Recognizability Embedding Enhancement for Very Low-Resolution Face Recognition and Quality Estimation

Jacky Chen Long Chai¹ Tiong-Sik Ng¹ Cheng-Yaw Low² Jaewoo Park¹ Andrew Beng Jin Teoh¹ ¹Yonsei University ²Institute for Basic Science
¹{jackyccl,ngtionsik,julypraise,bjteoh}@yonsei.ac.kr ²{chengyawlow}@ibs.re.kr

Figure 1. A deep face model pretrained on high-resolution face images introduces a cluster of unrecognizable instances (grey spike), dubbed unrecognizable identities (UIs) in [9]. (a) shows the bimodal distribution for a very low-resolution face dataset [6] based on distance against the UIs. Interestingly, a portion of hard-to-recognize faces (red peak) lie close to the UIs, indicating their low recognizability. (b) We propose to improve the recognizability of hard-to-recognize instances by pushing them away from the UIs center. Consequently, faces with higher recognizability indexes are further apart from UIs center in the embedding space. Our method not only induces more discriminative representations but also translates face quality into a measurable indicator that closely matches human cognition.

Abstract

Very low-resolution face recognition (VLRFR) poses unique challenges, such as tiny regions of interest and poor resolution due to extreme standoff distance or wide viewing angle of the acquisition devices. In this paper, we study principled approaches to elevate the recognizability of a face in the embedding space instead of the visual quality. We first formulate a robust learning-based face recognizability measure, namely recognizability index (RI), based on two criteria: (i) proximity of each face embedding against the unrecognizable faces cluster center and (ii) closeness of each face embedding against its positive and negative class prototypes. We then devise an index diversion loss to push the hard-to-recognize face embedding with low RI away from unrecognizable faces cluster to boost the RI, which reflects better recognizability. Additionally, a perceptibility attention mechanism is introduced to attend to the most recognizable face regions, which offers better explanatory and discriminative traits for embedding learning. Our proposed model is trained end-to-end and simultaneously serves recognizability-aware embedding learning and face quality estimation. To address VLRFR, our extensive evaluations on three challenging low-resolution datasets and face quality assessment demonstrate the superiority of the proposed model over the state-of-the-art methods.

1. Introduction

In real-world face recognition deployment scenarios, the pixel resolution of the detected face images is significantly deflated, due to extreme long-range distance and broad viewing angle of the acquisition devices, especially in surveillance applications. These tiny regions of interest are, in general, ranging from 16×16 to 32×32 pixels [60], thereby suffering from poor pixel resolution, in addition to unrestricted noises such as poor illumination conditions, non-frontal poses with awful angles, unconstrained facial expressions, blurriness, and occlusions [45]. It is noteworthy that these contaminated very low-resolution (VLR) face images undermine the overall performance of a face model trained with its high-resolution (HR) counterparts; therefore, there is a lack of generalizability to resolve the VLR face recognition (VLRFR) problem [6]. Apart from that, training of a VLRFR model often suffers from very limited representative face examples to extract meaningful
identity-specific patterns. These issues are further escalated due to ambiguous inter-class variations for the heavily distorted face instances with perceptually similar identities in particular \cite{40}. Whilst matching a probe to a gallery set of the same resolution (i.e. VLR to VLR) is still an open challenge, the resolution gap between galleries and probes triggers another problem in cross-resolution matching (typically HR galleries to VLR probes). Hence, the generalization performance of the prevalent deep learning models for VLRFR is still far from satisfactory.

As a whole, most existing works designated for VLRFR improve the face quality of the VLR instances based on an auxiliary set of HR face images \cite{28}. The generic operation modes are either in image domain (super-resolution, image synthesis) \cite{52,54,58}, embedding domain (resolution-invariant features, coupled mappings) \cite{33,44}, or at classifier level (transfer learning, knowledge distillation) \cite{13,14,20,39}. However, most of these state-of-the-art models require mated HR-VLR pairs of the same subject. This is unrealistic in practice as the HR-VLR pairs are often unavailable.

