International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02115-6

# Re-ID-leak: Membership Inference Attacks Against Person Re-identification

Junyao Gao[1] · Xinyang Jiang[2] · Shuguang Dou[1] · Dongsheng Li[2] · Duoqian Miao[1] · Cairong Zhao[1]

Received: 22 July 2023 / Accepted: 4 May 2024

**Abstract**

Person re-identification (Re-ID) has rapidly advanced due to its widespread real-world applications. It poses a significant risk of exposing private data from its training dataset. This paper aims to quantify this risk by conducting a membership inference (MI) attack. Most existing MI attack methods focus on classification models, while Re-ID follows a distinct paradigm for training and inference. Re-ID is a fine-grained recognition task that involves complex feature embedding, and the model outputs commonly used by existing MI algorithms, such as logits and losses, are inaccessible during inference. Since Re-ID models the relative relationship between image pairs rather than individual semantics, we conduct a formal and empirical analysis that demonstrates that the distribution shift of the inter-sample similarity between the training and test sets is a crucial factor for membership inference and exists in most Re-ID datasets and models. Thus, we propose a novel MI attack method based on the distribution of inter-sample similarity, which involves sampling a set of anchor images to represent the similarity distribution that is conditioned on a target image. Next, we consider two attack scenarios based on information that the attacker has. In the "one-to-one" scenario, where the attacker has access to the target Re-ID model and dataset, we propose an anchor selector module to select anchors accurately representing the similarity distribution. Conversely, in the "one-to-any" scenario, which resembles real-world applications where the attacker has no access to the target Re-ID model and dataset, leading to the domain-shift problem, we propose two alignment strategies. Moreover, we introduce the patch-attention module as a replacement for the anchor selector. Experimental evaluations demonstrate the effectiveness of our proposed approaches in Re-ID tasks in both attack scenarios.

**Keywords** Person re-identification · Membership inference attack · Privacy and security

## 1 Introduction

The deep learning model has made remarkable progress with wide applications, but it also exposes risks of leaking personal information from its training set (Fredrikson et al., 2015; Wu et al., 2016; Shokri et al., 2017). This is particularly concerning for sensitive tasks like person re-identification (Re-ID), which involves identifying a specific person in different images or video scenes. A Re-ID training set contains pedestrian images, and leaking information from it can cause serious social security and ethical risks. To address this issue, quantifying the information leakage of Re-ID data becomes necessary.

One common methodology to quantify the privacy risk of a trained model is using the attack success rate of membership inference (MI) attack (Shokri et al., 2017; Yeom et al., 2018; Salem et al., 2018; Long et al., 2018; Nasr et al., 2018b; Song et al., 2019; Chen et al., 2021). MI attack algorithm

Communicated by Segio Escalera.

✉ Cairong Zhao
zhaocairong@tongji.edu.cn

Junyao Gao
junyaogao@tongji.edu.cn

Xinyang Jiang
xinyangj@microsoft.com

Shuguang Dou
2010504@tongji.edu.cn

Dongsheng Li
dongsheng.li@microsoft.com

Duoqian Miao
dqmiao@tongji.edu.cn

[1] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

[2] Microsoft Research Asia, Shanghai 200232, China

Published online: 22 May 2024

infers whether a record belongs to the training set by some information of the target model and is generally described as a binary classification problem. However, most existing MI attack methods focus on the classification task, where the attacker infers the membership of a sample based on its corresponding model outputs such as logits or loss (Shokri et al., 2017; Yeom et al., 2018; Sablayrolles et al., 2019), as shown in Fig. 1.

In contrast, Re-ID follows a totally different training and inference paradigm. State-of-the-art (SOTA) Re-ID methods first extract visual features from each pedestrian image and then conduct recognition by retrieving images based on the relative similarity between image pairs. During training, SOTA Re-ID methods add an extra identity classifier after the feature extractor, which is not available during inference. As a result, the attacker generally only gets the feature embedding of individual images, while the commonly used logits or loss for MI attack on classification are not available in the Re-ID task. Moreover, compared to the general classification, Re-ID is a more challenging fine-grained recognition task, leading to a more complex and less discriminative feature distribution for MI attacks. Previous works (Nasr et al., 2018a; Sablayrolles et al., 2019) have also shown that feature embedding contains more information irrelevant to training data and does not characterize the training-test generalization gap well compared to logits and loss. Thus, MI attacks on Re-ID require new approaches and considerations to effectively assess the privacy risk of the model.

As a result, in this paper, we propose a novel methodology to quantify privacy risks associated with Re-ID training sets. We achieve this by finding a new set of features suitable for MI attacks on Re-ID, rather than relying on conventional model outputs such as features, logits, and loss. Unlike classification tasks that focus on the semantics of individual samples, Re-ID is a metric learning task that models the relative relationship between image pairs. Therefore, rather than examining individual image characteristics, we extensively analyze the inter-sample correlation between different images and study how the generalization gap of the Re-ID model affects the distribution of pair-wise similarity. Intuitively, the Re-ID model brings together images with the same identities in the training set while separating those with dissimilar identities (Oh Song et al., 2016; Duan et al., 2017; Ming et al., 2022). However, this may not generalize well to samples in the test set, resulting in an inter-image similarity distribution shift between the training and test sets. This intuition is supported by our formal analysis of optimal attack with preliminary experiments in Sect. 3. Our experiments reveal a noticeable difference between the statistical properties of the inter-sample similarity distribution of samples in the training and test sets, which is consistently observed across different Re-ID models and datasets. Based on this analysis, we introduce a novel MI attack method called *similarity distribution based MI attack* (SD-MI attack), which conducts membership inference by exploiting the relative correlation between image pairs. Using a set of sampled anchor images to represent the inter-sample similarity distribution conditioned on the target image, the membership of the target image is inferred by a neural network based on its similarity with the anchor images within the reference set.

Based on the adversarial knowledge that the attacker can obtain, we evaluate MI attack methods on two different attack scenarios to comprehensively describe the potential privacy risks in Re-ID tasks, as illustrated in Fig. 2. One of the attack scenarios is referred to as the "one-to-one", where the attacker has access to the dataset distribution, architecture, and parameters of the target Re-ID model (Yu et al., 2021; Sablayrolles et al., 2019; Song et al., 2019). In this scenario, the MI attack method is trained on a subset of the known target model's training and test datasets and evaluated on a another disjoint subset, which introduces the risk of overfitting and restricts its applicability to real-world scenarios. Therefore, we further consider a more realistic attack scenario called "one-to-any", where the attacker has no knowledge about the target Re-ID model and dataset. In this scenario, the attacker is allowed to possess their own auxiliary Re-ID model and dataset, which are easily obtained and distinct from the target Re-ID model and dataset. They train an MI attack method using this auxiliary model and dataset and then expect it to perform well on the unknown target model and dataset, for instance, a classic Re-ID model PCB (Sun et al., 2018) trained on an open

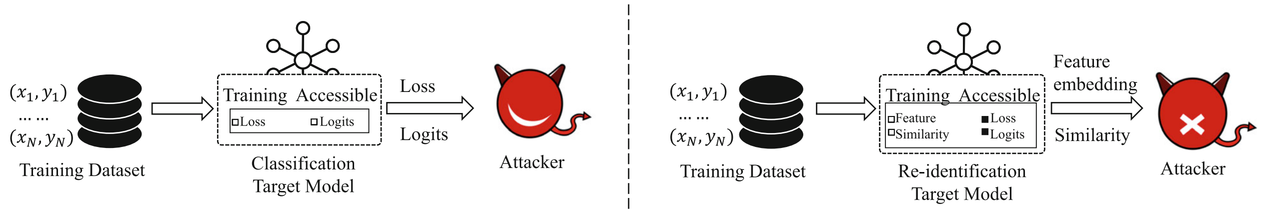


**Fig. 1** The different outputs for the classification model and Re-ID model under the black-box setting. For the classification model (left), an attacker can access the logits and loss both during and after the training process. However, for the Re-ID model (right), only similarity and feature embedding are accessible during inference, which is not suitable for existing classification-based MI attacks

source dataset Market1501 (Zheng et al., 2015) can be used to attack an unknown Re-ID model in online serving. This requires the attacker to identify and exploit vulnerabilities that are not specific to a particular system but rather exist across different Re-ID models and datasets, thus introducing two additional challenges. (1) *Cross-dataset*: Different Re-ID datasets exhibit significant shifts due to variations in scale, scene complexity, lighting conditions, viewpoint, pedestrians, and camera settings. (2) *Cross-model*: Different Re-ID models may have different architectures, training strategies, or data augmentation techniques, resulting in distinct representations in latent space and noticeable variations in similarity distributions. Hence, to tackle these challenges in the "one-to-any" attack scenario, we propose two alignment strategies. These strategies aim to align each anchor based on their similarity distribution descriptors, and standard normalization is employed to mitigate the domain shift problem.

