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

Jacky Chen Long Chai[1] Tiong-Sik Ng[1] Cheng-Yaw Low[2] Jaewoo Park[1] Andrew Beng Jin Teoh[1]
[1]Yonsei University [2]Institute for Basic Science
[1]{jackyccl,ngtiongsik,julypraise,bjteoh}@yonsei.ac.kr [2]{chengyawlow}@ibs.re.kr

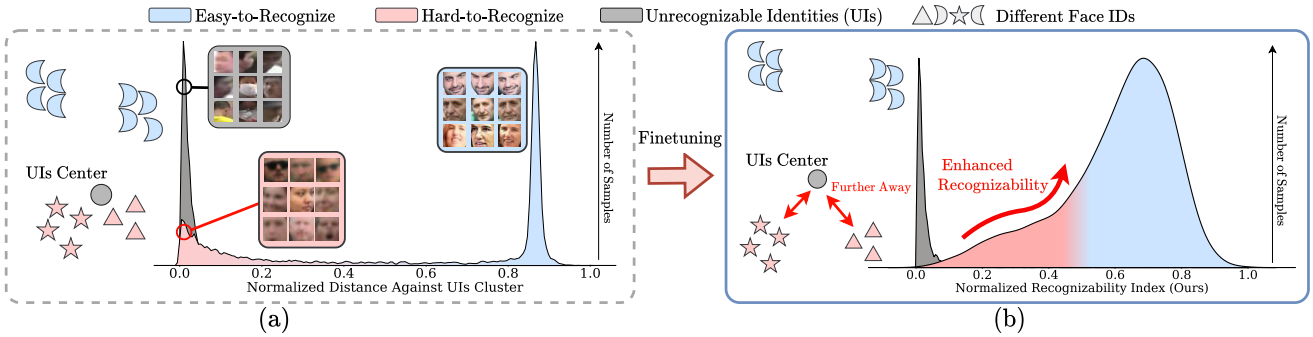


Figure 1. A deep face model pretrained on high-resolution face images introduces a cluster of unrecognizable instances (**grey spike**), dubbed *unrecognizable identities* (UIs) in [9]. (**a**) shows the bimodal distribution for a very low-resolution face dataset [6] based on distance against the UIs. 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 [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 [28]. The generic operation modes are either in image domain (*super-resolution, image synthesis*) [52, 54, 58], embedding domain (*resolution-invariant features, coupled mappings*) [33, 44], or at classifier level (*transfer learning, knowledge distillation*) [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* [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) [1, 18]. FIQA focuses either on (**i**) creating propositions to label the training data with face image quality scores and solve a regression problem [17, 18, 36], or (**ii**) linking the face embedding properties to FIQ scores [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 [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 [2].

Recently, [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 [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 dependence 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 [41, 54]. Several successors' works [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 mated pairs [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.

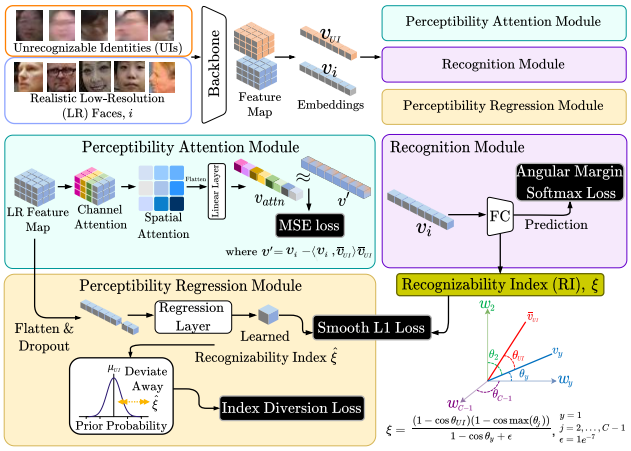


Figure 2. Our proposed model comprises three main modules: (i) *recognition module* learns the face embeddings v_i by optimizing the class prototypes for the recognition task; (ii) *perceptibility regression module* is designed to learn the recognizability index, $\hat{\xi}$ thereby enabling the recognizability prediction for any samples including the unseen ones; and (iii) *perceptibility attention module* performs channel and spatial-wise attention on the embeddings that approximate the projection away from the UIs cluster, v'. The RI is learned based on two criteria, as shown on the bottom right, where y is the target class, j is the non-target class across all C identities, and $\bar{v}_{UI}$ is the mean of UI cluster embeddings.

