UniFace: Unified Cross-Entropy Loss for Deep Face Recognition

Jiancan Zhou[1,2,3,†], Xi Jia[1,2,4,†], Qiufu Li[1,2,†], Linlin Shen[1,2,#], Jinming Duan[4,5]
[1]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[2]Computer Vision Institute, Shenzhen University
[3]Aqara, Lumi United Technology Co., Ltd.
[4]School of Computer Science, University of Birmingham, UK
[5]Alan Turing Institute, UK
zhoujiancan@foxmail.com;x.jia.1@cs.bham.ac.uk;{liqiufu,llshen}@szu.edu.cn;j.duan@bham.ac.uk

A. UCE Loss

This section elaborates on the derivations from the Softmax loss to our UCE loss. To encourage a similarity matrix $\mathcal{S}_{\text{sam-cla}}$ that is diagonally dominant in both its rows and columns.We expect a **unified threshold** t, such that

$$\begin{aligned}&\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}} \le t \le \cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}, \quad \text{and}\\ &\cos\theta^{(ji)}_{\boldsymbol{x},\text{w}} \le t \le \cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}, \quad \forall\, i,\ j,\ \text{with } j\neq i.\end{aligned} \tag{s1}$$

If we define the maximum angle between the features and their positive class proxy as θ_{pos} and the minimum angle between the features and their negative class proxies as θ_{neg}, that is

$$\theta_{\text{pos}} = \max\Big(\bigcup_{i=1}^{N}\{\theta^{(ii)}_{\boldsymbol{x},\text{w}} : \boldsymbol{x}^{(i)}\in\mathcal{F}_i\}\Big), \tag{s2}$$

$$\theta_{\text{neg}} = \min\Big(\bigcup_{i=1}^{N}\bigcup_{\substack{j=1\\ j\neq i}}\{\theta^{(ij)}_{\boldsymbol{x},\text{w}} : \boldsymbol{x}^{(i)}\in\mathcal{F}_i\}\Big), \tag{s3}$$

then, there exists a threshold t satisfying Eq. (s1) if and only if $\theta_{\text{pos}} \le \theta_{\text{neg}}$, and the unified threshold $t = \cos\theta_t$ is valid for any

$$\theta_t \in [\theta_{\text{pos}}, \theta_{\text{neg}}]. \tag{s4}$$

According to the analysis of the original softmax loss in the manuscript, a model $\mathcal{M}$ trained using the softmax loss cannot ensure $\theta_{\text{pos}} \le \theta_{\text{neg}}$. In order to alleviate this drawback, we design the **Unified Cross-Entropy (UCE)** by supposing that there exists a unified threshold $t = \cos\theta_t$ (i.e., $\theta_{\text{pos}} \le \theta_t \le \theta_{\text{neg}}$). Starting from the original softmax loss, we can have

$$\begin{aligned}&L_{\text{sl}}(\boldsymbol{X}^{(i)})\\ =&-\frac{1}{N}\sum_{k=1}^{N}\log\frac{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}+\sum_{j\neq i}\text{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}} &\text{(s5)}\\ =&-\frac{1}{N}\Big(\log\frac{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}+\sum_{j\neq i}\text{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}}\\ &+\sum_{\substack{k=1\\ k\neq i}}^{N}\log\frac{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}+\sum_{j\neq i}\text{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}}\Big) &\text{(s6)}\\ =&-\frac{1}{N}\Big(\log\frac{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}+\sum_{j\neq i}\text{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}}\\ &+\sum_{\substack{k=1\\ k\neq i}}^{N}\log\frac{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}+\text{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\text{w}}}+\sum\limits_{\substack{j\neq i\\ j\neq k}}\text{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\text{w}}}}\Big). &\text{(s7)}\end{aligned}$$