Subsequently, to better select anchor images in reference sets that accurately represent the similarity distribution, we introduce two specialized attention-based modules developed for two distinct attack scenarios. In the "one-to-one" attack scenario, we propose an anchor selector module capable of automatically selecting anchor images based on their feature embeddings. However, in the "one-to-any" attack scenario, the feature embeddings follow a clearly distinct distribution across different domains, unlike the similarity distribution. As a result, our alignment strategies cannot effectively align feature embeddings between different domains. Therefore, the attention weights learned from the auxiliary domain's feature embedding cannot be directly applied to a different target domain, which hinders the effective utilization of the anchor selector module. To address this limitation, we apply the patch-attention module to the aligned similarity vector to re-weight the anchors by modeling their inter-relationships. Our extensive experimental results demonstrate the superiority of our approaches over existing MI attack algorithms in both the "one-to-one" and "one-to-any" attack scenarios.

The contributions of our work are summarized as follows:

- We raise a rarely studied privacy risk of the training set in the Re-ID task, whose information leakage is quantified by our proposed MI attack algorithms. For the MI attack, We establish two attack scenarios, namely "one-to-one" and "one-to-any", which provide a comprehensive description of the potential privacy risks associated with Re-ID tasks.
- We propose the first MI attack algorithm on the Re-ID task, which exploits a target image's relative correlation with reference images.

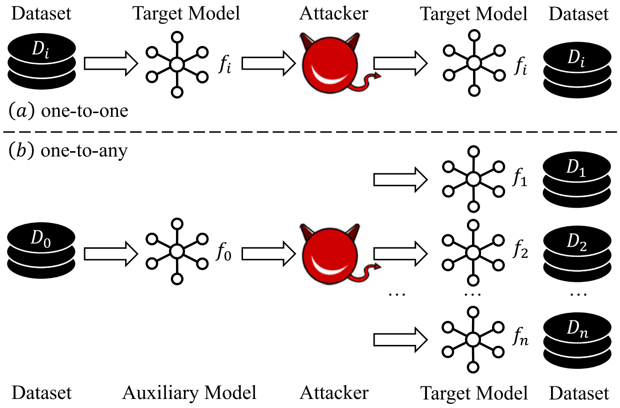


**Fig. 2** There are two MI attack scenarios in the Re-ID task: "one-to-one" (**a**) and "one-to-any" (**b**). In the "one-to-one" attack scenario, the attacker has access to the target Re-ID model and dataset, and they train and evaluate our MI method using the same known target Re-ID model and dataset. In the more realistic "one-to-any" attack scenario, the attacker lacks access to the target Re-ID model and dataset. They train MI methods on auxiliary datasets and expect them to perform well on the target Re-ID dataset and model

- We propose two alignment strategies to mitigate the domain-shift problem in the "one-to-any" attack scenario.
- We introduce two novel attention-based modules designed to effectively select the anchors that better represent the similarity distribution in two attack scenarios.
- Our proposed methods demonstrate superior performance compared to existing MI attack approaches on Re-ID models in both attack scenarios.

A preliminary version of this work was reported in Gao et al. (2023). Compared with our earlier study, the key differences are introduced: (1) We introduce a new "one-to-any" attack scenario to facilitate a more realistic discussion that has no access to target Re-ID model and dataset, accompanied by the corresponding results; (2) We propose two feature alignment strategies to mitigate the domain-shift problem in the "one-to-any" attack scenario; (3) We propose a novel patch-attention module for the new attack scenario, as the anchor selector module is not applicable; (4) We provide a detailed discussion and comparison of our new setting and method.

## 2 Related Works

### 2.1 Person Re-identification

Person re-identification aims to address the association and matching of target pedestrians across cameras and scenes by the features and similarities between pedestrians themselves. The existing approaches can be divided into two types depending on the architecture: CNN-based modeling and Transformer-based modeling.

*CNN-based Re-ID* Most CNN-based Re-ID models mostly extract local and global features to obtain discrimination information of the target person. Based on the methods to generate the local and global features, the models can be categorized into the following three categories.

(1) *Capturing features with multi-scale.* Yin et al. (2020) learn the local dynamic pose features and simultaneously quantify both motion and global visual cues to distinguish the different identities with similar appearance features. Zheng et al. (2019) reduces dependence on precise bounding boxes by blending local and global cues, and employs a dynamic training scheme to enhance identity representation. Zhu et al. (2020) proposed Viewpoint-Aware Loss with Angular Regularization (VA-reID) that projects features from various viewpoints onto a unified hypersphere, modeling identity and viewpoint-level distributions. Wang et al. (2018b) formulated the Re-ID task as a regression problem and constructed an identity regression space (IRS) by embedding different training person identities to solve the regression problem. Wang et al. (2018a) proposed the Multiple Granularities Network (MGN) that integrates global and local information at various granularities. Within each local branch of the MGN, the globally merged feature map is partitioned into distinct bands representing local regions, allowing independent learning of local feature representations.

(2) *Utilizing the attention mechanism.* Li et al. (2018) designed a lightweight attention network architecture called Harmonious Attention CNN (HAN) to learn the person invariant feature representation through hard-region soft-pixel-level attention. Yang et al. (2019) proposed an intra-attention network that aims to identify informative and discriminative regions within whole-body or body-part images. Chen et al. (2019) introduced the High-Order Attention (HOA) module, which effectively models and utilizes complex and high-order statistical information in the attention mechanism to capture subtle differences among pedestrians and generate discriminative attention proposals.

(3) *Partitioning the deep feature maps into pre-defined regions or parts.* Cheng et al. (2016) presented a multi-channel parts-based convolutional neural network that aims to jointly learn both the global full-body features and local body-part features of input persons. Sun et al. (2018) proposed a part-based convolutional baseline (PCB) that takes the entire image as input and divides the resulting feature map from the convolutional layer into $p$ uniformly sized parts to learn discriminative person features that are informed by different parts. Zhang et al. (2021) find different channels that activate responses for different body parts respectively and proposed an attention network with self or external guidance.

*Transformer-based Re-ID.* With the visual transformer (Dosovitskiy et al., 2020; Liu et al., 2021) demonstrating superior performance over CNN in more visual tasks, researchers are also focusing on the Person Re-ID with visual transformers. He et al. (2021) is the first to apply the pure transformer to Re-ID models (TransRe-ID), encoding an image as a sequence of patches and enhancing robust feature learning in the context of transformers using a novel jigsaw patch and side information embeddings module. In Sharma et al. (2021), researchers introduced a novel Locally Aware Transformer (LA-Transformer) that utilizes a strategy inspired by PCB to aggregate globally enhanced local classification tokens into an ensemble of $N$-classifiers, where $N$ represents the number of patches.

In this study, we employ CNN-based models with ResNet50 (He et al., 2016), MobileNetV2 (Sandler et al., 2018), and Xception (Chollet, 2017) backbones as our target models in "one-to-one" attack scenario. Moreover, we utilize **MGN** (Wang et al., 2018b), **HAN** (Li et al., 2018), **PCB** (Sun et al., 2018) and **TransRe-ID** (He et al., 2021) as our target models in "one-to-any" attack scenario, representing the diverse architectures of CNN-based Re-ID models and Transformer-based Re-ID models.

## 2.2 Membership Inference Attack

Membership inference (MI) attacks pose a significant challenge to researchers due to the high complexity of the training set and the target model, making it difficult for a theoretical analysis of why such attacks work. Recent research shows that the success rate of these attacks is mainly affected by the generalization gap of the target model. Shokri et al. (2017) and Sablayrolles et al. (2019) observe that the attack model is more likely to infer membership when the target model performs better on the training set than on the test set. Furthermore, Li et al. (2020) demonstrated experimentally that the generalization gap of the target model determines an upper bound on the success rate of MI attacks. While most MI attack issues are based on prediction vectors that directly relate to the generalization gap, such as loss and logits, our research conducts a thorough investigation of the distribution gap of similarities.

In metric embedding learning, Li et al. (2022) proposed a user-level MI attack based on the assumption that data from the same category forms a more compact cluster in the training set than in the test set. Our method differs from Li et al. (2022) by examining the similarity distribution over all sample pairs, including both intra- and inter-class similarity. Furthermore, our method does not require multiple samples for each identity. We establish the similarity distribution membership inference attack approach that describes the distribution gap between the training and test sets of a trained Re-ID model. This approach examines the similarity between the target image ($x_t$) and the anchor images to infer membership.

Shokri et al. (2017) proposed a method that trains a binary classifier to conduct membership inference on the classification model using logits as features. According to Hayes et al. (2017), for a generative adversarial network, the trained generator leads to more substantial confidence scores on the training set. In this paper, we introduce the similarity distribution membership inference attack approach that describes the distribution gap between the training and test sets of a trained Re-ID model. This approach examines the similarity between the target image ($x_t$) and the anchor images.

### 2.3 Domain Adaptation

Domain Adaptation is a technique that involves learning from a set of source domains to develop a high-performing model for an unseen target domain. Ganin et al. (2016) uses the gradient reverse layer and an adversarial loss to learn the domain-invariant representation. Recently, Gong et al. (2014) proposed unsupervised learning of a geodesic flow kernel to learn robust features that are resilient to the mismatch across domains. In addition, compared to feature space alignment, Bousmalis et al. (2017) learned a transformation in the pixel space from one domain to the other in an unsupervised manner.

