

Open-Set Biometrics: Beyond Good Closed-Set Models

Yiyang Su¹, Minchul Kim¹, Feng Liu¹, Anil Jain, and Xiaoming Liu¹

Michigan State University, East Lansing, MI 48824, USA
{suyiyang1, kimminc2, liufeng6, jain, liuxm}@msu.edu

Abstract. Biometric recognition has primarily addressed closed-set identification, assuming all probe subjects are in the gallery. However, most practical applications involve open-set biometrics, where probe subjects may or may not be present in the gallery. This poses distinct challenges in effectively distinguishing individuals in the gallery while minimizing false detections. While it is commonly believed that powerful biometric models can excel in both closed- and open-set scenarios, existing loss functions are inconsistent with open-set evaluation. They treat genuine (mated) and imposter (non-mated) similarity scores symmetrically and neglect the relative magnitudes of imposter scores. To address these issues, we simulate open-set evaluation using minibatches during training and introduce novel loss functions: (1) the *identification-detection* loss optimized for open-set performance under selective thresholds and (2) *relative threshold minimization* to reduce the maximum negative score for each probe. Across diverse biometric tasks, including face recognition, gait recognition, and person re-identification, our experiments demonstrate the effectiveness of the proposed loss functions, significantly enhancing open-set performance while positively impacting closed-set performance. Our code and models are available here.

Keywords: Open-set biometrics · Face recognition · Gait recognition · Person reID

1 Introduction

Biometrics, the study of unique physiological or behavioral characteristics for human identification, plays a pivotal role in applications such as access control, surveillance, and forensic analysis [52]. Each biometric task, from face recognition, gait recognition to person re-identification (reID), uses features tailored to its specific requirements [39, 55, 71].

In real-world applications, after test subjects are enrolled in the gallery, biometric systems inevitably encounter *non-mated* subjects, those not present in the gallery. Thus, it becomes imperative for such systems to demonstrate robust performance in the *open-set* framework with non-mated probes [16, 18, 28, 58]. The *closed-set* framework assumes all probe subjects are mated and predominantly focuses on *identification*, associating probes with the closest match in the

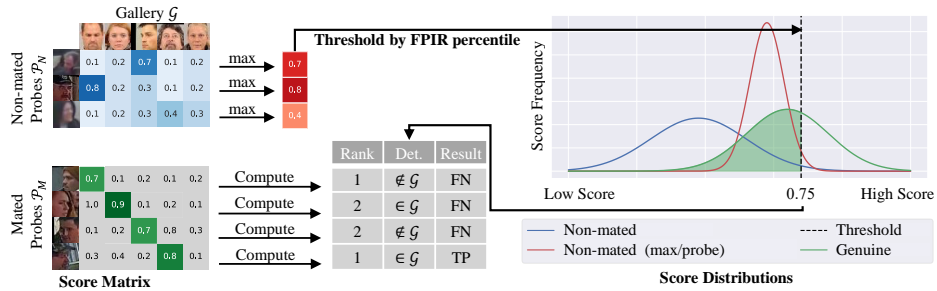


Fig. 1: A toy example for FNIR@FPIR calculation. The score matrix and distribution between non-mated probes and all gallery subjects are shown in blue. The maximum score for each non-mated probe is shown in red. A percentile of their distribution determines the threshold. A user-defined percentile (*e.g.*, 33%) sets the threshold to be 0.75 as 1 out of 3 (the 2nd) non-mated probes is incorrectly classified as “in the gallery”. The genuine scores of mated probes are shown in green. A mated probe is an FN if either of the conditions is met: (1) in detection (Det.), it is incorrectly classified as “not in the gallery” (\mathcal{G}) as the genuine score is below the threshold (the 1st mated probe); (2) its genuine score is not within a specified rank R , *e.g.*, 1, in the gallery (the 2nd mated probe). Hence, FNIR@33%FPIR is 75%.

gallery. However, the open-set framework introduces an additional challenge: *detection*, determining whether a probe subject is mated. For example, in airport surveillance, faces in CCTV videos are compared against a criminal watch list. The goal is to reduce the misidentification (false positives) of non-mated subjects such as airport staff and innocent passengers (improving detection), while simultaneously ensuring accurate recognition of known criminals (enhancing identification). For this reason, open-set biometrics is a harder and more general task than closed-set.

The biometric community has predominantly centered on the closed-set framework. While some face recognition benchmarks have open-set evaluation protocols [18, 19, 28], they are primarily used for evaluation rather than for developing tailored strategies to address the open-set challenges. In gait recognition and person reID, most benchmarks only provide official closed-set protocols, lacking open-set ones. To address this gap, we take the initiative to introduce open-set benchmarks and propose strategies to train models specifically for improving open-set performances while maintaining closed-set performance.

The false negative identification rate at a given false positive identification rate (FNIR@FPIR) is a *de facto* open-set metric [16]. We first compute the similarity score matrix between non-mated probes and the gallery, from which we extract the maximum score for each non-mated probe across the gallery, forming the maximum-per-probe non-mated score distribution. This distribution represents the likelihood that a non-mated probe has a match in the gallery. A threshold is set based on a user-defined percentile (*e.g.*, 1%) of this distribution. Mated probes with scores below this threshold are counted as false negatives (FNs), as they are incorrectly considered not to be in the gallery. Similar to the

closed-set scenario, a mated probe is also an FN if its gallery match doesn't rank among the gallery set within a user-specified rank R . FNIR is the proportion of mated probes classified as FNs. Fig. 1 depicts its calculation pipeline.

Unlike closed-set or verification metrics, FNIR@FPIR introduces unique challenges. It employs a threshold based on maximum scores for each non-mated probe, creating an asymmetry between genuine and imposter scores. SOTA biometric recognition methods achieve strong performance in closed-set scenarios leveraging triplet loss [20] and/or softmax cross-entropy based losses [9, 25, 29, 46, 60], both of which excel in reducing intra-person variance and increasing inter-person variance. However, they do not take into account the fact that the non-mated score distribution determines the threshold for the false negative.

