Uncertainty-Aware Face Embedding With Contrastive Learning for Open-Set Evaluation

Kyeongjin Ahn, Seungeon Lee, Sungwon Han, Cheng Yaw Low, and Meeyoung Cha

***Abstract*— While advances in deep learning have enabled novel applications in various fields, face recognition in open-set scenarios remains a complex task, owing to the challenges posed by the extensive volume of low-quality face images. We introduce a new approach for recognizing faces in unconstrained open-set settings by leveraging uncertainty-aware embeddings through contrastive learning. Our model, called UCFace, effectively regulates the contribution of each face image based on the *face uncertainty* derived from image quality as an inverse proxy. Face embeddings are reinterpreted as a probabilistic distribution within the embedding space, where the degree of sharpness (i.e., distribution concentration) reflects the underlying uncertainty and probability density is used as a similarity metric to facilitate contrastive learning. Experiments on a wide range of face datasets, including those with high, mixed, and real-world low-resolution face images, demonstrate that UCFace enhances open-set face recognition performance by integrating the aspect of uncertainty.**

***Index Terms*— Face recognition, low-resolution, uncertainty, contrastive learning, von Mises–Fisher distribution.**

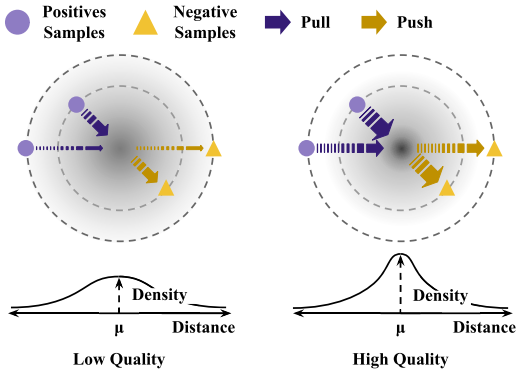


Fig. 1. Illustrating the uncertainty-aware contrastive learning objective with anchor probability distributions. When an anchor has *high uncertainty* (i.e., low-quality), its positive samples are weakly pulled towards the anchor and negative samples are pushed against the anchor. In contrast, when an anchor has *low uncertainty* (i.e., high-quality), its positive samples are strongly pulled toward the anchor and negative samples are strongly pushed against the anchor.

I. Introduction

FACE recognition (FR) has achieved remarkable breakthroughs with the help of deep learning, extensive labeled datasets, and scalable computing systems. Nevertheless, it remains a considerable challenge when compared to generic object recognition tasks. This is due to the complexity involved in distinguishing millions of unique identities, the scarcity of training samples for each identity, and the vast range of intra-class variations. Such variations include intractable poses and expressions, illumination conditions, occlusions, and other external factors such as low-resolution [1], [2], [3].

Recent face recognition work further delves into open-set recognition problems, which, unlike closed-set scenarios, include unknown identities within the test set to better simulate real-world deployment conditions. This requires the capacity to generalize knowledge learned from known identities to unknown identities. Various models have been proposed to refine the decision boundaries among unique identities in the latent space. For example, SphereFace [4] modifies the traditional softmax activation function with a margin term, leading to the emergence of angular margin-based softmax classifiers. These classifiers, which maximize the separation between samples with distinct identities using the angular margin principle, have shown promising results on open-set face recognition benchmark datasets, especially in high-resolution face images like LFW [5], CFP-FP [6], among others.

Softmax classifiers often suffer from discrepancies between training and inference settings. Specifically, these classifiers train the central representation (i.e., prototypes) of each identity in the training dataset, while they utilize the similarity between samples instead of prototypes during inference. This discrepancy between the training and inference settings can make classifiers vulnerable to misclassifying unknown testing identities as known training identities [7]. We refer to this scenario as the *open-set discrepancy challenge*.

Manuscript received 21 February 2024; revised 28 May 2024; accepted 24 June 2024. Date of publication 11 July 2024; date of current version 29 July 2024. This work was supported in part by the National Research Foundation under Grant RS-2022-00165347 and in part by the Institute for Basic Science in South Korea under Grant IBS-R029-C2. The associate editor coordinating the review of this article and approving it for publication was Dr. Naser Damer. *(Corresponding authors: Meeyoung Cha; Cheng Yaw Low.)*

Kyeongjin Ahn, Seungeon Lee, and Sungwon Han are with Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea.

Cheng Yaw Low is with the Institute for Basic Science (IBS), Daejeon 34126, South Korea (e-mail: chengyawlow@ibs.re.kr).

Meeyoung Cha was with IBS, Daejeon 34126, South Korea. She is now with KAIST, Daejeon 34141, South Korea, and also with the Max Planck Institute for Security and Privacy (MPI-SP), 44799 Bochum, Germany (e-mail: mia.cha@mpi-sp.org).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2024.3426973, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2024.3426973

Supervised contrastive learning (SCL) is a metric learning approach that addresses the challenge of open-set discrepancies. During training, it learns the degree of similarity among samples within the embedding space, which enhances the model's ability to generalize [8], [9]. SCL compares a target sample of interest, known as an *anchor*, against other samples to determine whether they share the same labels or differ. This comparison enforces the embeddings of an anchor and samples with the same label (i.e., positive samples) closer, while pushing the embeddings of dissimilar samples (i.e., negative samples) farther apart. This learning paradigm addresses the open-set discrepancy challenge by establishing a direct relationship between samples, focusing on learning from hard samples close to the class boundaries.

In this context, we investigate the potential benefit of integrating SCL into the open-set face recognition setting. We start by categorizing hard samples—which significantly affect SCL training—into two types, following the previous literature [10]. The first type is *unlearned samples*, which contain sufficient identity characteristics but remain unseen to the model during training. These samples enable the model to learn and improve over time. The second type is *low-quality samples*, which are typically noisy and lack essential identity information necessary for person identification. We observe that using SCL in face recognition can result in overfitting to noise, especially when the training set is dominated by low-quality samples with limited information for effective embedding learning [11]. This observation necessitates considering the uncertainty in low-quality samples. Several studies [12], [13] have addressed the uncertainty of noisy face images in face recognition by learning the standard deviation as an uncertainty index, but they do not provide explicit evidence of a correlation between the trained uncertainty index and image quality. Furthermore, because these methods are built upon softmax classifiers, they also inherit the open-set discrepancy challenge.

We propose UCFace, a new variant of metric learning that extends the existing SCL framework to address the open-set challenge while reducing the impact of noisy training signals from low-quality samples. Similar to [14] and [15], we leverage the feature norm as a reliable indicator of image quality and associate it with the concept of sample uncertainty [16]. This quality measure helps us infer the amount of information contained in each sample [17]. To handle the sample uncertainty, we define each anchor as a probability distribution in the representation space. By following the contrastive objective, we optimize the probability density of positive and negative samples relative to the anchor distribution. Fig. 1 visualizes our learning objective, which regulates the forces of pull or push among samples according to their probability densities within the anchor distribution. The proposed UCFace operates as an add-on to existing angular margin-based softmax losses for improved embedding learning that emphasizes more on meaningful samples. Our main contributions are as follows:

- We introduce a new metric learning approach, referred to as UCFace in this paper, to reduce the impact of noisy signals from low-quality samples for open-set face identification and verification tasks.
- UCFace encodes the concept of uncertainty into contrastive learning by interpreting an anchor as a probability distribution in the representation space and optimizing the probability density of positive and negative samples with respect to the anchor distribution.
- We demonstrate the effectiveness of UCFace in open-set deployment scenarios, achieving state-of-the-art performance not only on low-resolution face datasets but also on high- and mixed-resolution ones. Notably, in low-resolution cases, we observe remarkable performance improvement across most of our experiments.

We hope our findings are a step towards promising face recognition applications in low-quality images, e.g., forensics and security.