In this work, we found that our similarity distribution shift is consistently observed as the domain-invariant feature across various Re-ID models and datasets. Additionally, we propose two alignment strategies that align each anchor based on their similarity distribution descriptors and utilize standard normalization to mitigate the domain-shift problem in the "one-to-any" attack scenario.

## 3 Preliminary Analysis

### 3.1 Preliminaries

This paper focuses on the effectiveness of approaches commonly adopted by SOTA Re-ID models, which employ a softmax-based classifier as a loss function. Let $D$ be a Re-ID dataset consisting of images sampled from a data distribution $P(x)$ in the form of $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $x$ represents the pedestrian image and $y$ is the corresponding identity label. Previous methods (Zheng et al., 2016; Hu et al., 2017; Zhou et al., 2019; Wang et al., 2018b) first pass an image $x$ through a backbone network to extract high-dimensional features. These features are then fed into fully connected layers to classify $x$ based on its corresponding identity $y$. Then the model is trained using the cross-entropy loss function:

$$\mathcal{L}_{id} = \frac{1}{n} \sum_{i=1}^{n} \log(p(y_i|x_i)) \tag{1}$$

During the inference phase, Re-ID can be viewed as an image retrieval task aimed at identifying images with the same identity as the query image from a gallery. This is accomplished by excluding the identity classifier and utilizing the high-level feature extracted before the classifier to calculate the similarity between the query image and the gallery images. Subsequently, person re-identification is performed by sorting the images based on this similarity metric.

### 3.2 Optimal Membership Inference

We adopt the assumption from Sablayrolles et al. (2019) that models the posterior distribution of model parameters $\theta$ as follows:

$$P(\theta|(x_i, y_i, m_i)) \propto \exp\left(-\frac{1}{T}\sum_{i=1}^{n} m_i \mathcal{L}(\theta, x_i, y_i)\right) \tag{2}$$

The membership variable $m_i$ indicates whether a sample belongs to the test set ($m_i = 0$) or the training set ($m_i = 1$). Additionally, the temperature parameter $T$ controls the level of stochasticity in the model parameters $\theta$. By substituting the Re-ID loss function into Eq. 2, we obtain the posterior distribution of the model parameters for Re-ID:

$$\begin{aligned} P(\theta|(x_i, y_i, m_i) &\propto \exp\left(-\frac{1}{T}\sum_{i=1}^{n} m_i \mathcal{L}(\theta, x_i, y_i)\right) \\ &= \exp\left(-\frac{1}{T}\sum_{i=1}^{n} m_i \log P(y_i|x_i;\theta)\right) \\ &= \exp\left(-\frac{1}{T}\sum_{i=1}^{n} m_i \log \frac{d(x_i, a_{y_i})}{\sum_{j=1}^{k} d(x_i, a_j)}\right) \end{aligned} \tag{3}$$

The function $d(x_i, a_j)$ represents a similarity measurement in the Re-ID representation space. This function has multiple variants depending on the specific cross-entropy based Re-ID methods used, such as L2Softmax (Ranjan et al., 2017) and AngularSoftmax (Liu et al., 2016). The variable $a_j$ represents learned class centers that correspond to each identity, while $k$ denotes the total number of identities.

In accordance with Sablayrolles et al. (2019), considering the set of other samples and their membership denoted as

$\mathcal{T} = (x_i, y_i, m_i)_{i=1}^{n}$, the membership of the sample $x_1$ can be inferred as follows:

$$\begin{aligned}\mathcal{M}(\theta, x_1, y_1) &:= P(m_1 = 1|\theta, x_1, y_1)\\ &= E_{\mathcal{T}}\left[\sigma\left(s(x_1, y_1, \theta, P(\theta|\mathcal{T})) + \log\frac{P(m_1 = 1)}{1 - P(m_1 = 1)}\right)\right]\end{aligned} \tag{4}$$

where

$$\begin{aligned}s(x_1, y_1, \theta, P(\theta|\mathcal{T})) &= -\frac{1}{T}\log\frac{d(x_1, a_{y_1})}{\sum_{j=1}^{k} d(x_1, a_j)}\\ &- \log\left(\int_{\theta'} \exp\left(-\frac{1}{T}\log\frac{d(x_1, a_{y_1})}{\sum_{j=1}^{k} d(x_1, a_j)} P(\theta'|\mathcal{T})\right) d\theta'\right)\end{aligned} \tag{5}$$

From Eqs. 4 and 5, the second term of Eq. 5 represents the standard loss of $x_1$ under models that have not been trained with $x_1$ and can be interpreted as a threshold for membership inference (MI) attacks. If this term is computed or accurately approximated, the optimal membership inference depends solely on the relative similarity between the target sample $x_i$ and the identity centers $a_j$. However, as discussed in the introduction, these learnable identity centers are typically inaccessible to attackers. Consequently, since the Re-ID loss aims to minimize the distance between the training samples and their respective centers, it is intuitive to choose a set of proxy centers, referred to as anchor images in this paper, to approximate the learned centers using actual Re-ID dataset images and perform membership inference based on the sampled proxy centers. Our preliminary experiments in the next subsection confirm that there exists a noticeable and distinguishable distinction in the statistical properties of the similarity between the target image and randomly sampled anchor images in both the training and test sets.

### 3.3 Preliminary Experiments

*Experiment Configuration* The formal analysis in the last subsection demonstrated that the membership of a target image is contingent upon the relative similarity between the target image and the identity centers learned during Re-ID training. As identity centers are inaccessible for MI attacks, we propose sampling a set of reference images from the Re-ID dataset as proxy centers and investigating the impact of the training/test generalization gap on their similarities with the target image. Specifically, when considering a Re-ID model and its training dataset ($D_{train}$) and test dataset ($D_{test}$), we select a random subset comprising 10% of $D_{train}$ and $D_{test}$ as reference samples. Subsequently, we compute the Euclidean distance in the feature space of Re-ID model's feature extractor between the target samples from $D_{train}/D_{test}$ and these reference samples. We expect that the similarity distributions between the reference-$D_{train}$ and reference-$D_{test}$ pairings will exhibit significant differences.

**Table 1** The table presents the average discrepancy in average distance for all training and testing samples relative to the reference samples

| Dataset | MGN (%) | PCB (%) | HAN (%) |
|---|---|---|---|
| Market1501 | 10.87 | 8.66 | 8.39 |
| DukeMTMC | 15.06 | 4.66 | 8.12 |
| MSMT17 | 13.22 | 6.69 | 5.86 |

*Statistical Analysis* Upon analyzing the distance matrix, we observed that the individual pair-wise distances exhibit a high standard deviation and do not exhibit discernible patterns associated with membership. Consequently, various statistical properties of the overall distance distribution are compared between the distances from training samples and test samples to the reference samples.

First, for each reference sample, we calculate the average distance for both reference-$D_{train}$ and reference-$D_{test}$ pairings. Subsequently, we subtract the average distance of reference-$D_{train}$ pairings from the average distance of reference-$D_{test}$ pairings across all reference samples and calculate their average. As illustrated in Table 1, the discrepancies in average are consistently positive, predominantly ranging from 5 to 15%, signifying a discernible disparity in the similarity distribution among training and test set samples across diverse Re-ID models and datasets.

Next, we investigate the average and standard deviation distances for different reference samples on specific Re-ID datasets and models, as depicted in Fig. 3. In sub-figures (a), (e) and (i), the y-axis represents the average distance from each target sample in $D_{train}$ or $D_{test}$ to a specific reference sample, while the x-axis denotes different reference samples. As a result, we note a distinct separation between the average distances corresponding to $D_{train}$ and $D_{test}$. Typically, the average distance for target samples in $D_{train}$ is greater than that for samples in $D_{test}$.

Likewise, sub-figures (b), (f) and (j) illustrate the standard deviation of the distances between each target sample in $D_{train}/D_{test}$ and various reference samples in the respective Re-ID datasets and models, highlighting a more noticeable discrepancy between samples from $D_{train}$ and $D_{test}$ compared to the average distance.

In addition to examining the mean and deviation of distances based on each reference image, we also investigate the mean and standard deviation distribution across all reference images. This is represented by a cumulative distribution function, as illustrated in Fig. 3c, d, g, h, k, l. Notably, we observe that the cumulative distribution functions for samples in $D_{test}$ consistently lie above those for samples in $D_{train}$.

