Supplementary Materials

We offer additional resources in this section to enhance the understanding and reproducibility of this work. To summarize, our supplementary materials are presented as follows:

A. Theoretical Analysis

B. Benchmarking Datasets

C. Hyperparameter Analysis and Configuration

D. Stability Analysis

E. Model Extension

A. Theoretical Analysis

We provide the mathematical proofs for the conceptual principles that underlie SlackedFace, including the role of (1) slacked margin, and (2) regularization term. To simplify our analysis, we omit normalization procedures and subscript indexes.

A1. Slacked Margin

Assumption. A face example, either easy or hard, is ranked recognizable, if and only if two conditions are satisfied: (1) its embedding magnitude $\|\mathbf{z}\| > \tau_{\text{norm}}$, and the Cosine similarity between $\mathbf{z}$ and its true identity prototype $\cos\theta_y > \tau_{\text{cosine}}$.

Therefore, a high recognizability example has a large embedding magnitude and a large similarity score, while an unrecognizable example is deficient at both. We first show that the SlackedFace margin is proportional to these recognizability factors.

Proposition 1. *(Slacked margin) The slacked margin $m = \sigma^{1.0-\rho}$ is monotonically strictly increasing with respect to $\|\mathbf{z}\|$ and $\cos\theta_y$, if $\|\mathbf{z}\| < \tau$ and $\theta_y + m < \pi/2$.*

SlackedFace, hence, can accurately induce a margin that corresponds to recognizability, resulting in a model that produces a significantly large gradient update on recognizable examples and a small gradient update on unrecognizable examples during training.

Corollary. *(Gradient of SlackedFace Loss) The magnitude of SlackedFace loss gradient $\frac{\partial \mathcal{L}}{\partial \theta_y}$ on the target similarity angle θ_y is monotonically and strictly increasing with respect to $\|\mathbf{z}\|$ and $\cos\theta_y$, if $\|\mathbf{z}\| < \tau$ and $\theta_y + m < \pi/2$.*

Proof. Let $\mathcal{L} = -\log\exp(\cos(\theta_y+m))/[\sum_{k\neq y}\exp(\theta_k) + \exp(\theta_y+m)]$. Then,

$$\frac{\partial \mathcal{L}}{\partial \theta_y} = [1 - (\sum_{k\neq y}\exp(\cos\theta_k - \cos(\theta_y+m)))^{-1}]\sin(\theta_y+m). \tag{12}$$

Given $\theta_y + m < \pi/2$, $\sin(\theta_y+m)$ and $\sum_{k\neq y}\exp(\cos\theta_k - \cos(\theta_y+m))$ are strictly increasing with m. Similarly, the derivative of $\frac{\partial \mathcal{L}}{\partial \theta_y}$ with respect to m is also strictly increasing. Therefore, by the previous proposition, the gradient is also strictly increasing with embedding magnitude and cosine similarity. □

The above corollary demonstrates that *the SlackedFace margin is highly correlated with the recognizability of a face.* This means that SlackedFace induces a smaller margin for unrecognizable examples, and otherwise. As a result, SlackedFace can effectively promote empowered embedding learning based on face recognizability.

A2. Regularization Term

In accordance with (9), the SlackedFace loss incorporates a regularization term $\mathcal{L}_{\mathcal{R}}$ to maximize the embedding norm to an upper bound. We demonstrate that this improves learning stability and convergence by enforcing the model to focus on minimizing the relative angle between a face embedding and its true identity prototype.