In other words, conventional loss functions, while effective in closed-set scenarios, exhibit limitations in addressing the nuances of the FNIR@FPIR metric, which necessitate a handling of the maximum of imposter scores. These high-scoring non-mated probes disproportionately influence the threshold setting for false negatives, yet conventional loss functions often treat them equivalently to other scores. As a result, they face challenges in open-set scenarios. GaitBase [13], a SOTA gait recognition model, can only achieve a 61% FNIR@1%FPIR on the training set and 86% on the test set. This challenge extends to face recognition and person reID.

To address this challenge, we propose to create simulated open-set evaluation episodes during training, whose score distributions serve as proxies to those of the test set. Specifically, we introduce an *identification-detection* loss tailored to optimize the two core elements in open-set biometrics, *i.e.*, detection and identification. For detection, the model is trained to accurately pinpoint the exact probes whose mate is present in the gallery set; for identification, the model learns to establish associations between the gallery and probe samples of mated subjects. By integrating the aspects of detection and identification within training batches, we improve the model's efficacy in both domains, thus yielding enhanced testing performance. Further, we minimize similarity scores between non-mated subjects and the gallery, enabling a more comprehensive detection of mated subjects (see Fig. 3).

Empirically, we apply our proposed loss function to three biometric modalities, face, gait, and person reID, and show performance improvement in all tasks. While our focus is open-set biometrics, we also observe performance gain in closed-set search due to improved feature representations.

To summarize, our contributions include:

- ✓ We show that the existing loss functions, *e.g.*, triplet loss and softmax loss, do not treat genuine and imposter training samples in accordance with open-set evaluation. This inconsistency sheds light on why the open-set challenges have not been adequately addressed.
- ✓ We propose a novel loss function that leverages open-set mini-batches during training to optimize identification and detection performance and minimize open-set detection thresholds.

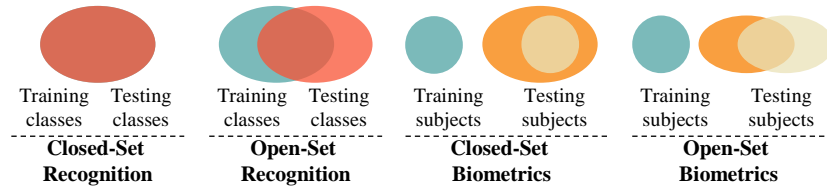


Fig. 2: Comparison between open-set recognition and biometrics. Closed-set recognition has the same set of classes for training and testing. Open-set recognition involves testing on unseen classes during training. On the other hand, in biometrics, the sets of training and testing subjects are always disjoint. In closed-set biometrics, all probe subjects are in the gallery, and in open-set biometrics, some probe subjects are not enrolled in the gallery. [Legends: ■ gallery subjects, ■ probe subjects]

- ✓ We experimentally demonstrate that our proposed loss function can effectively improve both closed- and open-set performance across diverse biometric modalities, including face, gait, and person reID.

2 Related Work

Open-Set Biometrics. While various biometric recognition tasks, such as face recognition [9, 24, 25, 29–31, 41, 43, 46, 50, 60], gait recognition [3, 4, 7, 10–14, 23, 26, 36–38, 44, 54, 61, 62, 68, 74–77], and person reID [5, 17, 21, 35, 36, 40, 42, 45, 48, 56, 63, 66], have witnessed significant progress, the majority of their focus remains on closed-set scenarios. Although EVM has been applied to open-set face recognition [18], EVM takes extracted features, instead of images/videos, as input. Our approach improves feature representation and EVM improves score calculation. While some face recognition datasets provide open-set evaluation protocols [28], open-set performance is often perceived as a consequence rather than a direct objective in advancing recognition models. Further, even the most recent gait recognition [68] and person reID [42] models are exclusively evaluated under closed-set settings due to the absence of open-set evaluation protocols.

Open-Set Recognition. While open-set biometrics is relatively underexplored, open-set recognition [1, 2, 6, 22, 32–34, 51, 59, 64, 67, 69, 72, 73, 78] has attracted attention in the literature. Open-set recognition distinguishes between known and unknown classes in image classification. In both open-set image recognition and open-set biometrics, the primary concern is how the models deal with instances that belong to unknown classes or non-mated probes. However, as illustrated in Fig. 2, in open-set recognition known classes come from the training set and the unknown classes refer to classes not used for training [15]. On the other hand, even in closed-set biometrics, all test subjects are typically not used during training and mated probes belong to test subjects enrolled in the gallery [27, 47]. Open-set biometrics differs from closed-set biometrics in that there are non-mated subjects are neither in the training set nor in the gallery [16, 18, 58]. In

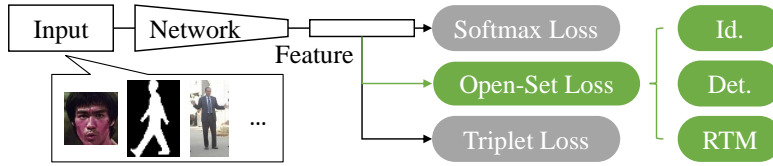


Fig. 3: Overview of our approach. We introduce an additional modality-agnostic loss function specifically designed for open-set biometrics, which simultaneously optimizes identification (Id.) and detection (Det.) and Relative Threshold Minimization (RTM) for open-set thresholds.

evaluation, open-set recognition models typically use closed-set recognition accuracy and AUROC for detecting classes not seen during training [15] and open-set biometrics uses FNIR@FPIR [16].

3 Method

3.1 Problem Formulation

Let $\mathcal{S}_{\text{test}}$ be the test set, $\mathcal{G} \subset \mathcal{S}_{\text{test}}$ represent the gallery set containing mated subjects, and $\mathcal{P} \subset \mathcal{S}_{\text{test}}$ denote the probe set consisting of test samples. For each probe sample $p_i \in \mathcal{P}$, there are two possible outcomes: (1) If p_i is mated, *i.e.*, its corresponding subject g_i is in \mathcal{G} , the biometric model must identify the correct subject. (2) If p_i is non-mated, the model must reject it. The open-set scenario poses a more challenging problem than the traditional closed-set one, as the model needs to recognize mated subjects while rejecting non-mated ones.