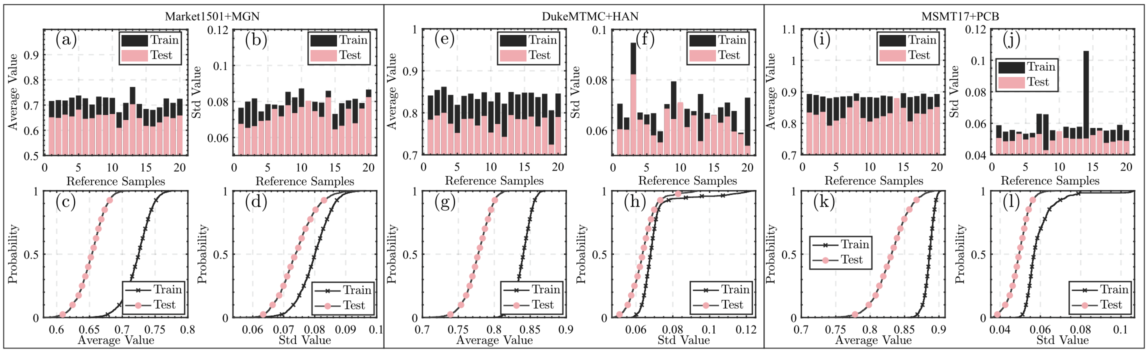


**Fig. 3** The average (**a**, **e**, **i**) and standard deviation (**b**, **f**, **j**) gap of distance from every reference sample to training target images or test target images and the cumulative density function of the average (**c**, **g**, **k**) and standard deviation (**d**, **h**, **l**) of the distance from all reference samples to training target images and test target images across Market1501+MGN,DukeMTMC+HAN and MSMT17+PCB

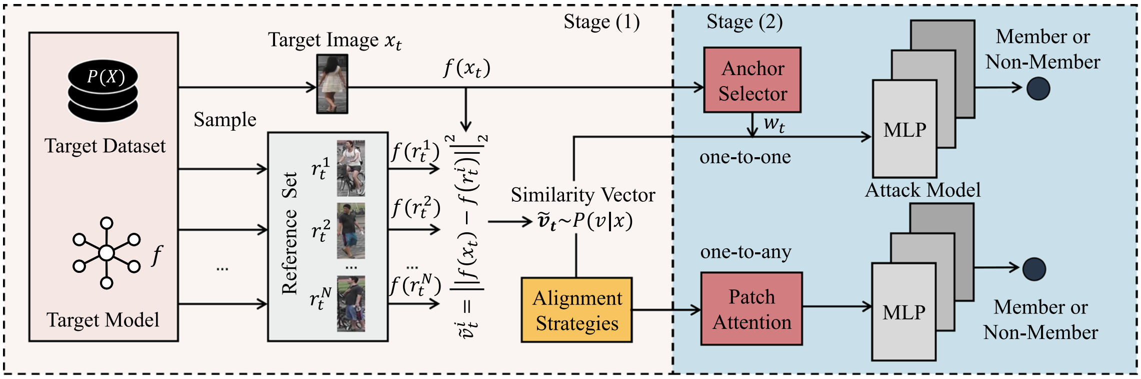


**Fig. 4** The two-stages pipeline of our black-box MI attack. First, for each target image $x_t$ we compute the similarity vector $\tilde{\boldsymbol{v}}_t$ with reference samples. Next, we feed the similarity vector $\tilde{\boldsymbol{v}}_t$ into the attack model to infer the membership of the target image $x_t$ (Two alignment strategies are applied to tackle the domain-shift problem in the “one-to-any” attack scenario). Furthermore, we propose two attention-based modules, namely anchor selector and patch-attention, to select useful anchor images from the limited reference set and improve the approximation of the similarity distribution

*Design Principles* In conclusion, our experiments reveal the existence of a noticeable shift in the similarity distribution between the training and test sets across diverse Re-ID datasets and models. This phenomenon occurs because various Re-ID models, employed on different Re-ID datasets, aim to minimize the distance between images of the same ID while maximizing the distance between images of other IDs in the latent space during training. Consequently, the similarity distribution shift among pedestrian images becomes more consistent and is less susceptible to different Re-ID models and datasets. This suggests that the similarity distribution between the target sample and a set of anchor images serves as an effective attack feature for membership inference in both “one-to-one” and “one-to-any” attack scenarios.

## 4 Proposed Method

We first provide a brief overview of the pipeline for our membership inference attack based on similarity distribution in two attack scenarios, as illustrated in Fig. 4. Our method consists of two main stages. In the first stage, we compute a similarity vector that represents the conditional distribution of similarity between the target image and other images in the dataset. The second stage involves conducting membership inference based on the similarity distribution using novel neural network structures. In the following two subsections, we will provide detailed explanations of our designs and implementations for each of the two stages.

### 4.1 Obtaining Similarity Vector

In line with the design principles, we infer the membership of a target image $x_t$ by examining its similarity with a set of anchors sampled from the target Re-ID data distribution $P(x)$.

*One-to-one Attack Scenario* Firstly, we begin by sampling a reference set $\boldsymbol{r_t} = [r_t^1, r_t^2, r_t^3, \ldots, r_t^N]$ from the Re-ID data distribution $P(x)$. Here, each $r_t^i \in P(x)$ is randomly sampled from the dataset distribution $P(x)$, and $N$ represents the number of images in the reference set. We calculate the $i$-th sampled distance $\tilde{v}_t^i$ of the similarity vector $\boldsymbol{\tilde{v}_t} = [\tilde{v}_t^1, \tilde{v}_t^2, \tilde{v}_t^3, \ldots, \tilde{v}_t^N]$ by computing the Euclidean distance between the target image $x_t$ and the $i$-th anchor image $r_t^i$ in the reference set:

$$\tilde{v}_t^i = \|f(x_t) - f(r_t^i)\|_2^2, \tag{6}$$

The function $f(\cdot)$ maps any input instance to its feature embeddings in the corresponding known model $f$. Specifically, we consider the feature embeddings $f(x_t)$ and $f(r_t^i)$ of the target image $x_t$ and the reference image $r_t^i$ as points in a $K$-dimensional Euclidean space. The sampled similarities from the known model $f$ and the dataset $P(x)$ are utilized to construct a similarity vector $\boldsymbol{\tilde{v}_t}$. In the "one-to-one" attack scenario, since we have knowledge of the dataset, parameters, and architecture of the target Re-ID model, we can effortlessly label a small portion of similarity vectors as member or non-member of the target Re-ID model and train our attack model using these samples. Finally, the remaining unlabeled samples from the target model are inputted into our attack model to predict the membership.

*One-to-any Attack Scenario* Conversely, in the "one-to-any" attack scenario, where we lack access to the target Re-ID model and dataset for training an attack model. To address this, we introduce an auxiliary model and dataset to train our attack model and evaluate the target sample of target model and dataset. Specifically, we randomly sample $N$ images from auxiliary model and dataset as the reference set $\boldsymbol{r'_t}$ and then compute the similarity vector $\boldsymbol{\tilde{v}'_t}$ for other samples. These samples are then labeled based on their corresponding membership in the auxiliary model, forming the dataset $A_r$. After training the attack model using $A_r$, we sample $N$ images in target model and dataset to construct the reference set $\boldsymbol{r_t}$, and compute $\boldsymbol{\tilde{v}_t}$ for target sample $x_t$. denoted as evaluation dataset $A_e$. It is important to note that the number of sampled images in $\boldsymbol{r'_t}$ and $\boldsymbol{r_t}$ should be equal to ensure the input dimensions of $\boldsymbol{\tilde{v}'_t}$ and $\boldsymbol{\tilde{v}_t}$ in attack model are identical.

Additionally, as mentioned in Sect. 3, although similarity distribution shift is prevalent in nearly all Re-ID models and datasets, the cross-dataset and cross-model challenges further amplify the domain-shift problem in the "one-to-any" attack scenario. Therefore, we propose two alignment strategies to mitigate the domain-shift problem between the training dataset $A_r$ and evaluation dataset $A_e$. Specifically, the random sampling strategy employed in the reference sets $\boldsymbol{r_t}$ and $\boldsymbol{r'_t}$ results in different similarity distributions between the $i$-th anchors $\tilde{v}_t^i$ and $\tilde{v}_t^{\prime i}$. Consequently, we sort each anchor in ascending order based on their respective standard deviation values. This alignment aims for a more uniform similarity distribution across the anchors by ensuring that the similarity distributions between the $i$-th anchors $\tilde{v}_t^i$ and $\tilde{v}_t^{\prime i}$ are as similar as possible in terms of their standard deviation value that provides a more informative representation of the similarity distribution compared to the mean value.