As far as face recognition is concerned, face recognizability (also known as face quality \cite{17,18}) can be deemed as a utility of how well a face image is for discrimination purposes. In other words, face quality is closely related to face recognition performance. Some works thus focus on predicting a face image’s suitability for face recognition, commonly known as Face Image Quality Assessment (FIQA) \cite{1,18}. FIQA focuses either on (i) creating propositions to label the training data with face image quality scores and solve a regression problem \cite{17,18,36}, or (ii) linking the face embedding properties to FIQ scores \cite{4,25,35,38,45}. The second approach shows better quality estimation, with the possible reason that the first approach is prone to mislabeling of ground truth quality \cite{35,45}. However, the second approach may not be optimal since the FIQ scores are estimated based on the embedding properties rather than through a learning process \cite{2}.

Recently, \cite{9} reported an intriguing observation that a deep learning-based face model induces an unrecognizable cluster in the embedding space. The cluster, known as unrecognizable identities (UIs), is formed by unrecognizable face examples, owing to diverse inferior quality factors, including VLR, motion blurred, poor illumination, occlusion, etc. Hence, these face examples with varying ground truths incline to lie close to the UIs, rather than their respective identity clusters. This observation inspires us to analyze the embedding distribution of the VLR face images against the UIs center. Interestingly, the extreme bimodal distribution in Fig. 1 discloses that a significant number of the VLR faces in TinyFace \cite{6}, i.e., a realistic VLR face dataset, are hard-to-recognize from the human perspective and therefore rendered next to the UIs cluster. We reckon that mining representative patterns from these hard-to-recognize faces is more meaningful for face recognition, in place of defining them as the elements of UIs. Apart from that, a more reliable objective quality metric is needed to better interpret each VLR face example in terms of its embedding recognizability for recognizability-aware embedding learning.

Instead of perceptual quality, this work aims to elevate the recognizability of every VLR face embedding. In a nutshell, we formulate a learning-based recognizability index (RI) with respect to the Cosine proximity of each embedding instance with (i) the UIs cluster, and (ii) the associated positive and negative prototypes. In the meantime, the index diversion (ID) loss is presented to detach the hard-to-recognize embeddings from the UIs cluster, alongside a perceptibility attention mechanism. We underline that embedding learning in the direction opposing the UIs contributes to a higher explanatory power whilst promoting inter-class separation, particularly for hard-to-recognize instances. For clarity, we summarize our contributions as follows:

- A novel approach is proposed to address the VLRFR, including VLR-VLR and HR-VLR matching, by leveraging the face recognizability notion in the embedding space to improve the hard-to-recognize instances.
- A robust learning-based face recognizability, dubbed RI, is put forward. RI relies on the face embeddings’ intrinsic proximity relationship against the UIs cluster, positive, and negative class prototypes.
- An index diversion (ID) loss is devised to enhance the RI for face embeddings. Further, we put forward a perceptibility attention mechanism to guide embedding learning from the most salient face regions.
- Our proposed model trained in an end-to-end manner not only renders a more discriminative embedding space for VLRFR but simultaneously serves recognizability-aware embedding learning and face recognizability estimation.

2. Related Work

**Very Low-Resolution Face Recognition.** Existing HR image dependency approaches for VLRFR can be categorized into image domain, embedding domain, and classifier. Under the image domain, the super-resolution (SR) model learns a mapping function to upscale VLR into HR images to improve faces’ visual quality \cite{41,54}. Several successors’ works \cite{19,52,53,58} relate recognition to visual quality. However, these works require mated HR-VLR pairs of the same subject to be available for embedding learning, which is unrealistic.

At both embedding domain and classifier levels, most knowledge distillation (KD) aimed to transfer the knowl-
edge of the HR domain to the VLR face model via a teacher-student network configuration [14]. To achieve cross-resolution distillation, [13] employed an additional network to bridge the teacher and student network, while [34] compared the face embeddings of teacher and student networks with a regression loss. On the other hand, distribution distillation loss [20], non-identity-specific mutual information [32] and implicit identity-extended augmentation [31] have also been explored to mitigate the performance gap between HR and VLR instances. [40] reconstructed an HR embedding to approximate the class variations of VLR instances while learning similar features for HR and VLR images. However, despite their high visual quality, some HR faces may not be recognized. In fact, [6, 29, 56] suggest that improving visual quality can undermine identity-specific traits important to the downstream recognition task. Our proposed method involves no auxiliary HR images. We aim to improve the hard-to-recognize instances based on the recognizability notion rather than visual quality.

