. Training proceeds by minimizing the margin-based loss over $\mathbf{W}$ and θ (cf. Eq. (4) in [5]):

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[-\log\frac{e^{r(\cos(m_1\vartheta_y+m_2)-m_3)}}{e^{r(\cos(m_1\vartheta_y+m_2)-m_3)}+\sum_{j=1,j\neq y}^{C}e^{r\cos\vartheta_j}}\right] \tag{1}$$

r-Radius von Mises Fisher Density. The r-Radius von Mises Fisher distribution (r-vMF) was first proposed in SCF [21], which generalizes von Mises Fisher density (vMF) into a support over a sphere of arbitrary radius r. Formally, an r-radius vMF is a distribution over a d-dimensional r-radius sphere $r\mathbb{S}^{d-1}$ whose density is given by

$$p(\boldsymbol{z}|\boldsymbol{\mu},\kappa) = \frac{\mathcal{C}_d(\kappa)}{r^d}\exp\left(\frac{\kappa}{r}\boldsymbol{\mu}^T\boldsymbol{z}\right), \mathcal{C}_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa)} \tag{2}$$

where $\mathcal{I}_{d/2-1}(\kappa)$ is the modified Bessel function of the first kind of order $(d/2-1)$. The distributional parameters $\boldsymbol{\mu}$ and κ are called the mean direction and concentration parameter, respectively. We refer readers to [21] for more details.

SCF. Li *et al.* [21] proposed Sphere Confidence Face (SCF) which represents each face image as an r-vMF distribution in the latent spherical space, where the mean direction and the concentration parameter are the functions of the face image. Specifically, the SCF optimization objective is to approximate a desired spherical Dirac delta u using the parameterized r-vMF v in the sense of KL divergence between these two probability distributions, i.e.

$$\min_{v} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathrm{KL}(u(\boldsymbol{z}|y)||v(\boldsymbol{z}|\boldsymbol{x}))] \tag{3}$$

where $u(\boldsymbol{z}|y) = \delta(\boldsymbol{z} - \mathbf{W}^T\boldsymbol{y})$ and $v(\boldsymbol{z}|\boldsymbol{x}) = r\text{-vMF}(\boldsymbol{z}; \boldsymbol{\mu}(\boldsymbol{x}), \kappa(\boldsymbol{x}))$. Here, $\boldsymbol{y}$ is the one-hot encoding of y. Consequently, it can be shown that the cosine distance to its corresponding class center (though implicit) is a monotonically increasing function of the optimal concentration parameter (see Theorem 2 in [21] for details). Empirically,

the learned concentration parameter $\kappa_{\boldsymbol{x}}$ reflects the quality of the given face image for recognition. For example, $\kappa_{\boldsymbol{x}}$ is large if the face image $\boldsymbol{x}$ exhibits more frontal face attributes, better lighting condition or higher resolution; otherwise $\kappa_{\boldsymbol{x}}$ is small.

Uncertainty Decomposition. Uncertainty learning is of vital importance to safety-critical decision-making scenarios. There are two main forms of uncertainty: epistemic uncertainty and aleatoric uncertainty. Epistemic uncertainty refers to the uncertainty of model parameters and can serve to determine how far a sample deviates from the training distribution [12, 23]. This type of uncertainty is commonly used for out-of-distribution detection. Aleatoric uncertainty, on the other hand, accounts for the irreducible uncertainty caused by the inherent noise present in data [19]. Therefore, it can be used for risk-controlled face recognition wherein ambiguous face images are rejected. Uncertainty decomposition should be of interest in face recognition community, however, we find existing uncertainty-aware face recognition methods [1, 18, 21, 27, 30] unable to achieve this. Our work aims to bridge this technical gap.

3. Proposed Method

We first propose feature alignment for face ensemble in open-set recognition and then generalize it into Bayesian Ensemble Averaging (BEA) through the lens of probabilistic modelling. BEA enables us to well explain the statistics of our proposed framework, connecting them with uncertainty estimation and decomposition theoretically, which benefits various face recognition settings, e.g. risk-controlled face recognition and out-of-distribution (OOD) detection of faceness.

3.1. Feature Alignment for Face Ensemble

In closed-set classification problems, the training set and the test set share a common label space. Hence, when averaging the predictive outputs of multiple ensemble members (i.e. mean ensemble), no alignment is needed in advance [20]. In open-set recognition (e.g. face recognition), however, the feature location of one identity in one ensemble member does not necessarily coincide with that of the same identity in another member. Therefore, mean ensemble in this case requires alignment before averaging.