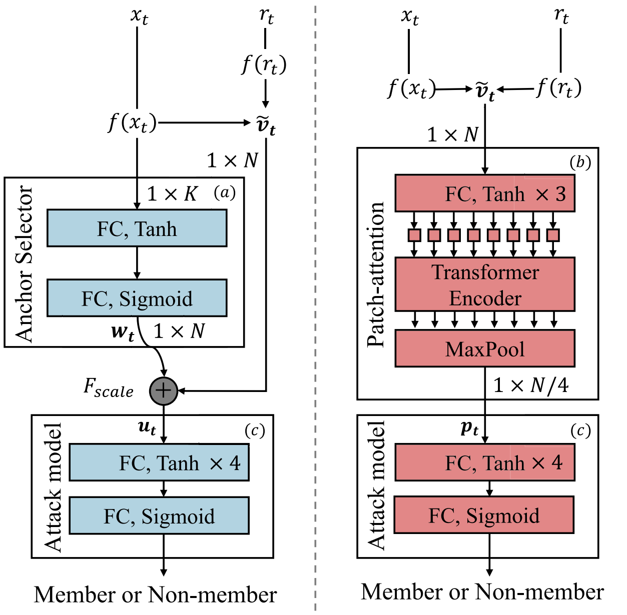


**Fig. 5** The specific architectures of our attack model (**c**), anchor selector module (**a**), and patch-attention module (**b**) in "one-to-one" (left) and "one-to-any" (right) attack scenarios

Subsequently, to reduce significant disparities in similarity distributions between different Re-ID models and datasets, we apply standard normalization on the similarity vectors $\boldsymbol{\tilde{v}_t}$ and $\boldsymbol{\tilde{v}'_t}$ in datasets $A_r$ and $A_e$:

$$V^i = \frac{V^i - \mu}{\sigma} \tag{7}$$

where $V^i$ represents the similarities between anchor image $i$ and all other images in $A_r$ or $A_e$, $\mu$ represents the mean value of $V_t^i$, and $\sigma$ refers to the standard deviation. After normalization, the similarity distribution of each anchor in the reference set is transformed to a numerical space with a mean of zero and a variance of one, aiming to reduce the domain shift.

### 4.2 Membership Inference Network

Figure 5 illustrates the model structure of our proposed membership inference network in two attack scenarios. It receives as input the similarity vector between a target image and a reference set of anchor images and produces a binary value that determines the membership.

*Attack Model* Consistent with prior research Shokri et al. (2017), Long et al. (2018), Salem et al. (2018), Yu et al. (2021), Chen et al. (2021), we employ a multi-layer perceptron (MLP) for membership inference. Our attack model for SD-MI attack consists of four hidden layers employing the Tanh activation function, and a binary classification output layer using the sigmoid activation function, as shown in Fig. 5c. We denote this approach as $M_{SD}$.

As mentioned in the previous section, Eq. 5 highlights the significance of selecting suitable reference images to achieve an improved approximation of the identity centers $a_j$ in Re-ID membership inference. Consequently, we propose two additional attention-based modules to enhance the representation of the similarity distribution in both the "one-to-one" and "one-to-any" attack scenarios.

*Anchor Selector Module* In the "one-to-one" attack scenario, we present the *anchor selector module* to choose relevant anchor images based on the content of the current image. This module assigns weights $w^i$ to the distances between the target image and various reference images. Figure 5a illustrates the utilization of the anchor selector $F_{as}$. It takes the high-dimensional Re-ID feature embedding $f(x_t)$ of the target image $x_t$ as its input. For the implementation of this module $F_{as}$, we employ a 2-layer MLP with a sigmoid activation function.

$$\boldsymbol{w_t} = F_{as}(f(x_t), \Theta) = \sigma(\Theta_2 \delta(\Theta_1 f(x_t))), \tag{8}$$

where $\delta$ represents the Tanh activation, $\Theta_1 \in \mathbb{R}^{K \times K}$ and $\Theta_2 \in \mathbb{R}^{N \times K}$. Then we rescale the weight vector $\boldsymbol{w_t}$ and the similarity vector $\tilde{\boldsymbol{v}}_{\boldsymbol{t}}$ as:

$$u_t^i = F_{scale}(w_t^i, \tilde{v}_t^i) = w_t^i \tilde{v}_t^i, \tag{9}$$

where $\boldsymbol{u_t} = [u_t^1, u_t^2, \ldots, u_t^N]$ is the input feature for attack model and $F_{scale}$ refers to a multiplication between the weight vector $\boldsymbol{w_t}$ and the similarity vector $\tilde{\boldsymbol{v}}_{\boldsymbol{t}}$. We refer to the SD-MI attack with *anchor selector module* as $M_{AS+SD}$.

*Patch-Attention Module* In the "one-to-any" attack scenario, the cross-model and cross-dataset challenges result in clearly distinct feature embedding distributions across different domains. This hampers the effective utilization of the anchor selector module. To address this limitation, we apply the novel *patch-attention module* to the aligned similarity vector to re-weight the anchors by modeling their inter-relationships in similarity distribution. Specifically, as depicted in Fig. 5b, this module initially maps the aligned similarity vector $\tilde{\boldsymbol{v}}_{\boldsymbol{t}}$ with $N$ dimensions into a latent representation with $2 \times N$ dimensions. Subsequently, the latent representation is partitioned into eight patches, with each patch representing the anchors with closely related similarity distributions.

$$[p_t^1, p_t^2, \ldots, p_t^8] = F_{patch}(\phi(\tilde{\boldsymbol{v}}_{\boldsymbol{t}})), \tag{10}$$

Here, $F_{patch}$ denotes the grouping operation, and $\phi(, \cdot, )$ is implemented as a 3-layer MLP with a tanh activation function. To effectively capture the relationships between similarity distributions at different levels, we utilize a transformer encoder with self-attention to weight the anchors among patches. Lastly, we perform max pooling on the eight patches to select the most salient anchor across the different patches:

$$\boldsymbol{p_t} = \max(T[p_t^1, p_t^2, \ldots, p_t^8]), \tag{11}$$

where $T$ represents the transformer encoder, and $\boldsymbol{p_t}$ refers to our final attack feature, which captures the prominent anchors with the most informative similarity distribution in the latent space. We denote the SD-MI attack with *patch-attention module* as $M_{PA+SD}$.

## 5 Experimental Setup

This section presents the configuration and implementation details of our experiments.

### 5.1 Datasets

We utilize three variant datasets for Re-ID: Market1501 (Zheng et al., 2015), DukeMTMC-Re-ID (Zheng et al., 2017) and MSMT17 (Wei et al., 2018). The Market1501 dataset consists of 1501 pedestrian classes, totaling 32,668 images captured by five high-resolution cameras and one low-resolution camera. We assign 751 pedestrian classes to the training set, while the remaining classes form the test set (gallery set). For evaluation, we select one image per pedestrian from the test set as a query to assess the Re-ID model. The DukeMTMC-Re-ID dataset comprises 16,522 training images from 702 pedestrians and 17,661 test images (gallery set) from 702 other pedestrians. The images are captured by eight static HD cameras located at Duke University. The query set is also selected from the gallery set. As a more realistic and larger Re-ID dataset, MSMT17 consists of 15 cameras capturing diverse scenes, and various weather conditions, and spanning multiple time periods. Ultimately, the dataset comprises a total of 126,441 images, featuring 4101 unique pedestrians. Out of these, 1041 pedestrians with a

**Table 2** We compared the performance of our proposed method with existing membership inference attack baselines in the "one-to-one" attack scenario on different Re-ID models trained on the Market1501 and DukeMTMC datasets

| Method | ResNet50 | | MobileNetV2 | | Xception | |
|---|---|---|---|---|---|---|
| | Market1501 (%) | DukeMTMC (%) | Market1501 (%) | DukeMTMC (%) | Market1501 (%) | DukeMTMC (%) |
| $M_{FE}$ | 80.1 | 80.5 | 74.9 | 72.7 | 78.5 | 76.1 |
| $M_{tloss}$ | 82.6 | 86.2 | 77.4 | 77.8 | 84.9 | 83.8 |
| $M_{U\_low}$ | 72.4 | 70.8 | 65.5 | 63.3 | 71.0 | 66.1 |
| $M_{U\_mid}$ | 78.6 | 77.4 | 71.0 | 69.3 | 76.9 | 72.3 |
| $M_{U\_high}$ | 82.9 | 81.9 | 74.0 | 72.6 | 79.6 | 75.9 |
| $M_{SD}$ (ours) | 87.0 | 88.7 | 80.6 | 81.4 | 89.7 | 90.6 |
| $M_{AS+SD}$ (ours) | **87.3** | **89.1** | **81.2** | **82.2** | **90.1** | **91.6** |

The highest performance is indicated in bold

combined image count of 32,621 were allocated to the training set, while the test set consisted of 3060 pedestrians with a collective image count of 93,820. For the query set, a random subset of 11,659 images was selected from the test set.

In the "one-to-one" attack scenario, the training dataset for the attack model consists of 2000 samples selected from the training set and 2000 samples from the test set of the target Re-ID model. Additionally, the evaluation dataset for the attack model is created by randomly sampling 6000 images from both the training set and the test set of the same target Re-ID model. In the "one-to-any" attack scenario, the entire dataset of the auxiliary Re-ID model is considered as the attack training dataset, while the dataset of the target Re-ID model is used as the attack evaluation dataset.

## 5.2 Target Models

In our experiments, we employ the CNN-based Re-ID models with different backbone networks, namely ResNet50, MobileNetV2, and Xception, as target models in the "one-to-one" attack scenario. Each model is trained on the Market1501 and DukeMTMC-Re-ID datasets. In addition, we utilize various architecture models for the "one-to-any" attack scenario, including CNN-based Re-ID models such as MGN, HAN, PCB, and a transformer-based Re-ID model TransRe-ID. Each target model is trained on the Market1501, DukeMTMC-Re-ID, and MSMT17 Re-ID datasets.