Since SlackedFace is updated by a gradient-based optimizer, the gradient update for the embedding vector is performed by $\mathbf{z} \leftarrow \mathbf{z} - \alpha \frac{\partial \mathcal{L}_{class}}{\partial \mathbf{z}}$ at a certain learning rate. Therefore, learning of embedding vector is determined by the gradient $\frac{\partial \mathcal{L}_{class}}{\partial \mathbf{z}}$ in (13). We analyze this gradient as follows:

Proposition 2. *For the classification loss $\mathcal{L}_{class}$,*

$$\frac{\partial \mathcal{L}_{class}}{\partial \mathbf{z}} = \frac{1}{\|\mathbf{z}\|} \cdot \sum_k \frac{\partial \mathcal{L}_{class}}{\partial \cos\theta_k} (\widehat{\mathbf{w}}_k - \cos\theta_k \widehat{\mathbf{z}}) \tag{13}$$

where $\widehat{\mathbf{w}}_k$ and $\widehat{\mathbf{z}}$ are obtained by normalizing $\mathbf{w}_k$ and $\mathbf{z}$, respectively.

The proposition indicates that the gradient $\frac{\partial L_{class}}{\partial \mathbf{z}}$ for an embedding update is *disentangled* to two terms, specifically, $\|\mathbf{z}\|^{-1}$ and $\sum_k \frac{\partial L_{class}}{\partial \cos\theta_k} (\widehat{w}_k - \cos\theta_k \widehat{\mathbf{z}})$, where the former depends on the embedding norm and the latter does not. Since the regularization term reduces the reciprocal of embedding magnitude, the regularizer makes the gradient less variant to the magnitude:

Corollary. *Minimizing $\mathcal{L}_r$ reduces the magnitude of the embedding gradient, thereby making the embedding gradient invariant to embedding magnitude.*

Overall, the regularization term serves two important roles: (1) preventing the overly large gradient update, which helps to *stabilize the training stage*. (2) *marking the embedding learning more depend on the relative angle between face embeddings and true identity prototypes*. Increasing the angle-dependency of the embedding learning improves the generalization of the corresponding cosine similarity metric for open-set applications [24].

B. Benchmarking Datasets

We further elaborate on our benchmarking datasets, including SCFace, TinyFace, and DroneFace, for performance evaluation under the open-set deployment scenario.

SCFace. Real-world face recognition systems enroll individuals using high-resolution (HR) mugshots, leaving unseen (test) face images unrestricted. Hence, SCFace includes a gallery set with a high-resolution (HR) mugshot per identity (ID), and three low-resolution (LR) probe sets, namely D1, D2, and D3, to simulate a real-world HR (gallery)-LR (probe) identification task. As a whole, these probe sets are compiled with examples captured at standoff distances of 4.20m, 2.60m, and 1.00m, respectively. In compliance with [19], we allocate the first 50 subjects (from ID 01 to 050) for training, while the other 80 subjects (from ID 051 to 130) are reserved for testing.

TinyFace. Contrary to SCFace, TinyFace is a large-scale LR face dataset with both an LR gallery and an LR probe set for an LR-LR identification task. Overall, it is a composition of 7,804 / 8,171 face images annotated with 2,570 / 2,569 ID labels in each training and testing set, respectively. On average, the pixel resolution of these examples is limited to only 20×16 pixels. It is worth noting that its gallery search space is interfered with 153,428 distractors of unknown identities to simulate a more challenging real-world scenario.

DroneFace. DroneFace, on the other hand, is only a test set for an HR-to-LR identification task (similar to SCFace). As a whole, it consists of 11 subjects with 1,364 examples detected from drone footage (at 1.5m to 5m high, and 2m to 17m away from the subjects) in the probe set and 2 frontal mugshots per ID as the enrolled templates in the gallery set. We evaluate the generalization performance on DroneFace using the SCFace-learned models.

Ad-Hoc Distractor Set. As both SCFace and DroneFace contain no distractors, we extend these datasets with an ad-hoc distractor set of 20,000 unknown examples randomly sampled from that of TinyFace.

We summarize the data distribution for each dataset in Table 3. On the other hand, we portray 10 hardest and easiest examples indexed by Norm and P-Norm in Figure 6.