A simple recipe for alignment is to train these ensemble members using a shared $\mathbf{W}$. Since $\mathbf{W}$ can be seen as a collection of class centers (each row of $\mathbf{W}$ corresponding to a class center embedding) [1, 21], a fixed shared $\mathbf{W}$ can automatically enforce the alignment throughout the training of all ensemble members. Given a data distribution $\mathcal{D}$, this can be achieved by first training one feature extractor f_{θ_1} along with its linear transform parameterized by $\mathbf{W}$ and subsequently training the rest of the feature extractors $f_{\theta_2}, ..., f_{\theta_n}$ with the same $\mathbf{W}$ (fixed throughout the training, critically). Consequently, this ensures that different ensemble members yield aligned spherical embeddings of a given face image. As different members are initialized using different random seeds before training, the resultant feature embeddings of a given face image $\boldsymbol{x}$ are complementary to one another, yielding a more robust feature embedding through averaging (as verified in Table 1):

$$\boldsymbol{z}_{\text{avg}} = r \cdot \frac{f_{\theta_1}(\boldsymbol{x}) + \cdots + f_{\theta_n}(\boldsymbol{x})}{||f_{\theta_1}(\boldsymbol{x}) + \cdots + f_{\theta_n}(\boldsymbol{x})||_2} \tag{4}$$

The detailed procedure is summarized in Algorithm 1. See Appendix A for more discussions.

Algorithm 1: Feature Alignment for Face Ensemble

Input: The distribution of face images (with labels) $\mathcal{D}$; The number of ensemble members n; A discrete distribution $\mathbb{P}(s)$ for random seed generation.
Output: The learned ensemble $\{f_{\theta_1}, ..., f_{\theta_n}\}$.

Draw random seeds $s_1, ..., s_n \overset{\text{iid}}{\sim} \mathbb{P}(s)$;
Initialize θ using random seed s_1;
$\{\theta_1, \mathbf{W}\} \leftarrow$ Train f_θ and the linear transform parameters $\mathbf{W}$ over $\mathcal{D}$ to minimize the margin-based loss (1);
for $i = 2$ *to* n **do**
  Train the rest of the members with the fixed $\mathbf{W}$
  Initialize θ_i using random seed s_i;
  $\theta_i \leftarrow$ Train f_θ with the fixed $\mathbf{W}$ over $\mathcal{D}$ to minimize the margin-based loss (1);
end
return $\{f_{\theta_1}, ..., f_{\theta_n}\}$, $\mathbf{W}$

Algorithm 2: Bayesian Ensemble Averaging

Input: The distribution of face images (with labels) $\mathcal{D}$; The ensemble $\{f_{\theta_1}, ..., f_{\theta_n}\}$; The shared $\mathbf{W}$.
for $i = 1$ *to* n **do**
  Train an SCF module $\kappa(\cdot)$ that is built upon f_{θ_i} by minimizing Eq. (3) over $\mathcal{D}$ with $\boldsymbol{\mu}(\cdot)$ fixed as $f_{\theta_i}(\cdot)$;
end
for *any* $\boldsymbol{x} \sim \mathcal{D}$ **do**
  Evaluate $\boldsymbol{\mu}_i$ and κ_i for all i in $\{1, ..., n\}$;
  Compute and cache $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\tilde{\kappa}_{\boldsymbol{x}}^{(n)}$ using Eq. (8);
end

3.2. Bayesian Ensemble Averaging (BEA)

Instead of treating all ensemble members equally as in Eq. (4) (i.e. mean ensemble), we take a further step to consider a weighted model averaging from a Bayesian perspective. As shown in Figure 1, our proposed method, termed as Bayesian Ensemble Averaging (BEA), merges ensemble members more flexibly. Unlike mean ensemble as in Eq. (4), BEA treats an ensemble in a *probabilistic* manner. Suppose a random seed s obeys a prescribed multinomial distribution $\mathbb{P}(s)$, then via Algorithm 1 the stochas-

ticity in seeds (randomness in network initialization) induces the stochasticity in θ which therefore follows an implicit distribution $p(\theta|\mathcal{D}_{\mathcal{X}})$. That being said, given a face image distribution $\mathcal{D}_{\mathcal{X}}$, the learnable parameters of the finite ensemble $\{\theta_1, ..., \theta_n\}$ pretrained over $\mathcal{D}_{\mathcal{X}}$ are the draws from the posterior distribution $p(\theta|\mathcal{D}_{\mathcal{X}})$. This probabilistic perspective brings an interesting implication: the corresponding deterministic embeddings obtained from the ensemble, $\{\boldsymbol{z}^{(i)} : \boldsymbol{z}^{(i)} = f_{\theta_i}(\boldsymbol{x}), i = 1, ..., n\}$, can therefore be seen as draws from an implicit posterior distribution $p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_n}, \boldsymbol{x})$.

However, n is typically too small to estimate p accurately. We remediate this issue by modifying the deterministic embedding $\boldsymbol{z}^{(i)}$ into a stochastic one. Formally, given an ensemble member f_{θ_i} and a face image $\boldsymbol{x}$, the embedding $\boldsymbol{z}$ is assumed to follow an r-vMF distribution, i.e., $(\boldsymbol{z}|f_{\theta_i}, \boldsymbol{x}) \sim r\text{-vMF}(\boldsymbol{z}; \boldsymbol{\mu}_i, \kappa_i)$, where $\boldsymbol{\mu}_i := f_{\theta_i}(\boldsymbol{x})$ and $\kappa_i := \kappa_{f_{\theta_i}}(\boldsymbol{x})$. The mean direction $\boldsymbol{\mu}_i$ indicates feature embeddings given by each feature extractor f_{θ_i}, and the concentration parameter κ_i is dependent on the given face image $\boldsymbol{x}$ and f_{θ_i}, capturing the image quality for recognition in a member-specific manner. This parameter can be readily available via an SCF uncertainty module [21] that is built upon each deterministic feature extractor, or using the MagFace loss [24] in place of Eq (1). In experiments, we choose SCF since it is a post-processing approach to face uncertainty learning which does not modify the given ensemble members (i.e., $\theta_1, ..., \theta_n$ are unchanged).

