Pattern Recognition 152 (2024) 110455

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Pairwise difference relational distillation for object re-identification

Yi Xie [a,b], Hanxiao Wu [c,d], Yihong Lin [e], Jianqing Zhu [b,*], Huanqiang Zeng [b]

[a] School of Future Technology, South China University of Technology, Guangzhou, China
[b] College of Engineering, Huaqiao University, Quanzhou, China
[c] College of Information Science and Engineering, Huaqiao University, Xiamen, China
[d] School of Computer Science and Artifical Intelligence, Wuhan University of Technology, Wuhan, China
[e] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

ARTICLE INFO

Keywords:
Knowledge distillation
Object re-identification

ABSTRACT

Most relationship knowledge distillation methods individually optimize pairwise similarities to improve the accuracy performance of a lightweight student network. However, this optimization approach may not be optimal for object re-identification (Re-ID) which prioritizes ranking. This is because it does not guarantee consistent ranking results between a lightweight student network and a large teacher network. For that, we propose a novel method called pairwise difference relational distillation (PDRD) for object Re-ID. First, we theoretically prove that minimizing the difference relationship between pairwise similarities resulting from student and teacher networks ensures consistent ranking results between the two networks. Second, based on this theoretical foundation, we combine non-linear activation functions on pairwise similarity discrepancies to create a non-linear pairwise difference rational knowledge loss function, which enhances knowledge transfer. Extensive experiments on four public datasets demonstrate that our method achieves state-of-the-art performance. For example, on Market-1501, using ResNet18 as a lightweight student network, our method acquires a rank-1 identification rate of 93.62%.

1. Introduction

Object re-identification (Re-ID), such as pedestrian Re-ID [1,2] and vehicle Re-ID [3,4], focuses on the retrieval of interest objects from massive traffic data, playing great potential in smart cities. However, pedestrians and vehicles are captured by different cameras installed at different city locations, making object Re-ID extremely challenging. In addition, Re-ID object methods [3,4] prefer to use complex networks to learn discriminative features from images to ensure accuracy, which causes low inference speed and thus limits practical applications.

Knowledge distillation (KD) [5–7] is an efficient network compression technology, which transfers knowledge from a complex teacher network to a simple student network to improve the accuracy of the student network. The original formulation of KD is logit distillation [5,8,9], which transfers posterior probability knowledge by constraining the soft logical value of the student network to be consistent with the teacher network. However, since posterior probability knowledge is only derived from a deep location of a teacher network, it could not carry intermediate knowledge from intermediate layers of the teacher network. Consequently, the features of the student network lack valuable fine-grained information. This limitation of being unable to utilize the intermediate knowledge of the teacher network hinders the student network from effectively performing the object re-identification. This is because a high accuracy object Re-ID model needs to utilize fine-grained features to distinguish pedestrian or vehicle images captured from different camera viewpoints.

In addition to logit distillation, feature distillation [10–13] is another common knowledge distillation method, which distills intermediate features to help a student network learn discriminative features as a teacher network. The feature distillation methods [10,14,15] minimize the distance between the intermediate features of the teacher network and the intermediate features of the student network to improve the accuracy of the student network. Unfortunately, because of the significant difference in model sizes between lightweight student and large teacher networks, features resulting from the two networks usually have different dimensions. For example, when the teacher network is ResNet101 [16] and the student network is ResNet18 [16], the intermediate layer feature dimension of the teacher network is four times that of the student network. As a result, due to different feature dimensions, it is difficult to measure the differences between these features directly using common distance metrics, such as cosine distances [17,18]

* Corresponding author.
E-mail addresses: ftyixie@mail.scut.edu.cn (Y. Xie), hxwu@whut.edu.cn (H. Wu), amcsyihonglin@gmail.com (Y. Lin), jqzhu@hqu.edu.cn (J. Zhu), zeng0043@hqu.edu.cn (H. Zeng).

https://doi.org/10.1016/j.patcog.2024.110455
Received 30 June 2023; Received in revised form 14 October 2023; Accepted 24 March 2024
Available online 27 March 2024

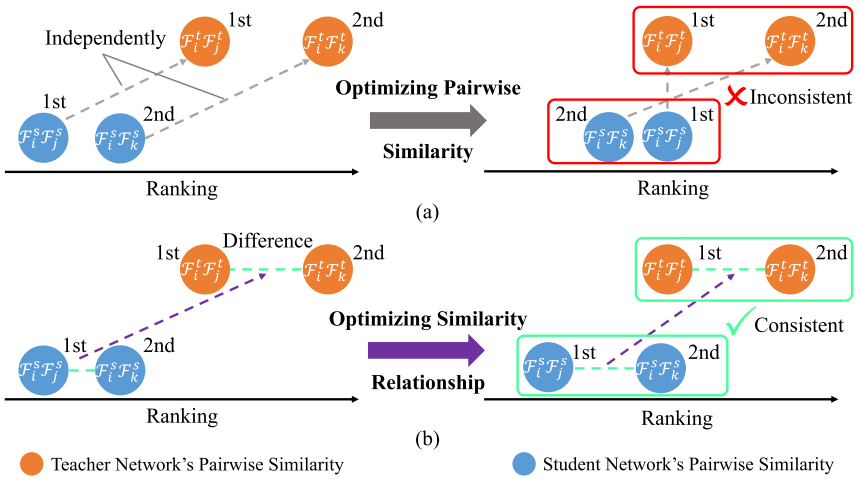

Fig. 1. Illustration of existing methods and our pairwise difference relational distillation (PDRD) method. (a) Existing methods independently optimize similarity pairs to produce inconsistent ranking results. (b) Our approach optimizes the difference relationship between similarity pairs to generate consistent ranking results. The superscript t represents a teacher and the superscript s represents a student. The subscripts i, j, and k denote sample numbers.

and Euclidean distances [15,19]. Most feature distillation methods [10,14,15] have to employ extra adaptation modules (e.g., 1×1 convolutional layers) to align the feature dimensions of the student and teacher networks. However, the usage of adapter modules for distillation in object Re-ID is not optimal because these modules are usually randomly initialized [10,15]. This random initialization can actually disrupt the feature learning process of the student network, as demonstrated in [20].

To effectively distill knowledge for object Re-ID, relational distillation focuses on ensuring that the relationship between the intermediate features of the student network is consistent with the relationship between the intermediate features of the teacher network. A widely adopted approach [11,17,18] is to minimize the distance between the similarity matrices of the pairwise samples in the student network and those in the teacher network to transfer pairwise similarity knowledge from the large teacher network to the lightweight student network, improving the accuracy performance of the student network. As illustrated in Fig. 1(a), similarity knowledge of paired samples is optimized independently. However, this one-to-one optimization is too rigid to effectively preserve knowledge in a lightweight student network with limited representation capacity. Consequently, the ranking results of the lightweight student network could not be consistent with those of a large teacher network. This limitation can restrict the potential for improving object Re-ID accuracy in a lightweight student network, as object Re-ID primarily focuses on the ranking of returned object images.

