Rank-in-Rank Loss for Person Re-identification

XIN XU and XIN YUAN, Wuhan University of Science and Technology, China
ZHENG WANG, Wuhan University, China
KAI ZHANG, Wuhan University of Science and Technology, China
RUIMIN HU, Wuhan University, China

Person re-identification (re-ID) is commonly investigated as a ranking problem. However, the performance of existing re-ID models drops dramatically, when they encounter extreme positive-negative class imbalance (e.g., very small ratio of positive and negative samples) during training. To alleviate this problem, this article designs a rank-in-rank loss to optimize the distribution of feature embeddings. Specifically, we propose a Differentiable Retrieval-Sort Loss (DRSL) to optimize the re-ID model by ranking each positive sample ahead of the negative samples according to the distance and sorting the positive samples according to the angle (e.g., similarity score). The key idea of the proposed DRSL lies in minimizing the distance between samples of the same category along with the angle between them. Considering that the ranking and sorting operations are non-differentiable and non-convex, the DRSL also performs the optimization of automatic derivation and backpropagation. In addition, the analysis of the proposed DRSL is provided to illustrate that the DRSL not only maintains the inter-class distance distribution but also preserves the intra-class similarity structure in terms of angle constraints. Extensive experimental results indicate that the proposed DRSL can improve the performance of the state-of-the-art re-ID models, thus demonstrating its effectiveness and superiority in the re-ID task.

CCS Concepts: • **Information systems → Information retrieval;**

Additional Key Words and Phrases: Person re-identification, metric learning, loss function

ACM Reference format:
Xin Xu, Xin Yuan, Zheng Wang, Kai Zhang, and Ruimin Hu. 2022. Rank-in-Rank Loss for Person Re-identification. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 2s, Article 130 (October 2022), 21 pages.
https://doi.org/10.1145/3532866

1 INTRODUCTION

Person **re-identification** (**re-ID**) is the task of retrieving a person of interest across images/videos captured from multiple non-overlapping cameras at different times or places [10, 50, 67, 70]. Existing re-ID models have achieved significant development with the help of feature representation learning [83] and loss function design [10, 68]. The former focuses on designing superior network architecture to extract discriminative person features [9]; while the latter tries to leverage a metric

This work was supported by National Nature Science Foundation of China (No. U1803262, 62171325, 62176191).
Authors' addresses: X. Xu, X. Yuan, and K. Zhang, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China; emails: {xuxin, xinyuan, zhangkai}@wust.edu.cn; Z. Wang and R. Hu, Wuhan University, Wuhan, Hubei 430072, China; emails: wangzwhu@whu.edu.cn, hrm1964@163.com.

1551-6857/2022/10-ART130 $15.00
https://doi.org/10.1145/3532866

loss function to optimize the feature distribution [23]. Both of these two paradigms aim at minimizing the intra-class feature distance among the samples with the same identity while maximizing the inter-class feature distance among the samples with different identities in the feature space.

Despite the success of feature representation learning, the re-ID datasets contain various influencing factors, such as person pose variations, illumination differences, occlusions, and so on. These factors pose great challenges to current feature representation learning based re-ID models. Recently, loss function design [14, 17, 46] has demonstrated its effectiveness to cope with these influencing factors. Therefore, this article focuses on the loss function design in person re-ID. It is worth noting that person re-ID is a sub-research area of the retrieval task [66] and is usually formulated as a ranking problem. Intuitively, it is natural to use the ranking loss to train the person re-ID models. However, the performance of these models drops dramatically, when they encounter extreme positive-negative class imbalance during training. To tackle this problem, **average precision** (**AP**) based loss functions [4, 6, 13, 42] are proposed recently. The AP-based loss functions have two advantages: (1) optimizing the AP directly provides consistency between training and evaluation objectives; (2) being robust to class imbalance. However, current AP-based loss functions mainly rank positive samples ahead of negative samples according to the distance in feature embedding space, the angle relationship among positive samples is ignored.

Moreover, we have noticed that the distribution of positive samples (i.e., intra-class sample distribution) is not considered in previous AP-based loss functions [4, 6, 13, 42]. Although these methods are effective in ranking positive samples ahead of negative samples according to the distance, the positive sample distribution has large randomness. Consequently, the distribution of positive samples may be compact or loose. In this case, these AP-based methods cannot guarantee to explicitly preserve the effective distribution of positive samples. Recently, Wang et al. [56] pointed out the importance of constraining the distribution of positive samples for the similarity structure. To explicitly preserve the intra-class similarity structure within each class, Wang et al. [56] introduced the ranked list loss to force the distance of a positive pair smaller than a threshold, which achieved the state-of-the-art performance. Therefore, it is necessary to consider the similarity structure within positive samples while ranking positive samples ahead of negative samples. Meanwhile, considering that the positive samples in feature space are not only constrained by the relationship of distance among them but also have the angle relationship dependence to each other. To preserve the similarity within positive samples, this article considers the angle relationship to constrain the distribution of positive samples.

Based on this motivation, in this article, we introduce a sorting operation among positive samples in AP Loss to consider the angle relationship through the similarity scores and design a rank-in-rank loss function to preserve a more effective similarity structure. Considering both ranking and sorting operations are non-differentiable and non-convex, it is difficult for them to perform the optimization with the commonly used gradient descent method directly. Therefore, we propose a ranking-based loss function called **Differentiable Retrieval-Sort Loss** (**DRSL**), which supervises the person re-ID model to rank each positive ahead of all negative samples according to the distance (Figure 1(a)) as well as to sort positive samples according to the angle, e.g., similarity score (Figure 1(b)). The proposed DRSL minimizes the distance between samples of the same category along with the angle between them and thus can deal with extreme positive-negative class imbalance during training through a sampling strategy in small batches. Moreover, DRSL does not need any hyper-parameter tuning and can be easily implemented into many re-ID models.

In summary, the main contribution of this article is threefold:

— To the best of our knowledge, we are among the first attempts to apply AP-based loss function to the re-ID task, and introduce the sorting operation to model intra-class similarity.

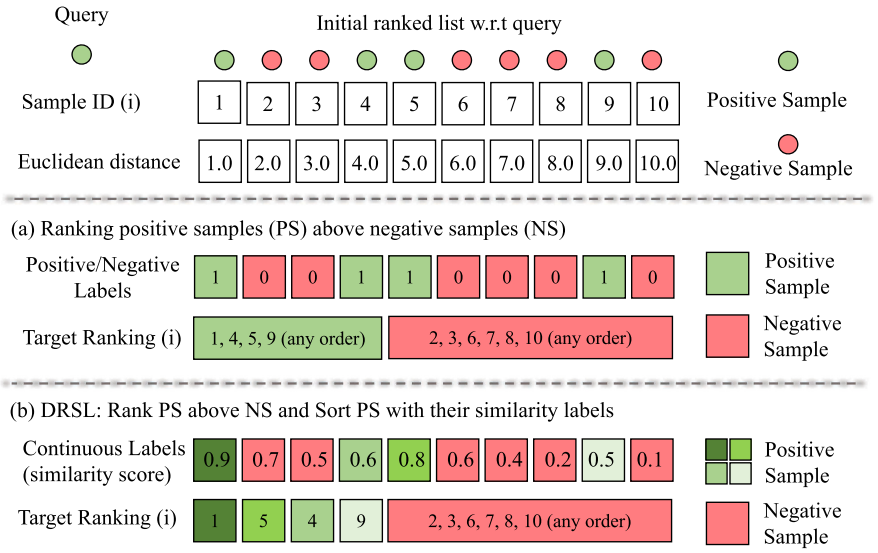

Fig. 1. Comparison with the ranking-based loss function. Two main components are included in the figure: (i) positive samples: green circle w.r.t the query; (ii) negative samples: red circle w.r.t the query. Given a query and its initial ranked list, (a) ranking positive samples ahead of negative samples through the distance delivers an effective strategy for model optimization. However, it neglects the angle relationship between positives reflected in similarity. (b) our proposed DRSL not only tackles the case of (a) but also considers the angle relationship (see Figure 2 (i)) by sorting positives w.r.t their corresponding continuous similarity scores. We replace the Heaviside step function with a differentiable approximate step function that overcomes the non-differentiable and non-convex nature of ranking and sorting operations. *Best viewed in color.*

- —The proposed DRSL defines a ranking objective between positive and negative samples as well as a sorting objective to prioritize positive samples w.r.t their similarity scores.
- —To cope with the non-differentiable nature of ranking and sorting operations, we propose a differentiable approximate step function to perform the optimization of automatic derivation and backpropagation during the training phase.

The rest of the article is organized as follows: Section 2 discusses related work, Section 3 briefly reviews loss functions commonly used in person re-ID and ranking problems, the proposed methodology is provided in Section 4, then Section 5 presents the experiments along with the discussion and analysis, finally Section 6 concludes this article.

2 RELATED WORK

In this section, we first review some relevant works for person re-ID in Section 2.1. Then, Section 2.2 describes some loss functions commonly used in metric learning and discusses and analyzes the widely used losses in person re-ID. Finally, Section 2.3 discusses the optimization of AP for vision tasks, such as object detection, information retrieval, and image retrieval.