II. Related Work

A. Angular Margin-Based Softmax Classifiers

Conventional CNN-based face models have achieved remarkable performance thanks to angular margin-based softmax losses, an extension of the traditional softmax loss function. The goal of these loss functions is to maximize the separation of inter-class embeddings based on a penalty margin for enhanced generalizability. Remarkable examples include CosFace [18] and ArcFace [19]. CosFace applies an additive cosine margin, while ArcFace employs an additive angular margin, each enhancing geometric properties in their respective approaches. Building upon these losses, MagFace [15] and AdaFace [14] replace the pre-determined margin with an adaptive one. Specifically, the margin's magnitude is adjusted based on the recognizability of each sample in the embedding space, emphasizing hard samples by assigning them a larger margin. Despite their improvements in discriminative power under open-set conditions, there is still room for further enhancement in generalizing to large-scale unseen identities, especially in practical face recognition scenarios.

B. Quality Estimation in Face Recognition

Apart from feature norms, there has been significant progress in face image quality assessment (FIQA) using other learning-based models. CR-FIQA [20] estimates face image quality (FIQ) by learning a model to determine the relative classifiability of face embeddings, considering both their prototype and their nearest negative prototype. On the other hand, FaceQAN [21] measures FIQ by exploiting adversarial noise from models trained using gradient descent. Similar to MagFace and AdaFace, we interpret FIQ using feature norms, which require no additional training and have been demonstrated to closely approximate the actual image quality [22].

C. Uncertainty in Face Recognition

The integration of uncertainty to enhance face representation learning has recently attracted considerable attention. PFE [12] is a pioneering work that models the uncertainty of a face image as a Gaussian distribution in the latent space. However, PFE employs a fixed representation vector encoded by a pre-learned model as the mean of the distribution $\boldsymbol{\mu}$, while

interpreting uncertainty through the standard deviation of that distribution $\boldsymbol{\sigma}$. The drawback of PFE is that it does not allow $\boldsymbol{\mu}$ to be learned together with $\boldsymbol{\sigma}$ in the training process. On the contrary, DUL [13] learns both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ simultaneously within the same network. This ensures the representation vector $\boldsymbol{\mu}$ to be readjusted according to the learning of uncertainty index $\boldsymbol{\sigma}$. Uncertainty modeling of existing face recognition models depends on the model's optimization process, without demonstrating a clear correlation between uncertainty and image quality. Therefore, these approaches cannot guarantee that the predicted uncertainty aligns with image quality, thereby failing to distinguish between two types of hard samples (i.e., unlearned samples and low-quality samples), especially in the early stages of training. For example, they may overlook important training signals from unseen samples or overfit low-quality samples. In contrast, we adopt feature norms that have been verified to correlate with image quality, which mitigates these concerns and enhances the stability of training.

D. Contrastive Learning in Face Recognition

Metric learning is another methodology for learning discriminative face representations by pulling samples with the same label closer together while pushing those with different labels further apart. FaceNet [23] proposes triplet loss to optimize the distances between an anchor and its corresponding positive and negative samples in the Euclidean space. Although earlier works, such as basic pairwise comparison, often compare a limited number of sample pairs, more advanced techniques like triplet loss effectively scale to handle larger and more complex datasets, particularly in face recognition.

Supervised contrastive learning (SCL) leverages ground-truth labels [8], alongside the InfoNCE loss objective [24] in representation learning. By increasing the mutual information of positive pairs through the InfoNCE loss, the SCL-based model effectively identifies the common features for multiple positive pairs based on the supervised signal. SCL shows robust performance in open-set scenarios, improving its ability to generalize to unknown identities [8], [9]. However, applying SCL directly to resolve the face recognition problems may result in overfitting when trained on very noisy images.

III. Methodology

A. Overview

We propose UCFace, a method using metric learning to address the open-set discrepancy challenge and effectively handle noisy data. Our model introduces the *inverse proxy of uncertainty* to extend the SCL paradigm and utilizes the feature norms of face representations as a reliable indicator of face quality. We consider each anchor as a probability distribution rather than a deterministic embedding vector, with the feature norm determines the distribution's sharpness. To measure similarity between the anchor and samples, we employ the logarithmic probability density of each sample within the anchor distribution. We optimize the InfoNCE-based contrastive objective based on this measure, which differs from the conventional use of cosine similarity in contrastive learning. Fig. 2 and Algorithm 1 demonstrate the entire training procedure for our model.

Algorithm 1 Uncertainty-Aware Contrastive Learning

Input: An embedding encoder f, a projection head g, a mini batch $\mathcal{B}$, an anchor sample i, a set of positive samples $P(i) \in \mathcal{B}$, a pre-determined temperature τ, an angular margin-based softmax loss L_C.
Output: Trained embedding encoder f

```
// Compute representation vector
```
$\mathbf{z}_i \leftarrow g \cdot f(i)$
```
// Transform into vMF distribution
```
$V_{\mathbf{z}_i}(\mathbf{x}) \leftarrow \mathbf{vMF}(\mathbf{x}; \hat{\mathbf{z}}_i, \|\mathbf{z}_i\|)$
```
// Compute uncertainty-aware contrastive loss
```
$\mathcal{L}_U \leftarrow 0$
for $p \in P(i)$ **do**
  $\mathcal{L}_U \leftarrow \mathcal{L}_U + \text{sim}(\hat{\mathbf{z}}_p, \mathbf{z}_i)/\tau$
  $\mathcal{L}_U \leftarrow \mathcal{L}_U - \log \sum_{j \in \{\mathcal{B} \setminus i\}} \exp(\text{sim}(\hat{\mathbf{z}}_j, \mathbf{z}_i)/\tau)$
end
$\mathcal{L}_U \leftarrow \frac{-1}{|P(i)|} \mathcal{L}_U$
```
// Combine contrastive loss as an add-on to
   angular margin-based softmax loss
```
$\mathcal{L}_{total} \leftarrow \mathcal{L}_C + \mathcal{L}_U$
Update weights of f via back-propagation

B. Preliminaries

We revisit the mathematical definitions of angular margin-based softmax losses inspired by SphereFace [4], namely CosFace [18], ArcFace [19], MagFace [15], and AdaFace [14]. These losses differ in terms of margin function, denoted by $h(\cdot)$, which describes how the margin penalizes the typical softmax loss. For a given sample i, the generic loss objective is expressed as follows:

$$\mathcal{L}_i = -\log \frac{\exp(h(\boldsymbol{\theta}_i^{(y)}, m, s))}{\sum_{j=1}^{K} \exp(h(\boldsymbol{\theta}_i^{(j)}, m, s))} \tag{1}$$

where K stands for the total number of identities in the training set. $\boldsymbol{\theta}_i^{(j)}$ indicates the angle between $\mathbf{r}_i$, i.e., the embedding vector of sample i encoded by the backbone model, and the identity prototype indexed by j. CosFace learns to maximize the decision margin m in the cosine space [18]. On the contrary, ArcFace optimizes the geodesic distance (the angular margin) between the sample and its prototype directly [19]. Given the ground-truth label y, the margin term m, and the scaling factor s, the margin functions for CosFace and ArcFace are defined as follows:

$$h(\boldsymbol{\theta}_i^{(j)}, m, s)_{\text{CosFace}} = \begin{cases} s(\cos \boldsymbol{\theta}_i^{(j)} - m) & j = y \\ s \cos \boldsymbol{\theta}_i^{(j)} & j \neq y \end{cases} \tag{2}$$

$$h(\boldsymbol{\theta}_i^{(j)}, m, s)_{\text{ArcFace}} = \begin{cases} s \cos(\boldsymbol{\theta}_i^{(j)} + m) & j = y \\ s \cos \boldsymbol{\theta}_i^{(j)} & j \neq y \end{cases} \tag{3}$$

MagFace focuses on image quality, using feature norms for optimization [15]. It employs the margin function of ArcFace, where the margin m is formulated by the feature norm $||\mathbf{r}||$. Then, it integrates an extra regularization term g_{mag} regulated