Given that object Re-ID is a task focused on ranking if the ranking results of a student network align with those of a teacher network, the student network would effectively serve to object Re-ID. To this end, we propose a pairwise difference relational distillation (PDRD) method. As shown in Fig. 1 (b), our PDRD pays attention to the difference relationship between pairwise similarities, i.e., pairwise similarity difference knowledge. Intuitively, this knowledge involves a pair of similarities, which solves the problem of independently optimizing similarity pairs in previous works [12,13,17]. Furthermore, we theoretically prove that minimizing the difference relationship between the pairwise similarities obtained from teacher and student networks ensures consistent ranking results. Finally, considering the complex knowledge transfer between student and teacher networks, especially when they have vastly different scales, we introduce a non-linear mechanism to enhance the transfer of pairwise similarity difference knowledge.

The main contribution of this paper lies in three aspects. (1) Theoretical analysis is constructed to show that minimizing the difference between pairwise similarity discrepancies of student and teacher networks is beneficial to consistent ranking results of student and teacher networks. (2) A non-linear pairwise difference relational knowledge (NPDRK) loss function is designed to better transfer pairwise difference knowledge from teachers to students. (3) Extensive experiments on four widely used object Re-ID datasets demonstrate that our approach acquires state-of-the-art performance.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes our approach. Section 4 presents experiments and analyses to validate the superiority of our approach. Section 5 elaborates on the application limitations of our method. Section 6 concludes this paper.

2. Related works

Knowledge distillation [5,21] can transfer knowledge from a large teacher network to a lightweight student network to help the student network acquire good performance. Therefore, knowledge distillation is a promising solution for efficient object re-identification. In what follows, we systematically review the research progress of knowledge distillation in recent years. Then, we discuss knowledge distillation applications in object re-identification.

2.1. Knowledge distillation

Logit distillation. Logit distillation methods [5,9,22] exclusively distill knowledge using output logits. For example, the initial knowledge distillation (KD) method [5] can be classified as a logit distillation approach. Other logit distillation methods improve knowledge distillation by introducing a mutual learning paradigm [23] or an additional teacher-assistant module [22]. Recently, Zhao et al. [9] presented a reformulation of conventional knowledge distillation (KD), introducing target-class knowledge distillation (TCKD) and non-target-class knowledge distillation (NCKD). Furthermore, they proposed decoupled knowledge distillation (DKD) to enhance the effectiveness and flexibility of TCKD and NCKD in their respective roles. However, since the output logit is from a deep location of a teacher network, it could not carry knowledge from intermediate layers of the teacher network, leaving room for improvement.

Feature Distillation. Feature distillation methods [14,15,19] usually minimize distances between intermediate features resulting from teacher and student networks, forcing the student network to mimic the intermediate feature output of the teacher network. For example, Komodakis et al. [19] distilled intermediate features using an attention transfer (AT) method that deals with spatial attention maps of intermediate features. However, since there are significant architectural

differences (e.g., different network depths and basic units) between a large teacher network and a lightweight student network, the feature dimension of a large teacher network could not match that of a lightweight student network, causing a huge challenge of readily transferring knowledge. Although feature distillation works [14,15,24] could apply adaption modules (e.g., 1×1 convolutional layers) to unify feature dimensions. However, adaption modules with randomly initialized [14] may disturb the student network's feature learning [20], limiting the student's convergence speed and accuracy performance.

Relational Distillation. To effectively transfer knowledge, relational distillation methods [11,17,25] recently attracted a lot of attention. Unlike feature distillation methods that minimize distances between intermediate features resulting from teacher and student networks, relational distillation focuses on distilling the correlation between intermediate features to transfer knowledge from a teacher network to a student network without adaption modules. For example, Tung et al. [17] proposed a similarity-preserving (SP) knowledge distillation method to transfer knowledge of pairwise relationships by encouraging the student network to mimic the pairwise similarities of the teacher network. Park et al. [25] proposed a relational knowledge distillation (RKD) method that optimizes the angles of sample triples to transfer triplet-wise angle relationship knowledge. Passalis et al. [6] proposed a probabilistic knowledge-transfer (PKT) method, which utilizes the data probability distribution as relationship knowledge to improve the accuracy performance of a student network. In summary, relational distillation methods have significant potential, as they provide an opportunity for intermediate layers of a teacher network to showcase a wealth of knowledge.

2.2. Knowledge distillation-based object Re-identification

Numerous knowledge distillation methods [8,11,18] have been applied to accelerate object Re-ID model. One common approach is to distill the knowledge of pairwise similarity from a large teacher network to a lightweight student network. For example, Xie et al. [11] designed a matching behavior difference learning (MBDL) method, which transfers knowledge of pairwise similarity variations from teachers to students by ensuring consistency between the matching behavior difference matrices of student and teacher networks. Wu et al. [12] proposed a flexible contextual similarity distillation (CSD) framework to transfer contextual pairwise similarity knowledge from teachers to students. Suma et al. [13] designed a dual-branch structure to transfer the knowledge of pairwise similarity from teachers to students in two distinct feature spaces. Furthermore, Wu et al. [26] proposed a framework to preserve monotonic similarity to transfer knowledge of monotonic mapping pairwise similarities from teachers to students. However, pairwise similarity knowledge only considers the distance between two samples, which is prone to result in inconsistent retrieval ranking between the student and teacher networks, limiting the lightweight student network's accuracy performance.

3. Method

3.1. Formulation and background

Assume that $\theta_s(\cdot)$ denotes a lightweight student network and $\theta_t(\cdot)$ denotes a large teacher network. Given a batch of n image samples $\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n]$, we respectively use teacher and student networks to extract features as follows:

$$\mathcal{F}_i^t = \theta_t(\mathcal{X}_i), \quad \mathcal{F}_i^s = \theta_s(\mathcal{X}_i), \quad i = 1, 2, \dots, n, \tag{1}$$

where $\mathcal{F}_i^t$ and $\mathcal{F}_i^s$ are features of i−th sample $\mathcal{X}_i$ via teacher and student networks, respectively. For knowledge distillation-based object re-identification, the student network should learn useful ranking knowledge from the teacher network, that is, given a query image, returning highly similar gallery images with the same identity to the query image to help itself produce accurate retrieval results. For that, how to constrain the retrieval ranking of the student network to be consistent with that of the teacher network is a crucial problem for knowledge distillation-based object re-identification. Without losing generality, the ranking knowledge distillation loss function L_{rank} can be formulated as follows:

$$L_{rank} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\sum_{k=1}^{n}\big(\mathcal{H}(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t) - \mathcal{H}(\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s)\big)^2\Big)^{\frac{1}{2}}, \tag{2}$$

where $\mathcal{F}_i^s\mathcal{F}_j^s$ denotes the inner product between $\mathcal{F}_i^s$ and $\mathcal{F}_j^s$. As both $\mathcal{F}_i^s$ and $\mathcal{F}_j^s$ are ℓ_2 normalized, $\mathcal{F}_i^s\mathcal{F}_j^s$ is the cosine similarity between $\mathcal{F}_i^s$ and $\mathcal{F}_j^s$. $\mathcal{H}(\cdot)$ is the Heaviside step function [27] as follows:

$$\mathcal{H}(t) = \begin{cases} 1 & t > 0, \\ 0, & t \le 0. \end{cases} \tag{3}$$

In what follows, we analyze the limitations of pairwise similarity knowledge when it is used to approximate L_{rank}. Then, we design a pairwise difference relational knowledge (PDRK) to approximate L_{rank} more appropriately. Finally, we give a complete ranking knowledge distillation-based object re-identification method.

