Neurocomputing 455 (2021) 111–124

Contents lists available at ScienceDirect

Neurocomputing

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

Hashing person re-ID with self-distilling smooth relaxation ☆

Hanyang Jin [a,1], Shenqi Lai [a,1], Guoshuai Zhao [b], Xueming Qian [a,c,*,2]

[a] SMILES LAB, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[b] School of Software Engineering, Xi'an JiaotongUniversity, Xi'an, Shaanxi 710049, China
[c] Key Laboratory for Intelligent Networks and Network Security , Ministry of Education, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China

ARTICLE INFO

Article history:
Received 21 December 2020
Revised 1 April 2021
Accepted 17 May 2021
Available online 20 May 2021
Communicated by Zidong Wang

Keywords:
Person re-ID
Deep hashing
Attribute learning
Knowledge distillation

ABSTRACT

Person re-identification (re-ID) has made substantial progress in recent years; however, it is still challenging to search for the target person in a short time. Re-ID with deep hashing is a shortcut for that but, limited by the expression of binary code, the performance of the hashing method is not satisfactory. Besides, to further speed up retrieval, researchers tend to reduce the number of feature bits, which will cause more performance degradation. In this paper, we design the attribute-based fast retrieval (AFR), which leverages the attribute prediction of the model trained in a binary classification manner tailor-made for hashing. The attribute information is also used to refine the global feature representation by an attribute-guided attention block (AAB). Then, to fully exploit deep feature to generate the hash codes, we propose a binary code learning method, named self-distilling smooth relaxation (SSR). In this method, a simple yet effective regularization is presented to distill the quantized knowledge in the model itself, thus mitigating the lack of semantic guidance in the traditional non-linear relaxations. We manually label attributes for each person in dataset CUHK03 and evaluate our method on four authoritative public benchmarks (Market-1501, Market-1501+500K, CUHK03, and DukeMTMC-reID). The experimental results indicate that with the SSR and AAB, we surpass all the state-of-the-art hashing methods. And compared with reducing the feature bits, the AFR strategy is more effective to save search time.

1. Introduction

Person retrieval, or person re-identification (re-ID), is aimed at finding targeted people from the non-overlapping cameras. The changes in lighting, pedestrian posture, and angles of view will incur evident changes in the appearance of the same person. With the rapid development of convolutional neural networks (CNN) and big data, deep learning methods [4,21,19,27,53,54,57] become the most common measures for feature extraction. However, in the retrieval process, directly calculating the similarity of raw high-dimensional features between gallery and query, is computationally intensive. And storing these floating-point features is also costly to memory resources. Hashing [8,7,55,56] is a technique to encode high-dimensional features into compact binary codes while preserving the similarity of images. In the hashing methods, fast image retrieval can be carried out by computing the Hamming distance between samples, which dramatically decreases computational costs.

Two factors could significantly affect the performance of hashing re-ID, namely, the feature representation and hash relaxation. The former refers to the ability of the network to form discriminative expression for each pedestrian, compared with floating-point features of non-hashing methods, hashing binary codes have an inherent disadvantage in expression ability. As to the hash relaxation, *i.e.*, in retrieval, for calculating the similarity metric in Hamming space, features must be quantified to discrete values such as −1, 0, or +1. In the training process, to avoid optimizing the non-differentiable loss function in Hamming space, neural network outputs are relaxed to binary-like real values. But, the process of learning binary-like codes is a shallow learning procedure and cannot fully exploit the representation, thus causes information-loss of features and compromises the performance [20].

There is another issue we concern in hashing re-ID: to further accelerate the retrieval, the most common way used is to reduce the bit-length of hash codes. But as we can see in these hashing

☆ This work was supported in part by the NSFC under Grant 61772407 and 61732008, and Microsoft Research.
* Corresponding author at: SMILES LAB, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China.
E-mail addresses: jhy0606@stu.xjtu.edu.cn (H. Jin), laishenqi@stu.xjtu.edu.cn (S. Lai), guoshuai.zhao@xjtu.edu.cn (G. Zhao), qianxm@mail.xjtu.edu.cn (X. Qian).
[1] Authors Hanyang Jin and Shenqi Lai contribute equally to this work.
[2] ORCID: 0000-0002-3173-6307

https://doi.org/10.1016/j.neucom.2021.05.059
0925-2312/

researches [16,22,3,2], as the number of bits decreases, the re-ID accuracy would sharply drop. Even worse, the relaxation and binary quantization in hashing inevitably lead to information-loss, making it even harder for hashing re-ID to achieve the same performance as non-hashing methods.