2.1 Person Re-identification

Person re-ID is a sub-task of image retrieval [50, 55, 63, 70], which has evolved from multi-target tracking [41, 54, 71]. It has achieved significant progress in recent years and becomes a hot research topic in the computer vision community. The objective of person re-ID is to learn a model with the robust discriminative ability to overcome real-world disturbances caused by variations in camera

parameter setting, viewing point, human pose, lighting, occlusions, and so on. Existing studies on person re-ID mainly focus on improving the discriminative ability of the re-ID model from three perspectives: (1) feature representation learning [9, 49], which extracts global feature, local feature, and auxiliary feature of pedestrian; (2) loss function design [18, 43], it is crucial to optimize the feature space distribution; (3) architecture design [7, 83], it aims at achieving better re-ID features.

For feature representation learning, various person re-ID models have been proposed to extract a global feature vector for each person in the early years. Wang et al. [52] proposed a joint learning framework that contains two parts, namely representation (SIR) and **cross-image representation** (**CIR**) to exploit the fine-grained cues. Zheng et al. [79] treated each person's **identity** (**ID**) as a class and used the person ID as the label to supervise the network training. This network was named **ID-discriminative Embedding** (**IDE**) model and later became the widely-used baseline in re-ID community [49, 81, 82]. Furthermore, some works introduced the attention mechanism to enhance feature learning [9, 27, 61, 65]. However, the problem of misalignment caused by occlusions, different human poses [48, 49], and so on, posed great challenges to the learning of global features. To tackle these challenges, several works attempted to learn part/region aggregated features (i.e., local features) by using human parsing/pose estimation or roughly horizontal division [20, 75]. Recently, both global and local features were considered simultaneously to obtain multi-granularity feature representation [2, 53]. Besides, the auxiliary feature learned from attribute [29], dense semantics [75], and pose [77] are usually used to reinforce the feature representation. Moreover, Wu et al. [62] attempted to learn discriminative and view-invariant representations [60] from limited labeled training samples due to the expensive cost of exhaustively labeling. It is worth noting that most of the above works directly used the network designed for image classification [16].

For the loss function design, the identity loss [79], verification loss [17], and triplet loss [46] are widely used loss functions in person re-ID. Generally, most of the existing re-ID models use identity loss and triplet-based loss for model training. The typical formulation of identity loss is the softmax cross-entropy usually used for classification tasks. While the triplet loss and its variants [14, 23, 72, 73] treat person re-ID as a ranking task. Besides, several works combine the verification loss with the identity loss to optimize the re-ID performance [17, 80]. Recently, the center loss [59] is introduced to improve the intra-class aggregability.

For the architecture design, many existing architectures [7, 28, 83] often adopt the modified ResNet-50 originally designed for image classification as the backbone. These architectures are made to overcome some problems in person re-ID, such as person pose variations, illumination differences, occlusions, and so on. Recently, some efforts designing the specific re-ID network architecture mainly focus on the two major concerns, e.g., accuracy [7, 21, 26] and efficiency [83]. To improve the accuracy, Li et al. [26] designed the **filter pairing neural network** (**FPNN**) to handle the problem of misalignment and occlusion. Chang et al. [7] proposed the **multi-Level factorisation net** (**MLFN**) composed of multiple stacked blocks to model latent factors at a specific level, which concentrates on modeling discriminative and view-invariant factors of person appearance at both high and low semantic levels. To increase the accuracy, similar to MobileNet [15] from the viewpoint of lightweight network design, Zhou et al. [83] raised the efficient small scale network called **omni-scale network** (**OSNet**) to achieve multi-scale feature learning. Notably, with the increasing interest in **neural architecture search** (**NAS**), Auto-ReID [40], and CDNet [24] are designed to provide an efficient and effective automated neural architecture design strategy.

All the above three aspects have a common objective, which is to extract robust and discriminative features. It is worth noting that one key issue during the design of these models is metric learning, which aims at optimizing the similarity or distance between feature vectors extracted from samples.

2.2 Metric Learning

The objective of metric learning is to construct an effective feature space that is robust to a given task. The main idea is to construct an effective feature embedding distribution to establish similarity or dissimilarity between feature vectors [14, 17, 33, 34, 46, 47, 56]. For example, Liu et al. [31] and Schroff et al. [46] imposed a margin between positive and negative samples to improve robustness. To explore richer structural relationships, N-pair [47] and Proxy NCA [33] built an (N+1)-tuple all together to jointly maximize multi-class distance. However, the above methods do not fully explore the information of all samples, i.e., only part of their information is used. To fully exploit sample information, Wang et al. [56] proposed Rank List loss to exploit the distribution structure of all samples (e.g., all positive and negative samples to query), which contained richer information than the above methods. Through continuous exploration and research in the past years, significant progress has been made in the field of metric learning.

For person re-ID, the widely used loss functions include softmax cross-entropy loss [49], contrastive loss [17], triplet loss [46], and center loss [59]. In the early years, the training process of person re-ID was treated as image classification, using softmax cross-entropy loss to optimize the similarity between samples and weight vectors. Meanwhile, person re-ID was also investigated as a person verification problem [57, 64], i.e., discriminating whether two-person images have the same identity or not. These works [1, 11, 80] optimize the pairwise relationship with the contrastive loss. In addition, the process of person re-ID can also be viewed as a retrieval ranking problem. The commonly used triplet loss in person re-ID imposes a distance ranking between the positive sample pair and the negative sample pair within a triplet. However, intra-class compactness in triplet loss is ignored. In order to solve the drawbacks of the triplet loss, Luo et al. [32] imposed the center loss to increase intra-class compactness. To improve the robustness of a metric distance under adversarial attacks, Liu et al. [30] developed a generative metric learning approach. The design and utilization of these loss functions have greatly facilitated the development of person re-ID.

Despite their remarkable success on many computer vision tasks, current loss functions for the re-ID task easily focus on optimizing the order of positive samples with low rank at the expense of high-rank examples. Besides, they ignore the importance of shifting ranking orders and are less informative. Therefore, most of current loss functions actually optimize a distance metric rather than a ranking metric (e.g., the AP commonly used in person re-ID).

2.3 Optimizing AP

The AP is an important evaluation metric in the computer vision community, e.g., object detection, information retrieval, and image retrieval. As a result, numerous works designed various loss functions to maximize the AP directly. Here, we revisit the typical and recent works as follows.

Object Detection. To alleviate the extreme foreground-background class imbalance problem, Chen et al. [12] replaced the classification task with a ranking task and adopted the AP-loss for the ranking problem. Inspired by this, Oksuz et al. [35] not only revealed the limitations of AP in the measurement of target detection tasks but also proposed a new metric, namely **Localisation-Recall-Precision** (**LRP**). Then, the **average Localisation-Recall-Precision** (**aLRP**) Loss was proposed both for classification and localization tasks [36]. Recently, Oksuz et al. [37] proposed **Rank & Sort** (**RS**) Loss to explicitly model positive-to-positive interactions.

Information Retrieval. The information retrieval community has made significant progress in optimizing AP [5, 8, 25, 39]. To closely match test-time performance measures, He et al. [22] proposed the tie-aware ranking metrics for hashing, aiming at directly optimizing ranking-based metrics such as AP and **Normalized Discounted Cumulative Gain** (**NDCG**). Similarly, rank fusion combined multiple sources of information into a single result set to boost retrieval performance [3].

Image Retrieval. There are many efforts [6, 42, 56] that mainly focus on the ordering of images to optimize AP. However, due to the non-differentiability and non-convexity of AP, its optimization is highly challenging. To cope with this challenge, recently introduced Listwise [42], FastAP [6], and Blackbox AP [44] linearly interpolated the non-differentiable (piecewise constant) function optimized by differentiable histogram binning [56] or error-driven update [13]. However, these methods can cause a similar problem with the triplet loss, i.e., optimizing the distance metric instead of the ranking metric. Motivated by this, Brown et al. [4] proposed Smooth-AP to provide a closer approximation to AP by the sigmoid function.

Although the above loss functions have achieved remarkable performance improvement, they ignore the angle relationship among intra-class samples which is essential for the similarity metric. In this article, we raise the intra-class angle relationship issue in the design of loss function to optimize the AP directly.

3 ANALYSIS OF LOSS FUNCTIONS

In this section, we review and analysis the loss functions commonly used in metric learning. Here, the training set is represented as $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_\mathcal{M}\}$ with $\mathcal{M}$ person images containing $\mathcal{N}$ classes, i.e., $\mathcal{N}$ identities. The training image $\mathcal{I}_i$ can be represented by the feature embedding $\mathcal{X}_i$ with annotated label $\mathcal{Y}_i$.