Recently, Li et al. [30] proposed a rival margin on the hardest non-target logits to maximize the separation against the nearest negative classes. However, low-quality face images with perceptually similar identities are close to each other. Therefore, enforcing separation for a low-quality face image only from the nearest non-target class hardly learns meaningful characteristics to differentiate the two identities. On the contrary, we strive to enlarge the overall inter-class dissimilarity by enhancing image recognizability.

**Face Image Quality Assessment (FIQA)** can be mainly divided into two categories. The first category is to solve a regression problem to assess the training images with FIQ scores [1, 2, 17, 18, 36, 51]. The FIQ scores include human quality annotation [1], the intra-class Euclidean distance between an instance and an ICAO [49] compliance instance [17, 18], the similarity between random positive paired [51], the discriminability on each instance [2] and the Wasserstein distance between intra and inter-class distributions [36]. The second category directly utilizes the intrinsic properties of face embeddings to estimate face quality without explicit regression learning. For instance, [45] defined the robustness of stochastic embeddings as FIQ score. However, the computational cost is significant as the assessed instance is required to pass through the network multiple times at different dropout rates. [38] and [35] relied on uncertainty variation and embedding norm of the face embeddings as the FIQ scores, respectively.

Aside from face quality, several works also explore classifiability according to face quality. Instead of using the fixed Gaussian mean as the face embedding [38], [4] proposed to learn a Gaussian mean alongside an uncertainty to reduce the adverse effects of noisy samples. In [21], the easy instances are first learned before the hard ones based on the adaptive margin angular loss. [48] defined hard samples and increased the weight of their negative cosine similarities with a preset constant. [35] assigned margins based on the embeddings’ norm. As a successor, Kim et al. [25] refined the decision on adaptive margin, which only emphasizes hard instances when necessary.

While most FIQA methods struggle to learn adaptive margins for competent quality-aware embedding [2], we strive to improve embedding learning based on the proposed RI. We demonstrate that RI characterizes the image quality better, and our model can extract meaningful semantics important to face recognition in Sec. 4.

### 3. Methodology

#### 3.1. Recognizability Index (RI) Formulation

A face model trains a discriminative embedding space by enforcing intra-class compactness and inter-class discrepancy. Interestingly, [36] discloses the relationship between face quality and recognition performance by computing the Wasserstein distances between intra-class and inter-class similarity distributions. Hence, it is suggested that face recognizability can be quantified by intra-class and inter-class similarity measures.

In our disposition, the recognizable instances are pushed closer to their positive prototype and further apart from negative prototypes upon convergence. Meanwhile, the hard-to-recognize instances can hardly be pulled toward their positive prototypes, and are usually surrounded by negative
ones. Among all negative prototypes, the nearest negative prototype is chosen for inter-class proximity estimation.

For each instance \( i \), the intra-class and the inter-class proximity of its L2-normalized embedding \( \hat{v}_i \) with respect to its positive prototype \( u_{yi} \) and the nearest negative prototype \( u_j, j \neq y_i \) across a set of \( C \) identities are as follows:

\[
d^i_I = 1 - \cos(\theta_{yi}) \quad d^i_N = 1 - \max_{j \in \{1, \ldots, C\} \setminus \{yi\}} \cos(\theta_j)
\]

where \( \theta_{yi} \) is the positive angle between \( \hat{v}_i \) and \( u_{yi} \), and \( \theta_j \) is the negative angle between \( \hat{v}_i \) and \( u_j \). On the other hand, feeding the unrecognizable faces to the model induces a UIs cluster in the embedding space [9]. Given the UIs cluster center, i.e., the average across all normalized UIs embeddings, \( \pi_{ui} \), the proximity between \( \hat{v}_i \) and \( \pi_{ui} \) is defined as:

\[
d^i_P = 1 - \cos(\theta_{ui})
\]