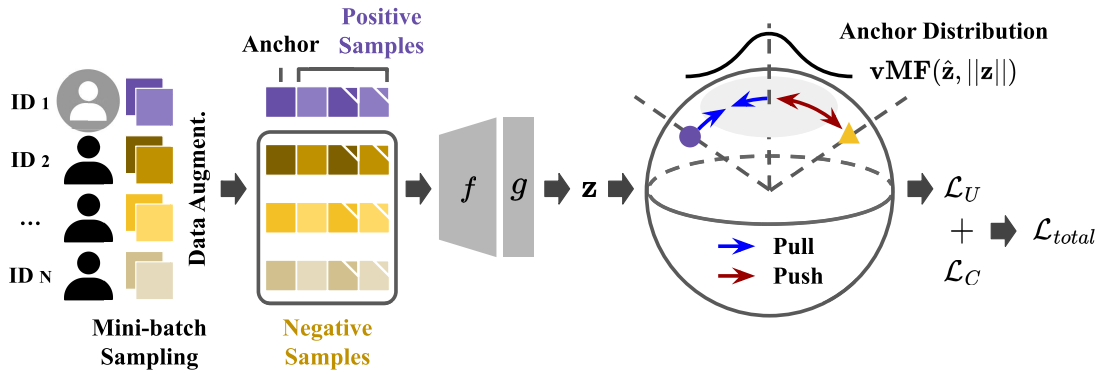


Fig. 2. Overall training pipeline for our model. Given a labeled face image and its augmented counterparts, we encode its representation vector $\mathbf{z}$ using an embedding encoder f and a projection head g. We map its L2-normalized representation vector, denoted by $\hat{\mathbf{z}}$, onto a von Mises-Fisher distribution, designating it as the anchor distribution. During the training stage, we optimize the probability density of both positive and negative samples about the anchor distribution based on the proposed uncertainty-aware contrastive loss $\mathcal{L}_U$. We obtain the final loss $\mathcal{L}_{total}$ by combining $\mathcal{L}_U$ with the softmax loss $\mathcal{L}_C$.

by a scaling factor λ_g with the generic loss objective $\mathcal{L}_i$.

$$m = \frac{u_m - l_m}{u_a - l_a}(||\mathbf{r}_i|| - l_a) + l_m \tag{4}$$

$$g_{mag} = \frac{1}{||\mathbf{r}_i||} + \frac{1}{u_a^2}||\mathbf{r}_i|| \tag{5}$$

$$\mathcal{L}_{mag} = \mathcal{L}_i + \lambda_g g_{mag} \tag{6}$$

where l_a, u_a, l_m, and u_m are hyperparameters that determine the relationship between margin and feature norm, and $\mathcal{L}_{mag}$ is the loss objective used in training of MagFace. While MagFace considers highly recognizable samples, it does not adequately emphasize hard training samples, which are helpful for learning discriminative features [14]. AdaFace overcomes this limitation by employing a sample-level adaptive process to handle varying sample difficulties [14]. It introduces two adaptive quality-based margin terms (i.e., angular margin g_{angle} and additive margin g_{add}). Its margin function is as follows:

$$\widehat{||\mathbf{r}_i||} = \left\lfloor \frac{||\mathbf{r}_i|| - \boldsymbol{\mu}_{\mathbf{r}}}{\boldsymbol{\sigma}_{\mathbf{r}}/h} \right\rceil_{-1}^{1} \tag{7}$$

$$g_{\text{angle}} = -m \cdot \widehat{||\mathbf{r}_i||}, \quad g_{\text{add}} = m \cdot \widehat{||\mathbf{r}_i||} + m \tag{8}$$

$$h(\boldsymbol{\theta}_i^{(j)}, m, s)_{\text{AdaFace}} = \begin{cases} s(\cos(\boldsymbol{\theta}_i^{(j)} + g_{\text{angle}}) - g_{\text{add}}) & j = y \\ s\cos\boldsymbol{\theta}_i^{(j)} & j \neq y \end{cases} \tag{9}$$

where $\widehat{||\mathbf{r}_i||}$ represents a feature norm normalized relative to the batch-wise moving mean $\boldsymbol{\mu}_{\mathbf{r}}$ and standard deviation $\boldsymbol{\sigma}_{\mathbf{r}}$, along with the concentration hyperparameter h. Accordingly, the loss function for the angular margin-based softmax classifiers [14], [15], [18], and [19] within the batch is given by:

$$\mathcal{L}_C = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i \tag{10}$$

where N refers to the number of samples in each training iteration.

We explore the characteristics exhibited by uncertainty-based models [12], [13] in the realm of face recognition. The uncertainty module from PFE [12] maximizes the Mutual Likelihood Score (MLS) between two genuine Gaussian distributions to optimize the objective function. A more detailed solution is provided below:

$$s(\mathbf{x}_i, \mathbf{x}_j) = -\frac{1}{2}\sum_{l=1}^{D}\left(\frac{(\boldsymbol{\mu}_i^{(l)} - \boldsymbol{\mu}_j^{(l)})^2}{\boldsymbol{\sigma}_i^{2(l)} - \boldsymbol{\sigma}_j^{2(l)}} + log(\boldsymbol{\sigma}_i^{2(l)} - \boldsymbol{\sigma}_j^{2(l)})\right) - H \tag{11}$$

$$\mathcal{L}_{PFE} = \frac{1}{|P|}\sum_{(i,j)\in P} -s(\mathbf{x}_i, \mathbf{x}_j) \tag{12}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ represent samples from the same group. $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\sigma}^{(l)}$ denote the feature embedding and the variance of the l^{th} dimension, respectively. H is a constant, P is the set of all positive pairs, and D represents the size of the embedding dimension.

DUL [13] defines the uncertainty-based embeddings by sampling from a Gaussian distribution to simultaneously learn the feature embedding ($\boldsymbol{\mu} = \mathbf{r}_i$) and variance ($\boldsymbol{\sigma}$).

$$\mathbf{r}_i^* \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \tag{13}$$

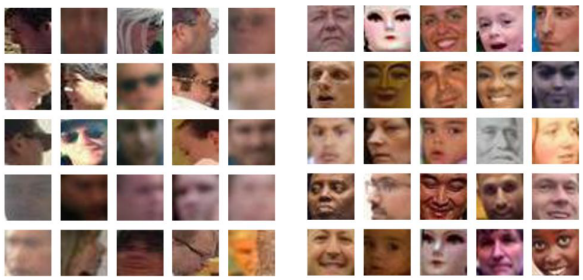
Then, it uses a reparameterization trick by adopting a noise ϵ sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

$$\mathbf{r}_i^* = \boldsymbol{\mu}_i + \epsilon\boldsymbol{\sigma}_i, \ \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{14}$$

Here, DUL utilizes the uncertainty-based stochastic embedding $\mathbf{r}_i^*$ to obtain $\theta_i^{(j)}$ in Eq. 1 instead of the deterministic embeddings $\mathbf{r}_i$. Then, it optimize the classifier loss (Eq. 10) with KL divergence regularization.

$$\mathcal{L}_{DUL} = \mathcal{L}_C + \lambda \cdot \mathbf{KL}\left[\mathcal{N}(\mathbf{r}_i^*|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)||\mathcal{N}(\epsilon|\mathbf{0}, \mathbf{I})\right] \tag{15}$$

Another advantage of modeling the uncertainty with distributions is the ability to fuse samples [12]. PFE introduces sample fusion by deriving the posterior probability distribution of multiple samples based on the individual probability distribution of each sample. This original work demonstrates the fusion process through a mathematical analysis of Gaussian distributions. Likewise, we showcase the fusion capability of UCFace by analyzing the von Mises-Fisher distribution in the Supplementary Material.



(a) Poorest-Quality (b) Best-Quality

Fig. 3. The relationship between feature norm and face image quality. Using a pre-trained ResNet-50 with VGGFace2, (a) the poorest-quality and (b) the best-quality face images in QMUL-TinyFace are sorted by the feature norm indexes of their representation vectors. We note that image quality improves from top-left to bottom-right.

C. Uncertainty Modeling With Image Quality

Although the traditional SCL learns a discriminative embedding space for downstream tasks, it is susceptible to overfitting caused by noisy patterns, especially for low-quality hard samples. Inspired by MagFace [15] and AdaFace [14], we utilize the feature norm, specifically L2-norm of the corresponding face representation vector as an image quality indicator. We use three low-resolution datasets for evaluation. Our results in Fig. 3 demonstrate a strong correlation between feature norm indexes and face recognizability as perceived by humans.