According to Eqs. (s2) - (s4), we can derive

$$\begin{aligned}&L_{\text{sl}}(\boldsymbol{X}^{(i)})\\ \le&-\frac{1}{N}\Big(\log\frac{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}+\sum_{j\neq i}\text{e}^{s\cos\theta_{\text{neg}}}}+\\ &\sum_{k\neq i}\log\frac{\text{e}^{s\cos\theta_{\text{pos}}}}{\text{e}^{s\cos\theta_{\text{pos}}}+\text{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\text{w}}}+\sum\limits_{\substack{j\neq i\\ j\neq k}}\text{e}^{s\cos\theta_{\text{neg}}}}\Big) &\text{(s8)}\\ \le&-\frac{1}{N}\Big(\log\frac{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}}{\text{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\text{w}}}+\sum_{j\neq i}\text{e}^{s\cos\theta_t}}+\\ &\sum_{k\neq i}\log\frac{\text{e}^{s\cos\theta_t}}{\text{e}^{s\cos\theta_t}+\text{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\text{w}}}+\sum\limits_{\substack{j\neq i\\ j\neq k}}\text{e}^{s\cos\theta_t}}\Big) &\text{(s9)}\end{aligned}$$

$$= \frac{1}{N}\Big(\log(1 + \sum_{j\neq i} \frac{\mathrm{e}^{s\cos\theta_t}}{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}}}}) + \sum_{k\neq i} \log(1 + \frac{\mathrm{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\mathrm{w}}}}{\mathrm{e}^{s\cos\theta_t}} + \frac{\sum\limits_{\substack{j\neq i\\ j\neq k}} \mathrm{e}^{s\cos\theta_t}}{\mathrm{e}^{s\cos\theta_t}})\Big) \tag{s10}$$

$$= \frac{1}{N}\Big(\log(1 + (N-1)\frac{\mathrm{e}^{s\cos\theta_t}}{\mathrm{e}^{s\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}}}}) + \sum_{k\neq i} \log(N - 1 + \frac{\mathrm{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\mathrm{w}}}}{\mathrm{e}^{s\cos\theta_t}})\Big) \tag{s11}$$

$$= \frac{1}{N}\Big\{ \log(1 + \mathrm{e}^{s\cos\theta_t - s\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}} + \log(N-1)}) + \sum_{k\neq i} \big[\log(N-1) + \log(1 + \frac{1}{N-1}\frac{\mathrm{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\mathrm{w}}}}{\mathrm{e}^{s\cos\theta_t}})\big]\Big\} \tag{s12}$$

$$= \frac{1}{N}\big[\log(1 + \mathrm{e}^{s\cos\theta_t - s\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}} + \log(N-1)}) + \sum_{k\neq i} \log(1 + \mathrm{e}^{s\cos\theta^{(ik)}_{\boldsymbol{x},\mathrm{w}} - s\cos\theta_t - \log(N-1)}) + (N-1)\log(N-1)\big] \tag{s13}$$

$$= \frac{1}{N}\big[\log(1 + \mathrm{e}^{-s\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}} + s\cos\theta_t + \log(N-1)}) + \sum_{j\neq i} \log(1 + \mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\mathrm{w}} - (s\cos\theta_t + \log(N-1))}) + (N-1)\log(N-1)\big]. \tag{s14}$$

We define the **UCE** loss $L_{\text{uce}}(\boldsymbol{X}^{(i)})$ as

$$L_{\text{uce}}(\boldsymbol{X}^{(i)}) = \log(1 + \mathrm{e}^{-s\cos\theta^{(ii)}_{\boldsymbol{x},\mathrm{w}} + \tilde{b}}) + \sum_{\substack{j\neq i\\ j=1}}^{N} \log(1 + \mathrm{e}^{s\cos\theta^{(ij)}_{\boldsymbol{x},\mathrm{w}} - \tilde{b}}), \tag{s15}$$

where $\tilde{b} = s\cos\theta_t + \log(N-1)$ is a constant to be learned.

B. Details of MFR Ongoing Testset

For evaluation of face verification performance, we adopt the ongoing online testing of ICCV-2021 Masked Face Recognition Challenge (MFR Ongoing)[1]. The MFR ongoing testing protocol contains not only several popular testsets, including LFW [3], CFP-FP [6], AgeDB [5], and IJB-C [4], but also its own testsets such as the Mask set, Children set, and Multi-Racial set (MR-All, containing 4 different racial faces: African, South Asian, East Asian, and Caucasian).