Next, we show that the ensemble posterior of interest, $p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_n}, \boldsymbol{x})$, can be determined analytically via Bayesian treatment (see Appendix B for the detailed derivation for Eq. (5)(6)(7)):

$$p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_n}, \boldsymbol{x}) \propto \frac{p(\boldsymbol{z}|f_{\theta_n}(\boldsymbol{x}))}{p(\boldsymbol{z})} p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_{n-1}}, \boldsymbol{x}) \tag{5}$$

where $p(\boldsymbol{z})$ is a prior that is not dependent on either the ensemble or the face image $\boldsymbol{x}$. By applying the recursive Eq. (5) for n times, one can show that

$$p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_n}, \boldsymbol{x}) \propto \prod_{i=1}^{n} \Big(r\text{-vMF}(\boldsymbol{z}; \boldsymbol{\mu}_i, \kappa_i) \Big) \tag{6}$$

Interestingly, close scrutiny of Eq. (6) suggests that the ensemble posterior $p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_n}, \boldsymbol{x})$ is also an r-radius vMF that 'averages' all the individual member distributions. Formally, it can be shown that

$$p(\boldsymbol{z}|f_{\theta_1}, ..., f_{\theta_n}, \boldsymbol{x}) = r\text{-vMF}(\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}, \tilde{\kappa}_{\boldsymbol{x}}) \tag{7}$$

where

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} = \frac{\kappa_1\boldsymbol{\mu}_1 + ... + \kappa_n\boldsymbol{\mu}_n}{||\kappa_1\boldsymbol{\mu}_1 + ... + \kappa_n\boldsymbol{\mu}_n||_2}, \tilde{\kappa}_{\boldsymbol{x}} = ||\kappa_1\boldsymbol{\mu}_1 + ... + \kappa_n\boldsymbol{\mu}_n||_2 \tag{8}$$

Here, subscripts denote the dependencies on $\boldsymbol{x}$. Eq. (8) depends on the number of members, n. To remove the dependency, alternatively, we consider a modified averaged term instead: $\bar{\kappa}_{\boldsymbol{x}}^{(n)} = ||\kappa_1\boldsymbol{\mu}_1 + ... + \kappa_n\boldsymbol{\mu}_n||_2/n$. We term the entire process as Bayesian ensemble averaging (BEA). Figure 1 illustrates its conceptual procedure, where BEA returns a single r-vMF$(\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}, \bar{\kappa}_{\boldsymbol{x}}^{(n)})$ that accounts for the expected positions and confidence by all the ensemble members, thereby yielding a more robust probabilistic embedding for feature matching. A rigorous treatment is summarized in Algorithm 2.

Note that our probabilistic view of ensemble members leads to BEA which is a generalization of the mean ensemble as in Eq. (4): when $\kappa_1, ..., \kappa_n$ are all identical, Eq. (8) reduces to Eq. (4). Moreover, this novel probabilistic view opens up the possibility of decomposing uncertainty into aleatoric uncertainty and epistemic uncertainty, the latter of which can be used for detecting out-of-distribution of faceness. Such decomposition enables our model to identify uncertainty sources, which boosts the performance of risk-controlled face recognition and OOD faceness detection. Relevant details will be presented in the next section.

3.3. Theoretical Analysis

In this section, we answer the following questions theoretically: (1) what does the proposed BEA statistic $\bar{\kappa}_{\boldsymbol{x}}^{(n)}$ represent and what is the principle behind it? (2) how is uncertainty decomposed via BEA?

3.3.1 BEA Statistic

We first present the r-vMF entropy and its monotonous decreasing property that will later be leveraged to show the main theoretical results regarding the BEA statistic $\bar{\kappa}_{\boldsymbol{x}}^{(n)}$. This property can be also of independent interest in the spherical density family [4].

Lemma 3.1. *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. The proof can be found in Appendix C. □

Relations to Aleatoric Uncertainty. Mathematically, prior works (e.g. [7]) have shown that aleatoric uncertainty can be quantified as the expected entropy of the posterior distribution, i.e. $\mathbb{E}_{p(\Theta|\mathcal{D}_{\mathcal{X}})}[\mathcal{H}[p(\boldsymbol{z}|\boldsymbol{x}, \Theta)]]$, where Θ is the model parameter. In our probabilistic treatment, $\Theta := \{\phi^*, \varphi^*\}$. Intriguingly, we theoretically find that our proposed BEA statistic is indicative of aleatoric uncertainty. Specifically, in the limit of infinite ensemble members, $\bar{\kappa}_{\boldsymbol{x}}^{(\infty)}$ monotonically correlates with aleatoric uncer-

tainty $\mathbb{E}_{p(\phi^*,\varphi^*|\mathcal{D}_\mathcal{X})}\left[\mathcal{H}\left[p(\boldsymbol{z}|\boldsymbol{x},\phi^*,\varphi^*)\right]\right]$ under mild conditions. The formal treatment can be found in 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*

$$\bar{\kappa}_{\boldsymbol{x}}^{(\infty)} := \lim_{n\to\infty} \frac{\|\kappa_1\boldsymbol{\mu}_1 + \ldots + \kappa_n\boldsymbol{\mu}_n\|_2}{n} \propto \kappa_{\varphi_0^*}(\boldsymbol{x}) \tag{9}$$