We interpret image quality of faces as an inverse proxy of uncertainty. For example, a high-quality face image, typically a frontal mugshot with meaningful identity attributes, reduces uncertainty and therefore contributes to better face recognizability. While most existing face recognition models encode each face image as a deterministic embedding vector, we depart from this convention by interpreting an anchor as a probabilistic distribution. Considering that SCL projects each sample representation to the unit hypersphere to avoid the divergence of training [25], we opt for a von Mises-Fisher (vMF) probability distribution as an anchor since it is a spherical analog of the normal distribution. Assuming $\mathbf{x}$ is an arbitrary unit vector, we represent the probability density function of the vMF distribution as follows:

$$\mathbf{vMF}(\mathbf{x}; \boldsymbol{\mu}, \kappa) = C(\kappa) \cdot \exp(\kappa \boldsymbol{\mu}^\top \mathbf{x}), \tag{16}$$

where $\boldsymbol{\mu}$ is the mean of the distribution with $\|\boldsymbol{\mu}\| = 1$, $\kappa \geq 0$ is a concentration parameter, and $C(\kappa)$ is a normalization constant.

The anchor sample i is encoded into the embedding vector $\mathbf{r}_i$ with dimension d_f using a backbone function $f(\cdot)$ such that $\mathbf{r}_i = f(i) \in \mathbb{R}^{d_f}$. This is followed by a projection head $g(\cdot)$ to map $\mathbf{r}_i$ into $\mathbf{z}_i = g(\mathbf{r}_i) \in \mathbb{R}^{d_g}$. To transform $\mathbf{z}_i$ into the vMF probability distribution, we set the mean direction vector $\boldsymbol{\mu}$ to the unit representation vector of the anchor sample through L2-normalization such that $\hat{\mathbf{z}}_i = \mathbf{z}_i / \|\mathbf{z}_i\|$. The concentration parameter κ, which controls the sharpness of distribution, is set to $\|\mathbf{z}_i\|$. When $\kappa = 0$, the distribution is uniform on the hypersphere, while it converges to a single point on the hypersphere for $\kappa = \infty$. Therefore, κ takes the key role of modeling the uncertainty of the anchor in the latent projection space [26]. For example, if a given sample has low recognizability, its feature norm $\|\mathbf{z}_i\|$ would be relatively small. Hence, κ is set to be small, resulting in a less concentrated probability distribution. Following this principle, the vMF distribution for the anchor sample i is given by:

$$V_{\mathbf{z}_i}(\mathbf{x}) = \mathbf{vMF}(\mathbf{x}; \hat{\mathbf{z}}_i, \|\mathbf{z}_i\|) \tag{17}$$

This results in a probability distribution on the vMF unit hypersphere that captures the uncertainty associated with the anchor sample. We stop the gradient of $\|\mathbf{z}_i\|$ to prevent the training stage from manipulating the feature norm.

D. Uncertainty-Aware Contrastive Learning

We reformulate SCL to mitigate the impact of noisy training signals from low-quality hard samples by utilizing probability distributions. In contrastive learning, for every positive pair in a given batch, each sample in the pair serves as both an anchor and a positive sample for the other. This symmetric property also holds for negative pairs, allowing us to focus on modeling only the anchor's uncertainty instead of considering all interactions among samples. Therefore, we treat the anchor as a probability distribution using deterministic embeddings for positive and negative samples. This approach maintains the capacity for modeling uncertainty but reduces computation complexity, making training more stable [27].

Representing an anchor as a probability distribution differs from the conventional SCL, which relies on the cosine similarity metric. This prompts the question of how to measure the similarity between a probability distribution (the anchor) and a feature vector (either a positive or negative sample). We leverage the key insight from probability theory and statistics [28], [29], [30] that the logarithm of a probability density function can be associated with a distance metric. Building upon this concept, we compute the similarity between an anchor distribution and a positive/negative sample as follows:

$$\text{sim}(\mathbf{z}_j, \mathbf{z}_i) = \log V_{\mathbf{z}_i}(\mathbf{z}_j) \tag{18}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the representation vectors of two samples for measuring similarity, while $\mathbf{z}_i$ serves as an anchor. The probability density of vMF monotonically decreases when the angular distance between the distribution's mean and the samples increases, similar to the behavior of cosine similarity. The logarithm of this value indicates the log-likelihood that the unit vector $\mathbf{z}_j$ is drawn from the probability distribution defined by the feature embedding $\boldsymbol{\mu}$ and the concentration κ, where a higher value indicates a greater likelihood, and vice versa. We assign the feature embedding $\boldsymbol{\mu}$ as $\hat{\mathbf{z}}_i$ and the concentration κ as $\|\mathbf{z}_i\|$, effectively adjusting the influence of every positive and negative pair based on the inverse of uncertainty. Therefore, the positive and negative pairs with lower uncertainty are estimated with higher confidence, whereas pairs with higher uncertainty are estimated with lower confidence. This emphasizes higher confidence pairs as more pronounced signals in contrastive learning between samples.

We revise the training objective of SCL to maximize the probability density for positive samples and minimize it for negative samples, given the anchor distribution. To ensure at least one positive sample for each anchor for a batch, we randomly select 2 samples for each of the N identities,

resulting in $2N$ samples. These samples are then augmented to yield a complete batch of $4N$ samples. Thus, every anchor is paired with 3 positive samples and $4(N-1)$ negative samples. For a given anchor i in a batch $\mathcal{B}$, we train our model using the following loss objective:

$$\mathcal{L}_U = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(\hat{\mathbf{z}}_p, \mathbf{z}_i)/\tau)}{\sum_{j \in \{\mathcal{B} \setminus i\}} \exp(\text{sim}(\hat{\mathbf{z}}_j, \mathbf{z}_i)/\tau)} \quad (19)$$

where $P(i)$ refers to a set of positive samples for the anchor i in batch $\mathcal{B}$, and $|P(i)|$ indicates its cardinality, where $|P(i)| = 3$ in our setting. The similarity between a positive sample and the anchor is represented by $\text{sim}(\hat{\mathbf{z}}_p, \mathbf{z}_i)$, while $\text{sim}(\hat{\mathbf{z}}_j, \mathbf{z}_i)$ denotes the similarity between the anchor and one of the samples within the batch $\mathcal{B}$. τ is a scalar temperature parameter.

The proposed uncertainty-aware contrastive learning paradigm, UCFace, serves as an add-on to an angular margin-based softmax classifier to enhance embedding learning. Let $\mathcal{L}_C$ be the angular margin-based softmax loss (Eq. 10) and $\mathcal{L}_U$ be the uncertainty-aware contrastive learning loss, the overall training loss $\mathcal{L}_{total}$ is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_C + \lambda \mathcal{L}_U \quad (20)$$

where λ is a weighting factor that regulates the contribution of $\mathcal{L}_U$ to $\mathcal{L}_{total}$.

IV. Experiments

A. Experimental Setup

1) Datasets: We evaluate UCFace and other relevant baseline models on three low-resolution face benchmark datasets for the open-set identification task: **QMUL-TinyFace** [1], **SCFace** [31], and **QMUL-SurvFace** [32]. In line with the conventional performance evaluation protocol, we also extend our experiments to high-resolution face datasets for the open-set verification task: **LFW** [5], **CALFW** [33], **CPLFW** [34], **CFP-FP** [6], and **AgeDB** [35], as well as mixed-resolution face datasets, **IJB-B** [36] and **IJB-C** [37]. Details of the benchmark datasets are in the Supplementary Material.

2) Baseline Models: We employ four baseline models learned with these classifiers in our analyses, including CosFace, ArcFace, MagFace, and AdaFace. The hyperparameter settings for each baseline model on low-resolution datasets are provided as follows: For CosFace, we assign the margin m as 0.3 and the scale s as 64.0, while for ArcFace, m is set to 0.6 and s is set to 64.0. In regards to MagFace, we set $(l_a, u_a, l_m, u_m, \lambda_g)$ to (10.0, 110.0, 0.45, 0.8, 20.0). Additionally, for AdaFace, we fix m to 0.4, s to 64.0, and h to 0.333. Meanwhile, the hyperparameter settings for each baseline model on high/mixed-resolution datasets are taken from the original papers.

