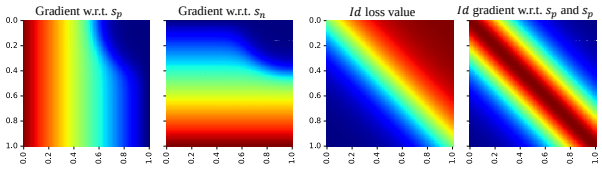
Open-Set Biometrics: Beyond Good Closed-Set Models — Supplementary Materials

Yiyang Su, Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu

Michigan State University, East Lansing, MI 48824, USA
{suyiyan1, kimminc2, liufeng6, jain, liuxm}@msu.edu

6 Comparison with Other Losses.

Id differs significantly from Circle Loss, MagFace, and InfoNCE, which push all genuine scores to 1 and imposter scores to 0. Id specifically targets hard-to-classify pairs, enhancing open-set identification by focusing on closely matched genuine and imposter scores. These approaches are compared in Fig. 7, emphasizing that, unlike existing loss, Id is well aligned with open-set challenges.



(a) Circle loss. **(b)** Id (Ours).

Fig. 7: Comparison between Id and Circle loss. The gradient of Id w.r.t. s_p and s_n have the same magnitude and opposite directions. [The color scale is the same as Fig. 4; x-axes: positive score s_p; y-axes: negative score s_n]

7 Additional Experiments

7.1 Implementation Details

For face recognition and person reID, we use the official implementation of AdaFace[1] and CAL[2], respectively. For gait recognition, we use the official implementation of GaitBase in OpenGait[3] and our own implementation of SwinGait [1] based on OpenGait. Tab. 7 lists the implementation details and hyperparameters in our experiments. Note that to ensure fair comparison in our experiments, we adhere to the original design of our baseline models. Specifically, face and person reID models (AdaFace [6] and CAL [3]) use cosine similarity, while gait models (GaitBase [2] and SwinGait [1]) use Euclidean distance.

[1] https://github.com/mk-minchul/AdaFace
[2] https://github.com/guxinqian/Simple-CCReID
[3] https://github.com/ShiqiYu/OpenGait

Table 7: Hyperparameters in our experiments. [Keys: MSLR=multi-step learning rate scheduler; SGDR=cosine annealing learning rate scheduler [7]; Eps=Epochs; Its=Iterations]

Module	Description	Symbol	AdaFace	GaitBase		SwinGait		CAL	
			IJB-S	Gait3D	GREW	Gait3D	GREW	MEVID	CCVID
Data	Input sequence length	—	1	10-50	20-40	10-50	20-40	8	8
	Input imagery height	—	112	64	64	64	64	256	256
	Input imagery width	—	112	44	44	44	44	128	128
Training	Batch Size	—	1024	128	128	128	128	32	32
	Number of subjects per batch	—	512	32	32	32	32	8	8
	Number of media per subject	—	2	4	4	4	4	4	4
	Optimizer	—	SGD	SGD	SGD	AdamW	AdamW	Adam	Adam
	Momentum	—	0.9	0.9	0.9	—	—	0.9	0.9
	Learning rate	—	0.1	0.1	0.1	3×10^{-4}	3×10^{-4}	3.5×10^{-4}	3.5×10^{-4}
	Weight decay	—	5×10^{-4}	5×10^{-4}	5×10^{-4}	2×10^{-2}	2×10^{-2}	5×10^{-4}	5×10^{-4}
	Learning rate scheduler	—	MSLR	MSLR	MSLR	SGDR	SGDR	MSLR	MSLR
	Learning rate milestones 1	—	12 Eps	20,000 Its	80,000 Its	—	—	40 Eps	40 Eps
	Learning rate milestones 2	—	20 Eps	40,000 Its	120,000 Its	—	—	80 Eps	80 Eps
	Learning rate milestones 3	—	24 Eps	50,000 Its	150,000 Its	—	—	120 Eps	120 Eps
	Half cosine cycle	—	—	—	—	60,000	150,000 Its	—	
Ours	Sigmoid temperature	α	6.0	6.0	9.0	6.0	9.0	6.0	6.0
	Sigmoid temperature	β	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	Sigmoid temperature	γ	6.0	6.0	9.0	6.0	9.0	6.0	6.0
	Ratio of non-mated subjects	p	25%	25%	25%	25%	25%	25%	25%
	RTM loss weight	λ	4.0	4.0	4.0	4.0	4.0	4.0	4.0

Table 8: Closed- and open-set gait recognition on Gait3D [8].