3.1 Loss Functions in Person re-ID

Contrastive Loss [17] aims at maximizing or minimize the relative pairwise distance between a pair of samples.

$$\mathcal{L}_{constractive} = \gamma_{i,j} d_{i,j} + \left(1 - \gamma_{i,j}\right) \left[m_{con} - d_{i,j}\right]_+, \tag{1}$$

where $d_{i,j} = ||\mathcal{X}_i - \mathcal{X}_j||_2$ is the Euclidean distance. $\gamma_{i,j}$ is the label indicator, e.g., $\gamma_{i,j} = 1$ when $\mathcal{X}_i$ and $\mathcal{X}_j$ are the same class, otherwise $\gamma_{i,j} = 0$. m_{con} means a margin parameter. $[x]_+$ is $max(x, 0)$.

Triplet Loss [46] has a series of triplets $\{\mathcal{X}_i, \mathcal{X}_j, \mathcal{X}_k\}$, where anchor $\mathcal{X}_i$ and positive $\mathcal{X}_j$ are the same class, $\mathcal{X}_k$ belongs to another class. The Triplet Loss is used to pull the distance $d_{i,j}$ while pushing the distance $d_{i,k}$ simultaneously to keep a fixed margin m_{tri}.

$$\mathcal{L}_{triplet} = \left[d_{i,j} - d_{i,k} + m_{tri}\right]_+. \tag{2}$$

Quadruplet Loss [14] also has a negative sample different from Triplet Loss, i.e., $\mathcal{X}_m$ and $\mathcal{X}_n$.

$$\mathcal{L}_{quadruplet} = \left[d_{i,j} - d_{i,m} + m_1\right]_+ + \left[d_{i,j} - d_{i,n} + m_2\right]_+, \tag{3}$$

where m_1 and m_2 correspond to the two margins.

Center Loss [59] can compensate for intra-class aggregability and inter-class distinguishability. The Center Loss effectively reflects the variation within the intra-class.

$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^{\mathcal{B}} ||\mathcal{X}_i - C_{\mathcal{Y}_i}||_2^2, \tag{4}$$

where $\mathcal{B}$ is the batch size number and $C_{\mathcal{Y}_i}$ represents the $\mathcal{Y}_i$th class center of feature embeddings in a mini-batch.

3.2 Ranking Based Loss Functions

N-pair [47] considers the influence of negative samples and classes. The N-pair attempts to find out a positive sample from $\mathcal{N}$ samples of $\mathcal{N}$ classes, i.e., one positive sample from the positive

class and $\mathcal{N}-1$ negative samples from the negative class (each negative class corresponds to one negative sample).

$$\mathcal{L}_{N-pair} = \log\left\{1 + \sum_{k\neq i, k\neq j} \exp\left(\mathcal{X}_i^\top \mathcal{X}_k - \mathcal{X}_i^\top \mathcal{X}_j\right)\right\}, \tag{5}$$

where $\mathcal{X}_i$ and $\mathcal{X}_j$ denote the anchor and positive sample, respectively. $\{\mathcal{X}_k, k \neq i, k \neq j\}$ are the negative samples.

Proxy-NCA [33] allocates a proxy to each class for the sampling problem. The proxy denotes the closest point to all samples with the same class label.

$$\mathcal{L}_{Proxy-NCA} = \log \frac{\exp\left(-d_{i,j}\right)}{\sum_{k\in\mathbf{K}} \exp\left(-d_{i,k}\right)}, \tag{6}$$

where $\mathbf{K}$ indicates the proxy set of negatives, $d_{i,j}$ and $d_{i,k}$ are the Euclidean distance with respect to the anchor and the proxies of positive and negative samples, respectively.

Lifted-Structure [34] pulls one positive sample as close as possible to the anchor sample while pushing all negative samples.

$$\mathcal{L}_{Lifited} = \frac{1}{2|\mathbf{P}|} \sum_{(i,j)\in\mathbf{P}} \left\{ d_{i,j} + \log\left\{ \sum_{(i,k)\in\mathbf{N}} \exp\left(m - d_{i,k}\right) + \sum_{(i,l)\in\mathbf{N}} \exp\left(m - d_{i,l}\right) \right\} \right\}_+, \tag{7}$$

where $\mathbf{P}$ and $\mathbf{N}$ represent the set of positive and negative pairs, respectively. $\{\cdot\}$ is the hinge function.

Ranked List [56] utilizes not only all negative samples, but also all positive ones.

$$\mathcal{L}_{RankedListed} = \lambda \sum_{k\neq i, k\neq j} \frac{w_{i,k}^n}{\sum_{k\neq i, k\neq j} w_{i,k}} L_{\mathrm{m}}\left(\mathcal{X}_i, \mathcal{X}_k\right) + \frac{1}{|\mathcal{P}|} \sum_{j\neq i} L_{\mathrm{m}}\left(\mathcal{X}_i, \mathcal{X}_j\right), \tag{8}$$

where λ constraints on the balance among positive and negative sets. $w_{i,k}^n = \exp\left(T \cdot (\alpha - d_{i,k})\right)$ and T are the temperature parameter that determines the weighted degree of negative examples. In addition, $L_{\mathrm{m}}(\mathcal{X}_i, \mathcal{X}_j) = (1 - y_{i,j})[\alpha - d_{i,j}]_+ + y_{i,j}[d_{i,j} - (\alpha - m)]_+$. Where $y_{i,j}$ is 1 if $\mathcal{Y}_i = \mathcal{Y}_j$; otherwise, its value is 0.

3.3 AP Based Loss Functions

Inspired by image retrieval, many works were proposed to directly optimize the AP through the ranking list of queries [4, 6, 42]. We review and analyze the representative loss functions as follows.S

FastAP [6] approximately optimizes AP via distance quantization to reduce the complexity and improve the efficiency.

$$\text{FastAP} = \sum_{z\in Z} \frac{F\left(z \mid \mathcal{R}^+\right) P\left(\mathcal{R}^+\right)}{F(z)} P\left(z \mid \mathcal{R}^+\right), \tag{9}$$

where $\mathcal{R}^+, (\mathcal{R}^-) \subset \mathcal{R}$ denotes the (non) neighbor set of query. $F(z) = P(\mathcal{Z} < z)$ denote the cumulative distribution function for $\mathcal{Z}$. $P(\mathcal{R}^+)$ denote prior probabilities. The finite set $Z = \{z_1, z_2, \ldots, z_L\}$ is quantized in the interval $[0, 2]$.

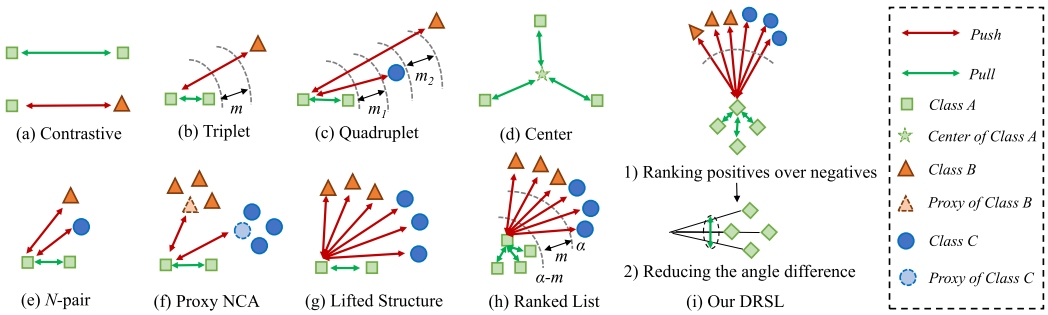

Fig. 2. Comparison among commonly used loss functions in re-ID model ((a)–(d)) and ranking-based loss functions ((e)–(h)). Different shapes (square, triangle, and circle) represent different classes. All shapes are feature embedding vectors in a min-batch, where dashed ones present proxies. It should be noting that all loss functions have the same objective, i.e., to pull the intra-class distance while pushing the inter-class distance. (a) The contrastive loss associates different pairs to separate different classes, but it limits discriminability for multiple classifications. (b) The triplet loss and (c) The quadruplet loss both associate the anchor with positive and negative samples without considering their hardness. (d) The purpose of the center loss is to increase the intra-class compactness. (e) The N-pair loss, (f) the proxy NCA loss, and (g) the lifted structure loss shares a common characteristic in that one positive sample and multiple negative samples from multiple classes are considered simultaneously. For the N-pair loss, there is only one sample from each class, whether positive or negative. The proxy NCA loss considers the relationship between the anchor and negative proxies rather than negative samples. Lifted Structure loss considers all negative samples. However, these three loss functions fail to utilize all data in the batch. (h) The ranked list loss not only exploits all negative examples but also makes use of all positive ones. (i) In addition to considering all positive and negative samples, our DRSL further explores the angle relationship among intra-class samples which is essential for the similarity metric. *Best viewed in color.*

Listwise [42] takes advantage of the recent listwise loss to use the histogram binning method to reformulate AP [51].

$$\mathrm{mAP_Q}(D, Y) = \frac{1}{B}\sum_{i=1}^{B} \mathrm{AP_Q}\left(d_i^{\top} D, Y_i\right), \tag{10}$$

where $\mathrm{AP_Q}$ is the quantized AP, and the loss function is defined as $\mathcal{L}_{Listwise} = 1 - \mathrm{mAP_Q}(D, Y)$.