The Mask set contains 13.9K positive pairs and 96.9M negative pairs (6.9K masked images and 13.9K non-masked images) of 6.9K identities. The Children set contains 157K images (in total 1.7M positive pairs and 24.7B negative pairs) of 14K identities. The Multi-Racial sets contain 1.6M images (in total 4.6M positive pairs and 2.6T negative pairs) of 242K identities.

C. Training Details

This section describes the detailed hyper-parameters used in training the face models on each dataset. Following [2], we use the customized ResNets as our backbone. All models are implemented using Pytorch and trained with the SGD optimizer (5e-4 weight decay and 0.9 momentum). Following [7, 2], the feature norm s in our UCE loss is fixated at 64 in all experiments.

CASIA-WebFace. The face models (using ResNet-50 as the backbone) are trained for 28 epochs with batch size 512 on the CASIA-WebFace, the learning rate is initialized to be 0.1 and decreased by a factor of 10 at the 16^{th} and 24^{th} epoch. In Ablation study, the margin m of $L_{\text{uce-m}}$, $L_{\text{uce-mb-}\lambda}$ and $L_{\text{uce-mb-}r}$ are set to be 0.45. The re-weighting λ of $L_{\text{uce-mb-}\lambda}$ is set to be 0.6, and sampling rate r of $L_{\text{uce-mb-}r}$ is set to be 0.5. In comparisons between different methods on MegaFace Challenge 1, we use m=0.4 and λ=0.7 for $L_{\text{uce-mb-}\lambda}$.

WebFace4M. We train the models (ResNet-50 as the backbone) for 20 epochs with batch size 1024, the learning rate is initialized as 0.1 and a polynomial decay (power=2) strategy is employed for the learning rate schedule. The final margin m in all the three losses (i.e., $L_{\text{uce-m}}$, $L_{\text{uce-mb-}\lambda}$ and $L_{\text{uce-mb-}r}$) are set to be 0.4. The re-weighting parameter λ of $L_{\text{uce-mb-}\lambda}$ is set to be 0.7, and sampling rate r of $L_{\text{uce-mb-}r}$ is set to be 0.4.

Glint360K. We train the models (ResNet-100) for 20 epochs with batch size 1024, the learning rate is initialized as 0.1 and a polynomial decay (power=2) strategy is employed for the learning rate schedule. We use m=0.4 and λ=0.7 for $L_{\text{uce-mb-}\lambda}$ on MegaFace Challenge 1.

WebFace42M. We train the models (ResNet-200) for 20 epochs with a batch size of 4096, we linearly warm up the learning rate from 0 to 0.4 within the first epoch. We then employ polynomial decay (power=2) for the rest 19 epochs. The final margin m of $L_{\text{uce-mb-}\lambda}$ and $L_{\text{uce-mb-}r}$ are set to be 0.4. The re-weighting parameter λ of $L_{\text{uce-mb-}\lambda}$ is set to be 0.7, and sampling rate r of $L_{\text{uce-mb-}r}$ is set to be 0.4.

D. Parameter Study

Though the proposed UCE loss does not contain any hyper-parameters, the improved marginal and balanced UCE losses introduce new parameters. We investigate the impact of different parameters of the two variants below.

Robustness Against Different Margins. We first study the impact of different margins of the marginal UCE loss. In A, when increasing the m from 0.2 to 0.6 with an interval of 0.05, the average performance of the original marginal

softmax loss and that of the Exclusive Regularization loss [8] first improves and then rapidly drops. For our marginal UCE loss, however, the performance is stably increased. To help clarify the differences between the three methods, we plot the changes in the average performance with increasing m in the left sub-figure of Fig. A, which suggests that our UCE loss is more robust and less sensitive to larger margins, while both the original marginal loss and the Exclusive Regularization loss [8] are not.

Effects of Different Balance Strategies. We then study different hyper-parameters for the proposed balanced UCE loss. We have two alternative ways to balance the proposed UCE loss, i.e., re-weight all the negative sample-to-class losses with λ or randomly sample the negative sample-to-class losses with a ratio of $r \times 100\%$. We examine different λ and r from 0.1 to 1.0 in B. It shows that proper adjustment of these parameters can improve the final performance, while a too-small value can lead to performance drops. To display the difference more clearly, we also plot the average results in A (the right sub-figure). It suggests that, with proper parameters, the balanced UCE loss can further improve the performance of the marginal UCE loss. Experimentally, the sampling strategy is better than the re-weighting strategy when r and λ are small, otherwise, the re-weighting strategy is better than the sampling strategy.