Proof. The proof can be found in Appendix D. □

Remark 1 (Significance of Proposition 3.2). Theorem 2 proposed in [21] suggests that κ estimated by a single SCF model is indicative of the cosine distance to its implicit unknown class center and therefore can be interpreted as confidence. However, this heuristic interpretation fails to provide a rigorous connection with aleatoric uncertainty. In stark contrast to SCF [21], our proposed theory establishes a mathematical relation between the proposed BEA statistic and aleatoric uncertainty, making itself a rigorous measure for confidence. Empirically, to verify our proposed theory, we plot the relation between aleatoric uncertainty and the BEA statistic as shown in Figure 2(a), which suggests that $\bar{\kappa}_{\boldsymbol{x}}^{(n)}$ indeed reflects confidence (the inverse of aleatoric uncertainty). Note that the aleatoric uncertainty is estimated using Monte Carlo method (see Appendix E for details).

Remark 2. In practice, the assumption that the conditional $p(\varphi^*|\phi^*,\mathcal{D}_\mathcal{X})$ is a point mass may not hold exactly, but empirically we find that the coefficient of variation* (`std/mean`) of $\kappa_1, ..., \kappa_n$ are quite small (see Figure 2(b)). This suggests that $p(\varphi^*|\phi^*,\mathcal{D}_\mathcal{X})$ is closer to $\delta(\varphi^* - \varphi_0^*)$ for some φ_0^*. Despite the non-exactness in practice, as shown in Figure 2(a), the aleatoric uncertainty of $\boldsymbol{x}$ still monotonically decreases as $\bar{\kappa}_{\boldsymbol{x}}^{(n)}$ grows.

3.3.2 Uncertainty Decomposition for Face Recognition

Uncertainty decomposition has been extensively explored for closed-set classfication problems, where it is shown that total uncertainty can be decomposed into epistemic uncertainty and aleatoric uncertainty [6, 7]:

$$\begin{aligned} &\underbrace{\mathcal{H}\left[\mathbb{E}_{p(\Theta|\mathcal{D}_\mathcal{X})}\left[p\left(\boldsymbol{z} \mid \boldsymbol{x},\Theta\right)\right]\right]}_{\text{Total Uncertainty } \mathcal{T}(\boldsymbol{x})} \\ = &\underbrace{\mathcal{I}\left[\boldsymbol{z},\Theta \mid \boldsymbol{x},\mathcal{D}_\mathcal{X}\right]}_{\text{Epistemic Uncertainty } \mathcal{E}(\boldsymbol{x})} + \underbrace{\mathbb{E}_{p(\Theta|\mathcal{D}_\mathcal{X})}\left[\mathcal{H}\left[p\left(\boldsymbol{z} \mid \boldsymbol{x},\Theta\right)\right]\right]}_{\text{Aleatoric Uncertainty } \mathcal{A}(\boldsymbol{x})} \end{aligned} \tag{10}$$

*In probability theory and statistics, the coefficient of variation is a standardized measure of dispersion of a probability distribution [10].

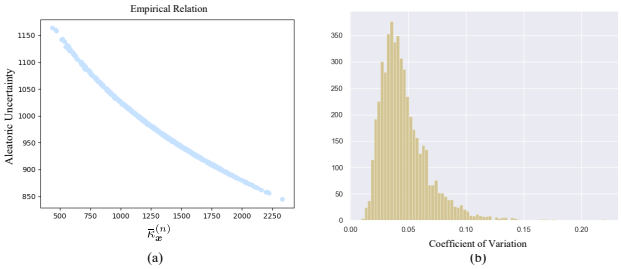

Figure 2. (a) Aleatoric uncertainty versus the BEA statistic. The aleatoric uncertainty is estimated using Monte Carlo method. The detailed estimation procedure is relegated to Appendix E. (b) The empirical distribution of the coefficient of variation of $\kappa_1, ..., \kappa_n$.

However, directly evaluating the epistemic uncertainty is intractable. Our probabilistic view of teachers can circumvent this difficulty by calculating total uncertainty $\mathcal{T}(\boldsymbol{x})$ and aleatoric uncertainty $\mathcal{A}(\boldsymbol{x})$ separately and performing subtraction to obtain epistemic uncertainty. Specifically, these two quantities can be approximated by

$$\mathcal{T}(\boldsymbol{x}) \approx \mathcal{H}\left[\frac{1}{n}\sum_{i=1}^{n} p(\boldsymbol{z}|\boldsymbol{x},\Theta^{(i)})\right] \tag{11}$$

$$\mathcal{A}(\boldsymbol{x}) \approx \frac{1}{n}\sum_{i=1}^{n} \mathcal{H}\left[p(\boldsymbol{z}|\boldsymbol{x},\Theta^{(i)})\right] \tag{12}$$

where n is the number of emsemble members and p is r-radius vMF distribution. Both of these quantities can be calculated thanks to our model assumption of r-radius vMF distribution, and therefore epistemic uncertainty can be obtained via subtraction. Epistemic uncertainty $\mathcal{E}(\boldsymbol{x})$ reflects the extent to which a given face image is distant from the data distribution the ensemble has seen so far. Therefore, in face recognition, we can use it as an indicator of faceness.