Model	Rank-1 Accuracy ↑	FNIR@1%FPIR ↓
SwinGait [1]	72.4	83.50±6.20
+Ours	75.3	79.30±6.74
+EVM [4]	73.5	81.08±8.00

7.2 Comparison with EVM

We integrate EVM into the SwinGait framework. As shown in Tab. 8, our approach surpasses the performance of EVM. It's noteworthy that EVM operates on extracted features rather than raw images or videos as input. Our method focuses on enhancing feature representation, while EVM enhances the score calculation process.

7.3 Additional Ablation Experiments

We conduct additional ablation experiments using ArcFace with a ResNet-18 backbone on subsets of the TinyFace and IJB-S datasets. The results, detailed in Tab. 9, validate the effectiveness of both IDL and RTM. These findings showcase our method's adaptability and robust performance across different baselines and biometric scenarios.

Table 9: Closed- and open-set face recognition performance. [Keys: R@1=rank-1 accuracy; 1%: FNIR@1%FPIR]

Model	TinyFace	Surv2Single		Surv2Book	
	R@1↑	R@1↑	1%↓	R@1↑	1%↓
ArcFace	64.22	53.73	68.19	55.34	68.37
ArcFace + IDL	64.27	53.81	67.23	55.29	66.99
ArcFace + RTM	64.67	53.84	67.68	55.82	**66.16**
ArcFace + IDL + RTM (Ours)	**64.70**	**55.26**	**66.70**	**56.51**	66.86

Table 10: Open-set evaluation of SwinGait-3D and SwinGait-3D with our loss functions. [Keys: Random=random sequences as gallery; Mean=mean feature as gallery]

Model	R	Gallery	FNIR@1%FPIR
SwinGait-3D [1]	20	Random	86.56±6.11
	1	Mean	83.50±6.19
	20	Mean	83.50±6.20
SwinGait-3D [1]+Ours	20	Random	82.87±6.72
	1	Mean	79.30±6.71
	20	Mean	79.30±6.74

7.4 Analysis on FNIR@FPIR

Effect of R. With smaller R, more mated probes become FNs due to failed identification. Yet, as shown in Tab. 10, the hyperparameter R has a marginal effect on FNIR@1%FPIR. The This suggests that the bottleneck in open-set biometrics is detection at high thresholds.

Evaluation Protocol. In gait recognition, each subject often possesses multiple gallery sequences. There are different strategies for electing a representative feature from these sequences for each gallery subject. As discussed in Sec. 4.1, we take the mean feature for each subject in the gallery in our experiments. To validate this design choice, we compare the open-set FNIR@1%FPIR on the Gait3D [8] dataset between the mean gallery feature and the gallery feature of a randomly chosen gait sequence. The results are presented in Tab. 10. That better results are observed with the mean feature indicates that the mean gallery features are more robust for open-set recognition when available.

Effect of q%. In evaluating open-set performance, q% represents the ratio of probe subjects that are not in the gallery set (Sec. 4.1). To compare the performance of SwinGait and SwinGait trained with our loss functions under different values of q%, we create 50 test splits. In each split, we randomly select 140 test subjects as the gallery set and randomly select 35, 90, 240, 360, 610 subjects from the remaining subjects as non-mated probes, corresponding to

Table 11: Open-set evaluation of SwinGait-3D [1] and SwinGait-3D with our loss functions. [Keys: Gallery Subj.=number of gallery subjects; NM Probe Subj.=number of non-mated probe subjects]

Gallery Subj.	NM Probe Subj.	q%	SwinGait	SwinGait+Ours
140	35	20%	60.36±11.83	57.50±10.40
140	90	40%	69.29±10.94	67.50±11.07
140	140	50%	72.73±9.87	71.64±10.41
140	240	60%	70.00±7.99	68.57±7.70
140	360	70%	69.29±6.27	67.14±6.62
140	610	80%	67.50±4.62	66.07±5.43

different q% values. Since these evaluation protocols use the same set of gallery subjects in each split, they are comparable to each other.