We also re-implement relevant uncertainty-based models, including PFE and DUL, using MobileFaceNet as a backbone model with the ArcFace classifier for more extensive analyses and comparisons. The hyperparameter settings for these models are as follows: For PFE, we set γ to $1e-4$ and β to -7.0. For DUL, the hyperparameter λ, which balances between classifier loss and KL-divergence loss, is set to 0.01.

For the baseline models, except for CosFace and ArcFace on low-resolution datasets, we follow the default settings reported in the original papers. For CosFace and ArcFace, the hyperparameters are experimentally selected. We provide the hyperparameter analysis for these two baseline models in the Supplementary Material.

3) Implementation Details: Our experiments utilize MobileFaceNet [38] and ResNet-50 [39] as a embedding encoder $f(\cdot) \in \mathbb{R}^{d_f}$, where $d_f = 1,024$. We interleave a projection head $g(\cdot) \in \mathbb{R}^{d_g}$ with a single fully-connected layer, where $d_g = 128$. We provide additional experiments on MobileFaceNet with $d_f = 512$, ResNet-50 with $d_f = 1,024$, and ResNet-100 [39] with $d_f = 1,024$ in our Supplementary Material. Note that all main experiments are conducted with backbone models with $d_f = 1,024$. Initially, we pre-train these backbone models using high-resolution face images from VGGFace2 [40] or MS1MV3 [41]. We then affix randomly initialized angular margin-based softmax classifiers to these backbone models to train the baseline models for performance comparison. A batch size of 64 is used for training the softmax classifiers (both baseline models and $\mathcal{L}_C$) in our training pipeline. For uncertainty-aware contrastive learning (i.e., $\mathcal{L}_U$), a batch size of 128 is used in each iteration, with $N = 32$ identities randomly selected to compose the batch. We adopt $\tau = 0.8$ as the default temperature setting in our experiments. The learning rate is set to $1e-3$ for MobileFaceNet and $1e-4$ for ResNet-50 and ResNet-100 with a decay ratio of 0.1 at every 10th epoch. For training the uncertainty module in PFE, we use a learning rate of 1e-4. All experiments are trained for 40 epochs using the Adam optimizer, and the internal dropout rate is 0.6.

4) Performance Evaluation: During the inference stage, we encode an unknown face image t into its embedding vector $\mathbf{r}_t = f(t)$. As in self-supervised contrastive learning, the projection head $g(\cdot)$ is withdrawn once the training is complete. For identification tasks presented by QMUL-TinyFace, SCFace, and QMUL-SurvFace, $\mathbf{r}_t$ is compared to each template in the enrolled gallery set, and its identity is inferred based on the highest cosine similarity score. Meanwhile, for verification tasks demonstrated by high-resolution and mixed-resolution datasets, the decision of whether the face images t_i and t_j belong to the same or different identities is made using any pair of face embeddings $\mathbf{r}_{t_i}$ and $\mathbf{r}_{t_j}$ based on a predefined empirical threshold. We perform experiments over five runs to demonstrate stability, each with randomly initialized model weights, and report averaged results. Detailed results and standard errors are described in the Supplementary Material.

We report performance in terms of the rank-1 identification rate (%) for QMUL-TinyFace and SCFace. Adhering to the pre-determined evaluation protocol, the overall performance for QMUL-SurvFace is computed based on the True Positive Identification Rate (TPIR) and the False Positive Identification Rate (FPIR) at FRIR=$\{0.01, 0.05, 0.1, 0.2\}$, estimated for the top-20 cosine similarity scores. For clarity, the best results are highlighted in bold, and the second-best results are underlined.

5) Hardware: Our server is equipped with an Intel Xeon Platinum 8268 CPU @ 2.90GHz, 376GB DRAM and 5

TABLE I

PERFORMANCE SUMMARY IN TERMS OF THE AVERAGE RANK-1 IDENTIFICATION RATE (%) AND THE AVERAGE TPIR20(%)@FPIR OVER FIVE EXPERIMENTAL RUNS FOR IDENTIFICATION TASKS ON LOW-RESOLUTION DATASETS. BASELINE MODELS USING **MOBILEFACENET** AS A BACKBONE MODEL ARE COMPARED, BOTH WITH AND WITHOUT UCFACE. THE BEST RESULTS ARE IN BOLD, THE SECOND-BEST RESULTS ARE UNDERLINED, AND '-' DENOTES VALUES CLOSE TO ZERO. THIS CONVENTION ALSO APPLIES TO OTHER TABLES IN THIS SECTION

Methods	Train Data	Fine-tune	Low-Resolution									
			QMUL-TinyFace		SCFace				QMUL-SurvFace			
			w/o distractor	w/ distractor	d_1	d_2	d_3	Avg.	0.2	0.1	0.05	0.01
ArcFace	VGGFace2		55.58	51.10	41.75	91.75	97.20	77.08	5.01	1.12	-	-
CosFace	VGGFace2	✓	73.16	67.05	83.30	97.30	96.90	92.50	28.51	23.53	19.33	13.15
CosFace + UCFace	VGGFace2	✓	73.49	67.98	85.55	97.45	97.00	93.33	30.17	25.67	20.68	14.42
ArcFace	VGGFace2	✓	73.39	67.36	83.60	97.15	96.80	92.52	26.73	21.80	19.90	13.16
ArcFace + UCFace	VGGFace2	✓	**74.02**	**68.59**	87.00	97.75	**98.75**	**94.50**	31.78	26.80	**22.14**	15.22
MagFace	VGGFace2	✓	73.33	67.52	86.55	97.35	97.50	93.80	29.85	24.80	19.01	13.87
MagFace + UCFace	VGGFace2	✓	73.58	68.05	**87.50**	**98.15**	97.25	94.30	31.24	**26.92**	20.46	**15.55**
AdaFace	VGGFace2	✓	73.16	67.32	85.80	97.35	97.50	93.55	29.10	24.15	19.44	13.42
AdaFace + UCFace	VGGFace2	✓	73.49	68.02	87.30	98.05	97.45	94.27	**31.79**	26.83	21.11	14.45

TABLE II

PERFORMANCE SUMMARY IN TERMS OF THE AVERAGE RANK-1 IDENTIFICATION RATE (%) AND THE AVERAGE TPIR20(%)@FPIR OVER FIVE EXPERIMENTAL RUNS FOR IDENTIFICATION TASKS ON LOW-RESOLUTION DATASETS. BASELINE MODELS USING **RESNET-50** AS A BACKBONE MODEL ARE COMPARED, BOTH WITH AND WITHOUT UCFACE

Methods	Train Data	Fine-tune	Low-Resolution									
			QMUL-TinyFace		SCFace				QMUL-SurvFace			
			w/o distractor	w/ distractor	d_1	d_2	d_3	Avg.	0.2	0.1	0.05	0.01
ArcFace	VGGFace2		57.32	53.11	44.50	93.25	97.10	78.23	7.83	2.04	-	-
CosFace	VGGFace2	✓	75.06	70.98	93.85	98.50	98.60	96.98	30.66	25.46	21.59	15.30
CosFace + UCFace	VGGFace2	✓	75.78	71.45	**94.35**	98.45	**98.85**	**97.22**	32.28	27.81	22.91	16.45
ArcFace	VGGFace2	✓	75.18	71.18	93.50	98.30	98.30	96.70	28.68	23.97	20.77	14.40
ArcFace + UCFace	VGGFace2	✓	**75.83**	**71.56**	93.85	**98.80**	98.70	97.11	32.74	27.33	22.41	16.29
MagFace	VGGFace2	✓	74.87	70.60	93.10	98.40	98.50	96.67	31.65	26.91	21.23	15.90
MagFace + UCFace	VGGFace2	✓	75.20	71.13	93.95	98.50	98.60	97.02	33.10	**28.65**	22.94	**16.93**
AdaFace	VGGFace2	✓	74.81	70.57	93.25	98.35	98.60	96.73	31.22	26.32	21.64	15.39
AdaFace + UCFace	VGGFace2	✓	75.16	71.13	93.90	98.65	98.65	97.07	**33.40**	28.24	**23.01**	16.67

NVIDIA A100 GPUs. Only a single GPU was used for each experiment.