3.2. Pairwise similarity knowledge limitation analysis

Recent works [12,13] design a pairwise similarity distillation loss function $L_{pairwise}$ for retrieval as shown in Eq. (4).

$$L_{pairwise} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\big(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^s\mathcal{F}_j^s\big)^2\Big)^{\frac{1}{2}}. \tag{4}$$

However, the pairwise similarity knowledge loss function could not ensure consistent retrieval rankings between a lightweight student network and a teacher network because of the pairwise restrictive nature. More details are discussed as follows.

From Eq. (2), we can find that the retrieval ranking is determined by the difference between two sign (i.e., Heaviside) computations. For that, to maintain consistency in the retrieval ranking of student and teacher networks, there are two solutions: (1) $\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s > 0$ and $\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t > 0$; (2) $\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s < 0$ and $\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t < 0$. These two solutions can be unified into Eq. (5), as follows:

$$\frac{\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s}{\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t} > 0. \tag{5}$$

Obviously, Eq. (5) involves the relationship between two pairs (i.e., i and j, i and k). As shown in Eq. (4), there are n^2 pairs optimized, but each pair is independent, which could not take into account the relationship between the two pairs. To make matters worse, it is extremely difficult for a lightweight student to generate the same similarity to that of a well-trained teacher network for each pair, thus there are usually significant disparities among optimizations of different pairs. Consequently, the pairwise similarity knowledge loss function optimization is hard to ensure that Eq. (5) holds, so it has limitations to approximate L_{rank}.

3.3. Pairwise difference relational knowledge

Eq. (5) can be rewritten as:

$$\begin{aligned}\frac{\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s}{\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t} &= 1 - \frac{\left(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t\right) - \left(\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s\right)}{\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t} \\ &= 1 + \frac{\left(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^s\mathcal{F}_j^s\right) - \left(\mathcal{F}_i^t\mathcal{F}_k^t - \mathcal{F}_i^s\mathcal{F}_k^s\right)}{\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t}.\end{aligned} \tag{6}$$

According to Eq. (6), once $(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^s\mathcal{F}_j^s) - (\mathcal{F}_i^t\mathcal{F}_k^t - \mathcal{F}_i^s\mathcal{F}_k^s) \to 0$, Eq. (6) could be tended to 1. Therefore, we design a pairwise difference

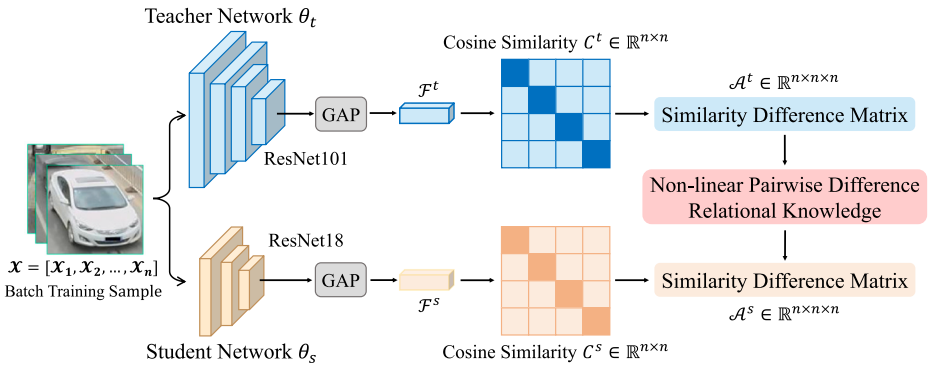

Fig. 2. The overall framework of the pairwise difference relational distillation (PDRD) for object Re-ID. Both teacher and student backbone networks are residual networks [16], but the teacher network is ResNet101 and the student network is ResNet18. GAP represents the global average pooling.

relational knowledge loss function L_{pdrk} to directly minimize $(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^s\mathcal{F}_j^s) - (\mathcal{F}_i^t\mathcal{F}_k^t - \mathcal{F}_i^s\mathcal{F}_k^s)$, as follows:

$$L_{pdrk} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\sum_{k=1}^{n}\big((\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^s\mathcal{F}_j^s) - (\mathcal{F}_i^t\mathcal{F}_k^t - \mathcal{F}_i^s\mathcal{F}_k^s)\big)^2\Big)^{\frac{1}{2}}. \tag{7}$$

Based on Eq. (7), mismatched signs between $(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^s\mathcal{F}_j^s)$ and $(\mathcal{F}_i^t\mathcal{F}_k^t - \mathcal{F}_i^s\mathcal{F}_k^s)$ would be greatly reduced, resulting in more consistent retrieval ranking between the student and teacher networks. Therefore, our pairwise difference relational knowledge would be better than the pairwise similarity knowledge where the student network has a low representation capacity.

In the practical knowledge distillation process, it is very difficult for a lightweight student network to approximate a well-trained large teacher network. Therefore, most researchers [7,15] believe that the transfer of knowledge between the student and teacher networks is a complex non-linear process. For that, we improve L_{pdrk} of Eq. (7) via non-linear activation functions to form a non-linear pairwise difference relational knowledge (NPDRK) loss function L_{npdrk}, which would further enhance the knowledge transferring from the teacher network to the student network, as follows:

$$L_{npdrk} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\sum_{k=1}^{n}\big(\sigma(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t) - \sigma(\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s)\big)^2\Big)^{\frac{1}{2}}, \tag{8}$$

where $\sigma(\cdot)$ is a non-linear activation function. In this paper, we consider three common non-linear activation functions, namely, Sigmoid [28], ReLU [29], and Mish [30], and the experiment analysis could be discussed in Section 4.5.2.

3.4. Complete object re-identification model

In the training phase, we adopt ResNet101 [16] as a teacher network and ResNet18 [16] as a student network to construct a pairwise difference relational distillation (PDRD) method for efficient object Re-ID, as shown in Fig. 2. In particular, the teacher network is frozen, while the student network is trainable.