Pedestrian attribute, containing detailed local descriptors, are often exploited as mid-level human semantic information of the pedestrian [26,14,28]. However, we find that after modifying attribute learning to binary classification settings, the attribute information can be conveniently used to coarsely screen pedestrians in the hashing scenarios. Different from person re-ID where the IDs in the training set and gallery set are non-overlapping, labels of the attribute in training and testing are identical. This motivates us to impose the attribute learning into a multi-task hashing model, and leverage the attribute prediction to provide effective description in different granularity from ID features.

Therefore, for the purpose of faster and stronger hashing person re-ID, in this paper, we present a self-distilled fast hashing re-ID framework. Specifically, we first propose an attribute-based fast retrieval (AFR) strategy to speed up the retrieval. In AFR, pedestrian attributes are learned in a binary classification manner, and serve as the coarse-grained identity information to cooperate with fine-grained features in the retrieval. Unlike most existing methods simply embedding attribute learning into the classification task. We adopt two types of attention mechanisms from aspects of spatial and channel in the attribute learning, to overcome the re-ID performance deterioration caused by heteroscedasticity (a mixture of different knowledge granularity and characteristics) learning problem [33]. Then, to fully exploit deep feature representations to generate the hash codes, we propose a binary code learning method, named self-distilling smooth relaxation (SSR). In this method, a simple yet effective regularization is presented to distill the quantized knowledge of the model itself, thus mitigating the lack of semantic guidance in the traditional non-linear relaxations. Moreover, since the SSR imposes a softer feature constraint than traditional relaxations, it can especially preserve the feature discrimination dividend brought by batch normalization (BN), which is a commonly used tool in the person re-ID. Finally, to make our evaluation more comprehensive, following the annotation standard on Market-1501 and DukeMTMC-reID, we manually label attributes for each person in dataset CUHK-03 and evaluate our framework on four datasets: Market-1501, Market-1501+500k, Duke-MTMC-reID, CUHK03.

The main contributions can be summarized as follows.

(1) We propose the attribute-based fast retrieval (AFR) strategy to speed up the retrieval for hashing person re-ID. With the binary attribute learning modification tailor-made for hashing re-ID, the AFR shortens the retrieval time more effectively than reducing the number of bits;
(2) We propose a binary code learning method for hashing re-ID named self-distilling smooth relaxation (SSR), which progressively distills a model's own knowledge to soften hard binary-like relaxation targets. With the refined semantics from the deep features, the SSR learns more discriminative capability of hash codes than traditional relaxation functions;
(3) The attribute information is also used to refine the global feature with a proposed attribute-guided attention block (AAB), which leverages the spatial properties and high-level semantic cues contained by attribute, to help to enhance the feature representation;
(4) To reach a more comprehensive evaluation, we manually annotate 27 attributes of pedestrians in the CUHK03 benchmark. Comprehensive experiments on four public benchmarks (Market-1501, Market-1501+500K, Duke-MTMC-reID, and CUHK03) confirm the efficacy of the proposed approach over state-of-the-art methods.

The rest of this paper is organized as follows. In Section 2, we review the related work. In Section 3, we describe the flow of our entire algorithm. In Section 4, we describe the proposed self-distilling smooth relaxation. The attribute-based fast retrieval is introduced in Section 5. Loss function is presented in Section 6. Experimental results are presented in Section 7. Finally, we conclude the paper in Section 8.

2. Related works

In this section, existing works on person re-ID are reviewed in Section 2.1. Since our work focuses on deep hashing re-ID and employs the pedestrian attributes, we also review related studies of hashing methods and attribute learning in Sections 2.2 and 2.3.