B. Feature Norm Analysis

We analyze feature norm distributions in Fig. 4 for QMUL-TinyFace, SCFace, and QMUL-SurvFace. The QMUL-SurvFace, the most challenging of the three, appears in the leftmost among the histogram distributions, indicating the smallest feature norms. Conversely, the SCFace with the high-quality image has relatively larger feature norms. The QMUL-TinyFace exhibits various feature norms, comprising samples of varying qualities. These observations underscore the significance that real datasets present unique feature norm distributions, indicating different difficulty levels in face recognition.

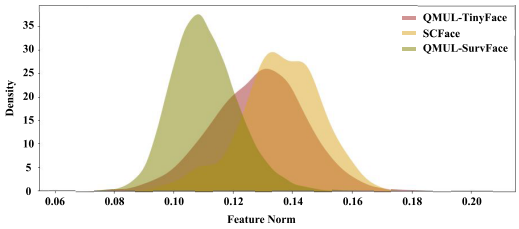


Fig. 4. Comparison of feature norm distributions for QMUL-TinyFace, SCFace, and QMUL-SurvFace. Considering feature norm as a metric for face image quality assessment, we observe that QMUL-SurvFace contains the most challenging samples, followed by QMUL-TinyFace and SCFace.

C. Comparison With Baseline Models

We compare our model to baseline models using the pre-trained MobileFaceNet and ResNet-50 as a backbone model. Table I shows that our model improves the generalizability of baseline models in QMUL-TinyFace, even in the presence of a large-scale distractor set comprising approximately 100K unknown low-resolution facial images. The models trained alongside UCFace prevail over baseline models in SCFace test sets. Our model shows the most substantial performance gain in the most challenging test set d_1 with all severely degraded face images. Furthermore, the improved performance on QMUL-SurvFace, recognized as the most challenging among low-resolution datasets, unequivocally demonstrates the effectiveness of our model. Table II exhibits a similar trend to Table I, with an overall improvement of approximately 2-3% on average. UCFace outperforms the comparison models across all results.

We evaluate its effectiveness not only on low-resolution datasets but also on datasets with varying resolutions. Table III presents the results from high-resolution and mixed-resolution datasets, demonstrating that our model performs on par with other baseline models across various resolutions, with a slight but noticeable improvement in most cases.

Table IV compares our model with other models that leverage the concept of uncertainty. Our model is effective on low-quality samples, particularly those classified as hard samples. This highlights the importance of estimating uncertainty-driven relationships among sample pairs, enriching the understanding of hard samples in embedding learning.

TABLE III

PERFORMANCE SUMMARY IN TERMS OF THE ACCURACY (%) WITH A SINGLE RUN FOR VERIFICATION TASKS ON HIGH-RESOLUTION AND MIXED-RESOLUTION DATASETS. BASELINE MODELS USING **RESNET-50** AS A BACKBONE MODEL ARE COMPARED, BOTH WITH AND WITHOUT UCFACE

Methods	Train Data	High-Resolution					Mixed-Resolution	
		LFW	CALFW	CPLFW	CFP-FP	AgeDB	IJB-B	IJB-C
ArcFace	MS1MV3	98.64	95.94	92.27	97.34	97.12	93.02	95.73
ArcFace + UCFace	MS1MV3	**98.67**	95.98	92.23	97.33	97.15	93.15	95.69
AdaFace	MS1MV3	98.62	95.97	92.30	97.37	97.19	93.92	95.87
AdaFace + UCFace	MS1MV3	98.61	95.97	92.29	97.36	97.23	**94.10**	95.94
ArcFace	VGGFace2	98.34	96.01	92.68	97.51	97.24	93.01	96.30
ArcFace + UCFace	VGGFace2	98.36	96.03	92.71	97.59	97.22	93.32	96.46
AdaFace	VGGFace2	98.36	96.06	**92.96**	97.62	**97.30**	93.21	96.52
AdaFace + UCFace	VGGFace2	98.37	**96.09**	92.94	**97.99**	97.29	93.43	**96.63**

TABLE IV

PERFORMANCE SUMMARY OVER FIVE EXPERIMENTAL RUNS FOR RELEVANT UNCERTAINTY-BASED BASELINE MODELS AND UCFACE WITH **MOBILEFACENET** AS A BACKBONE MODEL

Methods	Train Data	Fine-tune	QMUL-TinyFace		SCFace				QMUL-SurvFace			
			w/o distractor	w/ distractor	d_1	d_2	d_3	Avg.	0.2	0.1	0.05	0.01
ArcFace	VGGFace2		55.58	51.10	41.75	91.75	97.20	77.08	5.01	1.12	-	-
ArcFace	VGGFace2	✓	73.39	67.36	83.60	97.15	96.80	92.52	26.73	21.80	19.90	13.16
ArcFace + PFE	VGGFace2	✓	73.32	67.45	84.00	97.70	96.45	92.71	26.16	22.61	19.38	12.61
ArcFace + DUL	VGGFace2	✓	73.60	67.95	85.35	**98.00**	**98.10**	**93.81**	29.16	25.47	22.06	14.55
ArcFace + UCFace	VGGFace2	✓	**74.02**	**68.59**	**87.00**	97.75	**98.75**	**94.50**	**31.78**	**26.80**	**22.14**	**15.22**

TABLE V

PERFORMANCE SUMMARY OVER FIVE EXPERIMENTAL RUNS FOR ABLATION ANALYSIS WITH **MOBILEFACENET** AS A BACKBONE MODEL

Methods	Train Data	Fine-tune	QMUL-TinyFace		SCFace				QMUL-SurvFace			
			w/o distractor	w/ distractor	d_1	d_2	d_3	Avg.	0.2	0.1	0.05	0.01
Triplet	VGGFace2	✓	31.01	26.97	31.00	85.20	91.00	69.07	1.42	-	-	-
SCL	VGGFace2	✓	63.91	58.72	79.55	95.90	94.50	89.98	20.98	15.79	9.95	5.11
ArcFace + Triplet	VGGFace2	✓	69.21	64.77	81.70	96.45	97.00	92.38	27.33	20.98	13.48	9.66
ArcFace + SCL	VGGFace2	✓	72.59	67.02	86.05	97.40	96.95	93.47	30.17	25.02	18.90	13.86
ArcFace + UCFace	VGGFace2	✓	**74.02**	**68.59**	**87.00**	**97.75**	**98.75**	**94.50**	**31.78**	**26.80**	**22.14**	**15.22**

D. Component Analysis

Table V shows the ablation results and inspects the role of each model component. It reports performance after removing each component to reveal the effectiveness of each one. Applying contrastive learning alone to the target problem leads to substantial performance degradation in the d_1 test set, which confirms the vulnerability due to low-quality hard samples. Incorporating contrastive learning leads to improvement in low-quality performance, especially in the baseline model with SCL. Each component of our model plays a unique role in handling hard samples. This result validates our choice of the add-on approach as a design of the loss function.

We now report results on the sensitivity of hyperparameters. MobileFaceNet with SCFace and QMUL-SurvFace is used in the experiment. Table VI shows the effect of temperature parameter τ, where $\tau = 0.8$ resulted in the best performance. Therefore, we have adopted $\tau = 0.8$ as a default setting for all experiments. Table VII presents the results with varying batch sizes. 64 is the largest batch size that is available in our GPU, serving the best performance overall. Table VIII displays the results with varying parameter λ values. We observed an average difference of approximately 0.1 for each column, indicating the stability of the model regardless of the parameter λ. For all experiments, we consistently used $\lambda = 1.0$.