## 5.3 Baselines

*Feature based MI Attack* ($M_{FE}$) In order to validate the assumption that the feature embedding does not capture the train/test generalization gap as effectively as the direct model outputs, we employ a feature embedding-based MI attack method. This method involves feeding the Re-ID feature of the target image into the same MI backbone as $M_{SD}$, following the approach outlined in Nasr et al. (2018a).

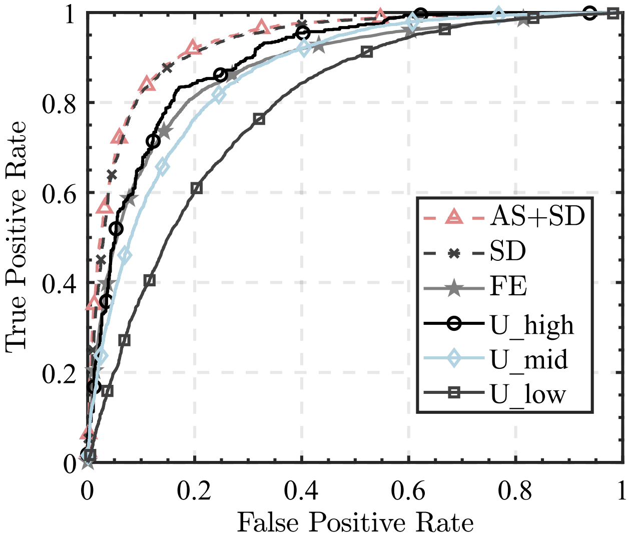


**Fig. 6** ROC curve of $M_{AS+SD}$, $M_{SD}$, $M_{U\_high}$, $M_{U\_mid}$, $M_{U\_low}$ and $M_{FE}$ on ResNet50 trained on Market1501

*Triple Loss based MI Attack* ($M_{tloss}$) This baseline approach follows the design of a state-of-the-art (SOTA) black-box metric-based MI attack (Sablayrolles et al., 2019), which infers membership by considering the target image's training loss and a manually determined threshold. Since the cross-entropy-based Re-ID losses are not directly accessible in our black-box setting, we compute the triplet loss (Schroff et al., 2015) based on the image features, serving as a surrogate loss function. More specifically, the triplet loss is computed by selecting the target image $x_t$ as the anchor image and treating all images with the same identity as positive samples, while sampling 100 images with different identities as negative samples.

*User-level MI Attack* ($M_U$) As a comparison baseline, we select the user-level MI attack (Li et al., 2022), which was originally designed for metric learning-based models. However, this method cannot be directly applied to instance-level MI attacks. Therefore, we adapt the original method by sam-

**Table 3** Performance comparison between our proposed method, $M_{PA+SD}$, and the baseline membership inference attack method, $M_U$, under the "one-to-any" attack scenario

| Source→Target | Method | Trans→HAN (%) | PCB→MGN (%) | HAN→PCB (%) | MGN→HAN (%) | Avg (%) |
|---|---|---|---|---|---|---|
| MSMT→Duke | $M_{U\_low}$(aligned) | 71.1 | 63.1 | 65.4 | 67.3 | 66.7 |
| | $M_{U\_mid}$(aligned) | 77.8 | 69.7 | 72.8 | 77.5 | 74.5 |
| | $M_{U\_high}$(aligned) | 82.9 | 77.7 | 77.4 | 82.3 | 80.1 |
| | $M_{SD}$(aligned) | 92.0 | 88.3 | **79.1** | 88.4 | 87.0 |
| | $M_{PA+SD}$ | **94.6** | **88.3** | 79.0 | **92.4** | **88.6** |
| MSMT→Market | $M_{U\_low}$(aligned) | 67.4 | 66.9 | 78.2 | 71.8 | 71.1 |
| | $M_{U\_mid}$(aligned) | 78.2 | 72.2 | 83.6 | 79.4 | 78.4 |
| | $M_{U\_high}$(aligned) | 82.1 | 77.1 | 87.5 | 83.9 | 82.7 |
| | $M_{SD}$(aligned) | 91.1 | 87.3 | 90.4 | 84.9 | 88.4 |
| | $M_{PA+SD}$ | **92.7** | **87.4** | **91.3** | **88.1** | **89.9** |
| Market→Duke | $M_{U\_low}$(aligned) | 80.4 | 72.0 | 71.4 | 80.7 | 76.1 |
| | $M_{U\_mid}$(aligned) | 88.5 | 79.4 | 77.6 | 87.6 | 83.3 |
| | $M_{U\_high}$(aligned) | 91.4 | 87.0 | **84.1** | 91.5 | 88.5 |
| | $M_{SD}$(aligned) | 92.3 | 89.4 | 78.7 | 88.4 | 87.2 |
| | $M_{PA+SD}$ | **94.8** | **89.6** | 79.0 | **94.1** | **89.4** |
| Avg. | $M_{U\_low}$(aligned) | 73.0 | 67.3 | 71.7 | 73.3 | 71.3 |
| | $M_{U\_mid}$(aligned) | 81.5 | 73.8 | 78.0 | 81.5 | 78.7 |
| | $M_{U\_high}$(aligned) | 85.5 | 80.6 | 83.0 | 85.9 | 83.7 |
| | $M_{SD}$(aligned) | 91.8 | 88.3 | 82.7 | 87.2 | 87.5 |
| | $M_{PA+SD}$ | **94.0** | **88.4** | **83.1** | **91.5** | **89.3** |

To ensure a fair comparison, we have adopted the alignment strategies used in $M_U$. All comparisons are conducted across multiple Re-ID models and datasets. Source→Target means auxiliary model and dataset to target model and dataset. The best-performing results are highlighted in bold

pling a set of images that share the same identity as the target image and computing the intra-class distance based on the sampled images. In order to examine the impact of the number of positive images on the performance of the user-level MI attack, we present three sets of results using different numbers of sampled images. More specifically, $M_{U\_low}$, $M_{U\_mid}$, and $M_{U\_high}$ refers to the user-level MI attack with two, four, and all positive images sampled for each target image, respectively. It is important to note that this method requires the attacker to possess the identity annotation of each pedestrian image and multiple positive images per identity, whereas our method does not have such a requirement.

In the "one-to-one" attack scenario, we employ $M_{FE}$, $M_{tloss}$, and $M_U$ as the baselines for comparison. Conversely, in the "one-to-any" attack scenario, where the domain shift exists between attack training and test set, we consider the method $M_{SD}$ and $M_U$ that utilizes our alignment strategies as the attack baselines.

### 5.4 Evaluation Metrics

In our experiments, we employ the attack success rate (ASR) (Shokri et al., 2017) as an evaluation metric, which is defined as the ratio of successful attacks, where members are correctly predicted as members and non-members as non-members, to all unknown attacks. We also generate Receiver Operating Characteristic (ROC) curves to evaluate the trade-off between the true positive rate and false positive rate of the compared methods.

## 6 Experiments

We compare the performance of the proposed method in MI attacks with several baselines on the Re-ID task in two attack scenarios. We also present an ablation study to investigate the impact of different components and hyperparameters on the performance of our method. Finally, we compare the performance of our approaches with SOTA methods on classification tasks.

### 6.1 Performance Comparison

*One-to-One Attack Scenario* Table 2 presents the ASR of our methods and the compared baselines when attacking Re-ID models with different backbones (ResNet50, MobileNetV2, and Xception) trained on different datasets (Market1501 and DukeMTMC) in "one-to-one" attack scenario. Firstly, we observe that our approach, $M_{SD}$, significantly outperforms existing baseline methods in both datasets and across

**Table 4** We compare the performance of the aligned $M_{SD}$ and the patch-attention based $M_{PA+SD}$ (given in $(\cdot)$) in the "one-to-any" attack scenario across four Re-ID models and three datasets in terms of ASR