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 $\widehat{\boldsymbol{v}}_i$ with respect to its positive prototype $\boldsymbol{w}_{y_i}$ and the nearest negative prototype $\boldsymbol{w}_j, j \neq y_i$ across a set of C identities are as follows:

$$d_i^P = 1 - \cos(\theta_{y_i}) \ , \ d_i^N = 1 - \max_{j \in \{1,\dots,C\} \setminus \{y_i\}} \cos(\theta_j) \quad (1)$$

where θ_{y_i} is the positive angle between $\widehat{\boldsymbol{v}}_i$ and $\boldsymbol{w}_{y_i}$ and θ_j is the negative angle between $\widehat{\boldsymbol{v}}_i$ and $\boldsymbol{w}_j$. 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, $\overline{\boldsymbol{v}}_{UI}$, the proximity between $\widehat{\boldsymbol{v}}_i$ and $\overline{\boldsymbol{v}}_{UI}$ is defined as:

$$d_i^{UI} = 1 - \cos(\theta_{UI_i}) \quad (2)$$

where θ_{UI_i} is the angle between $\widehat{\boldsymbol{v}}_i$ and $\overline{\boldsymbol{v}}_{UI}$. Since the proximity is in terms of Cosine distance, the instances closest to the UIs cluster center (computed with the smallest d_i^{UI}) 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), ξ_i as follows:

$$\xi_i = d_i^{UI} \frac{d_i^N}{d_i^P + \epsilon} \quad (3)$$

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 $\widehat{\xi}_i$. To match $\widehat{\xi}_i$ to ξ_i in Eq. (3), we apply the smooth L1 loss [15] as follows:

$$L_{L1} = \begin{cases} 0.5(\xi_i - \widehat{\xi}_i)^2/\beta & \text{if } |\xi_i - \widehat{\xi}_i| < \beta \\ |\xi_i - \widehat{\xi}_i| - 0.5 \times \beta & \text{otherwise} \end{cases} \quad (4)$$

where β is a threshold that switches between L1 and L2-losses. In the early training stage, the L1 loss computes consistent gradients to approximate $\widehat{\xi}$ towards ξ. 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 $\widehat{\xi}_i$ as:

$$div = \frac{\widehat{\xi}_i - \mu_{UI}}{\sigma_{UI}} \quad (5)$$

Since the range of $\widehat{\xi}_i$ is arbitrary, it is rescaled with respect to μ_{UI} and σ_{UI} of the UI distribution as:

$$\mu_{UI} = \frac{1}{K}\sum_{k=1}^{K} s_k \ , \ \sigma_{UI} = \sqrt{\frac{\sum_{k=1}^{K}(s_k - \mu_{UI})^2}{K-1}} \quad (6)$$

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:

$$L_{ID} = \max(0, \tau - div) \quad (7)$$

where τ is the confidence interval hyperparameter. The ID loss enforces a deviation of at least τ between $\widehat{\xi}_i$ and μ_{UI}. As a hard-to-recognize instance is associated with a small $\widehat{\xi}_i$, it induces a relatively large ID loss. We attempt to push the hard-to-recognize instances outside the designated τ by minimizing the ID loss. In other words, the ID loss is equivalent to enforcing a statistically significant deviation of μ_{UI} from the estimated $\widehat{\xi}_i$ of VLR instances in the upper tail. Note that although the ID loss enforces a confidence interval, it does not divert $\widehat{\xi}_i$ infinitely as the $\widehat{\xi}_i$ 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 $\widehat{\xi}_i$ 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 $\boldsymbol{v}_i$ and the L2-normalized UIs cluster center $\overline{\boldsymbol{v}}_{UI}$, we define the embedding projected in the direction away from the UIs cluster $\boldsymbol{v}'_i$ based on [9] as:

$$\boldsymbol{v}'_i = \boldsymbol{v}_i - \langle \boldsymbol{v}_i, \overline{\boldsymbol{v}}_{UI} \rangle \overline{\boldsymbol{v}}_{UI} \tag{8}$$

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 $\boldsymbol{v}'_i$ 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 $\boldsymbol{v}_i^{attn}$ at the same dimension as $\boldsymbol{v}'_i$. Lastly, we utilize the mean squared error loss and formalize the regression problem as:

$$L_{MSE} = \frac{1}{B} \sum_{i=1}^{B} (\boldsymbol{v}'_i - \boldsymbol{v}_i^{attn})^2 \tag{9}$$

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:

$$L_{total} = L_{cls} + \alpha L_{L1} + \beta L_{ID} + \gamma L_{MSE} \tag{10}$$

where we opt for ArcFace [8] as the classification loss, L_{cls}. α, β, and γ 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 (TPIR) (%) @ 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 in Table 1, 2, and 3[1]. The former task is relatively challenging as only the noisy VLR examples are provided for probe-to-gallery matching. However, the latter suffers from a severe resolution gap between the HR galleries and the VLR probes. We disclose that the proposed model is not only resistant to the resolution gap but also a viable solution to the downstream VLR-VLR task. On the other hand, we leave the column blank for KD and resolution-invariant methods without involving the VLR datasets for fine-tuning, seeing that these VLR datasets are relatively small-scale and therefore easily prone to over-fitting.