TABLE VI

PERFORMANCE ANALYSIS FOR THE HYPERPARAMETER TEMPERATURE τ WITH RESPECT TO $\tau = 0.1$ TO $\tau = 10.0$. NOTE THAT WE REPORT THE RANK-1 IDENTIFICATION RATE (%) FOR A SINGLE RUN ON **SCFACE** TEST SETS WITH **MOBILEFACENET** AS A BACKBONE MODEL

Method	SCFace				
	τ	d_1	d_2	d_3	Avg.
ArcFace + UCFace	0.1	83.50	97.50	98.00	93.00
	0.3	85.50	98.00	97.75	93.75
	0.5	86.25	98.25	97.25	93.92
	0.7	86.00	98.25	97.00	93.75
	0.8	86.75	98.25	97.00	94.00
	0.9	86.00	98.25	97.25	93.83
	1	87.00	97.75	98.75	94.50
	5	84.25	97.75	97.75	93.25
	10	83.25	97.75	98.25	93.08

E. Qualitative Analysis

We present 10 instances where the baseline model fails in Fig. 5, i.e., the model's predictions do not align with the ground-truth labels. These cases are sampled from test sets of SCFace, with the baseline model built on ResNet-50 backbone model with ArcFace classifier. We demonstrate that by integrating UCFace into the baseline model, these false predictions are rectified. Our findings show that the uncertainty-aware UCFace encodes invariant face representations, even when dealing with challenging samples.

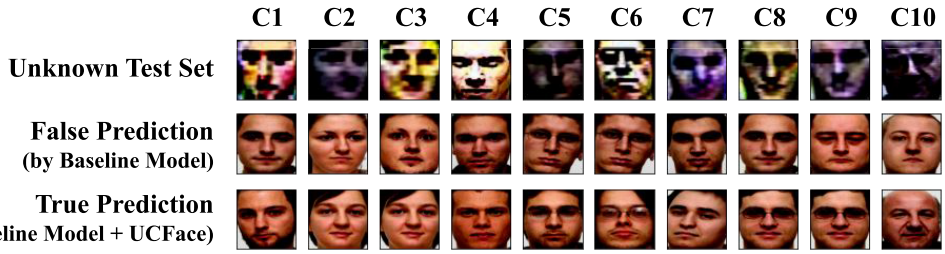


Fig. 5. Failure cases (C1 to C10) from the baseline model that are subsequently rectified using UCFace. The unknown test set in the top row shows example face images that the baseline model failed. The second row shows the gallery identity incorrectly predicted by the baseline model, while the third row shows the gallery identity correctly predicted after adding UCFace.

TABLE VII

PERFORMANCE ANALYSIS WITH VARYING BATCH SIZE FROM 8 TO 64. NOTE THAT WE REPORT THE RANK-1 IDENTIFICATION RATE (%) FOR A SINGLE RUN ON **SCFACE** TEST SETS USING **MOBILEFACENET** AS A BACKBONE MODEL

Method	SCFace				
	Batch	d_1	d_2	d_3	Avg.
ArcFace + UCFace	8	84.50	96.50	97.75	92.92
	16	88.00	96.00	96.50	93.50
	32	88.00	96.75	97.00	93.92
	64	87.00	97.75	98.75	94.50

TABLE VIII

PERFORMANCE ANALYSIS FOR THE HYPERPARAMETER λ WITH RESPECT TO $\lambda = 0.1$ TO $\tau = 10.0$. NOTE THAT WE REPORT THE PERFORMANCE IN TERMS OF THE TPIR20(%)@FPIR FOR A SINGLE RUN ON **QMUL-SURVFACE** TEST SETS WITH **MOBILEFACENET** AS A BACKBONE MODEL

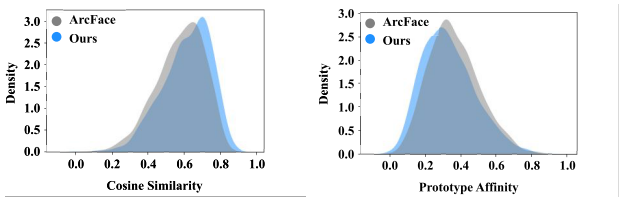
Method	λ	QMUL-SurvFace			
		0.2	0.1	0.05	0.01
ArcFace + UCFace	0.1	31.68	26.16	22.20	15.00
	0.5	31.75	26.19	22.28	15.09
	1.0	31.78	26.80	22.14	15.22
	5.0	31.88	26.21	22.31	15.20
	10.0	31.90	26.27	22.29	15.18

V. DISCUSSION

We discuss three challenges related to our model, especially when dealing with very noisy hard samples such as low-quality face images. We investigate open-set evaluation scenarios carried out on test sets of SCFace using the pre-trained MobileFaceNet backbone model with Arcface classifier.

A. Addressing the Open-Set Domain Discrepancy Challenge

Domain discrepancies in open-set scenarios lead to the misclassification of test samples as unknown identities and affect generalizability. We propose two metrics—intra-class compactness and prototype affinity—to assess the model's ability to generalize to unknown identities that are not present in the training set. Intra-class compactness refers to the degree to which test samples with the same identity cluster closely together, where a high value is desired. Prototype affinity represents the degree of closeness or similarity between test samples and identity prototypes discovered during training. A high affinity indicates that test samples closely resemble identity prototypes from the training set, whereas a low affinity represents a greater distinction between test samples and prototypes, which is desirable for recognizing unknown identities. We compute affinity using the cosine similarity between prototypes learned from the training set and the gallery templates in the test set.



(a) Intra-class Compactness (b) Prototype Affinity

Fig. 6. The histogram for (a) intra-class compactness and (b) prototype affinity between the baseline model and our model on **SCFace**. Intra-class compactness refers to how close test samples of the same identity cluster together, while prototype affinity represents the relative affinity of test samples to the identity prototype discovered in training. Our model shows higher intra-class compactness and lower prototype affinity than the baseline model, demonstrating its generalizability to unknown samples.

Fig. 6 depicts the histograms produced by the two similarity sets. Our model consistently outperforms the baseline model in terms of the intra-class compactness, as shown in Fig. 6a. This suggests that our model effectively captures the correct identity characteristics for unknown samples. Fig. 6b shows that our model has lower prototype affinity than the baseline model, implying that it is less likely to misclassify an unknown test sample as a known identity from the training set. These histograms validate the superior generalizability of our model in open-set scenarios.

B. Examining Vulnerability to Type 2 Hard Samples in SCL

To address this question, we investigate the model's confidence across different types of hard samples. The model's confidence is generally proportional to the predicted classification probability returned by the softmax classifiers. However, this principle is not applicable to open-set deployment scenarios. Inspired by the concept of margin of confidence [42], we define the model's confidence as the difference between the cosine similarity score of the matched gallery template and the highest cosine similarity score among all unmatched gallery templates. Intuitively, this refers to the first- and second-highest cosine similarity scores for the entire gallery set. The samples with the lowest confidence, as measured by pre-trained MobileFaceNet from d_3 and d_1 test sets, are selected as Type 1 and Type 2 hard samples, respectively. We then observe how the model's confidence changes during training for these selected hard samples.

Fig. 7 shows the confidence distributions across training epochs for SCL and our model on Type 1 and Type 2 hard samples. In the case of SCL, the confidence score for Type 2 samples consistently remains low, whereas the confidence for Type 1 samples gradually increases throughout the training process. This observation supports our conclusion that

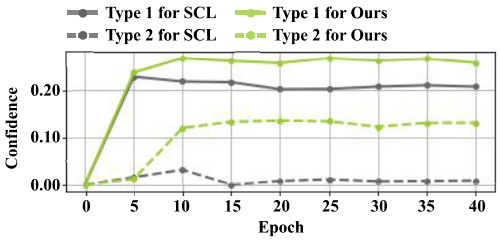


Fig. 7. Comparison of model's confidence between supervised contrastive learning (SCL) and our model for Type 1 and Type 2 hard samples. During training, SCL maintains low confidence for Type 2 hard samples, whereas our model's confidence increases for both Type 1 and Type 2 hard samples.