Smooth-AP [4] optimizes a smoothed approximation of AP.

$$\mathcal{L}_{Smooth-AP} = \frac{1}{m}\sum_{k=1}^{m} (1 - AP_k), \tag{11}$$

where m is the number of a retrieval set.

Figure 2 illustrates and compares the above loss functions widely used in person re-ID tasks and ranking tasks. The proposed DRSL is a ranking-based loss function, which uses all samples in a batch to directly optimize the ranking metric (e.g., AP) instead of the distance metric. Besides, the DRSL considers the intra-class angle relationship to preserve the similarity structure.

4 METHODOLOGY

Using a ranking-based loss function is attractive in many computer vision tasks, e.g., triplet-based loss in face recognition and image retrieval tasks. In recent years, some researchers [6, 42] have designed loss functions directly based on the metrics for performance measures. However, it is difficult for ranking operation due to its non-convexity and non-differentiability properties.

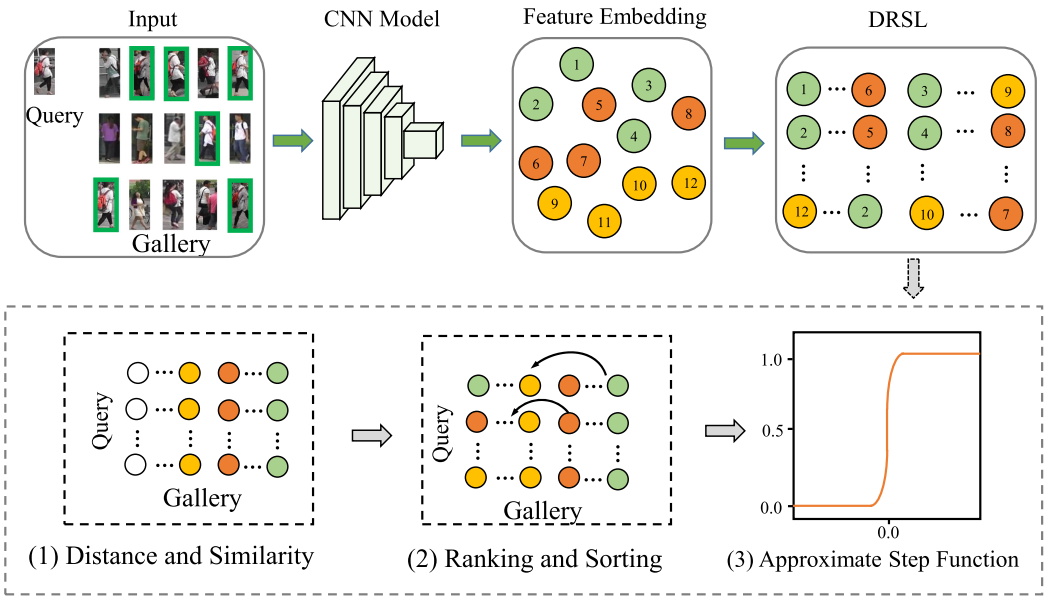

Fig. 3. The framework of the proposed DRSL. It simultaneously considers all the person images in a batch and optimizes the ranking and sorting relationships of these images in the feature space, thus directly improving the AP computed from these images.

Figure 3 shows the framework of the proposed DRSL. Below, the definition of the person re-ID task and some notations are firstly introduced. Then, the details of the **retrieval precision** (**RP**) loss and sort precision loss involved in the proposed DRSL are presented. Finally, we describe the differentiable approximate step function and provide the final formulation of **Differentiable Retrieval-Sort loss** (**DRSL**). Note that all operations and modifications are made on the loss function of the person re-ID model without changing the backbone architecture and other branches.

4.1 Person re-ID Task

During the person re-ID training process, given the training set $I = \{I_1, I_2, \ldots, I_N\}$ with N person images containing M classes, i.e., M person identities. Each image in the training set I can be used as a query for the person re-ID. For each query instance I_q, the rest of the training set I with $N-1$ images can be regarded as a gallery $\mathcal{G}_q$ w.r.t the query I_q and the gallery $\mathcal{G}_q$ can be divided into the positive $\mathcal{P}_q$ (e.g., the same identity with I_q) and negative $\mathcal{N}_q$ (e.g., a different identity to I_q) sets. Note that different query corresponds to different positive and negative set.

In our framework, each training person image $I_i \in I$ is represented by a feature embedding $\mathcal{X}_i$ with label $\mathcal{Y}_i \in Y, Y = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_M\}$. During training and testing phase, the output of CNN model in Figure 3 is a feature embedding set $(\mathcal{X}_q^1, \mathcal{X}_q^2, \ldots, \mathcal{X}_q^{N-1})$ for each query I_q. It is worth noting that for each sample $\mathcal{G}_q^j$ in the gallery $\mathcal{G}_q$ w.r.t the query I_q^i will be assigned a binary label $y_{i,j} \in \{0, 1\}$ depending on whether $\mathcal{G}_q^j$ is the same identity with I_q^i or not.

The person re-ID task indicates that positive samples should be ranked ahead of negative samples according to their distance. In this article, we directly optimize the evaluation metric (e.g., the AP evaluation metric in re-ID), thus providing the consistency between training and evaluation objectives.

4.2 Retrieval Precision Loss

AP is one of the most commonly used evaluation metrics in various computer vision tasks. In the proposed DRSL, the optimization is also conducted based on AP, and we call it RP. Next, transformation operations are required for our RP. For a query I_q^i, the distance set of all gallery $\mathcal{G}_q$ needs to be computed. In our design, we use the Euclidean distance (other distance metrics can also be used).

$$\mathcal{D}_{\mathcal{G}} = \left\{ d_q^{i,j} = ||\mathcal{X}_q^i - \mathcal{X}_q^j||_2, j = 1, \ldots, N-1 \right\}, \tag{12}$$

where $\mathcal{D}_{\mathcal{G}} = \mathcal{D}_P \cup \mathcal{D}_N$, and $\mathcal{D}_P = \{d_q^{i,\psi}, \forall \psi \in \mathcal{P}_q\}$, $\mathcal{D}_N = \{d_q^{i,\sigma}, \forall \sigma \in \mathcal{N}_q\}$ are the positive and negative distance sets, respectively, $\mathcal{X}_q^i$ refers to the query feature embedding, and $\mathcal{X}_q^j$ refers to gallery feature embedding. Then, the difference transformation transfers the distance $d_q^{i,j}$ to the difference form

$$\forall j,k, \quad z_{j,k} = -\left(d_q^{i,j} - d_q^{i,k}\right). \tag{13}$$

Then, we define $\Gamma(\cdot)$ as the primary term of the RP-loss

$$\Gamma_{j,k}(\mathbf{z}) = \frac{\mathcal{H}(z_{j,k})}{1 + \sum_{l \in \mathcal{P} \cup \mathcal{N}, l \neq k} \mathcal{H}(z_{j,l})} = \Gamma_{j,k}, \tag{14}$$

where $\mathcal{H}(\cdot)$ is the Heaviside step function:

$$\mathcal{H}(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}. \tag{15}$$

Next, the RP of query I_q^i can be formulated as

$$RP_q^i = \frac{1}{|\mathcal{D}_P|} \sum_{j \in \mathcal{D}_P} \frac{\mathcal{R}(j, \mathcal{D}_P)}{\mathcal{R}(j, \mathcal{D}_{\mathcal{G}})}, \tag{16}$$

where $\mathcal{R}(j, \mathcal{D}_P)$ and $\mathcal{R}(j, \mathcal{D}_{\mathcal{G}})$ refer to the rankings of the query I_q^i in $\mathcal{P}$ and $\mathcal{G}$, respectively. The ranking function $\mathcal{R}(\cdot, \cdot)$ can be formulated as

$$\mathcal{R}(j, \mathcal{D}) = 1 + \sum_{k \in \mathcal{D}, k \neq j} \mathcal{H}\left(z_{j,k}\right). \tag{17}$$

Finally, the RP-loss $\mathcal{L}_{RP}$ can be formulated as

$$\begin{aligned} \mathcal{L}_{RP} &= 1 - RP_q^i = 1 - \frac{1}{|\mathcal{D}_P|} \sum_{j \in \mathcal{D}_P} \frac{\mathcal{R}(j, \mathcal{D}_P)}{\mathcal{R}\left(j, \mathcal{D}_{\mathcal{G}}\right)} \\ &= 1 - \frac{1}{|\mathcal{D}_P|} \sum_{j \in \mathcal{D}_P} \frac{1 + \sum_{k \in \mathcal{D}_P, k \neq j} \mathcal{H}\left(z_{j,k}\right)}{1 + \sum_{k \in \mathcal{D}_{\mathcal{G}}, k \neq j} \mathcal{H}\left(z_{j,k}\right)} \\ &= \frac{1}{|\mathcal{D}_P|} \sum_{j \in \mathcal{D}_P} \sum_{k \in \mathcal{D}_N} \Gamma_{j,k} = \frac{1}{|\mathcal{D}_P|} \sum_{j,k} \Gamma_{j,k} \cdot y_{j,k} \end{aligned}. \tag{18}$$