TinyFace. We substantiate that our model outperforms other recent SoTAs designated for VLRFR by a remarkable margin, both with and without distractors (a summation of 153,428 face images of unknown identities in the gallery set). Specifically, we demonstrate that improving recognizability at the feature level is more meaningful than super-resolving the visual quality of the VLR face images, e.g. [6, 24, 39, 58]. With a quality-adapted margin for embedding learning, we discern that AdaFace [25] only readjusts the decision boundary - remaining the feature recog-

[1] In the "fine-tuned" column of these tables, we indicate ♯, †, and ✓ as fine-tuning on super-resolved, synthetic down-sampled VLR, and native VLR face images respectively.

Methods	Fine-Tuned	Rank-1 IR (%)	
		w/ dis	w/o dis
CSRI (ACCV19) [6]	♯	44.80	-
TURL (CVPR20) [39]		63.89	-
RIFR (T-BIOM20) [24]		70.40	-
VividGAN (TIP21) [58]	♯	47.16	-
MIND-Net (SPL21) [32]	✓	66.82	73.52
AdaFace (CVPR22) [25]		68.21	-
IDEA-Net (TIFS22) [31]	✓	68.13	-
AdaFace (Reproduce)	✓	71.38	75.67
Ours	✓	**73.06**	**77.22**

Table 1. IR Comparison on TinyFace using **ResNet-50**

Methods	Fine-Tuned	Rank-1 IR (%)			
		4.2m	2.6m	1.0m	Avg.
TCN (ICASSP10) [55]	✓	74.60	94.90	98.60	89.37
T-C (IVC20) [34]	†	70.20	93.70	98.10	87.33
FAN (ACCV19) [52]	✓	77.50	95.00	98.30	90.30
RAN (ECCV20) [11]	✓	81.30	97.80	98.80	92.63
DDL (ECCV20) [20]	✓	86.80	98.30	98.30	94.40
RIFR (T-BIOM20) [24]		88.30	98.30	98.60	95.00
MIND-Net (SPL21) [32]	✓	81.75	98.00	99.25	93.00
DSN (APSIPA21) [27]	✓	93.00	98.50	98.50	96.70
DRVNet (TPAMI21) [40]	✓	76.80	92.80	97.50	89.03
RPCL (NN22) [30]	✓	90.40	98.00	98.00	95.46
NPT (TPAMI22) [23]		85.69	99.08	99.08	96.61
IDEA-Net (TIFS22) [31]	✓	90.76	98.50	99.25	96.17
AdaFace (Reproduce)	✓	95.38	98.46	**99.84**	97.89
Ours	✓	**97.07**	**99.23**	99.80	**98.70**

Table 2. IR Comparison on SCFace using **ResNet-50**

Methods	Fine-Tuned	TPR(%)@FAR				TPIR20(%)@FPIR		
		0.3	0.1	0.01	0.001	0.3	0.2	0.1
CSRI (ACCV19) [6, 22]	♯	78.60	53.10	18.09	12.04	-	-	-
FAN (ACCV19) [52]		71.30	44.59	12.94	2.75	-	-	-
RAN (ECCV20) [11]		-	-	-	-	26.50	21.60	14.90
SST (ECCV20) [10]	✓	87.00	68.21	35.72	22.18	12.38	9.71	6.61
DSN (APSIPA21) [27]		75.09	52.74	21.41	11.02	-	-	-
DDAT (PR21) [22]		**90.40**	75.50	40.40	16.40	-	-	-
IDEA-Net (TIFS22) [11]		-	-	-	-	26.24	21.82	15.61
AdaFace (Reproduce)	✓	87.41	77.48	58.63	40.09	31.50	27.74	21.93
Ours	✓	90.21	**80.99**	**64.60**	**48.48**	**33.20**	**29.34**	**22.81**

Table 3. TPR(%)@FAR and TPIR20(%)@FPIR Comparison on SurvFace using **ResNet-50**

nizability unchanged.

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 downsampled 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 an estimated RI as shown in Eq. (4). It is believed that learning the 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-

Method	$\mathcal{L}_{cls}$	$\mathcal{L}_{ID}$	$\mathcal{L}_{MSE}$	$\mathcal{L}_{L1}$	Rank-1 IR (%)
Cross Entropy [43]	✓				68.884
NormFace [46]	✓				68.026
CosFace [47]	✓				70.306
MV-Softmax [48]	✓				70.547
CurricularFace [21]	✓				70.655
MagFace [35]	✓				70.467
AdaFace [25]	✓				70.359
Baseline (ArcFace) [8]	✓				70.333
I	✓	✓			71.298
II	✓		✓		71.540
III	✓			✓	71.674
Ours	✓	✓	✓	✓	**71.915**