Method	m	MR-All	IJB-C	LFW	CFP-FP	AgeDB
w/o UCE	0.2	38.22	84.67	99.18	96.25	93.41
	0.25	38.48	**85.75**	99.31	96.75	93.70
	0.3	41.66	85.60	99.41	96.87	94.11
	0.35	43.34	79.78	99.36	96.60	94.53
	0.4	**46.66**	59.04	99.43	96.82	**95.28**
	0.45	41.80	46.17	**99.50**	96.91	95.11
	0.5	31.76	36.85	99.46	**97.15**	94.86
	0.55	39.05	40.63	99.38	96.92	94.90
	0.6	31.67	33.39	99.45	96.84	95.21
w/o UCE + Regularization	0.2	34.20	84.13	99.33	96.18	93.65
	0.25	37.20	**86.39**	99.51	96.54	94.13
	0.3	37.00	86.29	99.50	96.58	94.38
	0.35	44.57	85.84	99.36	96.70	94.61
	0.4	45.27	78.21	99.25	**97.22**	94.98
	0.45	45.13	66.46	99.46	96.90	95.06
	0.5	**45.82**	53.10	**99.55**	96.88	**95.25**
	0.55	43.71	54.46	99.38	96.78	95.13
	0.6	35.69	38.73	99.48	96.60	94.80
with UCE	0.2	35.08	86.46	99.35	96.55	94.43
	0.25	41.17	87.31	99.41	96.52	94.30
	0.3	43.41	88.10	99.43	96.80	94.60
	0.35	44.45	88.51	99.48	97.17	94.78
	0.4	45.83	**88.97**	99.50	97.15	**95.20**
	0.45	**47.45**	88.65	**99.56**	**97.24**	94.71
	0.5	47.20	88.52	99.55	96.92	95.08
	0.55	46.48	88.57	99.50	97.00	94.61
	0.6	45.50	88.54	99.45	96.84	94.75

Table A. Impacts of different m.

E. Comparisons between BCE and UCE.

The binary cross entropy (BCE) loss is

$$L_{\text{bce}}(\boldsymbol{X}^{(i)}) = \log(1 + \mathrm{e}^{-(W_i^T \boldsymbol{x}^{(i)} + b_i)})$$

Method	λ	r	MR-All	IJB-C	LFW	CFP-FP	AgeDB
re-weighting	0.1	1.0	40.87	88.43	99.41	97.31	94.90
	0.2		42.42	88.42	99.51	96.97	94.61
	0.3		43.45	88.63	99.48	97.30	94.78
	0.4		45.67	88.75	99.50	96.97	95.00
	0.5		**48.75**	**89.03**	99.50	97.28	94.88
	0.6		48.54	88.96	99.55	**97.47**	**95.36**
	0.7		48.30	89.00	99.55	97.14	94.96
	0.8		47.41	88.88	99.43	97.12	95.03
	0.9		47.38	88.57	99.51	97.02	95.26
	1.0		47.45	88.65	**99.56**	97.24	94.71
sampling	1.0	0.1	44.91	88.03	99.36	97.20	95.03
		0.2	45.83	88.58	99.55	97.22	95.03
		0.3	48.01	88.47	99.40	97.27	94.95
		0.4	47.57	**88.99**	99.41	96.94	95.15
		0.5	**48.72**	88.94	99.30	97.20	94.95
		0.6	48.16	88.75	99.40	**97.30**	**95.28**
		0.7	47.85	88.61	99.46	97.24	95.21
		0.8	48.42	88.35	99.53	96.90	95.25
		0.9	46.79	88.92	99.43	97.08	95.26
		1.0	47.45	88.65	**99.56**	97.24	94.71

Table B. Different λ and r with m=0.45.