4.3 Sort Precision Loss

In order to supervise the CNN model to constrain the angle relationship between feature embeddings by considering the similarity scores of gallery w.r.t the query, we introduce another task that aims at sorting the distance set d_q^i in descending order w.r.t continuous labels $s_q^i \in [0, 1]$. s_q^i is formulated as follows:

$$\mathcal{S}_{\mathcal{G}} = \left\{ s_q^i = \left\langle \frac{\mathcal{X}_q^i}{\left\|\mathcal{X}_q^i\right\|} \cdot \frac{\mathcal{X}_q^j}{\left\|\mathcal{X}_q^j\right\|} \right\rangle, j = 1, \ldots, N-1 \right\}, \tag{19}$$

where $\mathcal{S}_{\mathcal{G}} = \mathcal{S}_P \cup \mathcal{S}_N$, and $\mathcal{S}_P = \{s_q^{i,\psi}, \forall \psi \in \mathcal{P}_q\}$, $\mathcal{S}_N = \{s_q^{i,\sigma}, \forall \sigma \in \mathcal{N}_q\}$ are the positive and negative similarity sets, respectively. Then, the Sort Precision loss can be defined as follows:

$$\mathcal{L}_{SP} = \frac{1}{|\mathcal{D}_P|} \sum_{j \in \mathcal{D}_P} \frac{\sum_{k \in \mathcal{D}_P} \mathcal{H}\left(z_{j,k}\right)\left(1 - s_q^{i,k}\right)}{\mathcal{R}\left(j, \mathcal{D}_{\mathcal{P}}\right)}. \tag{20}$$

For $j \in \mathcal{D}_P$, $\mathcal{L}_{SP}$ penalizes the positive samples with similarity scores larger than s_q^i by the average of their inverted labels, $1 - s_q^i$. Here, $\mathcal{L}_{SP}$ is essentially a measure of sorting errors.

4.4 Differentiable Approximate Step Function

The Heaviside step function described in Equation (15) is non-differentiable. Thus, it can not be optimized along with the person re-ID model in the training phase. To solve this problem, we propose to perform a step operation with an approximate step function:

$$\hat{\mathcal{H}}\left(x\right) = \frac{1}{1 + e^{-Tx}}, \tag{21}$$

where T is the temperature adjusting the sharpness. This approximate step function behaves similar to the Heaviside step function but is differentiable, thus it can be optimized along with the person re-ID model in the training phase.

4.5 Differentiable Retrieval-Sort Loss

The DRSL $\mathcal{L}_{DRSL}$ can be expressed as the sum of the RP loss $\mathcal{L}_{RP}$ and the sort precision loss $\mathcal{L}_{SP}$:

$$\mathcal{L}_{DRSL} = \mathcal{L}_{RP} + \beta \mathcal{L}_{SP}, \tag{22}$$

where β controls the balance between ranking and sorting operations. β is set to 0.0005 empirically.

5 EXPERIMENTS

In this section, we would like to comprehensively evaluate the effectiveness and superiority of our DRSL loss on three different datasets. Firstly, we describe the datasets and evaluation metrics in Section 5.1. Secondly, Section 5.2 represents the implementation details. Thirdly, Section 5.3 performs the ablation study. Next, Section 5.4 conducts the discussion and analysis, and Section 5.5 shows the visualization results. Finally, Section 5.6 discusses and analyzes our method.

5.1 Datasets and Evaluation Metrics

We assess the proposed DRSL on three large-scale person re-ID datasets that are commonly used in the person re-ID community. These datasets are: Market-1501, CUHK03, and MSMT17. The statistics of the three datasets are shown in Table 1. We have counted the number of person ID and images (Img) in the training set, as well as the number of queries (Query) and gallery (Gallery) in the testing set for each dataset. We use the **cumulative matching characteristic** (**CMC**) [19] and **mean average precision** (**mAP**) [78] as the evaluation metrics to measure the performance of re-ID models.

- **Market-1501** [78] is collected by a total of 6 different cameras at Tsinghua University. This dataset has 32,668 labeled bounding boxes of 1,501 identities acquired by the DPM detector. Specifically, the training set contains 12,936 pedestrian images with 751 identities; while the testing set contains 3,368 query images and 19,732 gallery images with the rest 750 identities. Currently, it is the most widely used dataset.
- **CUHK03** [26] is collected at Chinese University of Hong Kong via 10 (five pairs) different cameras. This dataset contains 14,096 images of 1,467 identities labeled by the human and

Table 1. Statistics of Training/Testing Set on Market-1501, CUHK03, and MSMT17 (# ID: Number of Person Identities, # Img: Number of Training Images, # Query: Number of Query, # Gallery: Number of Gallery)

Dataset	Training		Testing	
	# ID	# Img	# Query	# Gallery
Market-1501 [78]	751	12,936	3,368	19,732
CUHK03 [26]	767	14,733	2,800	10,660
MSMT17 [58]	1,041	30,248	11,659	82,161

the DPM detector. The training and testing sets contained 767 and 700 identities respectively. It contains 14,733 images for training, 2,800 query images and 10,660 gallery images for testing. We adopt the novel training/testing protocol proposed in [82], and we use the detected images for our evaluation in this article.

—**MSMT17** [58] contains 126,411 person images with 4,101 identities, where 1,041 identities are used for training and 3,060 identities are used for testing. Totally, 15 different cameras were used, including 12 outdoor cameras and three indoor cameras. All the bounding boxes of person images are acquired by the Faster R-CNN detector. Compared to Market-1501 and Duke-reID, it has more complicated scenarios.

5.2 Implementation Details

Our method is implemented using the Pytorch framework [38] based on BagTricks [32], and all the experiments are based on ResNet-50 backbone. For each training image, it is augmented by random horizontal flip and random erasing. And all input images are resized to 256×128 and padded with 10. We use a pre-trained ResNet-50 model (trained on ImageNet [16]) to initialize the parameters of the model. The Adam optimizer and global features are used in our experiments. The mini-batch size is 64, which contains 16 identities and four images for each identity. For all experiments, we set the epoch to 120 and T to 10. The initial learning rate $l(t)$ at epoch t is adjusted as follows.

$$l(t) = \begin{cases} 3.5 \times 10^{-4} \times \frac{t}{10} & t \leq 10 \\ 3.5 \times 10^{-4} & 10 < t \leq 40 \\ 3.5 \times 10^{-5} & 40 < t \leq 70 \\ 3.5 \times 10^{-6} & 70 < t \leq 120 \end{cases}. \tag{23}$$

5.3 Comparison with the State-of-the-art re-ID Models

We compare the proposed DRSL with 14 representative metric learning-based re-ID models. The results are shown in Table 2. Obviously, we can find that the proposed DRSL achieves consistent performance improvement on all datasets. We can see that the Baseline and Baseline† show a good performance on Market-1501 and MSMT17. Although our method has a very small performance degradation on Rank-1, it does have a significant improvement on mAP. It is a reasonable result for two reasons: (1) For simplicity, we add our loss function directly on Baseline without any fine-tuning; (2) Our method strives to optimize the overall mAP rather than Rank during the training process. Specifically, compared with the Baseline [32], we obtained 0.8% and 3.1% performance improvement in mAP on Market1501 and MSMT17, respectively. More importantly, We can achieve 11.5%/10.1% Rank-1/mAP performance improvement on CUHK03. Besides, with center loss, we can achieve significant performance improvement on three datasets due to the clustering property of

Table 2. Comparisons with the State-of-the-art Person Re-ID Models on Market1501, CUHK03 (Detected), and MSMT17

Method	Market-1501				CUHK03 (detected)				MSMT17			
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
IDE (R)+XQDA+k-reciprocal [82] CVPR'17	75.1	–	–	61.9	34.7	–	–	37.4	–	–	–	–
IDE (R)+KISSME+k-reciprocal [82] CVPR'17	77.1	–	–	63.6	–	–	–	–	–	–	–	–
IV-reID [80] TOMM'17	79.5	–	–	59.9	49.9	–	–	44.6	–	–	–	–
TriNet [23] arXiv'17	84.9	94.2	96.2	69.1	50.5	–	–	46.5	56.9	72.7	78.4	26.9
ECN (rank-dist) [45] CVPR'18	82.3	–	–	71.1	27.3	–	–	30.2	–	–	–	–
PSE+ ECN (rank-dist) [45] CVPR'18	90.3	–	–	84.0	27.3	–	–	30.2	–	–	–	–
AWTL [43] CVPR'18	89.5	–	–	75.7	–	–	–	–	–	–	–	–
PCB [49] ECCV'18	92.3	96.9	98.2	77.3	59.7	–	–	53.2	68.2	81.3	85.6	41.1
HPM [18] AAAI'19	94.2	97.5	98.5	82.7	63.9	79.7	86.1	57.5	–	–	–	–
CG+MB [74] CVPRW'19	92.6	97.4	–	78.3	59.9	–	–	53.3	–	–	–	–
CE-FAT [73] CVPRW'20	89.4	95.6	97.1	73.1	–	–	–	–	69.4	81.5	85.6	39.2
HCTL [76] TMM'20	93.8	97.8	98.6	81.8	–	–	–	–	74.3	84.7	87.9	43.6
EDL [69] Neurocomputing'21	93.3	–	–	83.6	65.9	–	–	63.6	–	–	–	–
CBDB [50] TCSVT'21	94.4	–	–	85.0	76.6	–	–	72.8	–	–	–	–
Baseline	94.1	**98.0**	**98.9**	85.4	59.6	79.3	86.4	57.7	72.3	84.2	87.8	48.3
Baseline + DRSL (our)	94.1	97.9	98.6	86.2	71.1	**84.4**	**90.1**	67.8	73.7	86.2	**89.6**	51.4
Baseline + DRSL (our) (Re-ranking)	**95.0**	97.4	98.2	**93.4**	**75.1**	84.2	89.7	**76.5**	**79.7**	**87.2**	89.5	**68.6**
Baseline†	94.2	**98.1**	**98.8**	85.9	60.4	79.4	87.0	58.6	72.0	83.9	87.5	48.2
Baseline† + DRSL (our)	93.9	97.7	98.6	86.3	72.0	85.6	92.0	69.1	73.1	85.7	89.3	51.0
Baseline† + DRSL (our) (Re-ranking)	**95.2**	97.4	98.2	**93.5**	**81.5**	**86.8**	**92.4**	**82.8**	**79.0**	**87.0**	**89.5**	**68.1**