where \( \theta_{ui} \) is the angle between \( \hat{v}_i \) and \( \pi_{ui} \). Since the proximity is in terms of Cosine distance, the instances closest to the UIs cluster center (computed with the smallest \( d^i_P \)) are referred to as hard-to-recognize instances, and vice versa. Since the proximity of each instance differs with respect to its positive and negative prototypes alongside the UIs cluster center, the association of Eq. (1) and Eq. (2) can serve to estimate the face quality. Given an imposed \( \epsilon = 1e^{-7} \) to avoid division by zero, we define a recognizability measure, dubbed Recognizability Index (RI), \( \xi_i \) as follows:

\[
\xi_i = \frac{d^i_P}{d^i_I} + \epsilon
\]

3.2. Perceptibility Regression Module

As our model serves recognizability-aware embedding learning and quality estimation simultaneously, we introduce a perceptibility regression module (yellow) as in Fig. 2. The input to the regression module is a flattened feature map learned from the backbone network, navigating through a dropout layer and a fully-connected regression layer to yield a learnable RI, denoted by \( \hat{\xi}_i \). To match \( \hat{\xi}_i \) to \( \xi_i \) in Eq. (3), we apply the smooth L1 loss [15] as follows:

\[
L_{L1} = \begin{cases} 
0.5(\xi_i - \hat{\xi}_i)^2 / \beta & \text{if } |\xi_i - \hat{\xi}_i| < \beta \\
|\xi_i - \hat{\xi}_i| - 0.5 \times \beta & \text{otherwise}
\end{cases}
\]

where \( \beta \) is a threshold that switches between L1 and L2-losses. In the early training stage, the L1 loss computes consistent gradients to approximate \( \xi \) towards \( \hat{\xi} \). When the regression error falls below the confidence interval, the L2 loss exhibits a smoother transition in the loss surface to facilitate convergence. In our experiments, we fix \( \beta = 0.75 \).

It is noteworthy that the RI from the regression module differs from ERS [9] in the following perspectives: (i) ERS does not consider the intra-class and inter-class proximities, and (ii) ERS requires the UIs cluster center to be known for recognizability score estimation, but our model simply withdraws the UIs cluster during the deployment. In other words, upon training completion, the regression module incorporates the proximity relation in Eq. (3) and allows the face quality prediction on any unknown samples without involving the UIs cluster.

3.3. Index Diversion Loss

Having RI obtained from the perceptibility regression module, the next following goal is to enhance the hard-to-recognize instances’ recognizability. We first model the RI distribution of the UIs cluster. According to the Central Limit Theorem [12], the Gaussian distribution is the most general distribution for fitting values derived from Gaussian or non-Gaussian random variables. Motivated by this, we can either estimate the mean and the variance of the Gaussian by forwarding the UIs through the perceptibility regression module or simply assume RI follows the standard Gaussian distribution with \( \mathcal{N}(0, 1) \). We opt for the latter as it has better interpretability and a more stable RI to resolve the hard-to-recognize instances.

Thus, we define the diversion of the estimated \( \hat{\xi}_i \) as:

\[
div = \frac{\hat{\xi}_i - \mu_{ui}}{\sigma_{ui}}
\]

Since the range of \( \hat{\xi}_i \) is arbitrary, it is rescaled with respect to \( \mu_{ui} \) and \( \sigma_{ui} \) of the UI distribution as:

\[
\mu_{ui} = \frac{1}{K} \sum_{k=1}^{K} s_k \quad \sigma_{ui} = \sqrt{\frac{\sum_{k=1}^{K} (s_k - \mu_{ui})^2}{K - 1}}
\]

where \( s_k \) denotes the RI of a random UI instance that is assumed to be i.i.d and follows \( \mathcal{N}(\mu, \sigma^2) \). Here, we set \( K = 5,000 \). In accordance with Eq. (5), the index diversion (ID) loss is formulated as:

\[
L_{ID} = \max(0, \tau - div)
\]

where \( \tau \) is the confidence interval hyperparameter. The ID loss enforces a deviation of at least \( \tau \) between \( \hat{\xi}_i \) and \( \mu_{ui} \). As a hard-to-recognize instance is associated with a small \( \hat{\xi}_i \), it induces a relatively large ID loss. We attempt to push the hard-to-recognize instances outside the designated \( \tau \) by minimizing the ID loss. In other words, the ID loss is equivalent to enforcing a statistically significant deviation of \( \mu_{ui} \) from the estimated \( \hat{\xi}_i \) of VLR instances in the upper tail. Note that although the ID loss enforces a confidence interval, it does not divert \( \hat{\xi}_i \) infinitely as the \( \hat{\xi}_i \) is still bounded to RI under the constraint of Eq. (4).