Although knowledge distillation does not care about the training overhead of the student network, it remains imperative to employ the non-linear pairwise difference relational knowledge (NPDRK) loss function L_{npdrk} efficiently to minimize the additional computational overhead. In what follows, we describe how to apply the NPDRK loss function for object Re-ID in detail.

First, we use a cosine distance function to compute pairwise similarity matrices $C^t \in \mathbb{R}^{n\times n}$ and $C^s \in \mathbb{R}^{n\times n}$ for a teacher network and a student network, respectively, as follows:

$$C^t = \text{Cosine}(\mathcal{F}^t, \mathcal{F}^t), \quad C^s = \text{Cosine}(\mathcal{F}^s, \mathcal{F}^s). \tag{9}$$

Second, we construct the teacher network's pairwise similarity difference matrix $\mathcal{A}^t \in \mathbb{R}^{n\times n\times n}$ and the student network's pairwise similarity difference matrix $\mathcal{A}^s \in \mathbb{R}^{n\times n\times n}$ as follows:

$$\mathcal{A}^t_{i,j,k} = C^t_{i,j} - C^t_{i,k}, \quad \mathcal{A}^s_{i,j,k} = C^s_{i,j} - C^s_{i,k}, \quad 1 \le i,j,k \le n. \tag{10}$$

Thirdly, we perform the non-linear pairwise difference relational knowledge (NPDRK) loss function L_{npdrk} by minimizing the $\mathcal{L}_2$ distance between the teacher network's pairwise similarity difference matrix $\mathcal{A}^t$ and the student network's pairwise similarity difference matrix $\mathcal{A}^s$, as follows:

$$\begin{aligned} L_{npdrk} &= \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\sum_{k=1}^{n}\big(\sigma(\mathcal{A}^t_{i,j,k}) - \sigma(\mathcal{A}^s_{i,j,k})\big)^2\Big)^{\frac{1}{2}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\sum_{k=1}^{n}\big(\sigma(\mathcal{F}_i^t\mathcal{F}_j^t - \mathcal{F}_i^t\mathcal{F}_k^t) - \sigma(\mathcal{F}_i^s\mathcal{F}_j^s - \mathcal{F}_i^s\mathcal{F}_k^s)\big)^2\Big)^{\frac{1}{2}}, \end{aligned} \tag{11}$$

where L_{npdrk} is placed in the final residual layer of a residual network. Furthermore, to reduce computing consumption, we employ global average pooling (GAP) to compress the features that result from the final residual layer of the residual network during distillation, as in previous KD work [11,17,31].

Finally, a student network's total loss function L_{total} is as follows:

$$L_{total} = L_{lsrce} + L_{triplet} + \alpha L_{npdrk}, \tag{12}$$

where L_{lsrce} is the cross-entropy loss function using the label smooth regularization and the label smooth constant is set to 0.1, as done in [32]; $L_{triplet}$ is the triplet loss using hard sample mining strategy, as done in [3] ; They are common configurations for image retrieval to guarantee a good baseline performance, as done in the previous object Re-ID works [11,31]; α is set to 2.0 to control the contribution of L_{npdrk}.

4. Experiments

In this section, we conduct experiments on three pedestrian Re-ID datasets (that is, Market-1501 [33], DukeMTMC-reID [34] and MSMT17 [35]) and a vehicle Re-ID dataset (that is, VeRi-776 [36]) to validate the superiority of PDRD. Regarding the accurate performance indices, the mean average precision (mAP) [33] and the rank-1 identification rate (R1) [36] are applied. The cosine distance is applied as a similarity metric to sort gallery images. In what follows, we first introduce the datasets and implementation details. Secondly, we conducted ablation experiments and compared PDRD with state-of-the-art methods. Finally, we analyze the influence of hyper-parameter for PDRD. We release our code on GitHub, and the link is: https://github.com/SCY-X/PDRD.

4.1. Datasets

Market-1501 [33] is a pedestrian Re-ID dataset captured by six cameras. It contains 32,668 pedestrian images of 1,501 identities. The training subset includes 12,936 images of 751 identities, while the test subset holds images of the rest 750 identities, i.e., 19,732 gallery pedestrian images and 3,368 query pedestrian images.

DukeMTMC-reID [34] also is a pedestrian Re-ID dataset, which comprised 36,411 pedestrian images of 1,404 identities. The training subset contains 16,522 images of 702 identities. The test subset does not overlap with the training subset and contains 2,228 query images of 702 identities and 17,661 gallery images of 1,110 identities.

MSMT17 [35] is the largest pedestrian retrieval database, consisting of 126,441 images belonging to 4,101 pedestrian identities. The database comprises data captured by 3 indoor cameras and 12 outdoor cameras. The training set encompasses 32,621 training images pertaining to 1,041 distinct identities. On the other hand, the test set comprises 11,659 query images and 82,161 gallery images, representing a total of 3,060 identities.

VeRi-776 [36] is a vehicle Re-ID data set captured from real traffic scenarios using 20 surveillance cameras. It consists of 776 subjects. The training subset contains 37,746 images of 576 subjects, and the rest 200 subjects are applied for the testing subset. Furthermore, the test subset includes a probe subset of 1,678 images and a gallery subset of 11,579 images.

4.2. Implementation details

The software tools are Pytorch 1.12 [37], CUDA 11.6, and Python 3.9. The hardware device is one GeForce RTX 3090 GPU 24G. The common training configuration of four datasets is as follows: (1) We use ImageNet pre-trained ResNet [16] as a backbone network and set the last stride of ResNet to 1. (2) The data augmentation includes z-score normalization, random cropping, random erasing [4,7,31], and random horizontal flip operations, as done in [7,11]. (3) The mini-batch size is set to 96, randomly selecting 16 identities from a training set and each identity contains 6 images. (4) The weight decay and momentum values are set to 5×10^{-4} and 0.9, respectively.

There are some different configurations on different datasets. For three pedestrian datasets, the special training configuration is as follows: (1) The image resolution is normally set to 256 × 128, as done in previous pedestrian Re-ID works [4,31]. (2) In the training phase, the learning rate adjustment strategy is as follows: (a) The initial learning rate is 2×10^{-3}. (b) The learning rate linearly increases by a factor of 10 in the first 10 epochs, as done in [38]. (c) The cosine annealing strategy is applied to adjust the learning rate in the 60th epoch to the 150th epoch. For the vehicle dataset (i.e., VeRi-776 [36]), the special training configuration is as follows: (1) The image resolution is normally set to 256 × 256, as done in [11,31]. (2) In the training phase, the learning rate adjustment strategy is as follows: (a) The initial learning rate is 1×10^{-3}. (b) The learning rate linearly increases by a factor of 10 in the first 10 epochs, as done in [38]. (c) The cosine annealing strategy is applied to adjust the learning rate in the 40th epoch to the 120th epoch.