The results in bold indicate the best performance. No results are listed as "–". Baseline means BagTricks [32] without center loss and Baseline† represents BagTricks with center loss. "Re-ranking" denotes the method with re-ranking operation in the testing phase. All methods rely on a standard convolutional backbone (generally ResNet-50).

Table 3. Impacts of Different Parameters: Temperature T, Size of Positive Set During Mini-batch Sampling $|\mathcal{P}|$, and Batch Size $\mathcal{B}$

T	Market-1501		CUHK03 (detected)	
	R1	mAP	R1	mAP
1	93.0	84.2	64.4	61.2
10	**94.4**	**86.2**	71.1	67.8
100	93.2	84.5	**73.3**	**69.9**
1000	87.6	72.7	49.4	47.5

$|\mathcal{P}| = 4$, $\mathcal{B} = 64$

$|\mathcal{P}|$	Market-1501		CUHK03 (detected)	
	R1	mAP	R1	mAP
4	**94.4**	**86.2**	71.1	67.8
8	93.9	86.1	73.3	69.9
16	93.8	85.3	**77.1**	**74.1**
32	90.6	77.5	52.4	52.2

$T = 10$, $\mathcal{B} = 64$

$\mathcal{B}$	Market-1501		CUHK03 (detected)	
	R1	mAP	R1	mAP
32	93.7	84.9	63.3	62.0
64	**94.4**	**86.2**	**71.1**	**67.8**
128	93.0	84.2	68.6	65.7
256	92.5	82.8	65.9	62.4

$T = 10$, $|\mathcal{P}| = 4$

The results in bold indicate the best performance.

center loss. In addition, we have also added the re-ranking performance of our method, which further improves performance significantly. It is obvious that our method impressively outperforms the state-of-the-art on three datasets when re-ranking operation is added.

5.4 Ablation Study

To investigate the impact of different hyper-parameter settings of the proposed DRSL, e.g., the temperature T, the size of the positive set $|\mathcal{P}|$, and batch size $\mathcal{B}$, we conducted ablation studies on Market-1501 and CUHK03. Note that we only changed one parameter at a time. The experimental results are shown in Tables 3 and 4, respectively. In addition, we included additional ablation studies on contrastive loss in Table 5.

Impact of temperature T. In Table 3, we can find that Rank-1 and mAP achieve the best performance when the value of T is 10 on Market-1501 and is 100 on CUHK03. The gradient of the approximate step function and AP's optimization achieved the tradeoff. The main reason lies in the fact that this value corresponds to a very small interval on the approximate step function,

Table 4. Ablation Study for the Impact of Two Components in our Method on the BagTricks Baseline

Method	$\mathcal{L}_{RP}$	$\mathcal{L}_{SP}$	Market-1501		CUHK03 (detected)	
			R1	mAP	R1	mAP
Baseline	✗	✗	94.1	85.4	59.6	57.7
Ours	✓	✗	94.1	86.0	68.9	66.6
	✗	✓	94.0	85.7	61.6	59.2
	✓	✓	**94.4**	**86.2**	**71.1**	**67.8**

We use ResNet-50 as our default backbone.
The results in bold indicate the best performance.

Table 5. Comparison with Some Methods on Contrastive Loss for re-ID

Method	Market-1501		CUHK03 (detected)	
	R1	mAP	R1	mAP
IDLA [1]	–	–	45.0	–
IV-reID [80]	79.5	59.9	49.9	46.6
DTL [11]	83.7	65.5	–	–
Baseline	94.1	85.4	59.6	57.7
Baseline + DRSL (our)	**94.4**	86.2	**71.1**	**67.8**
Baseline + C	94.1	85.2	61.9	62.4
Baseline + DRSL + C	93.9	**86.3**	67.4	65.9

Note that "C" denotes the contrastive loss, respectively. The results in bold indicate the best performance. No results are listed as "–".

which induces effective re-ranking gradients. In essence, the temperature T can be viewed as the inter-class margin, which helps to improve the generalizability of the model [46].

Impact of positive set $|\mathcal{P}|$. As shown in Table 3, $|\mathcal{P}|$ indicates the number of instances per person identity in a mini-batch. We can find that the Rank-1 and mAP achieve the best performance when $|\mathcal{P}| = 4$ on Market-1501 and $|\mathcal{P}| = 16$ on CUHK03. If $|\mathcal{P}|$ is larger, the number of classes sampled in a min-batch is smaller, and the sampling of hard negative samples becomes more difficult. In other words, the larger the number of classes in a min-batch, the closer it is to the real gallery setting, allowing each training iteration to enforce a more optimally structured feature space.

Impact of batch size $\mathcal{B}$. Table 3 shows that $\mathcal{B} = 64$ results in the highest Rank-1/mAP accuracy. Theoretically, larger batch sizes should result in better Rank-1/mAP because of more chance of sampling hard negative samples in the batch. However, the results in Table 3 did not meet the expectation. The main reason may lie in that a larger batch size brings more hard negative samples, but is not conducive to the optimization of triplet and softmax cross-entropy loss functions.

Impact of each Loss Function. We investigate how to make the $\mathcal{L}_{RP}$ and $\mathcal{L}_{SP}$ interact efficiently for joint optimization. The comparative results are illustrated in Table 4. (1) Compared to $\mathcal{L}_{RP}$, $\mathcal{L}_{SP}$ has very limited performance gains in Rank-1/mAP accuracy. This is because most of the intra-class samples share a similar structure, i.e., the angle differences between the feature vectors are small. $\mathcal{L}_{SP}$ plays an auxiliary optimization, and its weight β is set to 0.0005 in our experiments. (2) $\mathcal{L}_{DRSL}$ provides performance improvements compared to $\mathcal{L}_{SP}$ and $\mathcal{L}_{RP}$ separately.

Comparison with Contrastive Loss. Person re-ID could be investigated as a verification problem. Some works [1, 11, 80] have improved the performance of re-ID models by imposing the contrastive loss. To investigate whether the contrastive loss could improve the performance of

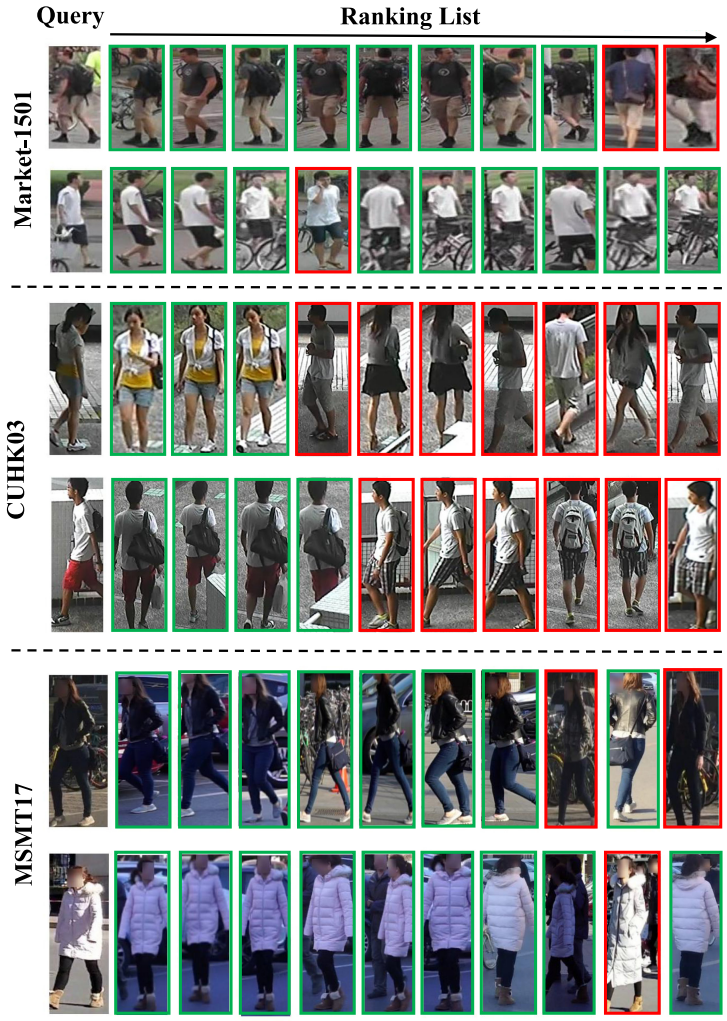

Fig. 4. Illustration of the retrieval ranking list of person samples in the top-10 on Market-1501, CUHK03, and MSMT17 datasets. The images in the first column are the query images. The retrieved images are sorted in descending order from left to right according to the similarity score. The correct matches are in the green boxes, and the false matching images are in the red boxes.