3.2 Revisiting Loss Functions for Biometrics

In this section, we compute gradients of two common loss functions in biometrics: softmax and triplet loss. Many loss functions are derived from them and share similar characteristics. For instance, center loss [65] can be analyzed similarly to softmax loss, and the logic of triplet loss can also be applied to circle loss [57].

Softmax loss. The softmax loss, often employed in closed-set classification, operates by computing the cross-entropy between the normalized logit of a sample and its corresponding ground truth class. Mathematically, for a sample i in a batch, the softmax loss $L_{\text{softmax}}(i)$ is:

$$L_{\text{softmax}}(i) = -\log \left(\frac{e^{z_{i,s}}}{\sum_t e^{z_{i,t}}} \right). \quad (1)$$

Here, $z_{i,s}$ denotes the logit for sample i belonging to subject s . This loss aims to minimize the negative log-likelihood of the true subject s for sample i . For one

sample, the gradient of the softmax loss *w.r.t.* the logit $z_{i,s}$ is given by:

$$\frac{\partial L_{\text{softmax}}(i)}{\partial z_{i,s}} = \frac{e^{z_{i,s}}}{\sum_t e^{z_{i,t}}} - \delta_{i,s}, \quad (2)$$

where $\frac{e^{z_{i,s}}}{\sum_{j=1}^N e^{z_{j,s}}}$ represents the probability that sample i belongs to subject s and $\delta_{i,c}$ is the Kronecker delta, which is 1 when $i = c$ (the true subject) and 0 otherwise. We visualize gradients of the softmax loss in Fig. 4(a).

However, it overlooks the relative magnitude of each imposter score within the batch. An imposter score may be deemed small by, *e.g.*, AdaFace [29] standards but still rank high among maximum-per-probe non-mated scores, raising the open-set threshold. Further, the softmax loss assumes that each sample can be assigned to one of the training subjects, which is not the case in open-set settings. Consequently, the softmax loss is not inherently designed to handle open-set biometrics.

Triplet loss. The triplet loss [53] is defined as:

$$L_{\text{tri}}(a, p, n) = \max(0, d(a, p) - d(a, n) + m). \quad (3)$$

Note that $d(a, p)$ is the distance between the anchor a and positive sample p , $d(a, n)$ is between a and the negative sample n , and m is a constant margin. The partial derivatives of $L(a, p, n)$ *w.r.t.* $d(a, p)$ and $d(a, n)$ are piecewise functions:

$$\frac{\partial L_{\text{tri}}(a, p, n)}{\partial d(a, p)} = \begin{cases} 1, & \text{if } d(a, p) + m > d(a, n) \\ 0, & \text{if } d(a, p) + m \leq d(a, n) \end{cases}; \quad (4)$$

$$\frac{\partial L_{\text{tri}}(a, p, n)}{\partial d(a, n)} = \begin{cases} -1, & \text{if } d(a, p) + m > d(a, n) \\ 0, & \text{if } d(a, p) + m \leq d(a, n) \end{cases}. \quad (5)$$

The magnitude of the gradient is constant regardless of the distance between two samples. This inability to differentiate samples in the gradient implies that all negative samples in the triplet are contributing equally, which is not reflective of the open-set scenario. One remedy often used with triplet loss is hard negative mining. However, it comes with a critical drawback of being susceptible to label noise [70]. Noisy and mislabeled samples are often considered hard negatives, which can make the loss function sensitive to noisy data when focusing on the hardest samples. Thus, we believe that a loss function that considers all negative samples while being able to adjust the gradient magnitude based on the anchor-negative distances would be effective in open-set biometrics.

Hence, we propose a novel loss function, shown in Fig. 3. It optimizes identification and detection at various thresholds and minimizes non-mated scores to reduce open-set thresholds. It can be applied to any biometric modality. Unlike existing loss functions, it is consistent with FNIR@FPIR in that it treats genuine and imposter scores asymmetrically and tackles large non-mated scores.

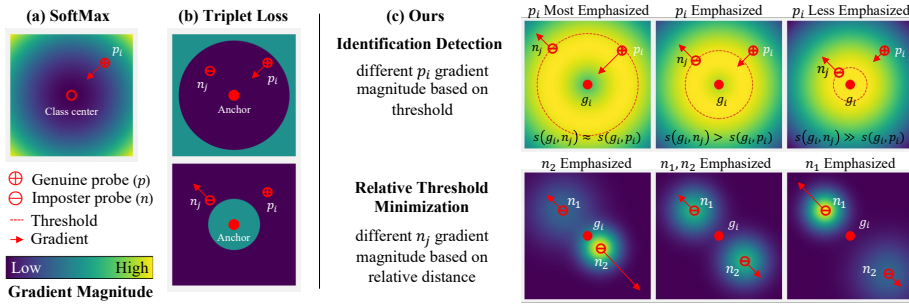


Fig. 4: Visualization of the gradient of different loss functions. The colors represent gradient magnitude and the arrows represent gradient directions. a) SoftMax loss pushes the sample toward the class center and away from the other classes, with gradual shifts in gradient magnitude. b) On the other hand, triplet loss has a hard cut gradient based on triplet samples. c) For S_r^{det} , as shown in the top row, the thresholds are determined by non-mated sample, n_j . p_i 's gradient has the greatest magnitude when it has a similar distance from the gallery g_i to n_j . For Relative Threshold Minimization, as shown in the bottom row, as n_2 moves away from the gallery, its gradient decreases. While n_1 remains at the same location, its gradient increases because it becomes closer to g_i than n_2 . The gradients *w.r.t.* genuine scores adapt to non-mated scores and the gradients *w.r.t.* non-mated scores are adaptive to other non-mated scores.

3.3 Our Approach

To design loss functions tailored for open-set biometrics, we create a training scenario mirroring the testing environment. During training, we partition each batch into gallery and probe sets to simulate open-set evaluation conditions. In this process, we randomly designate a certain percentage, denoted as $p\%$, of the subjects as mated subjects, while the rest are categorized as non-mated. For the mated subjects, we further randomly allocate their exemplars to the gallery and probe sets. Conversely, all exemplars of the non-mated subjects are placed in the probe set. This approach yields a more realistic and representative training scenario, preparing the biometric features to excel under open-set conditions. To avoid any potential notation ambiguity, we refer to the mated probe set created within each training batch as \mathcal{P}'_K , the non-mated probe set as \mathcal{P}'_U , and the gallery sets within each training batch as \mathcal{G}' .