4.3. Ablation

Table 1 presents ablation results on DukeMTMC-reID [34] and VeRi-776 [36] datasets to investigate the effectiveness of PDRD. "Teacher" and "Student" denote that we directly evaluate the accuracy performance of ResNet101 [16] and ResNet18 [16] without any knowledge distillation, respectively. "Student+$L_{pairwise}$" denotes that we evaluate the accuracy performance of ResNet18 received pairwise similarity knowledge (i.e., Eq. (4)). "Student+L_{pdrk}" denotes that we evaluated the accuracy performance of ResNet18 [16] received pairwise similarity difference knowledge by the PDRK loss function (i.e., Eq. (7)) "Student+L_{npdrk}" denotes that we evaluate the accuracy performance of ResNet18 [16] receiving pairwise similarity difference knowledge by the NPDRK loss function (i.e. Eq. (8)).

Table 1
Ablation studies. $L_{pairwise}$ is the pairwise similarity loss function (i.e., Eq. (4)). L_{pdrk} is the pairwise difference relational knowledge loss function (i.e., Eq. (7)). L_{npdrk} is the non-linear pairwise difference relational knowledge loss function (i.e., Eq. (8)). The larger the values of both mAP and R1, the better the performance.

Methmod	Teacher	Student	DukeMTMC-reID		VeRi-776	
			mAP (%)	R1 (%)	mAP (%)	R1 (%)
Teacher	–	ResNet101	78.45	88.82	80.20	95.77
Student	–	ResNet18	68.88	82.94	73.84	94.40
Student + $L_{pairwise}$	ResNet101	ResNet18	72.20	84.61	75.27	94.52
Student + L_{pdrk}	ResNet101	ResNet18	74.33	85.82	76.88	95.19
Student + L_{npdrk}	ResNet101	ResNet18	**74.85**	**86.22**	**77.60**	**95.23**

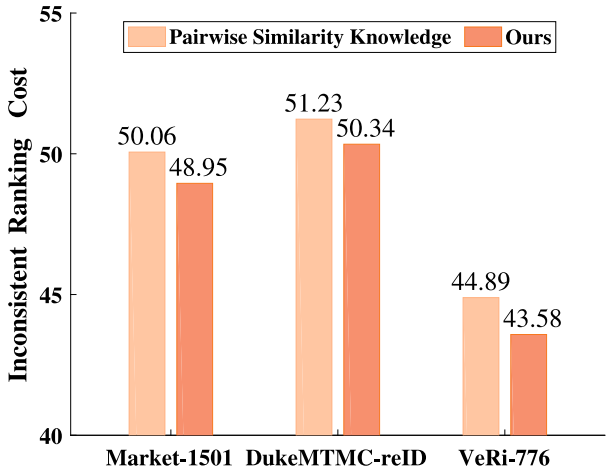

Fig. 3. The comparison result of the inconsistent ranking cost between pairwise similarity knowledge and our pairwise similarity difference knowledge on three datasets. The lower inconsistent ranking cost indicates a higher consistency between the ranking results of a student network and those of a teacher network. Our approach is better at maintaining this consistency.

DukeMTMC-reID [34]. From Table 1, we can find that the knowledge of the teacher network is helpful for the student network. First, compared to "Student", "Student+$L_{pairwise}$" acquires a higher accuracy performance of 3.22% mAP and 1.67% R1. This result demonstrates that pairwise similarity knowledge, as a widely used knowledge, is beneficial for a student network. Second, "Student+L_{pdrk}" achieves better results compared to "Student+$L_{pairwise}$" by 2.13% mAP and 1.21% R1. The comparison result demonstrates the superiority of our pairwise similarity difference knowledge over pairwise similarity knowledge. Finally, "Student+L_{npdrk}" surpasses "Student+L_{pdrk}" by 0.52% mAP and 0.40% R1, which shows that combining non-linear activation functions could better transfer pairwise similarity difference knowledge from teachers to students.

VeRi-776 [36]. From Table 1, we can find that the experimental ablation phenomenon on VeRi-776 [36] is consistent with that of DukeMTMC-reID [34]. For example, "Student+L_{pdrk}" obtains a improvement over "Student+$L_{pairwise}$" by 1.61% mAP and 0.67% R1. The result illustrates the superiority of pairwise similarity difference knowledge over pairwise similarity knowledge.

At last, in addition to the above re-identification directly related to performance comparison, we conduct an inconsistent ranking comparison to show our method's advantage in terms of objective function optimization. Specifically, in the inference phase, we first calculate the inconsistent ranking cost between the student and teacher networks according to Eq. (2). The experimental results on three datasets are presented in Fig. 3. From Fig. 3, we can find that using our pairwise similarity difference knowledge produces a lower inconsistent ranking

Table 2
Performance comparison on Market-1501 [33]. The larger the values of both mAP and R1, the better the performance.

METHOD[a]	TEACHEER	STUDENT	mAP (%)	R1 (%)	REFERENCE
FitNet [14] ∗	ResNet101	ResNet18	79.18	92.10	2015 ICLR
AT [19] ∗	ResNet101	ResNet18	81.81	92.90	2017 ICLR
FT [15] ∗	ResNet101	ResNet18	79.14	91.66	2018 NeurIPS
CCKD [10] ∗	ResNet101	ResNet18	79.13	91.89	2019 ICCV
SP [17] ∗	ResNet101	ResNet18	82.09	92.73	2019 ICCV
VID [39] ∗	ResNet101	ResNet18	78.27	90.83	2019 CVPR
RKD [25] ∗	ResNet101	ResNet18	82.12	92.43	2019 CVPR
PKT [25] ∗	ResNet101	ResNet18	82.02	93.14	2021 IEEE TNNLS
KDPE [24] ∗	ResNet101	ResNet18	78.11	91.36	2022 CVPR
CSD [12] ∗	ResNet101	ResNet18	81.64	92.07	2022 CVPR
PDRD	ResNet101	ResNet18	**84.76**	**93.62**	Ours

[a] The ∗ represents the result is re-implemented.

Table 3
Performance comparison on DukeMTMC-reID [34]. The larger the values of both mAP and R1, the better the performance.

METHOD[a]	TEACHEER	STUDENT	mAP (%)	R1 (%)	REFERENCE
FitNet [14] ∗	ResNet101	ResNet18	69.26	84.20	2015 ICLR
AT [19] ∗	ResNet101	ResNet18	73.91	86.27	2017 ICLR
FT [15] ∗	ResNet101	ResNet18	69.12	83.35	2018 NeurIPS
CCKD [10] ∗	ResNet101	ResNet18	68.30	83.80	2019 ICCV
SP [17] ∗	ResNet101	ResNet18	72.20	84.61	2019 ICCV
VID [39] ∗	ResNet101	ResNet18	68.42	83.03	2019 CVPR
RKD [25] ∗	ResNet101	ResNet18	71.76	84.92	2019 CVPR
PKT [25] ∗	ResNet101	ResNet18	72.44	85.37	2021 IEEE TNNLS
KDPE [24] ∗	ResNet101	ResNet18	68.24	82.45	2022 CVPR
CSD [12] ∗	ResNet101	ResNet18	70.13	84.56	2022 CVPR
PDRD	ResNet101	ResNet18	**75.63**	**87.52**	Ours

[a] The ∗ represents the result is re-implemented.