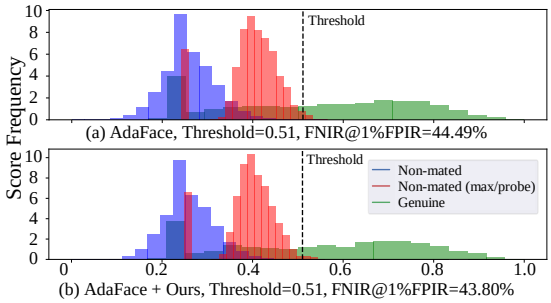
We present the detailed results of our analysis in Tab. 11. Notably, the gallery set size is relatively small, comprising only 140 subjects, yet the obtained results generally surpass those depicted in Table 4. Our observations reveal that despite training with only $p = 25\%$, our proposed approach consistently outperforms existing methods across various thresholds of q%. Moreover, we notice that as the ratio of non-mated subjects increases, the performance of both models initially decreases before showing improvement, indicating that very low q% values and very high q% values pose more challenging open-set scenarios.

7.5 Additional Visualizations

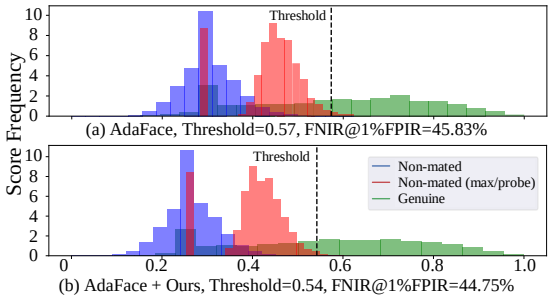
More Score Distributions. We visualize the genuine and non-mated score distributions of AdaFace [6] with and without our loss functions in Fig. 8. The official evaluation protocol of IJB-S [5] provides gallery 1 and gallery 2 for open-set evaluation. For gallery 1, our loss functions yield a comparable threshold but are able to reduce FNIR@1%FPIR by improving detection. For gallery 2, our loss functions significantly reduce the threshold, which leads to improved detection as well.

More FN Analyses. Similar to Fig 6, we visualize the face recognition FN breakdown on IJB-S [5] at different levels of FPIR in Fig. 9. At low FPIRs (Fig. 9a-Fig. 9c), detection is significantly harder than identification as almost all FNs have detection failures and a portion of FNs have both detection and identification failures.

Only at very high FPIR (*e.g.*, Fig. 9e and Fig. 9f) do a significant portion of FNs have only identification failures. The major failure shifts from detection to identification between 85% FPIR (Fig. 9d) and 90% FPIR (Fig. 9f). However, it is noteworthy that biometrics systems operating at such FPIRs are impractical for real-world use, as they lack the ability to effectively reject unknown subjects.



(a) Score distribution on IJB-S [5] gallery 1.