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

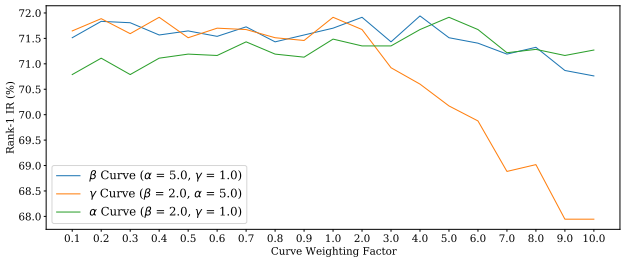


Figure 3. Ablation Studies for Hyperparameters α, β, and γ.

parameter by fixing the remaining weighting factors to be their optimal values in accordance to Table 5. We refer to Eq. (10) where α, β, γ correspond to the weighting of L_{L1}, L_{ID}, L_{MSE} respectively. Suppose that we attempt to study the various weighting factors of L_{L1} via manipulating the values of α, we fix $\gamma = 1$ and $\beta = 2$. The effects of α are shown as the α curve, and the same is applied to β and γ curves. Our ablations show that an overly large $\alpha(> 10)$ may experience difficulty in convergence [2], and 5 is the best choice. For β, the findings suggest a weighting factor slightly less than the α ensures better recognizability of the instances while still being upper bound by the devised RI. Lastly, we discover that L_{MSE} converges fast, and a higher γ hinders the overall model convergence, resulting in performance depreciation. A smaller γ is thus recommended and set to be 1 in our experiment.

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

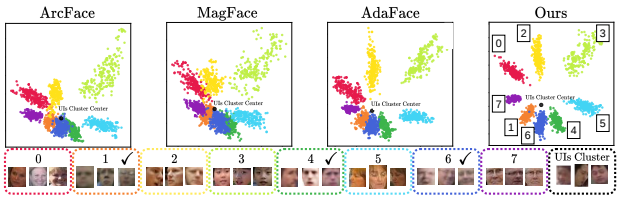


Figure 4. 2D embedding space for a toy problem with 8 identities (IDs). The IDs denoted with a "✓" contain hard-to-recognize instances, even from the human perspective. We visualize the UIs center as a reference index of recognizability, i.e., embeddings further away from the UIs center can better be recognized.

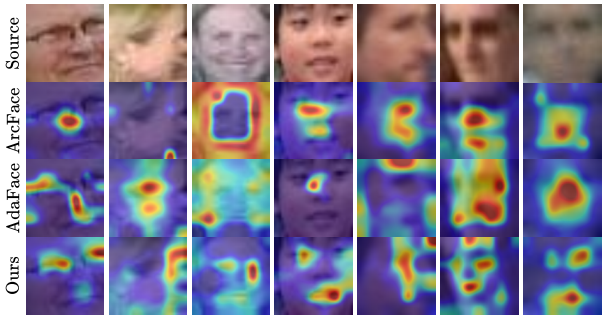


Figure 5. Class activation maps for several VLR face images generated based on ArcFace, AdaFace, and Ours.