To ensure robustness, when the metric space inhabited by the biometric features is defined under the Euclidean distance, we calculate the similarity scores between two features as:

$$s(p_i, g_j) = \frac{1}{1 + d(p_i, g_j)}, \quad (6)$$

where $d(\cdot, \cdot)$ represents the Euclidean distance. Alternatively, if the metric space is defined under the cosine distance, we simply adopt the cosine similarity.

To ensure that the model can generalize to open-set scenarios during testing, we propose loss functions that penalize the model for three types of errors during

training: (1) failing to detect a mated probe with a threshold τ , (2) failing to identify a mated probe within the top rank- r positions and (3) setting the threshold τ too high, leading to poor differentiation between mated and non-mated sample distributions.

(1) Detection. Detection is assessing whether the computed similarity score exceeds a predefined threshold τ . That is, for a given probe $p_i \in \mathcal{P}'_K$ and its corresponding gallery subject $g_i \in \mathcal{G}'$, we calculate a detection score as:

$$S_\tau^{det}(p_i, g_i) = \sigma_\alpha(s(p_i, g_i) - \tau), \quad (7)$$

where $\sigma_\alpha(x) = 1/(1 + \exp(-\alpha x))$ is the Sigmoid function with temperature hyperparameter α . The $S_\tau^{det}(p_i, g_i)$ score quantifies the model’s ability to detect if p_i and g_i belong to the same subject, given the threshold τ . A high similarity score, well above τ , indicates successful detection and $S_\tau^{det}(p_i, g_i)$ approaches 1. In contrast, if the model fails to detect that p_i and g_i belong to the same subject, $s(p_i, g_i)$ remains much smaller than τ , causing $S_\tau^{det}(p_i, g_i)$ to approach 0. Since the Sigmoid function has the highest gradient near 0, the term $s(p_i, g_i) - \tau$ makes the loss function focus more on the samples near the threshold τ .

During testing, the open-set detection threshold is determined by the non-mated score distribution, which we do not know during training. Thus, we sample τ from the non-mated scores in a batch. Specifically, we let

$$S^{det} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} S_\tau^{det}(p_i, g_i), \quad (8)$$

where $\mathcal{T} = \{s(n_j, g_i) | n_j \in \mathcal{P}'_U\}$ is the set of all non-mated scores. Fig. 4(c) illustrates the gradient of Det *w.r.t.* mated probes under different thresholds induced by various non-mated probes. The gradient is adaptable to different thresholds, increasing as τ decreases, even when $s(g, p)$ is constant. This adaptability is desirable in open-set settings, as it promotes mated probes to be more similar to their ground-truth gallery than non-mated probes. Note that the thresholds in Eq. (8) are only used during training. The detection thresholds during testing are automatically determined by the FNIR@FPIR metric as discussed in Sec. 1.

(2) Identification. It is not sufficient for a biometric model to distinguish mated subjects from non-mated ones based on their scores. Equally important is the model’s ability to accurately identify the correct subject among the gallery subjects. Therefore, we introduce an identification score $S^{id}(p_i, g_i)$ as:

$$S^{id}(p_i, g_i) = \sigma_\beta(1 - \text{sofrank}(p_i, g_i)), \quad (9)$$

where σ_β is a Sigmoid function with temperature β and *sofrank* is a differentiable function with a Sigmoid function σ_γ defined as:

$$\text{sofrank}(p_i, g_i) = \sum_{g_j \in \mathcal{G}'} \sigma_\gamma(s(p_i, g_j) - s(p_i, g_i)), \quad (10)$$

The softrank function accumulates values close to 1 for gallery subjects g_j more similar to the probe p_i than the correct match g_i and close to 0 otherwise. It essentially reflects the rank of the correct match g_i among all gallery subjects for probe p_i . The $S^{id}(p_i, g_i)$ is an indicator, approaching 1 if p_i has a softrank value less than 1, indicating successful identification.

To optimize open-set detection and identification simultaneously, the overall identification-detection loss $\mathcal{L}_{r,\tau}$ is formulated as the product of two components,

$$\mathcal{L}^{\text{IDL}} = -\frac{1}{|\mathcal{P}'_{\text{K}}|} \sum_{p_i \in \mathcal{P}'_{\text{K}}} S^{\text{det}}(p_i, g_i) \cdot S^{\text{id}}(p_i, g_i). \quad (11)$$

In other words, \mathcal{L}^{IDL} is only close to -1 if both $S^{\text{id}}(p_i, g_i)$ and $S^{\text{det}}(p_i, g_i)$ approach 1. Conversely, should either or both not approximate 1, \mathcal{L}^{IDL} converges to 0.

(3) Relative Threshold Minimization \mathcal{L}^{IDL} optimizes detection and identification under fixed thresholds. We also introduce a strategy to minimize the maximum scores for non-mated probes, discouraging the model from assigning high scores to non-mated probes. This helps mitigate the risk of false positive detections in open-set scenarios, enhancing overall robustness.

To achieve this, we minimize the weighted average of the similarities between the probe and each gallery subject, where the weights are obtained by applying the softmax function to each score. Mathematically, given a probe sample p_i and a set of gallery subjects $\mathcal{G}' = \{g_1, g_2, \dots, g_n\}$ with corresponding similarity scores s_1, s_2, \dots, s_n , we minimize the weighted average as follows:

$$\mathcal{L}^{\text{RTM}} = \frac{1}{\sum_{j=1}^n e^{s_j}} \sum_{j=1}^n e^{s_j} \cdot s_j, \quad (12)$$

where e^{s_j} is the softmax value for each similarity score.

Fig. 4 (c) shows gradients of \mathcal{L}^{RTM} *w.r.t.* non-mated probes. This gradient adapts to the relative position of non-mated probes, assigning higher gradients to samples closer to the gallery. Unlike softmax-based losses and the triplet loss, RTM consistently emphasizes imposter scores that rank higher compared to others within the batch to learn better feature representations for open-set biometrics. This property is desirable as it effectively minimizes the maximum score of each non-mated probe.