Since the ID loss is dedicated to differentiable learning of the recognizability measure, we argue that the statistical diversion of \( \hat{\xi}_i \) contributes to a more meaningful enhancement of recognizability in the embedding space, especially for the hard-to-recognize face images.
3.4. Perceptibility Attention Module

In [9], Deng et al. showed that diverting the embedding’s direction from the UIs cluster center results in face recognition improvement in the inference stage. Given the embedding vector of a VLR face instance \( v_i \) and the L2-normalized UIs cluster center \( \overline{v} \), we define the embedding projected in the direction away from the UIs cluster \( v'_i \) based on [9] as:

\[
v'_i = v_i - \langle v_i, \overline{v} \rangle \overline{v}
\]  

(8)

We conjecture that the embedding projection away from the UIs cluster center is beneficial to alleviate the model’s inadequacy to highlight the meaningful features when the face is obscure. Different from the conventional attention models for the classification tasks, we seek to explore the most salient feature representation with the greatest significance for recognizability by approximating \( v'_i \) through an attention module. Inspired by [50], we design a perceptibility attention module that attends to the spatial and channel dimension of the instances’ feature map sequentially. The attended feature maps are fed into a linear fully connected layer to learn an attended embedding \( v^{\text{attn}}_i \) at the same dimension as \( v'_i \). Lastly, we utilize the mean squared error loss and formalize the regression problem as:

\[
L_{\text{MSE}} = \frac{1}{B} \sum_{i=1}^{B} (v'_i - v^{\text{attn}}_i)^2
\]  

(9)

where \( B \) is the batch size. We argue that the projection of embeddings away from the UIs cluster can be deemed as RI-enhanced embeddings. This guides our model to attend to the parts of embedding that contain the richer interpretive contents important to recognition purposes. Therefore, introducing an attention module permits our model to attend to the most salient face regions, i.e., eyes, nose, etc., of a VLR face instance. In compliance with Eq. (4), (7) and (9), the overall loss function is expressed as follows:

\[
L_{\text{total}} = L_{\text{cls}} + \alpha L_{L1} + \beta L_{ID} + \gamma L_{\text{MSE}}
\]  

(10)

where we opt for ArcFace [8] as the classification loss, \( L_{\text{cls}} \), \( \alpha \), \( \beta \), and \( \gamma \) are the weighting factors for each loss term.

4. Experiments and Results

Datasets. To confront the real-world VLRFR problem, our experiments are conducted on three realistic LR face datasets, namely TinyFace [6], SurvFace [7], and SCFace [16] under an open-set evaluation protocol, given the identity labels are disjointed across training and testing sets. We assemble an unlabeled UI face dataset from two public person re-identification repositories by the MTCNN face detector [57], including LPW [42] and MARS [59]. Notably, most of the detected faces are unrecognizable, thereby facilitating the generation of an UIs cluster. The details of each dataset are provided in Sec. A of in supplemental material.

Experiment Settings. Our experiments utilize MobileFaceNet [5] and ResNet-50 [8] pretrained on the VGGFace2 [3] dataset as the representation encoder. For each dataset, we fine-tune these models using the respective training examples for performance evaluation. Our baseline model is the counterpart trained only with the ArcFace loss. We provide our experimental setup and other relevant settings in Sec. B of the supplemental material.

Evaluation Metrics. We summarize the overall performance in rank-1 identification rate (IR) (%) for TinyFace and SCFace. On the contrary, we report the positive identification rate (PPIR) (%) @ false positive identification rate (FPIR), and true positive rate (TPR) (%) @ false acceptance rate (FAR) (%) for SurvFace, due to the inclusion of non-mated face images in its testing set [7].