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

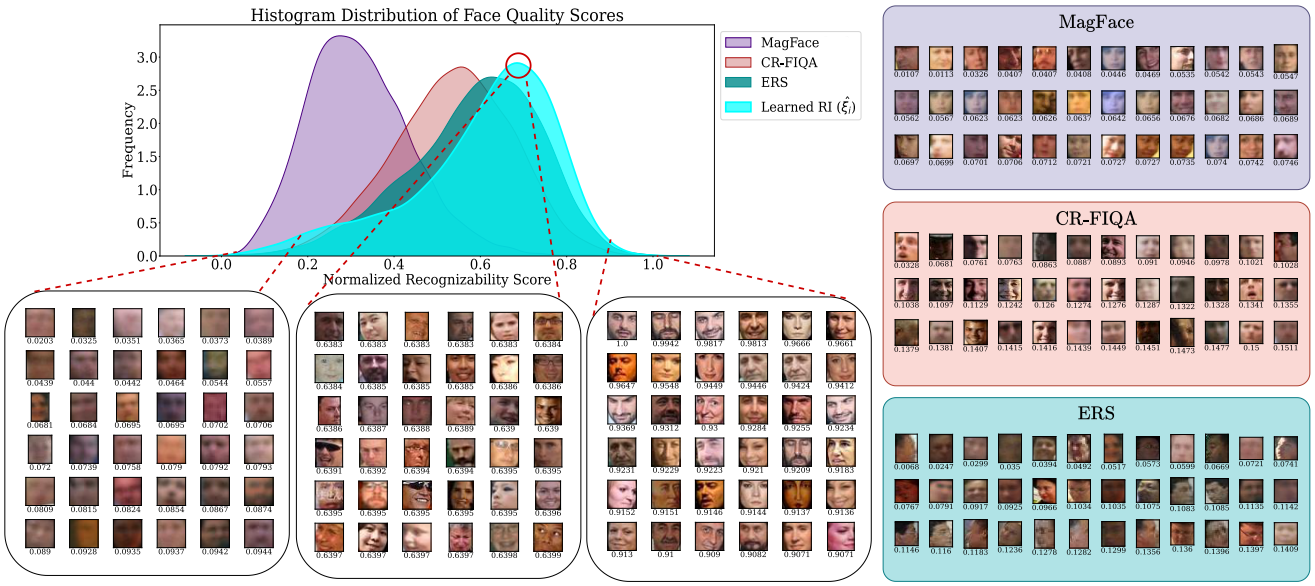


Figure 6. To facilitate direct comparison, all quality measures are normalized to ranges between 0 and 1. (**Left**) Visualization of RI sorted from the lowest scores (the poorest quality) to the highest (the best quality) in TinyFace [6] dataset. It is disclosed that the proposed RI characterizes the 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.

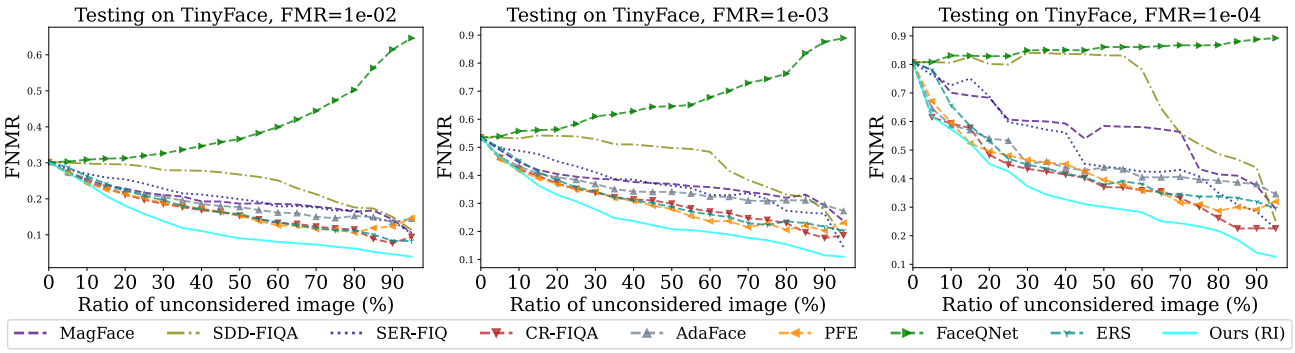


Figure 7. Whilst Fig. 6 shows that the learned RI can characterize face quality well, we further demonstrate in this figure that the learned RI is a reliable recognizability metric by means of analyzing false non-match rate (FNMR). In particular, FNMR decreases gradually when the ratio of unconsidered VLR face images increases (sampled based on the lowest RI). This indicates that the learn 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 $\widehat{\xi}_i$ outperforms SoTA in achieving stable and true rejection of unrecognized pairs in the rank of recognizability, especially when the FMR is at $1e^{-4}$.

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

[1] Lacey Best-Rowden and Anil K Jain. Learning face image quality from human assessments. *IEEE Transactions on Information Forensics and Security*, 13(12):3064–3077, 2018. 2, 3

[2] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. Cr-fiqa: Face image quality assessment by learning sample relative classifiability, 2021. 2, 3, 5, 7, 8

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 5

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020. 2, 3

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 5, 1

[6] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *Asian Conference on Computer Vision*, pages 605–621. Springer, 2018. 1, 2, 3, 5, 6, 7, 8

[7] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Surveillance face recognition challenge, 2018. 5, 1

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5, 7, 1, 3

[9] Siqi Deng, Yuanjun Xiong, Meng Wang, Wei Xia, and Stefano Soatto. Harnessing unrecognizable faces for improving face recognition, 2021. 1, 2, 4, 5, 8

[10] Hang Du, Hailin Shi, Yuchi Liu, Jun Wang, Zhen Lei, Dan Zeng, and Tao Mei. Semi-siamese training for shallow face learning. In *Proceedings of the European Conference on Computer Vision*, pages 36–53. Springer, 2020. 6

[11] Han Fang, Weihong Deng, Yaoyao Zhong, and Jiani Hu. Generate to adapt: Resolution adaption network for surveillance face recognition. In *European Conference on Computer Vision*, pages 741–758. Springer, 2020. 6

[12] Hans Fischer. *A history of the central limit theorem: from classical to modern probability theory*. Springer, 2011. 4

[13] Shiming Ge, Kangkai Zhang, Haolin Liu, Yingying Hua, Shengwei Zhao, Xin Jin, and Hao Wen. Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10845–10852, 2020. 2, 3

[14] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, 2018. 2, 3

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 4

[16] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011. 5, 1

[17] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet, 2020. 2, 3

[18] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 2, 3, 5

[19] Chih-Chung Hsu, Chia-Wen Lin, Weng-Tai Su, and Gene Cheung. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 28(12):6225–6236, 2019. 2