3.4. Knowledge Distillation from BEA

Note that calculating BEA is expensive, as it requires n inference of n ensemble members. This limits its practical use in space-constrained environments such as mobile and embedded devices. To accelerate the process, we expect to learn a parametrized distribution $q_{\phi,\varphi}(\boldsymbol{z}|\boldsymbol{x})$ that emulates the behavior of the ensemble posterior $p(\boldsymbol{z}|f_{\theta_1},...,f_{\theta_n},\boldsymbol{x})$. To this end, we propose to minimize the expected value of a general divergence $\mathbb{D}$ that measures some 'closeness' between p and q:

$$\min_{\phi,\varphi} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left[\mathbb{D}(p(\boldsymbol{z}|f_{\theta_1},...,f_{\theta_n},\boldsymbol{x})||q_{\phi,\varphi}(\boldsymbol{z}|\boldsymbol{x}))\right] \tag{13}$$

where $q_{\phi,\varphi}$ is a variational r-vMF density with the mean and the concentration parameter given by the parameterized functions $\boldsymbol{\mu}_\phi(\boldsymbol{x})$ and $\kappa_\varphi(\boldsymbol{x})$, respectively. These functions

can be instantiated using shallower neural networks as compared with f_{θ_i} (see Appendix F for the detailed comparison of model size). Mathematically, however, the learning objective (13) amounts to the minimization of the divergence $\mathbb{D}$ between two r-vMF distributions, which is generally intractable for optimization [8].

Nevertheless, noting that $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\bar{\kappa}_{\boldsymbol{x}}^{(n)}$ are both known for any $\boldsymbol{x}$ and any ensemble according to Eq. (8), we propose a tractable alternative to (13):

$$\min_{\phi,\varphi} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}} \left[\frac{1}{2} \left\| \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} - \boldsymbol{\mu}_{\phi}(\boldsymbol{x}) \right\|_2^2 + |\bar{\kappa}_{\boldsymbol{x}}^{(n)} - \kappa_{\varphi}(\boldsymbol{x})| \right] \tag{14}$$

Clearly, when this alternative loss (14) converges to zero for some minimizers ϕ^* and φ^*, we have $\boldsymbol{\mu}_{\phi^*}(\boldsymbol{x}) = \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\kappa_{\varphi^*}(\boldsymbol{x}) = \bar{\kappa}_{\boldsymbol{x}}^{(n)}$, leading to the global minimum of the original objective (13).

3.5. Inference

For face verification between a given pair of test face images $\boldsymbol{x}^a$ and $\boldsymbol{x}^b$, we follow SCF [21] and PFE [30] to employ the mutual likelihood score as the similarity measure:

$$s(\boldsymbol{x}^a, \boldsymbol{x}^b) = \log \frac{\mathcal{C}_d(\kappa^a) \cdot \mathcal{C}_d(\kappa^b)}{\mathcal{C}_d(\kappa^a \boldsymbol{\mu}^a + \kappa^b \boldsymbol{\mu}^b)} - d \log r \tag{15}$$

where $\boldsymbol{\mu}^a := \boldsymbol{\mu}_{\phi^*}(\boldsymbol{x}^a)$ and $\kappa^b := \kappa_{\varphi^*}(\boldsymbol{x}^b)$.

In the cases where each subject has more than one face images, it is desirable to obtain a compact representation aggregated from the multiple ones before face verification. For example, given two subjects A and B, each with a set of images $\{\boldsymbol{x}_{(l)}^{\cdot}\}$ (where "$\cdot$" can be either A or B, and l denotes the image index), our student model predicts the statistics $\boldsymbol{\mu}_{(l)}^{\cdot}$ and $\boldsymbol{\kappa}_{(l)}^{\cdot}$ for each, yielding the following pooled features by weighted averaging:

$$\boldsymbol{z}^A = \frac{\sum_l \kappa_{(l)}^A \boldsymbol{\mu}_{(l)}^A}{\sum_l \kappa_{(l)}^A}, \boldsymbol{z}^B = \frac{\sum_l \kappa_{(l)}^B \boldsymbol{\mu}_{(l)}^B}{\sum_l \kappa_{(l)}^B} \tag{16}$$

where $\boldsymbol{z}^A$ and $\boldsymbol{z}^B$ are the aggregated features for A and B, respectively. Then, face verification proceeds by calculating the cosine distance, i.e. $\cos\langle \boldsymbol{z}^A, \boldsymbol{z}^B \rangle$.