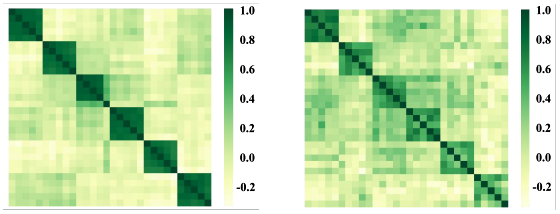


Fig. 8. The heatmap of cosine similarity for test sets of **SCFace**. (a) Samples from the d_3 test set are relatively higher in quality and therefore low uncertainty, resulting in a more condensed heatmap. (b) Samples from the d_1 test set are of relatively low quality and therefore high uncertainty, leading to a noisy heatmap.

SCL is susceptible to low-quality hard samples. Meanwhile, our model tackles this limitation by considering uncertainty, thereby enhancing the model's confidence for both Type 1 and Type 2 hard samples.

C. The Influence of Image Quality on Uncertainty in Representation Vectors

Samples with high uncertainty are likely to form more dispersed clusters even if they share the same identity. In contrast, samples with low uncertainty are closer to one another. For visualization, we randomly select six identities from each of d_1 and d_3 test sets of SCFace, with five samples per identity. Fig. 8 depicts the correlation heatmap of cosine similarity scores for all 30 selected samples from both test sets. The d_3 test set, which is presumed to have low uncertainty, shows a more condensed heatmap for samples with the same identity, as shown in Fig. 8a. On the contrary, the d_1 test set, which is presumed to have high uncertainty, displays a noisy and dispersed heatmap, as shown in Fig. 8b.

VI. Conclusion

In this paper, we introduced UCFace, an uncertainty-aware metric learning approach that encodes image quality into supervised contrastive learning for open-set face recognition. UCFace considers image quality as an inverse proxy of uncertainty and transforms each anchor's embedding into a probability distribution based on its estimated image quality. Subsequently, it refines the probability density of selected samples for the anchor distribution through the contrastive objective. Experimental evaluations using benchmark datasets of varying qualities demonstrate that UCFace improves baseline models in both open-set face identification and verification problems.

While our approach has shown improvements in practical open-set scenarios, there are numerous opportunities for future research. One potential direction is to replace the von Mises-Fisher distribution with alternatives such as the spherical t-distribution to enhance robustness against outliers, such as extremely noisy or incorrectly labeled facial images. Another direction is to represent both anchors and samples as probability distributions. This, however, involves assessing the similarity between distributions via sampling, which might impact the stability of the training process. Considering these factors, we hope that future research can broaden the scope of UCFace and enhance the capabilities of open-set face recognition methods.

References

[1] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Proc. 14th Asian Conf. Comput. Vis., (ACCV) Revis. Sel. Papers, Part III*, 2019, pp. 605–621.

[2] A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross, "Biometrics: A grand challenge," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, May 2004, pp. 935–942.

[3] Z. Zhu et al., "WebFace260M: A benchmark for million-scale deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2627–2644, Feb. 2023.

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.

[5] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, Oct. 2008, pp. 1–14.

[6] S. Sengupta, J. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[7] D. Miller, N. Sunderhauf, M. Milford, and F. Dayoub, "Class anchor clustering: A loss for distance-based open set recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3570–3578.

[8] P. Khosla et al., "Supervised contrastive learning," in *Proc. NIPS*, 2020, pp. 18661–18673.

[9] Y. Kodama, Y. Wang, R. Kawakami, and T. Naemura, "Open-set recognition with supervised contrastive learning," in *Proc. 17th Int. Conf. Mach. Vis. Appl. (MVA)*, Jul. 2021, pp. 1–5.

[10] R. Baldock, H. Maennel, and B. Neyshabur, "Deep learning through the lens of example difficulty," in *Proc. Adv. Neural. Inf. Process. Syst.*, vol. 34, 2021, pp. 10876–10889.

[11] S. Li, X. Xia, S. Ge, and T. Liu, "Selective-supervised contrastive learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 316–325.

[12] Y. Shi and A. Jain, "Probabilistic face embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6902–6911.

[13] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5710–5719.

[14] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18750–18759.

[15] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14225–14234.

[16] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Structural Saf.*, vol. 31, no. 2, pp. 105–112, Mar. 2009.

[17] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[18] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[20] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, "CR-FIQA: Face image quality assessment by learning sample relative classifiability," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5836–5845.

[21] Ž. Babnik, P. Peer, and V. Štruc, "FaceQAN: Face image quality assessment through adversarial noise exploration," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Sep. 2022, pp. 748–754.

[22] T. Schlett, C. Rathgeb, and O. Henniger, "Face image quality assessment: A literature survey," in *Proc. CSUR*, vol. 54, no. 10s, 2022, pp. 1–49.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[25] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. 37th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 119, Jul. 2020, pp. 9929–9939.

[26] J. Zhou, Y. Tang, B. Su, and Y. Wu, "Unsupervised embedding learning from uncertainty momentum modeling," 2021, *arXiv:2107.08892*.

[27] X. Hu, L. Chu, J. Pei, W. Liu, and J. Bian, "Model complexity of deep learning: A survey," *Knowl. Inf. Syst.*, vol. 63, pp. 2585–2619, Oct. 2021.

[28] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[29] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, Feb. 1975.

[30] T. M. Cover and J. A. Thomas, "Information theory and statistics," *Elem. Inform. Theory*, vol. 1, no. 1, pp. 279–335, 1991.

[31] M. Grgic, K. Delac, and S. Grgic, "SCface–surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, pp. 863–879, Feb. 2011.

[32] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," 2018, *arXiv:1804.09691*.

[33] T. Zheng, W. Deng, and J. Hu, "Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments," 2017, *arXiv:1708.08197*.

[34] T. Zheng and W. Deng, "Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Beijing Univ. Posts Telecommun., Beijing, China, Tech. Rep. 18-01, Feb. 2018. [Online]. Available: http://www.whdeng.cn/cplfw/?reload=true

[35] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 51–59.

[36] C. Whitelam et al., "IARPA Janus benchmark-B face dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 90–98.

[37] B. Maze et al., "IARPA Janus Benchmark-C: Face dataset and protocol," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 158–165.

[38] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2018, pp. 428–438.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[41] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi, "Lightweight face recognition challenge," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2638–2646.

[42] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. Int. Symp. Intell. data Anal. (IDA)*. Cham, Switzerland: Springer, 2001, pp. 309–318.

[43] P. C. Neto, A. F. Sequeira, J. S. Cardoso, and P. Terhörst, "PIC-score: Probabilistic interpretable comparison score for optimal matching confidence in single- and multi-biometric face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1021–1029.

[44] P. C. Neto et al., "OCFR 2022: Competition on occluded face recognition from synthetically generated structure-aware occlusions," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2022, pp. 1–9.

Kyeongjin Ahn received the Bachelor of Science (B.S.) degree in computer science engineering from the Hankuk University of Foreign Studies, South Korea, in 2019, and the master's degree in medical science from Yonsei University, Seoul, South Korea, in 2022. She is currently pursuing the Ph.D. degree with the School of Computing, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. Her research interests include deep representation learning and the application of deep learning methodologies to address social issues, leveraging large language models, and data analysis.

Seungeon Lee received the Bachelor of Science (B.S.) and master's degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the School of Computing. His research interests include explainable AI (XAI) and the application of deep learning methodologies, particularly large language models, in solving social problems.

Sungwon Han received the Bachelor of Science (B.S.) degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2019, where he is currently pursuing the integrated Ph.D. degree with the School of Computing. His research interests include deep representation learning, statistical analysis, and the development of responsible AI.

Cheng Yaw Low received the Ph.D. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2018. He is currently a Research Associate with the Data Science Group, Institute for Basic Science, Daejeon, South Korea. His research interests include computer vision and pattern recognition, with a specialization in biometric data, including face, periocular, fingerprint, and other modalities.

Meeyoung Cha is currently a Professor with Korea Advanced Institute of Science and Technology (KAIST) and the Scientific Director of the Max Planck Institute for Security and Privacy (MPI-SP), Germany. She studies the patterns of socially relevant information in various domains, including misinformation, poverty mapping, fraud detection, and long-tail content. She was a recipient of the Young Information Scientist Award in Korea and the two Test-of-Time Awards at AAAI ICWSM and ACM IMC.