2.1. Methods of person re-ID

The person re-ID approaches can be classified into two streams: seeking robust pedestrian feature representation [41,13,21,44,19,29] and learning discriminative distance metrics [1,23,42]. The first stream concentrates on building feature descriptions, which preserves the identity information in various postures across different cameras. The second stream generally designs proper metric distances to minimize the intra-class distance while maximizing the inter-class distance.

For feature representation, the literature [41] proposed a Siamese network using two sets of convolutional layers and fully connected layers, which jointly learned the colour features, texture and similarity metric. In [21], the authors proposed the multi-filters neural network for feature extraction on financial time series samples and price movement prediction task. Parted-based methods [13,29,44], which are designed to focus on local regions and capture fine-grained cues, are expected to be more effective and robust for feature representation. [13] employed human semantic parsing network to harness local visual cues. Semantic segmentation also helps to reduce the disturbance of complex background. [44] estimated the dense semantics first, which can be used to warp the original RGB image to the representation in UV space. [37] proposed an attention-based method for visible-infrared person re-ID. Method such as [40] also proposed a PurifyNet to tackle the label noise problem. For unsupervised learning of re-ID, [39] studied the unsupervised embedding learning problem by learning representation without using any category labels. In [29], the authors provided a compact network named part-based convolutional baseline (PCB) for person re-ID which horizontally divides the feature map into several stripes. For the sake of lower computation and memory cost, which is also the original purpose of the hashing technique, we use PCB as the basic network structure to construct our framework. For metric learning, in [1], the authors split the metric into independent colour and texture components. Liu et al. [23] built a method on reciprocal nearest neighbour search. Yu et al. [42] introduced an unsupervised re-ID method based on asymmetric clustering. In this paper, we adopt the hard example mining proposed in [10] to do the metric learning, which is a variant of the triplet loss.

2.2. Hashing methods

The traditional learning-based hashing methods usually deal with the hand-crafted image features, which can hardly obtain optimal hashing representations for semantic structures in images.

To effectively capture the semantic relationships between images, many researchers have drawn support from deep learning to hashing. Xia et al. [36] adopted a two-stage deep hashing algorithm CNNH. The first stage learned the approximate hash code, and the second stage learned the hashing functions and feature representations of an input image. In [46], Zhao et al. proposed Deep semantic ranking based hashing to preserve multi-level semantic similarity between multi-label images. Zhang et al. [43] posed hashing learning as a problem of regularized similarity learning. Similar frameworks such as DCH [32] and DSH [20] were also proposed and both achieved state-of-the-art performance. In DSH [20], the authors proposed a regularization relaxation method and gained encouraging performance for image retrieval. There are also hashing researches specifically designed for person re-ID. Based on discrete alternating optimization, [3] proposed a joint hashing person re-ID learning framework for subspace projection learning and binary coding. [22] proposed an adversarial binary coding to fit the feature distribution to the expected binary one by optimizing the Wasserstein distance. More recently, [16] focused on the consistency preservation of hash code and proposed a deep hashing framework for person re-ID, which improved robustness of both hash code and high-dimensional feature.

2.3. Attribute learning

Attributes are usually viewed as a mid-level semantic description for feature representation learning. Su et al. [28] considered multiple cameras as related tasks and learned a discriminative network by multi-task attribute learning. Khamis et al. jointly optimized the triplet loss for re-ID and attributes identity loss in [14]. In [26], fine-tuned CNN was embedded for attribute classification. Then, Lin et al. [18] and Chen et al. [6] followed the CNN based attribute learning approach and both achieved competitive results. Lin et al. [18] manually annotated the Market-1501 [47] and DukeMTMC-reID [49] datasets with attribute labels. For each ID, there is an adequate number of training samples for attribute learning, which the other attribute datasets do not have. In this paper, we follow the label criterion in [18] and annotate 27 attributes for another authoritative re-ID dataset CUHK03.