To evaluate the characterization of face image quality, we provide the Error versus Reject Curve (ERC) [2, 18, 36], where portions of low-quality face images are screened out with respect to quality indexes. This evaluation is assessed in terms of False Non-Match Rate (FNMR) [26] at a specific threshold for a fixed False Match Rate (FMR) [26].

4.1. Comparison with SoTA Methods

We compare the generalization performance on two open-set face identification tasks, VLR-VLR (TinyFace and SurvFace) and HR-VLR (SCFace), to the most recent SoTAs designated for VLRFR by a remark-

1In the “fine-tuned” column of these tables, we indicate \( \dagger \), \( \ast \), and \( \checkmark \) as fine-tuning on super-resolved, synthetic down-sampled VLR, and native VLR face images respectively.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Fine-Tuned</th>
<th>Rank-1 IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w/ dis</td>
<td>w/o dis</td>
</tr>
<tr>
<td>CSRI (ACCV19) [6]</td>
<td>✧ 44.80</td>
<td>-</td>
</tr>
<tr>
<td>TURL (CVPR20) [39]</td>
<td>63.89</td>
<td></td>
</tr>
<tr>
<td>RIFR (T-BIOM20) [24]</td>
<td>70.40</td>
<td></td>
</tr>
<tr>
<td>VividGAN (TIP21) [58]</td>
<td>✧ 47.16</td>
<td></td>
</tr>
<tr>
<td>MIND-Net (SPL21) [32]</td>
<td>✓ 66.82</td>
<td>73.52</td>
</tr>
<tr>
<td>AdaFace (CVPR22) [25]</td>
<td>✓ 68.21</td>
<td></td>
</tr>
<tr>
<td>IDEA-Net (TIFS22) [31]</td>
<td>✓ 68.13</td>
<td></td>
</tr>
<tr>
<td>AdaFace (Reproduce)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>✓ 73.06</td>
<td>77.22</td>
</tr>
</tbody>
</table>

Table 1. IR Comparison on TinyFace using ResNet-50

<table>
<thead>
<tr>
<th>Methods</th>
<th>Fine-Tuned</th>
<th>TPR(@FAR)</th>
<th>TPIR20(@FPIR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>CSRI (ACCV19) [6,22]</td>
<td>✧ 78.60</td>
<td>53.10</td>
<td>18.09</td>
</tr>
<tr>
<td>FAN (ACCV19) [52]</td>
<td>71.30</td>
<td>44.59</td>
<td>12.94</td>
</tr>
<tr>
<td>RAN (ECCV20) [11]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SST (ECCV20) [10]</td>
<td>✓ 87.00</td>
<td>68.21</td>
<td>21.41</td>
</tr>
<tr>
<td>DSN (APSIPA21) [27]</td>
<td>75.09</td>
<td>52.74</td>
<td>21.41</td>
</tr>
<tr>
<td>DDAT (PR21) [22]</td>
<td></td>
<td>90.40</td>
<td>75.50</td>
</tr>
<tr>
<td>IDEA-Net (TIFS22) [11]</td>
<td>✓ 87.41</td>
<td>77.48</td>
<td>58.63</td>
</tr>
<tr>
<td>AdaFace (Reproduce)</td>
<td>✓ 90.21</td>
<td>80.99</td>
<td>64.60</td>
</tr>
<tr>
<td>Ours</td>
<td>✓ 95.38</td>
<td>98.46</td>
<td>99.84</td>
</tr>
</tbody>
</table>

Table 2. IR Comparison on SCFace using ResNet-50

**SCFace.** For the HR-VLR evaluation, our model remains superior to other SoTAs over the three probe sets. This is attributed to: (i) the enhanced recognizability bridges the resolution gap between the HR and the VLR face features; and (ii) the attention module singles out the most salient regions from VLR and HR faces, resulting in better cross-resolution matching scores, especially for the face images captured from 4.2m (the largest standoff distances compared to 1.0m and 2.6m).