The **max** function is an alternative to the weighted average. Yet, simply minimizing the maximum score in training would not minimize other scores even if they are close to the maximum. Hence, the weighted average would make it easier for the trained model to generalize during testing.

3.4 Overall Loss

The overall objective is to minimize the combination of \mathcal{L}^{IDL} and \mathcal{L}^{RTM} ,

$$L = \mathcal{L}^{\text{IDL}} + \lambda \cdot \mathcal{L}^{\text{RTM}}. \quad (13)$$

where λ is the loss weight. Compared to existing loss functions, our loss function treats genuine and imposter scores differently. It aims to reduce open-set thresholds by pushing non-mated probes away from gallery while pulling genuine scores toward their corresponding galleries, similar to open-set evaluation. Moreover, the relative magnitude of non-mated scores is utilized by RTM to assign larger gradients to harder non-mated probes.

4 Experiments

4.1 Experimental Setup

Evaluation Protocols. In face recognition experiments, we use the official closed- and open-set evaluation protocols of IJB-S [28]. However, gait recognition and person reID datasets typically lack official open-set evaluation protocols, where some probe subjects are excluded from the gallery. Naïvely removing gallery subjects often results in highly unstable outcomes. This instability arises because threshold calculation heavily relies on a small fraction (1%) of non-mated probes due to the limited number of probes in these datasets. For instance, in a dataset with 500 non-mated probes, the threshold is computed from the highest scores of only 5 probes.

To ensure robust open-set evaluation, we create N ($= 50$) random sets of non-mated probes and compute FNIR@1%FPIR for each set. In each set, $q\%$ $= 21.5\%$ subjects are designated as non-mated. We report the median value and the standard deviation from N trials. The choice of median over mean is deliberate to avoid undue influence from extreme values and provide a more stable metric.

Evaluation Metrics. We evaluate in both open- and closed-set settings. In closed-set settings, we use the Cumulative Matching Curve (CMC) at rank-1; for open-set, we use FNIR@FPIR at 1% and 0.1% FPIR and rank $R = 20$.

Datasets. For face recognition, we train on WebFace4M [80] and test on IJB-S [28]. Our experiments in gait recognition involve the Gait3D [77] and GREW [79] datasets, while for person reID, we employ the MEVID [8] and CCVID [17] datasets. Further details are provided in Tab. 1.

One technical caveat is that although FNIR@1%FPIR implicitly assumes that each subject only appears once in the gallery, a typical gait recognition or person re-id dataset contains more than one gallery template per subject. To

Table 1: Summary of datasets used in our experiments. [Keys: Mod.=Modality; Img.=Open-set image recognition; Subj.=the number of subjects ; Media=the number of images for face/image recognition datasets and the number of sequences/tracklets for gait and video-based person reID datasets]

Dataset	Year	Mod.	Train		Test	
			Subj.	Media	Subj.	Media
IJB-S [28]	2018	Face	-	-	202	> 10M
Gait3D [77]	2022	Gait	3,000	18,940	1,000	6,369
GREW [79]	2021	Gait	20,000	102,887	6,000	24,000
MEVID [8]	2023	ReID	104	6,338	55	1,754
CCVID [17]	2022	ReID	75	948	151	1,908

address this issue, we take the mean feature over all gallery templates for each subject. This benefits the models compared to randomly sampling one gallery template (details are in the Supp.).

Implementation Details. We implement our proposed approach in PyTorch [49]. For fair comparisons, we follow the experiment setup of our baselines. For details on the hyperparameters, please refer to the Supp.

4.2 Ablation Studies

We conduct our ablation studies on the Gait3D [77] dataset using GaitBase [13] as the baseline. In the ablation studies, we hold out 20% of subjects in the training set as the validation set.

Effect of Det, Id and RTM.

We ablate the effect of Det, Id and RTM in Tab. 2. RTM consistently enhances both closed- and open-set performance. Id further improves open-set performance but has a minor impact on rank-1 accuracy. Meanwhile, Det significantly enhances both closed- and open-set performance. In summary, each component contributes to its effectiveness.

Table 2: Ablation study results. [Keys: ✓=Kept; ✗=Removed; R@1=rank-1 accuracy]

Id	Det	RTM	α, γ	β	λ	R@1	FNIR@1%	FPIR
✗	✗	✗	—	—	—	64.6	86.43±5.44	
✗	✗	✓	6.0	0.2	4.0	65.1	85.81±5.78	
✗	✓	✓				64.4	85.16±5.13	
✓	✓	✓	3.0	0.2	4.0	63.9	85.22±4.73	
						9.0	64.2	84.08±6.28
✓	✓	✓	6.0	0.1	4.0	63.1	85.10±5.67	
						0.4	63.8	85.92±5.45
✓	✓	✓	6.0	0.2	2.0	64.8	83.12±5.94	
						8.0	61.6	85.67±5.15
✓	✓	✓	6.0	0.2	4.0	64.8	82.99±4.50	

Effect of Hyperparameters.

We ablate the effect of hyperparameters α, β, γ , and λ in Tab. 2. Across different hyperparameter settings, our approach can consistently improve open-set performance compared to the baseline, GaitBase [13]. Also, while the closed-set rank-1 accuracy is more sensitive to the hyperparameters, our approach can achieve slightly higher rank-1 accuracy with the best set of tested hyperparameters.

4.3 Comparison with State of the Art

Across biometric modalities, existing methods exhibit competitive closed-set recognition performance, as shown in Tab. 3 for face, Tab. 4 for gait, and Tab. 5 for person reID. In particular, in gait recognition, GaitBase [13] and SwinGait [12] showcase strong closed-set performance with a notable rank-1 accuracies of greater than 60% and 70%, respectively, for in-the-wild gait datasets Gait3D [77] and GREW [79]. However, in open-set scenarios, they face challenges, as evidenced by GaitBase’s high FNIR of 86.43% at 1% FPIR in Gait3D.