The advantages of our proposed BEA-KD are two-fold. Firstly, during inference, one does not have to run all the ensemble members to obtain features; nor is it necessary to perform BEA, which is computationally expensive when n is relatively large. Instead, a one-pass evaluation of $q_{\phi^*,\varphi^*}(\boldsymbol{z}|\boldsymbol{x})$ suffices (practically, $\boldsymbol{\mu}_{\phi^*}$ and κ_{φ^*}). Secondly, as compared with existing KD methods, the proposed $q_{\phi^*,\varphi^*}(\boldsymbol{z}|\boldsymbol{x})$ inherits the uncertainty estimation power of BEA, yielding not only the expected feature embedding $\boldsymbol{\mu}_{\phi^*}(\boldsymbol{x})$ for an individual data point $\boldsymbol{x}$ but also the confidence $\kappa_{\varphi^*}(\boldsymbol{x})$ for it. Consequently, as we will show later through experiments, the confidence provides uncertainty measure for the feature embedding, which benefits safety-sensitive settings such as face recognition.

Table 1. Ensemble performance comparison.

	CFPFP	CPLFW	IJB-B		IJB-C	
			1e-5	1e-4	1e-5	1e-4
Single	98.4	93.4	86.0	92.6	91.0	94.8
Mean Ensemble	98.8	93.7	88.3	93.9	92.7	95.7
BEA	**99.0**	**94.3**	**89.7**	**94.8**	**93.5**	**96.4**

4. Experiments

In this section, we answer the following research questions (RQs) via empirical studies:

RQ1: **(Performance)** How is the performance of the proposed BEA and BEA-KD compared to the ensemble and SOTA KD methods?

RQ2: **(Uncertainty Estimation)** Can our proposed framework capture aleatoric uncertainty and epistemic uncertainty, for risk-controlled face recoginition and faceness detection, respectively?

RQ3: **(Ablation Study)** How is the effectiveness of each component of our approach?

4.1. Implementation Details

We consider two experimental settings to demonstrate the effectiveness of our proposed framework: (1) ResNet12 [15] is employed as the student to distill knowledge from ResNet34; (2) MobileFaceNet [2] is employed as the student to distill knowledge from ResNet34. Throughout the experiments, we use the following notation for shorthand: `[S]-KD-[n][T]`, where `[S]` is the placeholder for the student network, `[T]` is the placeholder for the teacher and n is the number of the teachers being used. This notation specifies that model `[S]` is employed to distill knowledge from an ensemble of n teachers `[T]`'s. Note that `[T]` can be either a deterministic model or a stochastic model (modified from a deterministic one by training SCF or MagFace).

Before being fed into the models, normalized face crops are generated using five detected facial points, which yields (112×112) face images. Deterministic embeddings are pretrained using Residual Networks [15] as backbones and ArcFace [5] as the loss functions. The embedding network $\boldsymbol{\mu}_{\phi}(\cdot)$ maps face images into a 512-dimensional hyperspherical space ($d = 512$). Following [5], we set the hypersphere radius r to 64 and choose the angular margin 0.5. The module $\kappa_{\varphi}(\cdot)$ is instantiated using a fully convolutional network: `[CONV-ReLU-BN](×3)-AvgPool-[FC-ReLU](×2)-FC-exp`, where `CONV` denotes convolutional layers, `BN`

Table 2. State-of-The-Art Comparison and Ablation Study. Following the [S]-KD-[n][T] notation, we use $n = 5$ teachers for distillation.

Settings	Model	LFW	CFP-FP	CPLFW	IJB-B		IJB-C	
					1e-5	1e-4	1e-5	1e-4
[S]: ResNet12; [T]: ResNet34	TAKD [25]	99.58	93.49	89.45	76.49	87.19	82.22	89.77
	DGKD [31]	99.43	95.71	90.80	77.96	88.48	84.07	91.13
	MEAL [29]	99.45	92.24	88.57	72.59	84.33	78.34	87.12
	AE-KD [9]	99.58	95.34	90.26	77.91	88.20	83.38	90.98
	Hydra [32]	99.53	95.47	90.98	77.58	88.58	84.51	91.01
	CA-MKD [37]	99.56	95.17	90.50	76.83	88.11	83.31	90.57
	Eff-KD [11]	99.50	91.10	88.30	71.92	84.34	78.04	86.90
	BEA-KD (Ours)	**99.71**	**96.17**	**91.48**	**81.64**	**90.28**	**87.02**	**92.73**
	ResNet12-KD-1ResNet34	99.54	95.38	91.29	77.91	88.49	84.58	90.98
	ResNet12-KD-1ResNet34 + SCF	99.71	95.77	91.04	80.55	89.23	85.43	91.39
	ResNet12-KD-nResNet34	99.50	95.60	90.96	78.92	88.74	84.42	91.32
	ResNet12-KD-nResNet34 + SCF	99.69	95.95	91.20	81.01	89.36	84.93	91.71
	ResNet12-KD-1SCF(ResNet34)	99.60	95.85	91.28	80.93	89.32	84.73	91.55
[S]: MobileFaceNet; [T]: ResNet34	TAKD [25]	99.56	97.10	91.67	80.29	89.44	86.24	91.85
	DGKD [31]	99.65	96.71	92.31	80.59	88.83	86.65	92.10
	MEAL [29]	99.65	96.84	91.60	80.46	89.43	86.25	91.93
	AE-KD [9]	99.58	97.08	91.68	80.22	89.16	86.24	92.05
	Hydra [32]	99.50	96.86	91.97	80.64	89.60	86.39	91.96
	CA-MKD [37]	99.60	96.95	91.10	81.85	89.74	87.19	92.01
	Eff-KD [11]	99.48	96.95	91.55	80.71	89.57	86.96	92.13
	BEA-KD (Ours)	**99.70**	**97.23**	**92.93**	**83.45**	**91.27**	**89.23**	**93.40**
	MobileFaceNet-KD-1ResNet34	99.50	96.85	92.02	80.86	89.72	86.82	92.16
	MobileFaceNet-KD-1ResNet34 + SCF	99.59	96.99	92.36	82.21	90.19	87.62	92.41
	MobileFaceNet-KD-nResNet34	99.62	97.08	92.08	80.77	90.09	87.02	92.37
	MobileFaceNet-KD-nResNet34 + SCF	99.68	97.09	92.77	82.72	91.05	88.54	93.13
	MobileFaceNet-KD-1SCF(ResNet34)	99.65	96.93	92.05	82.94	90.60	88.25	93.10