Table 4
Performance comparison on MSMT17 [35]. The larger the values of both mAP and R1, the better the performance.

METHOD[a]	TEACHEER	STUDENT	mAP (%)	R1 (%)	REFERENCE
FitNet [14] ∗	ResNet101	ResNet18	39.44	67.59	2015 ICLR
AT [19] ∗	ResNet101	ResNet18	46.78	71.93	2017 ICLR
FT [15] ∗	ResNet101	ResNet18	40.70	67.73	2018 NeurIPS
CCKD [10] ∗	ResNet101	ResNet18	40.64	68.56	2019 ICCV
SP [17] ∗	ResNet101	ResNet18	46.11	71.25	2019 ICCV
VID [39] ∗	ResNet101	ResNet18	40.34	67.92	2019 CVPR
RKD [25] ∗	ResNet101	ResNet18	44.54	70.35	2019 CVPR
PKT [25] ∗	ResNet101	ResNet18	46.73	72.17	2021 IEEE TNNLS
KDPE [24] ∗	ResNet101	ResNet18	39.12	66.87	2022 CVPR
CSD [12] ∗	ResNet101	ResNet18	44.06	70.79	2022 CVPR
PDRD	ResNet101	ResNet18	**47.99**	**72.25**	Ours

[a] The ∗ represents the result is re-implemented.

Table 5
Performance comparison on VeRi-776 [36]. The larger the values of both mAP and R1, the better the performance.

METHOD[a]	TEACHEER	STUDENT	mAP (%)	R1 (%)	REFERENCE
UMTS [40]	ResNet50	ResNet50	75.9	**95.8**	2020 AAAI
FitNet [14] ∗	ResNet101	ResNet18	74.17	94.40	2015 ICLR
AT [19] ∗	ResNet101	ResNet18	75.27	94.87	2017 ICLR
FT [15] ∗	ResNet101	ResNet18	74.42	94.10	2018 NeurIPS
CCKD [10] ∗	ResNet101	ResNet18	74.48	95.23	2019 ICCV
SP [17] ∗	ResNet101	ResNet18	75.27	94.52	2019 ICCV
VID [39] ∗	ResNet101	ResNet18	74.50	94.58	2019 CVPR
RKD [25] ∗	ResNet101	ResNet18	76.46	95.23	2019 CVPR
PKT [25] ∗	ResNet101	ResNet18	75.91	94.99	2021 IEEE TNNLS
KDPE [24] ∗	ResNet101	ResNet18	73.63	94.64	2022 CVPR
CSD [12] ∗	ResNet101	ResNet18	75.23	95.17	2022 CVPR
PDRD	ResNet101	ResNet18	**77.60**	95.23	Ours

[a] The ∗ represents the result is re-implemented.

cost than using pairwise similarity knowledge, which demonstrates that our method could better keep the ranking results of the student network consistent with that of the teacher network.

4.4. Comparison with the state-of-the-art

To comprehensively compare our method with state-of-the-art KD methods, we conduct comparisons on four public datasets. Tables 2, 3, 4, and 5 show comparisons on Market-1501 [33], DukeMTMC-reID [34], MSMT17 [35], and VeRi-776 [36] datasets, respectively. To ensure a fair comparison, we re-implement existing KD methods by using the same training tricks as our method. Both mAP and R1 are obtained using the student network during the inference phase. The comparison analyzes are discussed below.

Market-1501. As shown in Table 2, using the same teacher network (i.e., ResNet101 [16]) and the same student network (i.e., ResNet18 [16]), PDRD acquires the first place in accuracy performance among all compared methods. In particular, compared to RKD [25] with the second highest mAP, PDRD has a 2.64% larger mAP and a 1.19% higher R1. Additionally, compared to pairwise similarity knowledge distillation methods like SP [17] and CSD [12], PDRD's mAP and R1 are 2.67% larger and 0.89% higher than that of SP [17]. PDRD's mAP and R1 are 3.30% larger and 1.55% higher than that of CSD [12]. The above comparisons justify the advantage of PDRD on Market-1501 [33].

DukeMTMC-reID. As shown in Table 3, PDRD demonstrates favorable performance among all compared KD methods. Under the condition of using the same teacher network (i.e., ResNet101 [16]) and the same student network (i.e., ResNet18 [16]), PDRD achieves a 1.72% larger mAP and a 1.25% higher R1 than the second-place method (i.e., AT [19]). Furthermore, PDRD surpasses PKT [6] by 3.19% mAP and 2.15% R1, establishing a substantial performance advantage over all other competing approaches. These remarkable results show that PDRD is a state-of-the-art method for DukeMTMC-reID [34].

MSMT17. The comparisons of our PDRD and other state-of-the-art methods on the MSMT17 dataset are shown in Table 4. From Table 4, one can see that our PDRD method achieves the best performance among the compared methods, that is, 47.99% mAP and 72.25% R1. Compared to the excellent PKT [6] method that obtains 46.73% mAP and 72.17% R1, the PDRD method gets a 1.26% higher mAP. Furthermore, compared to newly published methods, our PDRD method exceeds KDPE [24] by 8.87% mAP and 5.38% R1, and surpasses CSD [12] by 3.93% mAP and 1.46% R1. It should be mentioned that MSMT17 is a large-scale dataset with severe challenges, so the superior performance of our PDRD method is a solid demonstration of its effectiveness.

VeRi-776. Comparisons on VeRi-776 [36] are shown in Table 5. First, PDRD obtains the highest mAP (i.e., 77.60%) and the second R1 (i.e., 95.23%) by using ResNet18 [16] as the student network. Although the R1 of PDRD is slightly lower (i.e., 0.57%) than that of UMTS [40], it should be noticed that that this comparison is unfair to PDRD because UMTS [40] is a self-distillation method that uses a large student network (i.e., ResNet50 [16]). Second, compared to pairwise relationship knowledge distillation methods (e.g., SP [17] and CSD [12]), PDRD has an obvious performance advantage. For example, PDRD surpasses SP [17] by 2.33% mAP and 0.71% R1, and exceeds CSD [12] by 2.37% mAP.

4.5. Hyper-parameter analysis

4.5.1. Influence of NPDRK loss function

We assign different values of α (see Eq. (12)) to investigate the influence of the NPDRK loss function on the accuracy performance of a student network. The experimental results are presented in Fig. 4.