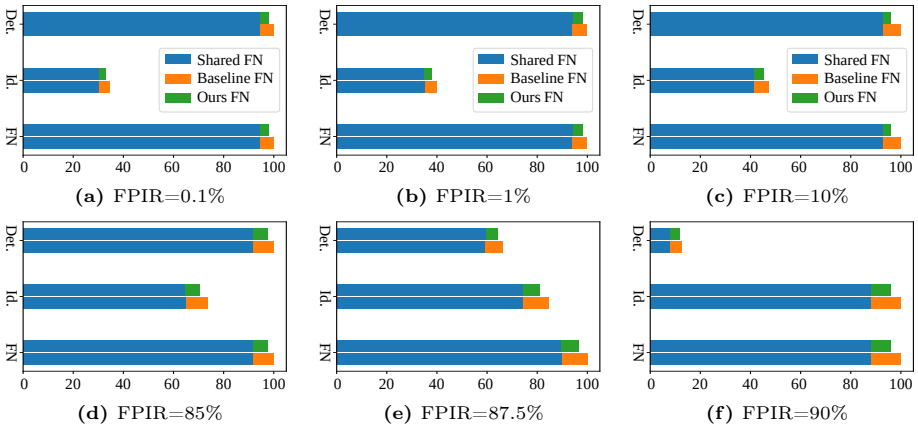
(b) Score distribution on IJB-S [5] gallery 2.

Fig. 8: Face recognition and person reID score distributions.

Exemplars. We visualize the detections of two mated and two non-mated probes in Fig. 10. This illustrates that our loss functions can boost both the correct detection of mated subjects and the rejection of non-mated subjects, both of which translate directly to better open-set performance.

References

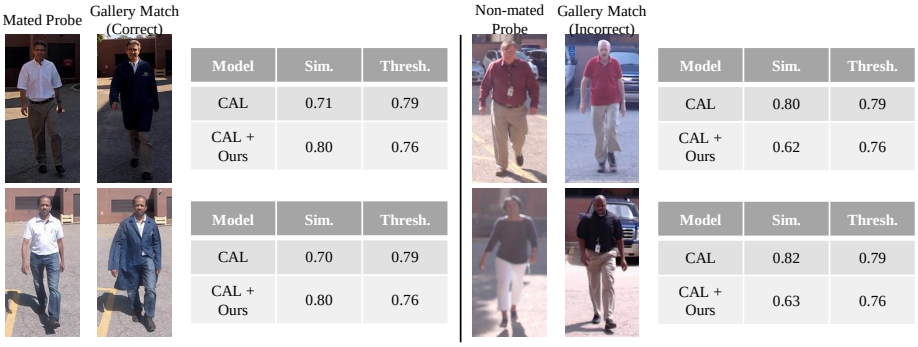
1. Fan, C., Hou, S., Huang, Y., Yu, S.: Exploring deep models for practical gait recognition. arXiv preprint arXiv:2303.03301 (2023)
2. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: CVPR (2023)
3. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: CVPR (2022)
4. Gunther, M., Cruz, S., Rudd, E.M., Boult, T.E.: Toward open-set face recognition. In: CVPR Workshops (2017)
5. Kalka, N.D., Maze, B., Duncan, J.A., O'Connor, K., Elliott, S., Hebert, K., Bryan, J., Jain, A.K.: IJB-S: IARPA janus surveillance video benchmark. In: BTAS (2018)
6. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: CVPR (2022)
7. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)



(a) FPIR=0.1% **(b)** FPIR=1% **(c)** FPIR=10%

(d) FPIR=85% **(e)** FPIR=87.5% **(f)** FPIR=90%

Fig. 9: Open-set FNs due to detection, identification, and either. Blue indicates FNs that are shared by the baseline, AdaFace [6], and our approach. Orange indicates the FNs of the baseline only. Green indicates the FNs of our approach.



Model	Sim.	Thresh.
CAL	0.71	0.79
CAL + Ours	0.80	0.76

Model	Sim.	Thresh.
CAL	0.80	0.79
CAL + Ours	0.62	0.76

Model	Sim.	Thresh.
CAL	0.70	0.79
CAL + Ours	0.80	0.76

Model	Sim.	Thresh.
CAL	0.82	0.79
CAL + Ours	0.63	0.76

Fig. 10: Four probe sequences in CCVID [3]. On the left, two mated probes are shown. They are incorrectly rejected by CAL [3] but correctly detected with our approach. Two non-mated are shown on the right. They failed to be rejected by CAL, which is corrected by our loss functions.

8. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., Mei, T.: Gait recognition in the wild with dense 3d representations and a benchmark. In: CVPR (2022)