refers to batch normalization layers, and $\exp$ the exponent nonlinearity to force the positiveness of concentration values. The total training epoch is set to 26. The batch size is set to 2048. Experiments are carried out using 8 Tesla V100 32GB GPUs.

4.2. Training Set and Benchmarks

All models are trained using WebFace260M [5]. WebFace260M training data is a new million-scale face dataset that contains 2 million identities ($C = 2 \times 10^6$) with 42 million face images. Models are evaluated on standard benchmarks, including LFW [16], CFP-FP [28], CPLFW [38], IJB benchmarks [35].

4.3. Comparison with Mean Ensemble (RQ1)

This section empirically verifies the effectiveness of BEA as a flexible generalization of mean ensemble. As shown in Table 1, BEA clearly outperforms mean ensemble, which suggests that BEA acts as a better ensemble model from which knowledge can be distilled.

4.4. Risk-controlled Face Recognition (RQ2)

Risk-controlled face recognition (RC-FR) is a benchmark where we expect a face recognition system to be able to reject face images with low confidence or high uncertainty for safety reasons. In experiments, we take all images from a given dataset and sort them by confidence in descending order or by uncertainty in ascending order [30]. Then, we filter out a proportion of images with low confidence or high uncertainty, and use the rest of the images for face verification. According to Proposition 3.2, our Bayesian treatment naturally allows for the quantification of aleatoric uncertainty along with each predictive feature embedding, which can be used as confidence for sorting in RC-FR. As shown in Figure 3, our proposed method surpasses all the other methods in RC-FR.

4.5. Comparison with State-of-The-Art (RQ1)

To demonstrate the effectiveness of our proposed approach BEA-KD, we compare its performance with the state-of-the-art KD methods: TAKD [25], DGKD [31], MEAL [29], AE-KD [9], Hydra [32], CA-MKD [37] and

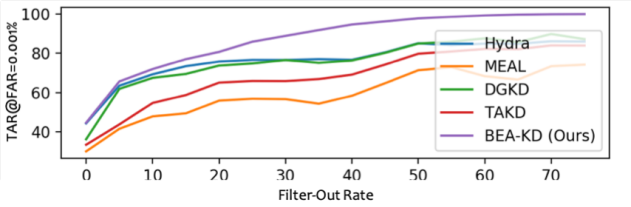

Figure 3. Risk-controlled face recognition on IJB benchmarks.

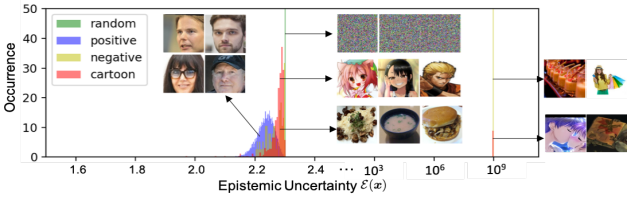

Figure 4. Out-of-distribution detection of faceness.

Eff-KD [11]. By clear margins, our proposed approach surpasses these models across all benchmarks under different settings (see Table 2). Noticeably, these KD methods are designed for the closed-set classification problem and aim to distill logits in Δ^{C-1} by minimizing $\mathcal{L}_{\text{KD}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\text{KL}\left(p^{\mathcal{T}}(y|\boldsymbol{x})||p^{\mathcal{S}}(y|\boldsymbol{x})\right)\right]$, where $p^{\mathcal{T}}(y|\boldsymbol{x})$ denotes the probability vector given by the teacher models and $p^{\mathcal{S}}(y|\boldsymbol{x})$ the probability vector by the student. The inferior performance of these KD methods can be attributed to the following reasons. First, through $\mathcal{L}_{\text{KD}}$, one can see the failure of these models when applied in large-scale face recognition where C is as large as millions. Since these models are designed for closed-set classification KD, such a large scale in face identities results in sparse probabilities in both $p^{\mathcal{S}}(y|\boldsymbol{x})$ and $p^{\mathcal{T}}(y|\boldsymbol{x})$, which makes it hard for amenable optimization. Second, the teacher classifier's predictive outputs may exhibit overconfidence as shown in [3, 13, 17]: the predictive softmax outputs tend to be large for wrong classes when the inputs are unseen, which makes the KD process unreliable. Third, the feature embeddings given by these models are deterministic, which cannot address the Feature Ambiguity Dilemma [30]. In contrast, our proposed approach operates in the feature space whose dimensionality is typically smaller than C (the dimensionality of the label space) in million-scale face recognition. Moreover, our proposed approach is designed for open-set recognition KD and can provide confidence measure that provably relates to aleartoric uncertainty, leading to the superiority of our method.

