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**

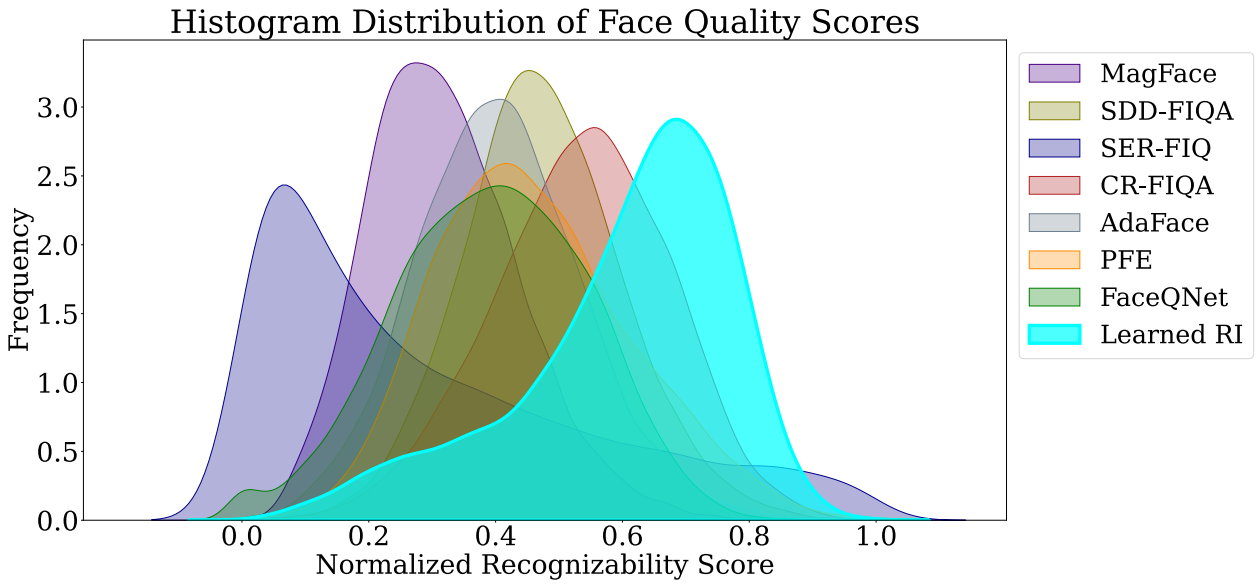
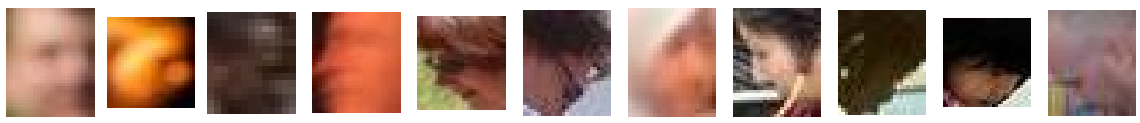


Figure 9. Quality score distributions for state-of-the-art metrics. In comparison, the score distribution for our learned RI is left-skewed. This indicates that the recognizability for majority of the VLR instances are enhanced, leaving minority of the hard-to-recognize VLR instances at lower scores.



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].