our method, we also add the contrastive loss to our method. The quantitative results are shown in Table 5. By adding the contrastive loss, the baseline and our method have similar performance on Market-1501, and the baseline achieves some performance improvement (i.e., 2.3%/4.7% Rank-1/mAP) on CUHK03. However, the introduction of the contrastive loss significantly decreases the performance of our method on CUHK03 (i.e., 3.7%/1.9% Rank-1/mAP performance degradation). Meanwhile, the performance of our method (Baseline + DRSL) is better than that of our method added with contrastive loss (Baseline + DRSL + C). Therefore, the contrastive loss can improve the performance of the baseline on CUHK03; while it decreases the performance of our method on CUHK03. It should be noted that our method achieves the best performance, as shown in Table 5. The key of our method is to globally optimize the ranking of all samples, while the

contrastive loss is to optimize the local pairwise relationships of the samples. In this case, as the performance of the baseline is already high on Market-1501, it is difficult to achieve the performance improvement using the contrastive loss. However, due to the challenging nature of CUHK03 (detected), the performance of the baseline still leaves much room for the performance improvement. Therefore, it is reasonable that using the contrastive loss or our method could improve the performance on CUHK03. Also, due to the local optimization of the contrastive loss that may lead the model to be sub-optimal, it makes sense that adding the contrastive loss to our method would lead the performance to degrade on CUHK03.

5.5 Visualization Results

In Figure 4, the visualization results of our method are shown on Market-1501, CUHK03, and MSMT17 datasets, respectively. Figure 4 presents some retrieval ranking lists of person samples. The images in the first column are the query images. The retrieval of person images is sorted in descending order according to similarity score (from left to right in the ranking list). As we can see, most ground-truth candidate person images are correctly retrieved. Although our method retrieves some false candidate cases that are in the second row on Market-1501 and are in the eighth and ninth columns on MSMT17, we find that it is a reasonable prediction since the person with similar clothes is similar to the query. Our proposed method yields 95.0%, 75.1%, and 79.7% Rank-1, 93.4%, 76.5%, and 68.6% mAP on Market-1501, CUHK03, and MSMT17, respectively. And our method outperforms the state-of-the-arts.

5.6 Discussion and Analysis

In the above experimental results, there are still some points that need to be further investigated. The proposed DRSL outperforms all representative metric learning-based methods on three person re-ID datasets, i.e., Market-1501, CUHK03, and MSMT17, with limited performance improvement. The main reason lies in that the current performance of the baseline has arrived at a bottleneck with global features. Discriminative features need to be combined to assist further performance improvement. In addition, due to the challenging nature of CUHK03 and MSMT17, these two datasets still have many issues worth investigating to improve performance.

6 CONCLUSIONS

In this article, we proposed a ranking-based loss function DRSL to optimize person re-ID models. Different from existing ranking-based loss functions, the proposed DRSL not only ranks positive samples ahead of negative samples but also sorts positive samples w.r.t. their similarity scores to preserve useful similarity structure. With DRSL, we utilized a simple and loss-value-based heuristic to constrain the distance between positive samples and negative samples as well as to maintain the intra-class similarity metric. Besides, the discussion and analysis of the proposed DRSL are provided to illustrate its effectiveness. Experimental results on three commonly used person re-ID datasets validate that the proposed DRSL consistently improves performance and significantly simplifies the training process.

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K. Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3908–3916. DOI: https://doi.org/10.1109/CVPR.2015.7299016

[2] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. 2020. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognition* 98 (2020), 107036. DOI: https://doi.org/10.1016/j.patcog.2019.107036

[3] Rodger Benham, Joel Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting search performance using query variations. *ACM Transactions on Information Systems* 37, 4 (2019), 1–25. DOI: https://doi.org/10.1145/3345001

[4] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. 2020. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *Proceedings of the European Conference on Computer Vision*. 677–694. DOI: https://doi.org/10.1007/978-3-030-58545-7_39

[5] Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. In *Proceedings of the Advances in Neural Information Processing Systems*. 193–200.

[6] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. 2019. Deep metric learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1861–1870. DOI: https://doi.org/10.1109/CVPR.2019.00196

[7] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. 2018. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2109–2118. DOI: https://doi.org/10.1109/CVPR.2018.00225

[8] Olivier Chapelle, Quoc Le, and Alex Smola. 2007. Large margin optimization of ranking measures. In *Proceedings of the NIPS Workshop: Machine Learning for Web Search*.

[9] Guangyi Chen, Tianpei Gu, Jiwen Lu, Jin-An Bao, and Jie Zhou. 2021. Person re-identification via attention pyramid. *IEEE Transactions on Image Processing* 30 (2021), 7663–7676. DOI: https://doi.org/10.1109/TIP.2021.3107211

[10] Guangyi Chen, Yuhao Lu, Jiwen Lu, and Jie Zhou. 2020. Deep credible metric learning for unsupervised domain adaptation person re-identification. In *Proceedings of the European Conference on Computer Vision*. 643–659. DOI: https://doi.org/10.1007/978-3-030-58598-3_38

[11] Haoran Chen, Yaowei Wang, Yemin Shi, Ke Yan, Mengyue Geng, Yonghong Tian, and Tao Xiang. 2018. Deep transfer learning for person re-identification. In *Proceedings of the 2018 IEEE 4th International Conference on Multimedia Big Data*. IEEE, 1–5.

[12] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. 2019. Towards accurate one-stage object detection with ap-loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5119–5127. DOI: https://doi.org/10.1109/CVPR.2019.00526

[13] Kean Chen, Weiyao Lin, Jianguo Li, John See, Ji Wang, and Junni Zou. 2021. AP-loss for accurate one-stage object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2021), 3782–3798. DOI: https://doi.org/10.1109/TPAMI.2020.2991457

[14] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: A deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 403–412. DOI: https://doi.org/10.1109/CVPR.2017.145

[15] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. 2021. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5270–5279.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. DOI: https://doi.org/10.1109/CVPR.2009.5206848

[17] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 994–1003. DOI: https://doi.org/10.1109/CVPR.2018.00110

[18] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. 2019. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8295–8302. DOI: https://doi.org/10.1609/aaai.v33i01.33018295

[19] Douglas Gray, Shane Brennan, and Hai Tao. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*. Citeseer, 1–7.

[20] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3642–3651. DOI: https://doi.org/10.1109/ICCV.2019.00374

[21] Yiluan Guo and Ngai-Man Cheung. 2018. Efficient and deep person re-identification using multi-level similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2335–2344. DOI: https://doi.org/10.1109/CVPR.2018.00248

[22] Kun He, Fatih Cakir, Sarah Adel Bargal, and Stan Sclaroff. 2018. Hashing as tie-aware learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4023–4032. DOI: https://doi.org/10.1109/CVPR.2018.00423

[23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. arXiv:1703.07737. Retrieved from https://arxiv.org/abs/1703.07737.

[24] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. 2021. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6729–6738. DOI: https://doi.org/10.1109/CVPR46437.2021.00666

[25] Kehuang Li, Zhen Huang, You-Chi Cheng, and Chin-Hui Lee. 2014. A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4503–4507. DOI: https://doi.org/10.1109/ICASSP.2014.6854454

[26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 152–159. DOI: https://doi.org/10.1109/CVPR.2014.27

[27] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294. DOI: https://doi.org/10.1109/CVPR.2018.00243

[28] Wei Li, Xiatian Zhu, and Shaogang Gong. 2020. Scalable person re-identification by harmonious attention. *International Journal of Computer Vision* 128, 6 (2020), 1635–1653. DOI: https://doi.org/10.1007/s11263-019-01274-1

[29] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognition* 95 (2019), 151–161. DOI: https://doi.org/10.1016/j.patcog.2019.06.006

[30] Deyin Liu, Lin Wu, Richang Hong, Zongyuan Ge, Jialie Shen, Farid Boussaid, and Mohammed Bennamoun. 2022. Generative metric learning for adversarially robust open-world person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 1, 1 (2022), 1–20. DOI: https://doi.org/10.1145/3522714

[31] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 212–220. DOI: https://doi.org/10.1109/CVPR.2017.713

[32] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. 2019. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia* 22, 10 (2019), 2597–2609. DOI: https://doi.org/10.1109/TMM.2019.2958756

[33] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. 2017. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*. 360–368. DOI: https://doi.org/10.1109/ICCV.2017.47

[34] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4004–4012. DOI: https://doi.org/10.1109/CVPR.2016.434

[35] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. 2018. Localization recall precision (LRP): A new performance metric for object detection. In *Proceedings of the European Conference on Computer Vision*. 504–519. DOI: https://doi.org/10.1007/978-3-030-01234-2_31

[36] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. 2020. A ranking-based, balanced loss function unifying classification and localisation in object detection. In *Proceedings of the Advances in Neural Information Processing Systems*. 15534–15545.