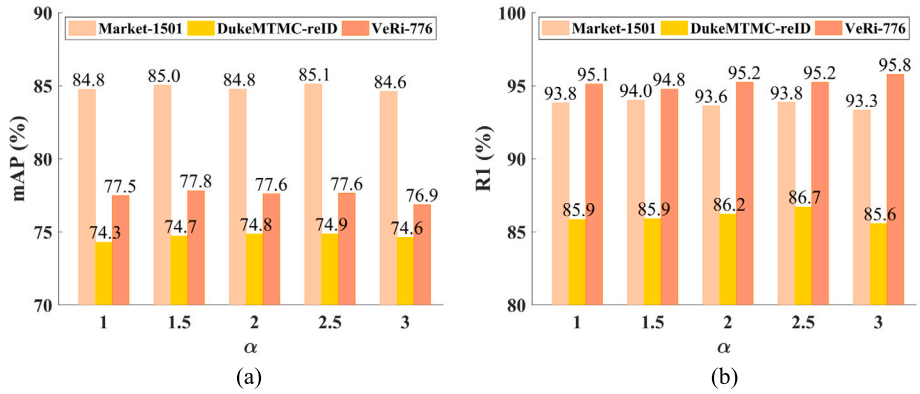

Fig. 4. The influence of NPDRK loss function weights α on (a) mAP and (b) R1 performance. The performance is relatively insensitive to the variation of α on each dataset.

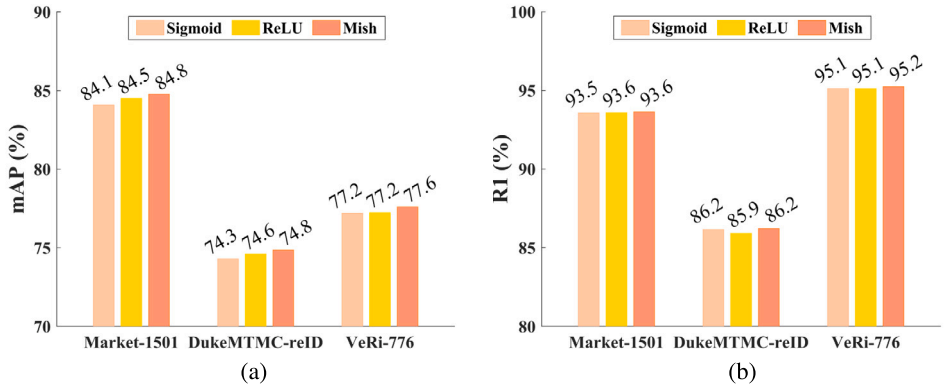

Fig. 5. The impact of changing different non-linear activation functions (i.e., Sigmoid, ReLU and Mish [30]) on (a) mAP and (b) R1 performance. The Mish activation function has better performance than the Sigmoid and ReLU activation functions.

From Fig. 4, we can observe that changing α has little effect on PDRD's performance. For example, on Market-1501 [33], the mAP fluctuates only by 0.5% when adjusting α between 1.0 and 3.0. This observation illustrates that the NPDRK loss function is certainly stable by changing α. This stability could be mainly attributed to the meticulous design of the NPDRK loss function, which effectively captures activation information by pairwise differences between samples.

4.5.2. *Impact of non-linear activation function choices*

In this experiment, we test different non-linear activation functions σ (in Eq. (8)) on three datasets. The results are shown in Fig. 5.

From Fig. 5, we can see that the worst performance among student networks occurs when the Sigmoid activation function [28] is used. The adverse performance of using a Sigmoid could be attributed to the tendency of the Sigmoid to experience gradient saturation. Although the ReLU activation function [29] is known for its ease of optimization, it is not optimal because it could encounter a significant loss of negative information when the similarity difference falls within the range of −2.0 to 2.0. Therefore, we recommend using the Mish activation function [30], which preserves negative information well to outperform both Sigmoid and ReLU in most scenarios.

5. Limitation discussion

Although PDRD could ensure consistent ranking results between lightweight student and large teacher networks, it has certain limitations. (1) PDRD may not be the ideal approach to self-distillation. This is because self-distillation primarily involves large-scale student networks with high representational capacity, which tend to generate ranking results highly similar to those of the teacher network. Therefore, the effectiveness of PDRD may be limited in self-distillation scenarios. (2) PDRD may only be effective for ranking-oriented tasks, such as image retrieval, gait recognition, and face recognition, where accurate ranking results are crucial. This is because the foundation of PDRD lies in minimizing the disparity between pairwise similarity discrepancies of teacher and student networks, thus guaranteeing consistent ranking results. In summary, while PDRD offers consistent ranking results for various network sizes, it may not be the optimal choice for self-distillation and is best suited for tasks that prioritize ranking.

6. Conclusion and future work

In this paper, we propose a pairwise difference relational distillation (PDRD) method to transfer pairwise similarity difference knowledge from a large teacher network to a lightweight student network, which effectively improves the student network's object re-identification accuracy performance. We analyze the limitation of commonly used pairwise similarity knowledge in previous works, i.e., pairwise similarity knowledge could not well guarantee consistent ranking of student and teacher networks. Furthermore, we theoretically prove that minimizing the difference between pairwise similarity discrepancies of student and teacher networks is beneficial to solving the limitation,

raising the pairwise difference rational knowledge (PDRK) loss function. Additionally, we reinforce PDRK via a non-linear mechanism to form a non-linear pairwise difference rational knowledge (NPDRK) loss function for object re-identification. Extensive experiments conducted on four public object Re-ID datasets show that our method yields state-of-the-art performance.

In this paper, we focus on improving the accuracy of the lightweight student network by reducing inconsistent ranking situations between teacher and student networks. But, the representational capacity gap between the lightweight student network and the large teacher network remains unresolved. In the future, we will investigate how to increase the representational capacity of a lightweight student network in the distillation phase to fill the representational capacity gap between the lightweight student network and the large teacher network.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported in part by the National Key R&D Program of China under the Grant of 2021YFE0205400, in part by the High-level Talent Innovation and Entrepreneurship Project of Quanzhou City under the Grant of 2023C013R, in part by the Natural Science Foundation for Outstanding Young Scholars of Fujian Province under the Grant of 2022J06023, in part by the National Natural Science Foundation of China under the Grant of 61976098, in part by the Project funded by the Fujian Provincial Natural Science Foundation under the Grant of 2023J02022, in part by the High-level Introduced Talent Team Project of Quanzhou City under the Grant of 2023CT001.