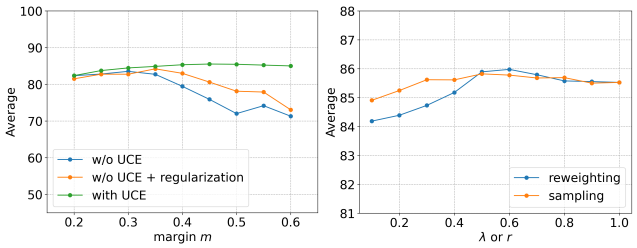

Figure A. Left: impacts of different m of the compared losses, our marginal UCE loss stably improves the performance with larger m. Right: impacts of different λ and r of our balanced UCE loss on the average results of MFR ongoing testset.

$$+ \sum_{\substack{j \neq i \\ j=1}}^{N} \log(1 + \mathrm{e}^{W_j^T \boldsymbol{x}^{(i)} + b_j}) \tag{s16}$$

$$= \log\left(1 + \mathrm{e}^{-(s \cos \theta_{\boldsymbol{x},\mathrm{w}}^{(ii)} + b_i)}\right) + \sum_{\substack{j \neq i \\ j=1}}^{N} \log\left(1 + \mathrm{e}^{s \cos \theta_{\boldsymbol{x},\mathrm{w}}^{(ij)} + b_j}\right). \tag{s17}$$

Though the UCE loss (Eq. (s15)) is similar to BCE loss in Eq. (s17), there are several key differencs between them. Firstly, UCE loss is designed from the objective of an explicit unified threshold t to constrain the similarity of both positive and negative sample-to-class pairs, while BCE loss does not have such explicit constraints. Secondly, we derive the UCE loss from softmax loss, and we present the relationship between the unified threshold t and bias $\tilde{b} = s \cos \theta_t + \log(N-1)$ with a clear mathematical derivation, we then evaluate that the t is in line with the expectation of face verification with a qualitative illustration in Fig. 3 (c). Lastly, we systematically compare the UCE loss and BCE loss on a large benchmark dataset, where we compare (1) a standard BCE loss assigning respective biases for different classes (in Table C), and (2) a simple modification

$m=0$	MR-All	IJB-C	$m=0.45$	MR-All	IJB-C
BCE ($\boldsymbol{b}=0$)	15.67	0.23	BCE ($\boldsymbol{b}=0$)	42.61	79.81
BCE ($\boldsymbol{b}$)	18.91	69.69	BCE ($\boldsymbol{b}$)	45.35	83.88
UCE	**19.59**	**74.80**	UCE	**47.45**	**88.65**

Table C. BCE (w/wo $\boldsymbol{b}$) achieve lower performance than UCE.

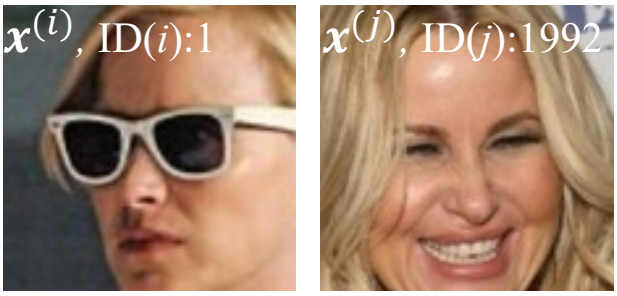

Figure B. Image & ID

of BCE loss excluding any biases, implying bias $b=0$ (in Supplementary). The experimental results show continuous improvements by UCE loss over the two naive variants of BCE loss.

Table C compares the two different settings of BCE loss, with 1) respective $\boldsymbol{b}$s and 2) a constant zero bias $\boldsymbol{b}=0$. The results (trained on CASIA-WebFace, testing on MFR ongoing) show that our UCE loss outperforms the two BCE losses with large margins.

F. Discussion

1. An example to explain the motivation.

As depicted in Fig. 2(a), we claimed that for a correctly classified sample/feature $\boldsymbol{x}^{(i)}$ using a model trained with sample-to-class loss (such as softmax L_{sl}), its similarity with its class proxy W_i might be smaller than the similarity between W_i and a sample $\boldsymbol{x}^{(j)}$ from a different identity. We believe this issue generally exists, even when all samples are correctly labeled.