**SurvFace.** Being the most challenging VLR face dataset for the VLR-VLR deployment scenario, SurvFace evaluates both open-set identification and verification tasks in the presence of 141,736 unmated distractors as a part of the probe set. We disclose that our model significantly outperforms other SoTAs under the most rigorous settings, i.e., TPR@FAR=0.001 and TPIR20@FPIR=0.01. While [11,22,52] are trained based on synthetic LR images down-sampled from HR-paired counterparts for visual quality improvement, we underline that none of these SoTAs resolves the VLRFR by means of refining feature recognizability.

### 4.2. Ablation Analysis

**Effect of Each Loss Component.** We present in Table 4 an ablation analysis to explore the effect of each loss term using MobileFaceNet on TinyFace without gallery distractors. Compare to our baseline model (trained using only ArcFace), the inclusion of ID loss in Baseline I discloses that learning to enhance the recognizability of the hard-to-recognize instances offers a performance improvement close to 1.0%. Baseline II, on the other hand, reveals that the perceptibility attention module allows the most salient characteristics to be attended, resulting in at least 1.2% of performance gain. Meanwhile, Baseline III demonstrates that our model benefits from learning an RI based on estimated RI and softmax prototypes simultaneously with the RI prompts our model to encode the embedding recognizability at each training step. As a result, the learned RI can be viewed as an index of the model’s confidence corresponding to the classifiability of any face image, including the unknown instances. Overall, our model trained with all loss terms outperforms the baselines and two most relevant SoTAs, i.e., MagFace [35], and AdaFace [25]. An important reason is that these SoTAs are reliant on the embedding’s norm that does not always convey face recognizability, particularly for the VLR face images. This is substantiated by our empirical proofs in the following section.

**Intuition of Hyperparameter Selection.** In Fig. 3, we conduct the study of each weighting factor on each hyper-
yielding a well-separable (therefore a more discriminative) feature space to resolve the downstream VLRFR task.

In Fig. 4, the effect of ID loss reveals a greater negatively-skewed distribution of RI than other SoTAs (extended version in Fig. 9), indicating an improvement in recognizability, notably the hard-to-recognize instances. We discern that the proposed RI is a robust indicator in characterizing the recognizability notion - the highest and the lowest RI reflect the best-quality and the poor-quality face images, respectively. We also provide in Fig. 6 the face images with the lowest quality scores estimated by other competing scores. Through the comparison, RI is deemed proportional to the human cognitive level in terms of recognizability for the VLR face images under extreme poses, illuminations, occlusions, and others in the wild conditions.

**Error versus Rejection Curve (ERC).** To further substantiate the RI’s robustness, we perform a verification task on TinyFace [6] by randomly sampling 20,888 positive and 50,000 negative pairs from its gallery and probe sets to show

<table>
<thead>
<tr>
<th>Method</th>
<th>$L_{L1}$</th>
<th>$L_{ID}$</th>
<th>$L_{MSE}$</th>
<th>$L_{L1}$</th>
<th>Rank-1 IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Entropy [43]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>68.884</td>
</tr>
<tr>
<td>NormFace [46]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>68.026</td>
</tr>
<tr>
<td>CosFace [47]</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>70.306</td>
</tr>
<tr>
<td>MV-Softmax [48]</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>70.547</td>
</tr>
<tr>
<td>CurricularFace [21]</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>70.655</td>
</tr>
<tr>
<td>MagFace [35]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70.467</td>
</tr>
<tr>
<td>AdaFace [25]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70.359</td>
</tr>
<tr>
<td>Baseline (ArcFace) [5]</td>
<td>I</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>71.298</td>
</tr>
<tr>
<td>I</td>
<td>II</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>71.540</td>
</tr>
<tr>
<td>II</td>
<td>III</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>71.674</td>
</tr>
<tr>
<td>III</td>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>71.915</td>
</tr>
</tbody>
</table>

Table 4. Ablation Analysis for each Loss Term on TinyFace without Distractors using MobileFaceNet.