References

[1] Y. Xie, H. Wu, J. Zhu, H. Zeng, Distillation embedded absorbable pruning for fast object re-identification, Pattern Recognit. 152 (2024) 110437.
[2] Y. Huang, S. Lian, H. Hu, AVPL: Augmented visual perception learning for person Re-identification and beyond, Pattern Recognit. 129 (2022) 108736.
[3] F. Shen, J. Zhu, X. Zhu, Y. Xie, J. Huang, Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification, IEEE Trans. Intell. Transp. Syst. (2021) 1–12.
[4] F. Shen, Y. Xie, J. Zhu, X. Zhu, H. Zeng, Git: Graph interactive transformer for vehicle re-identification, IEEE Trans. Image Process. 32 (2023) 1039–1051.
[5] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: Conference on Neural Information Processing Systems Workshops, 2015.
[6] N. Passalis, M. Tzelepi, A. Tefas, Probabilistic knowledge transfer for lightweight deep representation learning, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2030–2039.
[7] Y. Xie, H. Zhang, X. Xu, J. Zhu, S. He, Towards a smaller student: Capacity dynamic distillation for efficient image retrieval, in: IEEE/CVF Computer Vision and Pattern Recognition, 2023, pp. 16006–16015.
[8] Y. Xie, F. Shen, J. Zhu, H. Zeng, Viewpoint robust knowledge distillation for accelerating vehicle re-identification, EURASIP J. Adv. Signal Process. 2021 (1) (2021) 1–13.
[9] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang, Decoupled knowledge distillation, in: IEEE/CVF Computer Vision and Pattern Recognition, 2022, pp. 11953–11962.
[10] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, Z. Zhang, Correlation congruence for knowledge distillation, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 5007–5016.
[11] Y. Xie, J. Zhu, H. Zeng, C. Cai, L. Zheng, Learning matching behavior differences for compressing vehicle re-identification models, in: IEEE International Conference on Visual Communications and Image Processing, 2020, pp. 523–526.
[12] H. Wu, M. Wang, W. Zhou, H. Li, Q. Tian, Contextual similarity distillation for asymmetric image retrieval, in: IEEE/CVF Computer Vision and Pattern Recognition, 2022, pp. 9489–9498.
[13] P. Suma, G. Tolias, Large-to-small image resolution asymmetry in deep metric learning, in: Winter Conference on Applications of Computer Vision, 2023, pp. 1451–1460.
[14] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, in: International Conference on Learning Representations, 2015.
[15] J. Kim, S. Park, N. Kwak, Paraphrasing complex network: Network compression via factor transfer, in: Conference on Neural Information Processing Systems, 2018, pp. 2765–2774.
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE/CVF Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[17] F. Tung, G. Mori, Similarity-preserving knowledge distillation, in: IEEE/CVF International Conference on Computer Vision, 2019, pp. 1365–1374.
[18] A. Porrello, L. Bergamini, S. Calderara, Robust re-identification by multiple views knowledge distillation, in: European Conference on Computer Vision, 2020, pp. 93–110.
[19] N. Komodakis, S. Zagoruyko, Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer, in: International Conference on Learning Representations, 2017.
[20] K. Yue, J. Deng, F. Zhou, Matching guided distillation, in: European Conference on Computer Vision, 2020.
[21] Y. Tian, D. Krishnan, P. Isola, Contrastive representation distillation, in: International Conference on Learning Representations, 2020.
[22] S.I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, H. Ghasemzadeh, Improved knowledge distillation via teacher assistant, in: AAAI Conference on Artificial Intelligence, 2020, pp. 5191–5198.
[23] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: IEEE/CVF Computer Vision and Pattern Recognition, 2018, pp. 4320–4328.
[24] R. He, S. Sun, J. Yang, S. Bai, X. Qi, Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability, in: IEEE/CVF Computer Vision and Pattern Recognition, 2022, pp. 9161–9171.
[25] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: IEEE/CVF Computer Vision and Pattern Recognition, 2019, pp. 3967–3976.
[26] H. Wu, M. Wang, W. Zhou, H. Li, A general rank preserving framework for asymmetric image retrieval, in: International Conference on Learning Representations, 2023.
[27] B. Davies, Integral transforms and their applications, 41 (2002).
[28] A.A. Waoo, B.K. Soni, Performance analysis of sigmoid and relu activation functions in deep neural network, in: Intelligent Systems, Springer, 2021, pp. 39–52.
[29] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Conference on Neural Information Processing Systems, 2012, pp. 1106–1114.
[30] D. Misra, Mish: A self regularized non-monotonic activation function, in: British Machine Vision Conference, 2020.
[31] Y. Xie, H. Wu, F. Shen, J. Zhu, H. Zeng, Object re-identification using teacher-like and light students, in: British Machine Vision Conference, 2021.
[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: IEEE/CVF Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[33] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: IEEE/CVF International Conference on Computer Vision, 2015, pp. 1116–1124.
[34] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: IEEE/CVF International Conference on Computer Vision, 2017, pp. 3754–3762.
[35] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: IEEE/CVF Computer Vision and Pattern Recognition, 2018, pp. 79–88.
[36] X. Liu, W. Liu, H. Ma, H. Fu, Large-scale vehicle re-identification in urban surveillance videos, in: International Conference on Multimedia & Expo, 2016, pp. 1–6.
[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Conference on Neural Information Processing Systems, 2019, pp. 8024–8035.
[38] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, large minibatch sgd: Training imagenet in 1 hour, 2017, arXiv preprint arXiv:1706.02677.
[39] S. Ahn, S.X. Hu, A. Damianou, N.D. Lawrence, Z. Dai, Variational information distillation for knowledge transfer, in: IEEE/CVF Computer Vision and Pattern Recognition, 2019, pp. 9163–9171.

[40] X. Jin, C. Lan, W. Zeng, Z. Chen, Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification, in: AAAI Conference on Artificial Intelligence, 2020, pp. 11165–11172.

Yi Xie received the B.S. degree in Communication Engineering from Guangdong Polytechnic Normal University, Guangzhou, China, in 2019 and the M.S. degree in Computer Science and Technology from Huaqiao University, Quanzhou, China, in 2022, respectively. He is currently pursuing the D.Eng. degree with the School of Future Technology, South China University of Technology. His current research interests include knowledge distillation and image network.

Hanxiao Wu eceived the B.S. degree in Internet of Thing Engineering, Huaqiao University, Quanzhou, China, in 2020 and received the M.S. degree in Information and Communication Engineering, Huaqiao University, Quanzhou, China, in 2023. Now, she is pursuing PhD degree in Computer Science and Technology, Wuhan University of Technology. Her research interests are object re-identification and multi-modal learning.

Yihong Lin received the B.S. degree in Mathematics and Applied Mathematics from South China University of Technology, Guangzhou, China, in 2022. He is currently pursuing the M.S. degree with the School of Computer Science and Engineering, South China University of Technology. His current research interests include image generation and 3D reconstruction.

Jianqing Zhu received the Ph.D. degree in Computer Applications Technology from Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is now a Professor at Huaqiao University, Quanzhou, China. His research interests include computer vision and pattern recognition.

Huanqiang Zeng received the Ph.D. degree in Electrical Engineering from Nanyang Technological University, Singapore. He is now a Professor at Huaqiao University, China. His research interests are in the areas of image processing and video coding, pattern recognition, and computer vision.