[37] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. 2021. Rank & sort loss for object detection and instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3009–3018. DOI: https://doi.org/10.1109/ICCV48922.2021.00300

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems*. 8024–8035.

[39] Tao Qin, Tie-Yan Liu, and Hang Li. 2010. A general approximation framework for direct optimization of information retrieval measures. *Information Retrieval* 13, 4 (2010), 375–397. DOI: https://doi.org/10.1007/s10791-009-9124-x

[40] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. 2019. Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3750–3759. DOI: https://doi.org/10.1109/ICCV.2019.00385

[41] Ragesh Kumar Ramachandran, Nicole Fronda, and Gaurav Sukhatme. 2021. Resilience in multi-robot multi-target tracking with unknown number of targets through reconfiguration. *IEEE Transactions on Control of Network Systems* 8, 2 (2021), 609–620. DOI: https://doi.org/10.1109/TCNS.2021.3059794

[42] Jerome Revaud, Jon Almazan, Rafael S. Rezende, and Cesar Roberto de Souza. 2019. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5107–5116. DOI: https://doi.org/10.1109/ICCV.2019.00521

[43] Ergys Ristani and Carlo Tomasi. 2018. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6036–6046. DOI: https://doi.org/10.1109/CVPR.2018.00632

[44] Michal Rolínek, Vít Musil, Anselm Paulus, Marin Vlastelica, Claudio Michaelis, and Georg Martius. 2020. Optimizing rank-based metrics with blackbox differentiation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7620–7630. DOI: https://doi.org/10.1109/CVPR42600.2020.00764

[45] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 420–429. DOI: https://doi.org/10.1109/CVPR.2018.00051

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823. DOI: https://doi.org/10.1109/CVPR.2015.7298682

[47] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the Advances in Neural Information Processing Systems*. 1857–1865. Retrieved from https://proceedings.neurips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html.

[48] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision*. 402–419. DOI: https://doi.org/10.1007/978-3-030-01264-9_25

[49] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*. 480–496. DOI: https://doi.org/10.1007/978-3-030-01225-0_30

[50] Hongchen Tan, Xiuping Liu, Yuhao Bian, Huasheng Wang, and Baocai Yin. 2021. Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 14, 8 (2021), 1–12. DOI: https://doi.org/10.1109/TCSVT.2021.3061412

[51] Evgeniya Ustinova and Victor S. Lempitsky. 2016. Learning deep embeddings with histogram loss. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 29. 4170–4178. http://papers.nips.cc/paper/6464-learning-deep-embeddings-with-histogram-loss.

[52] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. 2016. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1288–1296. DOI: https://doi.org/10.1109/CVPR.2016.144

[53] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM International Conference on Multimedia*. 274–282. DOI: https://doi.org/10.1145/3240508.3240552

[54] Meng Wang, Richang Hong, Xiao-Tong Yuan, Shuicheng Yan, and Tat-Seng Chua. 2012. Movie2comics: Towards a lively video content presentation. *IEEE Transactions on Multimedia* 14, 3–2 (2012), 858–870. DOI: https://doi.org/10.1109/TMM.2012.2187181

[55] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. 2012. Multimodal graph-based reranking for web image search. *IEEE Transactions on Image Processing* 21, 11 (2012), 4649–4661. DOI: https://doi.org/10.1109/TIP.2012.2207397

[56] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M. Robertson. 2019. Ranked list loss for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5207–5216. DOI: https://doi.org/10.1109/CVPR.2019.00535

[57] Zheng Wang, Xin Yuan, Toshihiko Yamasaki, Yutian Lin, Xin Xu, and Wenjun Zeng. 2020. Re-identification= retrieval+ verification: Back to essence and forward with a new metric. arXiv:2011.11506. Retrieved from https://arxiv.org/abs/2011.11506.

[58] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 79–88. DOI: https://doi.org/10.1109/CVPR.2018.00016

[59] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 499–515. DOI: https://doi.org/10.1007/978-3-319-46478-7_31

[60] Lin Wu, Richang Hong, Yang Wang, and Meng Wang. 2019. Cross-entropy adversarial view adaptation for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 7 (2019), 2081–2092. DOI: https://doi.org/10.1109/TCSVT.2019.2909549

[61] Lin Wu, Yang Wang, Junbin Gao, Meng Wang, Zheng-Jun Zha, and Dacheng Tao. 2020. Deep coattention-based comparator for relative representation learning in person re-identification. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2020), 722–735. DOI: https://doi.org/10.1109/TNNLS.2020.2979190

[62] Lin Wu, Yang Wang, Hongzhi Yin, Meng Wang, and Ling Shao. 2019. Few-shot deep adversarial learning for video-based person re-identification. *IEEE Transactions on Image Processing* 29 (2019), 1233–1245. DOI: https://doi.org/10.1109/TIP.2019.2940684

[63] Pengyu Xie, Xin Xu, Zheng Wang, and Toshihiko Yamasaki. 2021. Unsupervised video person re-identification via noise and hard frame aware clustering. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6. DOI: https://doi.org/10.1109/ICME51207.2021.9428200

[64] Xin Xu, Lei Liu, Xiaolong Zhang, Weili Guan, and Ruimin Hu. 2021. Rethinking data collection for person re-identification: Active redundancy reduction. *Pattern Recognition* 113 (2021), 107827. DOI: https://doi.org/10.1016/j.patcog.2021.107827

[65] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. 2019. Attention driven person re-identification. *Pattern Recognition* 86 (2019), 143–155. DOI: https://doi.org/10.1016/j.patcog.2018.08.015

[66] Xun Yang, Xiaoyu Du, and Meng Wang. 2020. Learning to match on graph for fashion compatibility modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 287–294. DOI: https://doi.org/10.1609/aaai.v34i01.5362

[67] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3 (2017), 1–23. DOI: https://doi.org/10.1145/3089249

[68] Xun Yang, Meng Wang, and Dacheng Tao. 2017. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing* 27, 2 (2017), 791–805. DOI: https://doi.org/10.1109/TIP.2017.2765836

[69] Zhao Yang, Jiehao Liu, Tie Liu, Yuanxin Zhu, Li Wang, and Dapeng Tao. 2021. Equidistant distribution loss for person re-identification. *Neurocomputing* 455 (2021), 255–264. DOI: https://doi.org/10.1016/j.neucom.2021.04.070

[70] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. 2022. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2022), 2872–2893. DOI: https://doi.org/10.1109/TPAMI.2021.3054775

[71] Wei Yi, Ye Yuan, Reza Hoseinnezhad, and Lingjiang Kong. 2020. Resource scheduling for distributed multi-target tracking in netted colocated MIMO radar systems. *IEEE Transactions on Signal Processing* 68 (2020), 1602–1617. DOI: https://doi.org/10.1109/TSP.2020.2976587

[72] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. 2018. Hard-aware point-to-set deep metric for person re-identification. In *Proceedings of the European Conference on Computer Vision*. 188–204. DOI: https://doi.org/10.1007/978-3-030-01270-0_12

[73] Ye Yuan, Wuyang Chen, Yang Yang, and Zhangyang Wang. 2020. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 354–355. DOI: https://doi.org/10.1109/CVPRW50498.2020.00185

[74] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li. 2019. In defense of the classification loss for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1526–1535. DOI: https://doi.org/10.1109/CVPRW.2019.00194

[75] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2019. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 667–676. DOI: https://doi.org/10.1109/CVPR.2019.00076

[76] Cairong Zhao, Xinbi Lv, Zhang Zhang, Wangmeng Zuo, Jun Wu, and Duoqian Miao. 2020. Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. *IEEE Transactions on Multimedia* 22, 12 (2020), 3180–3195. DOI: https://doi.org/10.1109/TMM.2020.2972125

[77] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1077–1085. DOI: https://doi.org/10.1109/CVPR.2017.103

[78] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124. DOI: https://doi.org/10.1109/ICCV.2015.133

[79] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1367–1376.

[80] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 1 (2017), 1–20. DOI: https://doi.org/10.1145/3159171

[81] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*. 3754–3762. DOI: https://doi.org/10.1109/ICCV.2017.405

[82] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3652–3661. DOI: https://doi.org/10.1109/CVPR.2017.389

[83] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2021. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–14. DOI: https://doi.org/10.1109/TPAMI.2021.3069237

Received October 2021; revised February 2022; accepted April 2022