Fig. B presents two such images. After training, the similarity between $\boldsymbol{x}^{(i)}$ and its own proxy W_i is the highest (0.3162) among all class proxies, i.e., $\boldsymbol{x}^{(i)}$ is correctly classified, while there is a negative sample $\boldsymbol{x}^{(j)}$ whose similarity with W_i (0.3191) is even higher.

In this example, both the two images are correctly labeled. We did statistics on the samples correctly classified by ResNet50 trained with L_{sl} in Fig. 4(a), which achieved 99.92% classification accuracy, and found that 7.28% of these correctly classified samples have similar problems.

2. Comparing the marginal UCE and margin-penalty-based softmax losses.

The core idea of our UCE L_{uce} is incorporating the unified threshold t, instead of directly imposing margin-penalty on the distance/similarity of positive or negative sample pairs. Built on L_{uce}, the marginal UCE $L_{\text{uce-m}}$ introduces a simple cosine margin, which is similar to CosFace, but different from the angular margin in ArcFace. CurricularFace and ArcNegFace, on the other hand, adopt more sophisticated margin-penalty strategies (angular margin with adaptive curriculum learning and hard negative mining).

	m	MR-A	IJBC4	IJBC5
CosFace	0.35	43.34	79.78	38.82
CosFace	0.4	46.66	59.04	13.73
CosFace	0.5	31.76	36.85	4.11
ArcFace	0.35	42.88	84.82	61.31
ArcFace	0.4	42.06	76.24	33.26
ArcFace	0.5	45.59	60.31	17.20

Table D. CosFace vs ArcFace under different m.

Moreover, as depicted in Fig. 1, the marginal softmax losses (CosFace, ArcFace, CurricularFace) are imposed on each individual sample to ensure a margin between the positive and negative sample-to-class pairs, for this particular sample only, while our t in UCE aims to achieve such a separability for all samples.

The UCE is different from other adaptive margin-penalty softmax loss, in both motivations and methodology and can thus achieve global separability between all positive and negative pairs, across all samples.

3. Complexity of UCE's formula.

Our L_{uce} has two terms for incorporating the unified threshold t and hence is slightly more complex than softmax loss, it however only increases $N-1$ logarithmic operations ($\mathcal{O}(N)$) and remains efficient.

4. Sensitivity of hyperparameters.

Our marginal and balanced UCE ($L_{\text{uce-m}}$ and $L_{\text{uce-mb}}$) need hyperparameter tuning, and so does the SOTA CosFace, ArcFace, and Partial FC. We have shown that our $L_{\text{uce-m}}$ and $L_{\text{uce-mb}}$ are robust to the changes of the margin in Fig. 3.

5. Margins of ArcFace and CosFace in Table 3.

Firstly, their incorporated margins are different. ArcFace uses the angular margin, while CosFace uses the cosine margin. Secondly, the performance varies with the values of m, the results in Table 3 are obtained using the recommended values (0.35 for CosFace and 0.5 for ArcFace) from their papers. Table D lists more results for them under different m, which also suggests that the marginal softmax losses are sensitive to the changes of m, as shown in Fig. 3(a).

6. Changes of t during training.

As shown in Fig. C, the t first rapidly increased to 0.3108 from the random initial value 0.2061, it then gradually decreased to 0.2893 and finally stabilized at 0.2928 after a fluctuation caused by the changes of learning rate at the 16^{th} epoch.

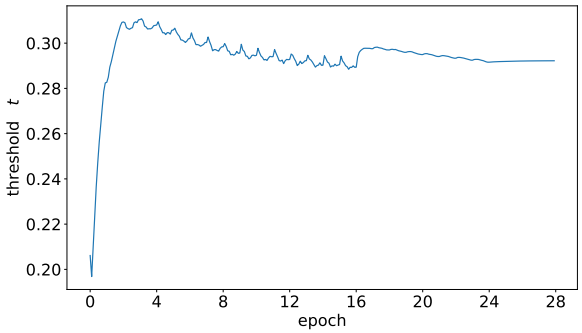

Figure C. t.

References

[1] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021.

[2] 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.

[3] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[4] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[5] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[6] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[7] 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.

[8] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.