Datasets	Desc.	Train. Set	Test. Set			Eval. Protocol
			Gallery	Probe	Distract.	
SCFace	# IDs	50	80	80	-	HR-to-LR
	# Imgs.	800	80	1,200	20,000*	
TinyFace	# IDs	2,570	2,569	2,569	-	LR-to-LR
	# Imgs.	7,804	4,443	3,728	153,428	
DroneFace	# IDs	-	11	11	-	HR-to-LR
	# Imgs.	-	22	1,364	20,000*	

Table 3: Data distribution for our benchmarking datasets, including TinyFace, SCFace, and DroneFace. Note that "*" refers to a random distractor set of 20,000 LR face images sampled from that of TinyFace.

C. Hyperparameter Analysis and Configuration

Compared to other static margin-based softmax losses that involve two primary hyperparameters, i.e., the scaling factor s and the margin term m, training a SlackedFace model requires tuning two additional hyperparameters, specifically the degree of margin relaxation η in (8), and the regularization weighting factor λ in (9). Other default hyperparameters are the Sigmoid steep slope $\Lambda = 6.0$ in (3), the generic upper bound for the embedding norm $\tau = 10^2$ in (4), and the regression transition parameter $\gamma = 0.5$ in (10). Accordingly, we analyze in Table 4 the effect of η and λ, in addition to m. These are the key parameters to determine the generalization power of the SlackedFace models, achieving optimal performance with and without distractors in the most difficult SCFace D1 probe set. We summarize our hyperparameter configuration in Table 5.

D. Stability Analysis

To assess the model's stability against random initializations, we train the SlackedFace models and the comparing instances using 5 random seeds, i.e., 0, 1, 42, 1234, and 2023. Our experimental results in Table 6 reveal that the SlackedFace models exhibit the ideal robustness to multiple random initializations, underlining that SlackedFace is a reliable alternative to other non-static margin-based softmax losses.

Hyperparams.	Setting	Ori. / Ext. SCFace			
		D1	D2	D3	Mean
Effects of η ($m = 0.5$; $\lambda = 0.1$)	0.025	89.25 / 59.25	**98.50** / 92.50	**98.75** / 90.25	95.50 / 80.67
	0.05	89.50 / 60.25	98.25 / **93.00**	98.25 / 90.25	95.33 / 81.17
	0.10	**90.00** / 60.75	98.25 / 92.50	98.25 / **90.75**	**95.50** / **81.33**
	0.15	89.25 / **61.00**	98.00 / 92.25	98.25 / 89.50	95.17 / 80.92
	0.20	88.50 / 58.00	98.00 / 92.00	98.25 / 90.25	94.92 / 80.08
Effects of λ ($m = 0.5$; $\eta = 0.1$)	0.0	88.75 / 59.75	**98.50** / **93.00**	97.75 / 90.25	95.00 / 81.00
	0.05	89.75 / 61.00	98.25 / 92.25	98.50 / 90.00	**95.55** / 81.08
	0.10	**90.00** / 60.75	98.25 / 92.50	98.25 / **90.75**	95.50 / 81.33
	0.20	89.25 / **62.25**	98.25 / **93.00**	**98.75** / 89.75	95.42 / **81.67**
	0.50	88.25 / 61.50	**98.50** / 92.25	98.25 / 90.00	95.00 / 81.25
Effects of m ($\eta = 0.1$; $\lambda = 0.1$)	0.40	**90.00** / 59.75	98.25 / 90.75	**98.50** / 89.25	95.58 / 79.92
	0.45	**90.00** / 60.00	**98.50** / 92.00	**98.50** / 89.75	**95.67** / 80.58
	0.50	**90.00** / 60.75	98.25 / 92.50	98.25 / **90.75**	95.50 / **81.33**
	0.55	89.00 / **60.75**	98.25 / 92.50	**98.50** / 90.50	95.25 / 81.25
	0.60	88.75 / 61.00	**98.50** / **93.25**	**98.50** / 89.75	95.25 / **81.33**

Table 4: Hyperparameter analyses for SlackedFace (using pre-trained MobileFaceNet as the embedding encoder).