| Source→Target | MGN→Trans | PCB→Trans | HAN→Trans | Avg |
|---|---|---|---|---|
| Market→Duke | 58.3% (62.4%) | 60.6% (63.1%) | 65.1% (64.7%) | 61.3% (**63.4%**) |
| Market→MSMT | 54.6% (54.0%) | 56.2% (58.9%) | 59.0% (63.3%) | 56.6% (**58.7%**) |
| Duke→Market | 59.8% (61.2%) | 53.3% (59.3%) | 61.1% (61.7%) | 58.1% (**60.7%**) |
| Duke→MSMT | 54.4% (54.5%) | 52.0% (54.2%) | 58.5% (60.8%) | 55.0% (**56.5%**) |
| MSMT→Market | 57.7% (57.9%) | 61.1% (61.2%) | 60.7% (61.2%) | 59.8% (**60.1%**) |
| MSMT→Duke | 57.6% (60.3%) | 58.6% (58.4%) | 64.0% (65.6%) | 60.1% (**61.4%**) |
| Source→Target | Trans→MGN | PCB→MGN | HAN→MGN | Avg |
| Market→Duke | 89.3% (90.1%) | 89.4% (89.6%) | 89.9% (89.8%) | 89.5% (**89.8%**) |
| Market→MSMT | 76.9% (77.7%) | 77.7% (77.8%) | 78.9% (78.8%) | 77.8% (**78.1%**) |
| Duke→Market | 86.7% (87.0%) | 78.0% (82.6%) | 87.8% (87.7%) | 84.2% (**85.8%**) |
| Duke→MSMT | 76.7% (77.4%) | 72.5% (74.1%) | 77.8% (78.7%) | 75.7% (**76.7%**) |
| MSMT→Market | 86.3% (86.9%) | 87.3% (87.4%) | 86.4% (87.1%) | 86.6% (**87.1%**) |
| MSMT→Duke | 89.1% (90.0%) | 88.3% (88.3%) | 89.1% (89.6%) | 88.8% (**89.3%**) |
| Source→Target | Trans→PCB | MGN→PCB | HAN→PCB | Avg |
| Market→Duke | 79.0% (79.5%) | 75.5% (78.4%) | 78.7% (79.0%) | 77.7% (**78.9%**) |
| Market→MSMT | 86.9% (87.6%) | 85.3% (85.9%) | 88.9% (89.1%) | 87.0% (**87.5%**) |
| Duke→Market | 90.6% (90.9%) | 88.6% (90.2%) | 90.7% (91.3%) | 90.0% (**90.8%**) |
| Duke→MSMT | 87.8% (87.5%) | 84.3% (86.3%) | 88.0% (89.0%) | 86.7% (**87.6%**) |
| MSMT→Market | 90.1% (90.7%) | 87.7% (89.6%) | 90.4% (91.3%) | 89.4% (**90.5%**) |
| MSMT→Duke | 79.7% (79.7%) | 68.9% (75.1%) | 79.1% (79.0%) | 75.9% (**77.9%**) |
| Source→Target | Trans→HAN | MGN→HAN | PCB→HAN | Avg |
| Market→Duke | 92.3% (94.8%) | 88.4% (94.1%) | 94.5% (94.8%) | 91.7% (**94.5%**) |
| Market→MSMT | 77.9% (81.3%) | 76.6% (76.6%) | 81.1% (81.3%) | 78.5% (**79.7%**) |
| Duke→Market | 91.8% (92.8%) | 88.7% (91.6%) | 84.5% (88.9%) | 88.3% (**91.1%**) |
| Duke→MSMT | 80.2% (80.6%) | 78.2% (80.2%) | 69.8% (74.6%) | 76.1% (**78.4%**) |
| MSMT→Market | 91.1% (92.7%) | 84.9% (88.1%) | 92.7% (92.6%) | 89.6% (**91.1%**) |
| MSMT→Duke | 92.0% (94.6%) | 88.4% (92.4%) | 94.0% (94.0%) | 91.5% (**93.7%**) |

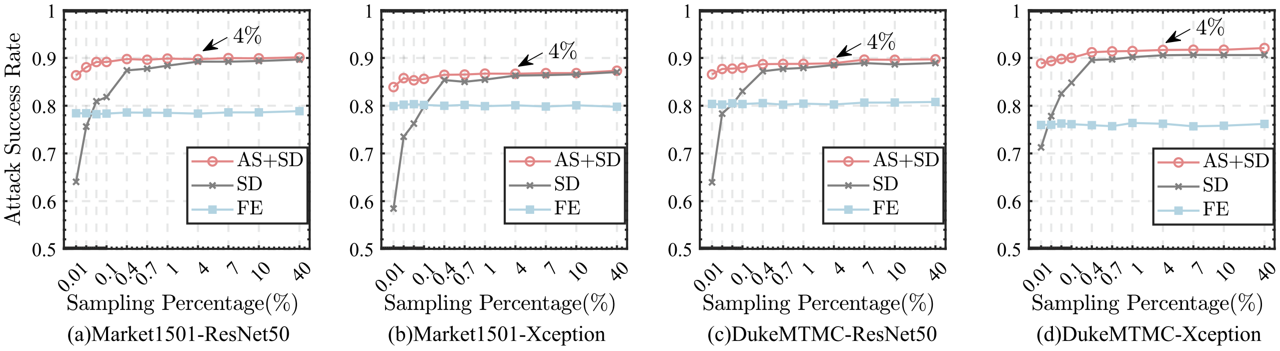
Source→Target means auxiliary model and dataset to target model and dataset. The highest performance is indicated in bold

all three Re-ID backbones, demonstrating the effectiveness of leveraging the relative similarity between samples for membership inference. By introducing an anchor selector, $M_{AS+SD}$ achieves the highest ASR, underscoring the importance of selecting appropriate anchors for different images.

Additionally, we notice that $M_{FE}$ achieves a lower ASR compared to other methods, further validating the assumption that feature embedding contains additional information unrelated to the training data. Moreover, individual feature embedding provides less informative evidence regarding training set membership compared to methods that consider inter-sample similarities. The user-level method $M_U$ also outperforms feature-based methods on Market1501, highlighting the significance of inter-sample relationships. However, this method solely considers the correlation among positive samples, resulting in inferior performance compared to $M_{SD}$ and $M_{AS+SD}$. Furthermore, we observe that the performance of $M_U$ is significantly influenced by the number of positive samples in each class. Specifically, upon examining the results of experiments conducted with $M_{U\_high}$, $M_{U\_mid}$, and $M_{U\_low}$, we observe a decrease in ASR as the number of positive images per identity decreases. In conclusion, compared to our method, the user-level MI attack imposes stricter requirements on background knowledge of the target images, including the identity annotation and a large number of positive images for each identity.

Moreover, we compare our methods and the compared baselines using ResNet50 trained with Market1501 in terms of the ROC curve, as depicted in Fig. 6. Our methods, $M_{AS+SD}$ and $M_{SD}$, outperform the other methods, achieving the highest *Area Under Curve (AUC)* values of 0.935 and 0.930, respectively.

*One-to-Any Attack Scenario* We compare our proposed method $M_{PA+SD}$ with other aligned methods in a more realistic "one-to-any" attack scenario by presenting the attack results across different Re-ID models and datasets in Table 3.



(a)Market1501-ResNet50 (b)Market1501-Xception (c)DukeMTMC-ResNet50 (d)DukeMTMC-Xception

**Fig. 7** Attack success rate with different sampling percentages of sampled anchor images in the reference set for methods $M_{FE}$, $M_{SD}$ and $M_{AS+SD}$ on ResNet50 and Xception backbones trained on Market1501 and DukeMTMC datasets

Firstly, it is important to note that in the absence of our alignment strategies, $M_{SD}$ and $M_U$ will consistently yield poor performance due to the domain-shift problem between the auxiliary and target Re-ID model.

Furthermore, we observed that our method $M_{SD}$ mostly achieves a higher ASR than the method $M_U$ when we apply the alignment strategies. This finding further supports our earlier assertion in Sect. 3 that the similarity distribution shift exhibits the domain-invariant features present in nearly all Re-ID models and datasets that will easily and effectively expose the privacy and security vulnerabilities of real-world Re-ID models.

Moreover, our novel patch-attention-based method $M_{PA+SD}$ surpasses the aligned method $M_{SD}$ in the majority settings, especially in the Avg. column, as shown in Table 4. This highlights the effectiveness of our patch-attention module, which models the inter-relationship of similarity distribution and selects the most crucial anchors in the latent space.

## 6.2 Ablation Study

*Reference Set Sampling* Based on the formal analysis in Sect. 3, we have determined the importance of selecting appropriate reference images as proxy centers to approximate the learned identity centers. In general, with a sufficiently large reference set, it is always possible to find samples that are close enough to the identity center. However, when the reference set is small, there may not be enough samples to accurately approximate the identity centers.

Consequently, Fig. 7 illustrates the impact of the percentage of sampled anchor images in the reference set on the success of the attacks conducted by $M_{SD}$ and $M_{AS+SD}$. Our observations reveal that the attack success rate (ASR) of $M_{SD}$ significantly decreases as the percentage of sampled reference images decreases. Conversely, when the number of anchors is low, $M_{AS+SD}$ surpasses $M_{SD}$ by incorporating an additional anchor selector that assigns higher importance weight to appropriate anchor images. Notably, even when only 4% of the images are sampled, $M_{AS+SD}$ achieves the upper-bound performance, thereby emphasizing the significance of selecting suitable reference images to approximate the identity anchors.

**Table 5** We evaluate $M_{PA+SD}$ for various numbers of latent space dimensions (L-D) and patch dimensions (P-D) in the MGN→Trans and Market→Duke attack setting, where $N$ represents the dimensions of the similarity vector

| L-D | P-D | | |
|---|---|---|---|
| | $N/2$ (%) | $N/4$ (%) | $N/8$ (%) |
| $4N$ | 60.9 | 61.8 | 61.7 |
| $2N$ | 61.5 | **62.4** | 61.5 |
| $1N$ | 61.5 | 61.9 | 61.6 |

The highest performance is indicated in bold

*Dimension in Latent Space and Patch* We conducted ablation experiments on the dimensions of the latent space representation and each patch. As shown in Table 5, it can be observed that the attack achieves optimal performance when the similarity vectors are mapped to a latent space of $2 * N$ dimensions and each patch has a $N/4$ dimension.