[20] Yuge Huang, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. Improving face recognition from hard samples via distribution distillation loss. In *European Conference on Computer Vision*, pages 138–154. Springer, 2020. 2, 3, 6

[21] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2020. 3, 7

[22] Qianfen Jiao, Rui Li, Wenming Cao, Jian Zhong, Si Wu, and Hau-San Wong. Ddat: Dual domain adaptive translation for low-resolution face verification in the wild. *Pattern Recognition*, 120:108107, 2021. 6

[23] Syed Safwan Khalid, Muhammad Awais, Zhenhua Feng, Chi Ho Chan, Ammarah Farooq, Ali Akbari, and Josef Kittler. Npt-loss: Demystifying face recognition losses with nearest proxies triplet. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 6

[24] Syed Safwan Khalid, Muhammad Awais, Zhen-Hua Feng, Chi-Ho Chan, Ammarah Farooq, Ali Akbari, and Josef Kittler. Resolution invariant face recognition using a distillation approach. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):410–420, 2020. 5, 6

[25] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022. 2, 3, 5, 6, 7, 1

[26] Eric P Kukula, Mathias J Sutton, and Stephen J Elliott. The human–biometric-sensor interaction evaluation method: Biometric performance and usability measurements. *IEEE Transactions on Instrumentation and Measurement*, 59(4):784–791, 2010. 5

[27] Shun-Cheung Lai and Kin-Man Lam. Deep siamese network for low-resolution face recognition. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1444–1449. IEEE, 2021. 6

[28] Pei Li, Loreto Prieto, Domingo Mery, and Patrick Flynn. Face recognition in low quality images: A survey, 2018. 2

[29] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019. 3

[30] Peiying Li, Shikui Tu, and Lei Xu. Deep rival penalized competitive learning for low-resolution face recognition. *Neural Networks*, 2022. 3, 6

[31] Cheng-Yaw Low and Andrew Beng-Jin Teoh. An implicit identity-extended data augmentation for low-resolution face representation learning. *IEEE Transactions on Information Forensics and Security*, 17:3062–3076, 2022. 3, 6

[32] Cheng-Yaw Low, Andrew Beng-Jin Teoh, and Jaewoo Park. Mind-net: A deep mutual information distillation network for realistic low-resolution face recognition. *IEEE Signal Processing Letters*, 28:354–358, 2021. 3, 6

[33] Ze Lu, Xudong Jiang, and Alex Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(4):526–530, 2018. 2

[34] Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99:103927, 2020. 3, 6

[35] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 2, 3, 6, 7, 8

[36] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7670–7679, 2021. 2, 3, 5

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 7

[38] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 2, 3

[39] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6817–6826, 2020. 2, 5, 6

[40] Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Derivenet for (very) low resolution image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 6

[41] Maneet Singh, Shruti Nagpal, Mayank Vatsa, Richa Singh, and Angshul Majumdar. Identity aware synthesis for cross resolution face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–488, 2018. 2

[42] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 5, 1

[43] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *Advances in Neural Information Processing Systems*, 27, 2014. 7

[44] Veeru Talreja, Fariborz Taherkhani, Matthew C Valenti, and Nasser M Nasrabadi. Attribute-guided coupled gan for cross-resolution face recognition. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2019. 2

[45] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5651–5660, 2020. 1, 2, 3

[46] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 7

[47] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 7

[48] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12241–12248, 2020. 3, 7

[49] Andreas Wolf. Portrait quality (reference facial images for mrtd). *Version: 0.06 ICAO, Published by authority of the Secretary General*, 2016. 3

[50] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 5

[51] Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. Inducing predictive uncertainty estimation for face verification. In *British Machine Vision Conference*, 2020. 3

[52] Xi Yin, Ying Tai, Yuge Huang, and Xiaoming Liu. Fan: Feature adaptation network for surveillance face recognition and normalization. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 6

[53] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2926–2943, 2019. 2

[54] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408, 2016. 2

[55] Juan Zha and Hongyang Chao. Tcn: Transferable coupled network for cross-resolution face recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3302–3306. IEEE, 2019. 6

[56] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *Proceedings of the European conference on computer vision (ECCV)*, pages 183–198, 2018. 3

[57] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 5, 1

[58] Yang Zhang, Ivor W Tsang, Jun Li, Ping Liu, Xiaobo Lu, and Xin Yu. Face hallucination with finishing touches. *IEEE Transactions on Image Processing*, 30:1728–1743, 2021. 2, 5, 6

[59] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 868–884. Springer, 2016. 5, 1

[60] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011. 1

Supplemental Materials

A. Realistic Low-Resolution Datasets

We elaborate the face datasets involved in our experiments in this section. These include three realistic VLR datasets for benchmarking under the open-set evaluation protocol, i.e. Tinyface, SCface, and SurvFace, alongside a UI face dataset for recognizability index (RI) learning.