4.6. Ablation Study (RQ3)

This section demonstrates the effectiveness of each contributing components of our BEA-KD approach via ablation study. Specifically, we compare BEA-KD with its all ablated variants under each of the two settings described in Section 4.1. Under the first setting, we consider the following variants: Single ResNet12, ResNet12-KD-1ResNet34, ResNet12-KD-1ResNet34+SCF, ResNet12-KD-5ResNet34, ResNet12-KD-5ResNet34+SCF, ResNet12-KD-1SCF(ResNet34). The other setting is the same as the first except for the change of the student network (from ResNet12 to MobileFaceNet). Detailed descriptions of these models are summarized in Appendix G. The empirical results (Table 2) show that BEA-KD outperforms all its variants. We also investigate the compression rate of BEA-KD. Detailed analysis on this issue is relegated to Appendix F.

4.7. Faceness Detection (RQ2)

As a further qualitative exploration, this section investigates the effect of epistemic uncertainty on faceness detection. Our probabilistic perspective makes it possible to evaluate the epistemic uncertainty of a face image (using Eq. (10)), which can be interpreted as faceness of a given image. In this experimental setting (face versus non-face images), we consider three types of out-of-distribution images: `random`, `negative` and `cartoon`. Face images are in-distribution data denoted by `positive`; `random` refers to as images of white Gaussian noise; `negative` refers to image patches from LAION-400M* that are considered as false positive by SOTA face detectors [14]; `cartoon` refers to cartoon face images collected from iCartoonFace†. As shown in Figure 4, the epistemic uncertainty of face images is clearly below that of other OOD (i.e. non-face) images.

5. Conclusions

In this paper, we have presented Bayesian Ensemble Averaging (BEA) as a generalization of mean ensemble through the lens of probabilistic modelling. The proposed BEA can capture aleatoric uncertainty and epistemic uncertainty theoretically and empirically. We have also proposed BEA-KD, a single smaller (hence efficient) model that inherits the power of uncertainty estimation from BEA yet reduces the high computational cost of BEA inference. Consequently, BEA-KD outperforms SOTA KD methods through extensive experiments. We believe BEA-KD could serve as a strong baseline and inspire further advances in open-set KD research.

Acknowledgment. This work was supported in part by NUS ODPRT Grant R252-000-A81-133. The work was completed while Shen Li was with NUS-IDS under the sponsorship of Google PhD Fellowship 2021.

*https://laion.ai/projects/
†https://github.com/luxiangju-PersonAI/iCartoonFace

References

[1] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020. 3

[2] 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. 6

[3] 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. 8

[4] Nicola De Cao and Wilker Aziz. The power spherical distribution. *arXiv preprint arXiv:2006.04437*, 2020. 4

[5] 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. 2, 6, 7

[6] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Uncertainty decomposition in bayesian neural networks with latent variables. *arXiv preprint arXiv:1706.08495*, 2017. 2, 5

[7] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018. 2, 4, 5

[8] Tom Diethe. A note on the kullback-leibler divergence for the von mises-fisher distribution. *arXiv preprint arXiv:1502.07104*, 2015. 6

[9] 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. 1, 7

[10] B Everitt. The cambridge dictionary of statistics cambridge university press. *Cambridge, UK Google Scholar*, 1998. 5

[11] 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. 1, 7, 8

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 3

[13] 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. 8

[14] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021. 8

[15] 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. 6

[16] 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. 7

[17] 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. 8

[18] Roman Kail, Kirill Fedyanin, Nikita Muravev, Alexey Zaytsev, and Maxim Panov. Scaleface: Uncertainty-aware deep metric learning. *arXiv preprint arXiv:2209.01880*, 2022. 3

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 3

[20] 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, 3

[21] Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15629–15637, 2021. 2, 3, 4, 5, 6

[22] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. Face model compression by distilling knowledge from neurons. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 1

[23] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 3

[24] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 4

[25] 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. 7

[26] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 2

[27] Chingis Oinar, Binh M Le, and Simon S Woo. Kappaface: Adaptive additive angular margin loss for deep face recognition. *arXiv preprint arXiv:2201.07394*, 2022. 3

[28] 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. 7

[29] 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. 1, 7

[30] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6902–6911, 2019. 3, 6, 7, 8

[31] 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. 7

[32] 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. 1, 7

[33] Mengjiao Wang, Rujie Liu, Nada Hajime, Abe Narishige, Hidetsugu Uchida, and Tomoaki Matsunami. Improved knowledge distillation for training fast low resolution face recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[34] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 325–342. Springer, 2020. 1

[35] 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. 7

[36] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2017. 1

[37] 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. 1, 7

[38] 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. 7