## 6.3 Evaluation on Classification

To investigate the performance of our proposed method in tasks beyond Re-ID, we apply it to the classification task and compare its effectiveness with several state-of-the-art MI attack methods in the “one-to-one” attack scenario. We choose CIFAR10 (Krizhevsky, 2009) as our benchmark dataset and evaluate our method on various target models, namely ResNet18 (He et al., 2016), ResNet50 (He et al., 2016), VGG19 (Simonyan & Zisserman, 2014), and GoogLeNet (Szegedy et al., 2015). The target models are trained using the SGD optimizer with a learning rate of 0.1, 200 epochs, and $l_2$ regularization with a weight of 0.0005. The comparison methods consist of the logits-based MI attack $M_{logits}$ (Shokri et al., 2017; Salem et al., 2018), which utilizes the output logits in the attack neural network, the

**Table 6** Performance comparison between the proposed method and existing membership inference attack baselines on different classification models trained on CIFAR10 in terms of attack success rate

| Model | $M_{loss}$ (%) | $M_{logits}$ (%) | $M_{FE}$ (%) | $M_{AS+SD}$ (%) |
|---|---|---|---|---|
| ResNet18 | 78.7 | 78.5 | 78.6 | 79.0 |
| ResNet50 | 69.1 | 67.9 | 66.9 | 68.3 |
| VGG19 | 63.9 | 63.8 | 63.6 | 63.6 |
| GoogLeNet | 67.8 | 65.9 | 63.9 | 66.9 |

feature-based method $M_{FE}$, and the loss-based MI attack $M_{loss}$ (Sablayrolles et al., 2019), which performs MI based on the classification loss using a manually defined threshold.

As shown in Table 6, our algorithm $M_{AS+SD}$ achieves a similar ASR to the previous state-of-the-art algorithm $M_{loss}$ on most target models and a higher ASR on ResNet18. This result demonstrates that the inter-sample similarity also provides sufficient information about the generalization gap between the training and test sets in the classification task.

## 7 Conclusion

This paper highlights a rarely explored privacy risk associated with the training data of person re-identification. Membership inference attacks can quantify the information leakage from Re-ID data. However, Re-ID is a fine-grained recognition task with complex feature embedding, and model outputs commonly utilized by existing MI methods, such as logits and losses, are not accessible during inference. Consequently, this paper conducts formal and empirical analyses to uncover a new set of features for Re-ID MI attacks, namely the inter-sample similarity of image pairs. Therefore, a novel membership inference attack method is proposed in order to quantify the information leakage of the Re-ID dataset by leveraging the inter-sample correlation among pedestrian images. We analyze two attack scenarios "one-to-one" and "one-to-any" to comprehensively understand the privacy risks in Re-ID tasks. In the more realistic "one-to-any" scenario, we introduce two alignment strategies to mitigate the domain-shift problem. Furthermore, in both scenarios, we propose the attention-based module that accurately selects anchors representing the similarity distribution. Our proposed method achieves superior performance compared to existing MI attack approaches when applied to Re-ID models.

**Acknowledgements** This work was supported by the National Natural Science Fund of China (62076184, 61976158, 61976160, 62076182, 62276190), in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University), in part by Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700).

**Data Availability** The datasets used during and analyzed during the current study are available in the following public domain resources: https://www.cs.toronto.edu/kriz/cifar.html; https://zheng-lab.cecs.anu.edu.au/Project/project_reid.html; https://www.pkuvmc.com/dataset.html; The models and source data generated during and analyzed during the current study are available from the corresponding author upon reasonable request.

## References

Bousmalis, K., Silberman, N., Dohan, D., et al. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3722–3731).

Chen, B., Deng, W., & Hu, J. (2019). Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 371–381).

Chen, M., Zhang, Z., Wang, T., et al. (2021). When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security* (pp. 896–911).

Cheng, D., Gong, Y., Zhou, S., et al. (2016). Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1335–1344).

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929

Duan, Y., Lu, J., Feng, J., et al. (2017). Deep localized metric learning. *IEEE Transactions on Circuits and Systems for Video Technology, 28*(10), 2644–2656.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security* (pp. 1322–1333).

Ganin, Y., Ustinova, E., Ajakan, H., et al. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research, 17*(1), 1–35.

Gao, J., Jiang, X., Zhang, H., et al. (2023). Similarity distribution based membership inference attack on person re-identification. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 14,820–14,828).

Gong, B., Grauman, K., & Sha, F. (2014). Learning kernels for unsupervised domain adaptation with applications to visual object recognition. *International Journal of Computer Vision, 109*(1), 3–27.

Hayes, J., Melis, L., Danezis, G., et al. (2017). Logan: Membership inference attacks against generative models. arXiv:1705.07663

He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, S., Luo, H., Wang, P., et al. (2021). Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15,013–15,022).

Hu, H. M., Fang, W., Zeng, G., et al. (2017). A person re-identification algorithm based on pyramid color topology feature. *Multimedia Tools and Applications, 76*(24), 26,633-26,646.

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*. Master's Thesis, University of Tront.

Li, G., Rezaei, S., & Liu, X. (2022). User-level membership inference attack against metric embedding learning. arXiv:2203.02077

Li, J., Li, N., & Ribeiro, B. (2020). Membership inference attacks and defenses in supervised learning via generalization gap. arXiv:2002.12062.

Li, W., Zhu, X., & Gong, S. (2018). Harmonious attention network for person re-identification. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 2285–2294).

Liu, W., Wen, Y., Yu, Z., et al. (2016). Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd international conference on international conference on machine learning—Volume 48. JMLR.org, ICML'16* (pp. 507–516).

Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10,012–10,022).

Long, Y., Bindschaedler, V., Wang, L., et al. (2018). Understanding membership inferences on well-generalized learning models. arXiv:1802.04889

Ming, Z., Zhu, M., Wang, X., et al. (2022). Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing, 119*(104), 394.

Nasr, M., Shokri, R., & Houmansadr, A. (2018a). Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE symposium on security and privacy (SP)* (pp. 1–15).

Nasr, M., Shokri, R., & Houmansadr, A. (2018b). Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security* (pp. 634–646).

Oh Song, H., Xiang, Y., Jegelka, S., et al (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4004–4012).

Ranjan, R., Castillo, C. D., & Chellappa, R. (2017). L2-constrained softmax loss for discriminative face verification. arXiv:1703.09507

Sablayrolles, A., Douze, M., Schmid, C., et al. (2019). White-box vs black-box: Bayes optimal strategies for membership inference. In *International conference on machine learning, PMLR* (pp. 5558–5567).

Salem, A., Zhang, Y., Humbert, M., et al. (2018). Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv:1806.01246

Sandler, M., Howard, A., Zhu, M., et al. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Sharma, C., Kapil, S. R., & Chapman, D. (2021). Person re-identification with a locally aware transformer. arXiv:2106.03720

Shokri, R., Stronati, M., Song, C., et al. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3–18). IEEE.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556

Song, L., Shokri, R., & Mittal, P. (2019). Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security* (pp. 241–257).

Sun, Y., Zheng, L., Yang, Y., et al. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)* (pp. 480–496).

Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 1–9).

Wang, G., Yuan, Y., Chen, X., et al. (2018a). Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on multimedia* (pp. 274–282).

Wang, H., Zhu, X., Gong, S., et al. (2018b). Person re-identification in identity regression space. *International Journal of Computer Vision, 126*, 1288–1310.

Wei, L., Zhang, S., Gao, W., et al. (2018). Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 79–88).

Wu, X., Fredrikson, M., Jha, S., et al. (2016). A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th computer security foundations symposium (CSF)* (pp. 355–370). IEEE.

Yang, F., Yan, K., Lu, S., et al. (2019). Attention driven person re-identification. *Pattern Recognition, 86*, 143–155.

Yeom, S., Giacomelli, I., Fredrikson, M., et al. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. In*2018 IEEE 31st computer security foundations symposium (CSF)* (pp. 268–282). IEEE.

Yin, J., Wu, A., & Zheng, W. S. (2020). Fine-grained person re-identification. *International Journal of Computer Vision, 128*, 1654–1672.

Yu, D., Zhang, H., Chen, W., et al. (2021). How does data augmentation affect privacy in machine learning? In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10,746–10,753).

Zhang, S., Chen, D., Yang, J., et al. (2021). Guided attention in CNNs for occluded pedestrian detection and re-identification. *International Journal of Computer Vision, 129*, 1875–1892.

Zheng, F., Deng, C., Sun, X., et al. (2019). Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8514–8522).

Zheng, L., Shen, L., Tian, L., et al. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision* (pp. 1116–1124).

Zheng, L., Yang, Y., & Hauptmann, A. G. (2016). Person re-identification: Past, present and future. arXiv:1610.02984

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision* (pp. 3754–3762).

Zhou, K., Yang, Y., Cavallaro, A., et al. (2019). Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3702–3712).

Zhu, Z., Jiang, X., Zheng, F., et al. (2020). Aware loss with angular regularization for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 13,114–13,121).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.