TinyFace. TinyFace [6] is a composition of 7,804 and 8,171 VLR face images annotated with 2,570 and 2,569 identity labels in each training and testing set, respectively. On average, the input resolution for each VLR face image is of 20×16 pixels. The gallery search space is interfered with 153,428 distractors of unknown identities to simulate a more challenging open-set VLR-VLR identification scenario.

SurvFace. SurvFace [7] is the largest surveillance face dataset for VLR-VLR identification and verification tasks evaluated in presence of unmated distractors (probe images without a matched gallery ID). In a nutshell, it contains 463,507 VLR face images contributed by 15,573 subjects with an average resolution of 24×20 pixels. For the VLR-VLR identification task, it is partitioned with 220,890 face images for 5,319 subjects in the training set, whilst the testing set consists of 242,617 face images for 10,254 subjects (including 141,736 unmated probe images from 4,935 subjects). On the other hand, it is sampled with 5,319 matched and unmatched pairs for evaluating the VLR-VLR verification task.

SCFace. Unlike TinyFace and SurvFace, SCFace [16] is a small-scale face dataset with HR and VLR face images for the HR-VLR identification task. Overall, each of the 130 subjects is provided with a single HR mugshot (as a gallery template), alongside 15 VLR face images captured from three standoff distances, i.e. 4.20m (D1), 2.60m (D2) and 1.00m (D3). Our experiments allocate the face images for the first 50 subjects (from ID *001* to *050*) for training, excluding the corresponding HR gallery templates. The remaining subjects (from ID *051* to *130*) are probed with respect to all 130 gallery templates.

UI Face Dataset. To elicit an UIs cluster for RI learning, we assemble an ad-hoc UI face dataset with a summation of 11,707 unlabeled VLR face images. We single out these VLR face images from two person re-identification datasets (independent from the three benchmarking datasets), i.e. LPW [42] and MARS [59], by the MTCNN face detector [57]. As illustrated in Fig. 8, these VLR face images are close to unrecognizable from the image quality perspective.

B. Experiment Settings

Given face images of low pixel resolutions in TinyFace, SurvFace, SCFace, and the UI dataset, we rescale all the images into 112×112 pixels through bi-cubic interpolation. The rescaled images are forwarded to the pretrained MobileFaceNet [5], and ResNet-50 [8] for embedding learning in the training stage, followed by feature extraction in the inference stage. Since the validation set is not available for all datasets for hyperparameter tuning, we sample a random subset with definite horizontal flip and extreme downsampled counterparts by 16×16 pixels from the respective training sets. For consistent batch-wise statistics, we suspend the pre-trained batch-normalization (BN) layers from learning - applicable to all the BN layers after every 2D convolutional layer. Our model is trained with an Adam optimizer and augmented instances, including random horizontal flip, random rotation (in the range of ±10 degree), and rescaling (by 64×64, 100×100). Our implementation is exercised using a machine with two NVIDIA 2080Ti GPUs in the PyTorch framework.

Figure 8. Sample face images in our UI dataset.

In Table 5, we provide a complete description of the ResNet-50 hyperparameters used for all three datasets defined in Sec. A, namely TinyFace, Survface, and SCFace. We refer all modules mentioned in Table 5 to Fig. 2 to aid the readability. For MobileFaceNet, the learning rate for backbone, classifier head and perceptibility attention module are changed to $1e^{-4}$, $1e^{-3}$ and $1e^{-2}$ respectively. The others remain the same unless otherwise specified. These hyperparameters are determined based on the validation sets sampled from the corresponding training set.

C. Extended Visualization

An extended visualization of Fig. 6 is shown in Fig. 9.

For further exploration, we examine the ERS scores and RI with additional hard-to-recognize instances in Fig. 10. We demonstrate that RI is proportional to the human cognitive level over ERS for the VLR face images with extreme poses, illuminations, occlusions, and others in the wild conditions.

On top of Fig. 5, we also provide an extended visualization of class activation maps rendered based upon varying softmax-based losses in Fig. 11, including the most recent SoTA - AdaFace [25]. While some methods can capture certain parts of a face, our method provides more consistent attention to the salient facial features and is more capable of handling multiple poses, especially when the face is facing sideways or not aligned to the center of an image.