![Figure 3. Ablation Studies for Hyperparameters $\alpha$, $\beta$, and $\gamma$.](image)

![Figure 4. 2D embedding space for a toy problem with 8 identities (IDs) showing hard-to-recognize instances, even from the human perspective.](image)

**Face Recognizability-Aware Embeddings.** We conduct a toy experiment with 5 easy-to-recognize (IDs 0,2,3,5,7) and 3 hard-to-recognize (IDs 1,4,6) face examples to examine the embedding space learned by SoTAs and our model in Fig. 4. The UIs center is visualized as a reference index of recognizability, where a poor recognizability (hard-to-recognize) instance is essentially projected close to the UIs center. It is discerned that the competing models are incapable of separating the hard-to-recognize clusters from the UIs center. On the contrary, our model is inclined to divert the hard-to-recognize instances from the UIs center, yielding a well-separable (therefore a more discriminative) embedding space to resolve the downstream VLRFR task.

**Effect of Perceptibility Attention Module.** We portray the Class Activation Maps [37] obtained from ArcFace [8], Adaface [25], and our method for comparison. Interestingly, we observe from Fig. 5 that attending to the embeddings away from the UIs center focuses more on the salient face features, i.e., eyes, nose, and mouth, even when the face images appear to be obscure. We provide the extended heatmap visualizations in Fig. 11 of supplemental material.

**4.3. Face Image Quality Assessment (FIQA)**

**Learned RI from Perceptibility Regression Module.** In Fig. 6, the effect of ID loss reveals a greater negatively-skewed distribution of RI than other SoTAs (extended version in Fig. 9), indicating an improvement in recognizability, notably the hard-to-recognize instances. We discern that the proposed RI is a robust indicator in characterizing the recognizability notion - the highest and the lowest RI reflect the best-quality and the poor-quality face images, respectively. We also provide in Fig. 6 the face images with the lowest quality scores estimated by other competing scores. Through the comparison, RI is deemed proportional to the human cognitive level in terms of recognizability for the VLR face images under extreme poses, illuminations, occlusions, and others in the wild conditions.

**Error versus Rejection Curve (ERC).** To further substantiate the RI’s robustness, we perform a verification task on TinyFace [6] by randomly sampling 20,888 positive and 50,000 negative pairs from its gallery and probe sets to show
Figure 6. To facilitate direct comparison, all quality measures are normalized to ranges between 0 and 1. **(Left)** Visualization of RI sorted from the lowest scores (the poorest quality) to the highest (the best quality) in TinyFace [6] dataset. It is disclosed that the proposed RI characterizes the face quality better, closely simulating the human cognitive level. **(Right)** The VLR face images estimated with the lowest quality scores based on MagFace [35], CR-FIQA [2] and ERS [9]. In place of hard-to-recognize instances, we observe that several high-quality face instances are mistakenly assigned with low-quality scores, particularly MagFace and CR-FIQA.

Figure 7. Whilst Fig. 6 shows that the learned RI can characterize face quality well, we further demonstrate in this figure that the learned RI is a reliable recognizability metric by means of analyzing false non-match rate (FNMR). In particular, FNMR decreases gradually when the ratio of unconsidered VLR face images increases (sampled based on the lowest RI). This indicates that the learned RI well characterizes the recognizability of the VLR face images, such that the most recognizable VLR face images are learned with the highest RIs (corresponding to the right of the left-skewed RI distribution) for FNMR evaluation.

ERC in Fig. 7. The learned $\hat{\xi}$ outperforms SoTA in achieving stable and true rejection of unrecognized pairs in the rank of recognizability, especially when the FMR is at $10^{-4}$.

5. Conclusion

This paper addresses the problem arising from the hard-to-recognize faces in VLR images. Rather than treating these faces as UIs, we take a principled recognizability notion to characterize the recognizability of each image with a robust indicator, specifically the recognizability index (RI). The recognizability of an instance can thus be adjusted based on RI. Interestingly, attending to the embeddings projected away from the UIs cluster provides more explanatory power to the model to highlight the facial features more precisely. We evaluate the proposed method trained in an end-to-end manner on three VLR datasets and achieve SoTA for both VLRFR and FIQA.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NO. NRF-2022R1A2C1010710) and Institute for Basic Science (IBS-R029-C2) Korea.
References