Table 3: Closed- and open-set face recognition on two official protocols of IJB-S [28]. [Keys: R@1=rank-1 accuracy; 0.1%=FNIR@0.1%FPIR; 1%: FNIR@1%FPIR]

Model	Surveillance-Single			Surveillance-Booking		
	R@1 ↑	0.1% ↓	1% ↓	R@1 ↑	0.1% ↓	1% ↓
AdaFace [29]	68.44	45.18	38.37	68.84	45.98	37.80
AdaFace+Ours	69.92	44.30	36.73	70.00	43.48	36.17

It is evident that the inclusion of our loss functions consistently yields improvements in both closed- and open-set performance across three modalities. In face recognition, we boost the rank-1 accuracy of AdaFace [29] to 69.92% (+1.48%) in Surveillance-to-Single. In the challenging open-set scenario, our losses further demonstrate their effectiveness by reducing FNIR@0.1%FPIR from 45.98% to 43.48% in Surveillance-Booking.

In gait recognition, when applied to Swin-Gait [12] backbone, our losses improve the rank-1 accuracy to 75.3% (+2.9%) on Gait3D, and FNIR@1%FPIR decreases to 79.30% (-4.20%) on Gait3D. In person reID, integrating our loss with CAL [17] reaches a rank-1 accuracy of 84.8% (+2.2%) and a FNIR of 30.81% (-3.08%) on MEVID [8]. We conduct the independent two-sample t-test to assess the statistical significance of performance

improvements in open-set scenarios. For instance, our approach outperforms CAL on the CCVID [17] dataset, with a p-value of 0.052%.

Collectively, these findings emphasize the consistent effectiveness of our proposed loss functions in enhancing both closed- and open-set performance across a spectrum of biometric tasks, ultimately improving the models’ ability to handle open-set challenges, while maintaining strong closed-set performance.

Efficiency. We present the floating point operations (FLOPs) of training baselines with and without our loss functions in Tab. 6. It is expected that incorporating our loss into baselines will increase the computational load in training, but the increase is very minor. For example, when SwinGait [12] is used, our loss functions only incur 46.9% of the FLOPs of the triplet loss, which is negligible compared to the model. Moreover, our approach does not incur any computation overhead during testing while improving closed- and open-set performance.

Table 4: Closed- and open-set gait recognition on Gait3D [77] and GREW [79]. SMPLGait [77] takes SMPL parameters as additional input. SwinGait-3D [12] uses our own implementation. [Keys: Rank@1=rank-1 accuracy]

Model	Rank@1 ↑		FNIR@1%FPIR ↓	
	Gait3D	GREW	Gait3D	GREW
GaitSet [4]	36.7	48.4	96.05±2.30	81.17±1.45
GaitPart [14]	28.2	47.6	96.88±1.80	88.43±1.26
GaitGL [38]	29.7	41.5	94.97±1.24	80.55±1.11
SMPLGait [77]	46.3	—	93.18±2.58	—
GaitBase [13]	64.6	60.1	86.43±5.44	77.67±1.60
+Ours	64.8	60.1	82.99±4.50	76.55±1.54
SwinGait-3D [12]	72.4	79.3	83.50±6.20	58.32±1.79
+Ours	75.3	80.4	79.30±6.74	57.29±1.83

Table 5: Closed- and open-set person reID on MEVID [8] and CCVID [17].

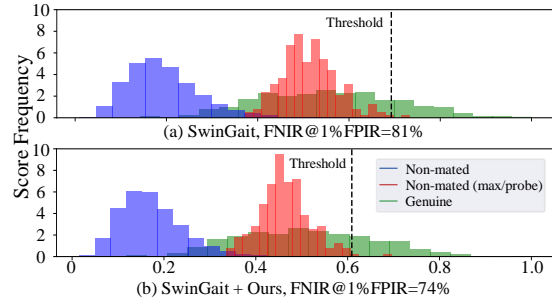
Model	Dataset	Rank-1 Accuracy \uparrow	FNIR@1%FPIR \downarrow
AGRL [66]	MEVID [8]	48.4	66.67 \pm 9.57
TCLNet [21]	MEVID [8]	48.1	70.34 \pm 11.68
PSTA [63]	MEVID [8]	48.2	77.11 \pm 11.72
CAL [17]	MEVID [8]	52.5	66.95 \pm 11.59
+Ours	MEVID [8]	55.1	64.80 \pm 10.71
CAL [17]	CCVID [17]	82.6	33.89 \pm 7.65
+Ours	CCVID [17]	84.8	30.81 \pm 8.02

Table 6: FLOPs in training different baseline models with and without our approach. We report the mean FLOPs across a minibatch. Our approach incurs no additional FLOP during testing.

Modality	Model	FLOPS +Triplet	+Ours
Face	AdaFace [29]	24.19G	— 503K
Gait	GaitBase [13]	35.43G	524.3K 245.8K
Gait	SwinGait [12]	120G	491.5K 230.4K
Body	CAL [17]	32.75G	4.09K 15.38K

4.4 Visualizations

Score Distributions. Fig. 5 visualizes the non-mated, maximum non-mated score for each probe, and genuine scores of SwinGait [12], with and without our loss. Our introduced RTM effectively reduces the distribution of maximum-per-probe scores, shifting it towards lower values. After linearly normalizing all scores between 0 and 1, the threshold at 1% FPIR is 0.61 with our loss (Fig. 5 (b)) compared to 0.69 without ours (Fig. 5 (a)). Moreover, our loss does not significantly alter the rightmost portion of the genuine score distribution, enhancing detection. This analysis offers insights into how our loss changes score distributions, and through which improves open-set performance.

**Fig. 5:** Score distributions on the test set of Gait3D [77] dataset. Ours (b) effectively reduces the threshold at 1% FPIR while only shifting the genuine score distribution marginally to the left compared to SwinGait [12] (a). Both distributions are shifted and scaled horizontally so that the min/max scores are within [0, 1].