TinyFace			SCFace			Survface		
Mini-Batch Size		64	Mini-Batch Size		64	Mini-Batch Size		100
# Epoch		15	# Epoch		6	# Epoch		30
Learning Rate	Backbone	$1e^{-5}$	Learning Rate	Backbone	$1e^{-5}$	Learning Rate	Backbone	$1e^{-5}$
	RM[1]	$1e^{-3}$		RM[1]	$1e^{-3}$		RM[1]	$5e^{-4}$
	PAM[2]	$1e^{-3}$		PAM[2]	$1e^{-3}$		PAM[2]	$1e^{-3}$
	PRM[3]	$1e^{-4}$		PRM[3]	$1e^{-4}$		PRM[3]	$1e^{-4}$
Learning Rate Decay		0.1 / 12^{th} epoch	Learning Rate Decay		0.1 / 12^{th} epoch	Learning Rate Decay		0.1 / 6^{th} epoch
Dropout (Backbone)		0.2	Dropout (Backbone)		0.2	Dropout (Backbone)		0.4
Dropout (PRM[3])		0.9	Dropout (PRM[3])		0.9	Dropout (PRM[3])		0.9
Weight Decay		$1e^{-4}$	Weight Decay		$1e^{-4}$	Weight Decay		$1e^{-3}$
s, m		64, 0.45	s, m		64, 0.45	s, m		64, 0.45
α, β, γ		5, 2, 1	α, β, γ		5, 2, 1	α, β, γ		5, 2, 1

[1] Recognition Module
[2] Perceptibility Attention Module
[3] Perceptibility Regression Module

Table 5. Hyperparameter Configuration for **ResNet-50**

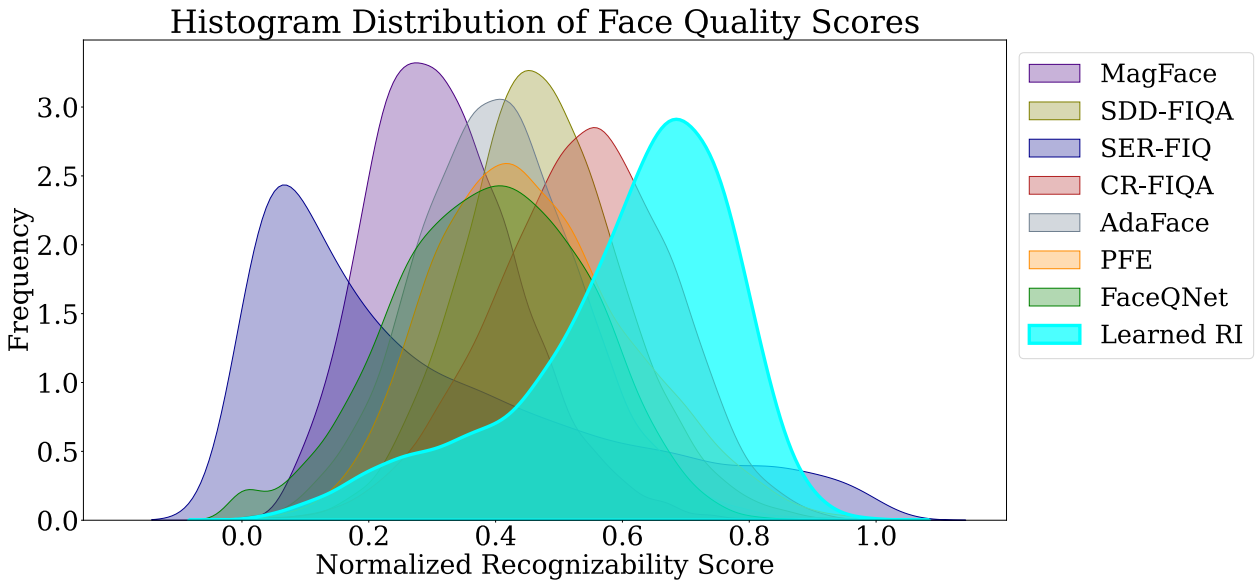
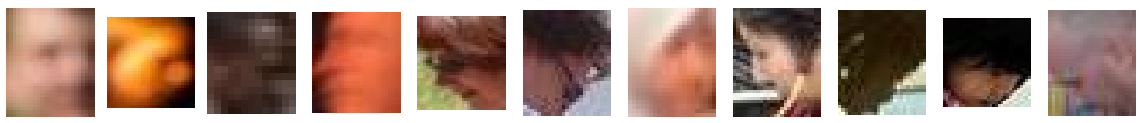


Figure 9. Quality score distributions for state-of-the-art metrics. In comparison, the score distribution for our learned RI is left-skewed. This indicates that the recognizability for majority of the VLR instances are enhanced, leaving minority of the hard-to-recognize VLR instances at lower scores.



ERS	0.4044	0.4080	0.3381	0.3483	0.4402	0.2919	0.2725	0.2884	0.4189	0.4828	0.3201
Ours	0.1799	0.1035	0.1478	0.1479	0.1720	0.0979	0.0954	0.2155	0.1923	0.2104	0.1950

Figure 10. Comparison of recognizability between ERS [9] and our proposed RI for VLR face images captured under various unconstrained surveillance scenarios. Note that the lower the value, the harder the VLR face image is to be recognized.

Figure 11. Class Activation Maps (extension to Fig. 5) for our RI, ArcFace [8] (Baseline), Cross Entropy, MagFace [35], and AdaFace [25].