	Hyperparameters	SCFace		TinyFace
		MobileFaceNet	ResNet50	ResNet50
Basic	Mini Batch Size	32	32	32
	# Epochs (Fast-HC + End-to-End)	8 + 32	8 + 32	8 + 32
	Learning Rate	$1e^{-03}$	$1e^{-03}$	$1e^{-03}$
	Learning Rate Decay	0.1 / 8 epochs	0.1 / 4 epochs	0.1 / 6 epochs
	Weight Decay	$1e^{-04}$	$1e^{-04}$	$1e^{-04}$
	Dropout Rate	0.6	0.8	0.8
SlackedFace (Default)	Sigmoid Steep Slope Λ in (3)	6.0		
	Upper Bound for Norm τ in (4)	100		
	Regress. Transition Parameter. γ in (10)	0.5		
SlackedFace (Fine-Tuned)	Scaling s, Margin m in (8)	60, 0.50		
	Slacked Margin Degree η in (8)	0.10		
	Reg. Weighting Factor λ in (9)	0.10		

Table 5: Overall hyperparameter configuration in our experiments. We set the learning rate for the pre-trained backbone (inclusive of the embedding MLP) to 0.1x of the softmax classifier to prevent the prior knowledge from being distorted with noises in LR face examples.

Face Models	Ori. / Ext. SCFace			
	D1	D2	D3	Mean
ElasticFace	95.75±0.35 / 78.45±0.97	99.50±0.25 / 94.95±0.62	99.55±0.27 / 97.90±0.45	98.27±0.11 / 90.43±0.41
MagFace	95.80±0.62 / 77.65±1.24	99.50±0.18 / 94.65±0.58	99.60±0.22 / 98.10±0.38	98.30±0.24 / 90.13±0.59
AdaFace	95.90±0.55 / 77.95±0.65	**99.55±0.11** / 95.75±0.40	99.95±0.11 / 97.95±0.27	98.47±0.14 / 90.55±0.10
SlackedFace	**96.50±0.50** / **79.30±0.48**	99.50±0.18 / 96.00±0.59	**100.0±0.00** / **98.65±0.29**	**98.67±0.12** / **91.32±0.15**
SlackedCosFace	96.20±0.21 / 78.45±1.05	99.40±0.22 / **96.10±0.74**	**100.0±0.00** / 98.55±0.37	98.53±0.13 / 91.03±0.55

Table 6: Performance comparison for SlackedFace and SoTAs (using pre-trained ResNet50 as the embedding encoder) over 5 random initializations, in terms of averaged rank-1 identification rate (%) and standard deviation. Note that SlackedCosFace is an extended variant to be disclosed in Section E of this supplementary material.

E. Extension of SlackedFace

Aside from ArcFace (reported in our manuscript), we extend SlackedFace based on CosFace, termed SlackedCosFace in this section, for further exploration. Likewise, we substitute the static margin m in $\mathcal{T}(\theta_j, s, m)_{\text{CosFace}}$ (2) with a slacked margin term $\delta(m)$ as follows:

$$\mathcal{T}(\theta_j, s, m)_{\text{SlackedCosFace}} = s(\cos\theta_{y_i} - \delta(m)) \;\;,\;\; \delta(m) = m + \eta\,\hat{\sigma}_i \tag{14}$$

We also disclose in Table 6 that SlackedCosFace performs on a par with the ArcFace-learned

Figure 6: A collection of 10 hardest and easiest (unseen) test examples indexed by Norm and proposed P-Norm for each benchmarking dataset.

counterpart. More importantly, we demonstrate that the overall performance of the proposed SlackedFace models (both learned with ArcFace and CosFace) surpass other SoTAs by a significant margin in resolving open-set LR face identification tasks, especially in the most challenging D1 probe set with distractors.