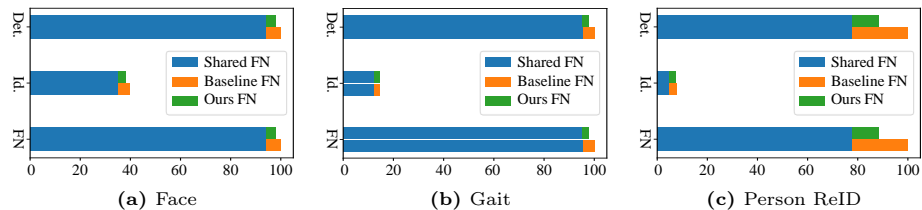


Fig. 6: Open-set FNs due to detection, identification, and either. The average across the 50 open-set evaluation protocols is plotted. Blue indicates FNs that are shared by the baseline, SwinGait [12], and our approach. Orange indicates FNs of the baseline only. Green indicates FNs of our approach.

False Negatives Breakdown. As illustrated in Fig. 6, across different biometric and generic tasks, almost all of the false negatives are failed detection, *e.g.* the 1st mated probe in Fig. 1. A small portion of false negatives have both failed identification and detection, *e.g.* the 3rd mated probe in Fig. 1. When we apply our loss functions, the number of FNs due to detection is reduced and the number of FNs due to identification is comparable to that of the baseline. These results highlight that our approach can improve open-set detection without sacrificing identification in various settings.

5 Conclusions

Our work addresses the critical gap in open-set biometrics, a practical generalization of the well-studied closed-set biometrics. We introduce novel I-DL and RTM losses aiming to improve biometric models’ open-set performance. Through extensive experiments across diverse modalities, including face, gait, and person reID, we demonstrate the efficacy of our approach. We not only advance the state of the art in the challenging open-set domain but also enhance closed-set recognition capabilities. We advocate for increased attention to the open-set problem within the biometric community.

Limitations While our paper makes significant strides in addressing open-set challenges in biometrics through loss functions, our exploration does not extensively explore dedicated architectures for this specific context. Future research endeavors may benefit from investigating tailored architectures that complement and enhance the effectiveness of our proposed loss functions.

Potential Societal Impacts. The use of biometric datasets in our work containing personally identifiable images prompts a crucial discussion on privacy and ethical implications. As biometric technologies become more prevalent, it is imperative to ensure that the collection of such datasets is aligned with ethical standards and privacy regulations.

Acknowledgements. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

1. Bendale, A., Boulton, T.E.: Towards open set deep networks. In: CVPR (2016)
2. Cen, J., Luan, D., Zhang, S., Pei, Y., Zhang, Y., Zhao, D., Shen, S., Chen, Q.: The devil is in the wrongly-classified samples: Towards unified open-set recognition. In: ICLR (2023)
3. Chai, T., Li, A., Zhang, S., Li, Z., Wang, Y.: Lagrange motion analysis and view embeddings for improved gait recognition. In: CVPR (2022)
4. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: AAAI (2019)
5. Chen, C., Ye, M., Jiang, D.: Towards modality-agnostic person re-identification with descriptive query. In: CVPR (2023)
6. Chen, G., Peng, P., Wang, X., Tian, Y.: Adversarial reciprocal points learning for open set recognition. IEEE TPAMI (2021)
7. Cui, Y., Kang, Y.: Multi-modal gait recognition via effective spatial-temporal feature fusion. In: CVPR (2023)
8. Davila, D., Du, D., Lewis, B., Funk, C., Van Pelt, J., Collins, R., Corona, K., Brown, M., McCloskey, S., Hoogs, A., et al.: MEVID: multi-view extended videos with identities for video person re-identification. In: WACV (2023)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
10. Dou, H., Zhang, P., Su, W., Yu, Y., Li, X.: Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In: ECCV (2022)
11. Dou, H., Zhang, P., Su, W., Yu, Y., Lin, Y., Li, X.: Gaitgci: Generative counterfactual intervention for gait recognition. In: CVPR (2023)
12. Fan, C., Hou, S., Huang, Y., Yu, S.: Exploring deep models for practical gait recognition. arXiv preprint arXiv:2303.03301 (2023)
13. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: CVPR (2023)
14. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: CVPR (2020)
15. Geng, C., Huang, S., Chen, S.: Recent advances in open set recognition: A survey. IEEE TPAMI (2020)
16. Grother, P., Ngan, M., Hanaoka, K.: Face recognition vendor test (FRVT) part 2: Identification (2019), <https://www.nist.gov/document/nistir827120190911pdf>
17. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: CVPR (2022)
18. Gunther, M., Cruz, S., Rudd, E.M., Boulton, T.E.: Toward open-set face recognition. In: CVPR Workshops (2017)

19. Günther, M., Hu, P., Herrmann, C., Chan, C.H., Jiang, M., Yang, S., Dhamija, A.R., Ramanan, D., Beyrer, J., Kittler, J., et al.: Unconstrained face detection and open-set face recognition challenge. In: IJCB. IEEE (2017)
20. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
21. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Temporal complementary learning for video person re-identification. In: ECCV (2020)
22. Huang, H., Wang, Y., Hu, Q., Cheng, M.M.: Class-specific semantic reconstruction for open set recognition. IEEE TPAMI (2022)
23. Huang, X., Zhu, D., Wang, H., Wang, X., Yang, B., He, B., Liu, W., Feng, B.: Context-sensitive temporal feature learning for gait recognition. In: ICCV (2021)
24. Huang, Y., Shen, P., Tai, Y., Li, S., Liu, X., Li, J., Huang, F., Ji, R.: Improving face recognition from hard samples via distribution distillation loss. In: ECCV (2020)
25. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: CVPR (2020)
26. Huang, Z., Xue, D., Shen, X., Tian, X., Li, H., Huang, J., Hua, X.S.: 3d local convolutional neural networks for gait recognition. In: ICCV (2021)
27. Jain, A.K., Deb, D., Engelsma, J.J.: Biometrics: Trust, but verify. IEEE Transactions on Biometrics, Behavior, and Identity Science (2021)
28. 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)
29. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: CVPR (2022)
30. Kim, M., Liu, F., Jain, A.K., Liu, X.: Cluster and aggregate: Face recognition with large probe set. In: NeurIPS (2022)
31. Kim, M., Su, Y., Liu, F., Jain, A., Liu, X.: Keypoint relative position encoding for face recognition. In: CVPR (2024)
32. Koch, T., Riess, C., Köhler, T.: Lord: Leveraging open-set recognition with unknown data. In: ICCV (2023)
33. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. In: ICCV (2021)
34. Kuchibhotla, H.C., Malagi, S.S., Chandhok, S., Balasubramanian, V.N.: Unseen classes at a later time? no problem. In: CVPR (2022)
35. Lee, G., Lee, S., Kim, D., Shin, Y., Yoon, Y., Ham, B.: Camera-driven representation learning for unsupervised domain adaptive person re-identification. In: ICCV (2023)
36. Li, W., Hou, S., Zhang, C., Cao, C., Liu, X., Huang, Y., Zhao, Y.: An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In: CVPR (2023)
37. Liang, J., Fan, C., Hou, S., Shen, C., Huang, Y., Yu, S.: Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In: ECCV (2022)
38. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: ICCV (2021)
39. Liu, F., Ashbaugh, R., Chimmitt, N., Hassan, N., Hassani, A., Jaiswal, A., Kim, M., Mao, Z., Perry, C., Ren, Z., et al.: Farsight: A physics-driven whole-body biometric system at large distance and altitude. In: WACV (2024)
40. Liu, F., Kim, M., Gu, Z., Jain, A., Liu, X.: Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In: ICCV (2023)
41. Liu, F., Kim, M., Jain, A., Liu, X.: Controllable and guided face synthesis for unconstrained face recognition. In: ECCV (2022)

42. Liu, F., Kim, M., Ren, Z., Liu, X.: Distilling CLIP with dual guidance for learning discriminative human body shape representation. In: CVPR (2024)
43. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SpheroFace: Deep hypersphere embedding for face recognition. In: CVPR (2017)
44. Ma, K., Fu, Y., Zheng, D., Cao, C., Hu, X., Huang, Y.: Dynamic aggregated network for gait recognition. In: CVPR (2023)
45. Mekhazni, D., Dufau, M., Desrosiers, C., Pedersoli, M., Granger, E.: Camera alignment and weighted contrastive learning for domain adaptation in video person reid. In: WACV (2023)
46. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: MagFace: A universal representation for face recognition and quality assessment. In: CVPR (2021)
47. Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., Zhang, D.: Biometrics recognition using deep learning: A survey. *Artificial Intelligence Review* (2023)
48. Ni, H., Li, Y., Gao, L., Shen, H.T., Song, J.: Part-aware transformer for generalizable person re-identification. In: ICCV (2023)
49. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
50. Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., Deng, W.: SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *TCSVT* (2023)
51. Ren, Z., Su, Y., Liu, X.: ChatGPT-powered hierarchical comparisons for image classification. In: NeurIPS (2023)
52. Ross, A., Banerjee, S., Chen, C., Chowdhury, A., Mirjalili, V., Sharma, R., Swearingen, T., Yadav, S.: Some research problems in biometrics: The future beckons. In: ICB (2019)
53. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
54. Shen, C., Fan, C., Wu, W., Wang, R., Huang, G.Q., Yu, S.: Lidargait: Benchmarking 3d gait recognition with point clouds. In: CVPR (2023)
55. Shen, C., Yu, S., Wang, J., Huang, G.Q., Wang, L.: A comprehensive survey on deep gait recognition: algorithms, datasets and challenges. *arXiv preprint arXiv:2206.13732* (2022)
56. Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: CVPR (2023)
57. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: CVPR (2020)
58. Tabassi, E., Watson, C., Fiumara, G., Salamon, W., Flanagan, P., Cheng, S.L.: Performance evaluation of fingerprint open-set identification algorithms. In: IJCB (2014)
59. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need. In: ICLR (2022)
60. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: CosFace: Large margin cosine loss for deep face recognition. In: CVPR (2018)
61. Wang, L., Liu, B., Liang, F., Wang, B.: Hierarchical spatio-temporal representation learning for gait recognition. In: ICCV (2023)
62. Wang, M., Guo, X., Lin, B., Yang, T., Zhu, Z., Li, L., Zhang, S., Yu, X.: DyGait: Exploiting dynamic representations for high-performance gait recognition. In: ICCV (2023)
63. Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D.: Pyramid spatial-temporal aggregation for video-based person re-identification. In: ICCV (2021)

64. Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., Huang, Q.: OpenAUC: Towards AUC-oriented open-set recognition. In: *NeurIPS* (2022)
65. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: *ECCV* (2016)
66. Wu, Y., Bourahla, O.E.F., Li, X., Wu, F., Tian, Q., Zhou, X.: Adaptive graph representation learning for video person re-identification. *TIP* (2020)
67. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: Openood: Benchmarking generalized out-of-distribution detection. In: *NeurIPS* (2022)
68. Ye, D., Fan, C., Ma, J., Liu, X., Yu, S.: Biggait: Learning gait representation you want by large vision models. In: *CVPR* (2024)
69. Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., Naemura, T.: Classification-reconstruction learning for open-set recognition. In: *CVPR* (2019)
70. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In: *CVPR* (2020)
71. Zahra, A., Perwaiz, N., Shahzad, M., Fraz, M.M.: Person re-identification: A retrospective on domain specific open challenges and future trends. *Pattern Recognition* (2023)
72. Zhai, Y., Liu, Z., Wu, Z., Wu, Y., Zhou, C., Doermann, D., Yuan, J., Hua, G.: Soar: Scene-debiasing open-set action recognition. In: *ICCV* (2023)
73. Zhang, H., Li, A., Guo, J., Guo, Y.: Hybrid models for open set recognition. In: *ECCV* (2020)
74. Zhang, Z., Tran, L., Liu, F., Liu, X.: On learning disentangled representations for gait recognition. *IEEE TPAMI* (2020)
75. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J., Wang, N.: Gait recognition via disentangled representation learning. In: *CVPR* (2019)
76. Zheng, J., Liu, X., Gu, X., Sun, Y., Gan, C., Zhang, J., Liu, W., Yan, C.: Gait recognition in the wild with multi-hop temporal switch. In: *ACMMM* (2022)
77. 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)
78. Zhou, D.W., Ye, H.J., Zhan, D.C.: Learning placeholders for open-set recognition. In: *CVPR* (2021)
79. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: *ICCV* (2021)
80. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., et al.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: *CVPR* (2021)