3. General pipeline

The general pipeline of the proposed framework is shown in Fig. 1. During training, each pedestrian image is first fed into the backbone network. The architectures after the second layer of the backbone are copied for generating a pair of feature maps, which are used to construct the global–local feature representation of teacher branch and student branch. The student branch will distill binary semantic knowledge from the teacher branch as the training proceeds. In each branch, the global feature is refined by attribute-guided attention block (AAB) to emphasize meaningful pedestrian body parts. The binary attribute learning is conducted on the global description vector, and both global and local vectors are imposed with the self-distilling smooth relaxation (SSR), ID classification learning, and triplet hard example mining. In the testing, predicted attribute probabilities and ID features are quantified to hash codes, which are then used to construct the attribute-based fast retrieval (AFR).

4. Quantized semantic self-distillation

In this section, we describe the global–local feature extraction, and the self-distilling smooth relaxation.

4.1. Global–local representation

We use a classification network as the backbone to rapidly build the retrieval framework, and employ the ResNet [9] as the backbone considering its superior performance and relatively concise structure. Besides, for the sake of lower computation and memory cost, which is also the motive of hashing technique, we choose a simple yet strong method part-based convolutional baseline (PCB) [29] as our basic network to obtain the rough local representation. As shown in the training process in Fig. 1, we use two branches to generate the teacher binary descriptor (the upper branch) and student relaxed descriptor (the lower branch). In each branch, the feature map is first horizontally cut into 6 stripes as local features. Based on it, we impose the global representation refined by AAB. With a 1×1 convolution (which is not shown in Fig. 1 for brevity) to reduce the dimensionality, we get the teacher binary descriptor $\boldsymbol{S}^t = [\boldsymbol{S}_L^t, \boldsymbol{S}_G^t]$, which consists of local features $\boldsymbol{S}_L^t = [\boldsymbol{s}_{L1}^t, \boldsymbol{s}_{L2}^t, \ldots, \boldsymbol{s}_{L6}^t]$ and global features $\boldsymbol{S}_G^t = [\boldsymbol{s}_G^t]$. Similarly, the student descriptor is noted as $\boldsymbol{S}^s = [\boldsymbol{S}_L^s, \boldsymbol{S}_G^s]$.

4.2. Self-distilling smooth relaxation

In the hashing methods, binary feature descriptors are sorted by Hamming distance in retrieval. Still, in the training process, the binary values cannot be directly optimized with the standard back-propagation algorithm. While due to the discrepancy between Euclidean space and Hamming space, the real-valued learned features are suboptimal for retrieval. Thus, relaxation, namely, appending a binary-like hidden layer before the classifiers, is commonly used in the training process. These relaxations appear as non-linear activation functions such as *Sigmoid*, *Tanh* or *Hardsigmoid*, whose outputs are values from 0 to $+1$, or -1 to $+1$. In the testing end, relaxed features are quantized to binary code for retrieval.

As shown in Fig. 2, we preserve one batch of trained features $\boldsymbol{S}^t$ of the images, and plot the distributions of $\boldsymbol{S}^t$ under four types of relaxations. We can see that as these saturated non-linear functions work, rigid distribution constraints restrict the distributions of these feature values into a more narrow scope than original features. However, we think these constraints also narrow the expression space of features and cause information loss. One basic assumption of our algorithm is that the learning of binary-like features should not only be constrained by rigid distribution regularization, but also supervised by more semantic information. Based on this assumption, we develop a self-distilling smooth relaxation. As shown in Fig. 1, the teacher descriptor $\boldsymbol{S}^t$ is first quantized to hashing binary codes without relaxation of any nonlinear function. For each element s in $\boldsymbol{S}^t$, the quantization is performed as:

$$s' = \begin{cases} 1, s > 0 \\ -1, s \leqslant 0, \end{cases} \tag{1}$$

After quantifying the teacher descriptors $\boldsymbol{S}^t$ into $\mathbb{S}^t$, we transfer the binary semantic knowledge from $\mathbb{S}^t$ to the student descriptor $\boldsymbol{S}^t$, which is then used for constructing identity representation of pedestrian image in retrieval. In this way, we soft the optimization target from rigid distribution regularizing to binary semantic learning, and the real-valued hash codes are encouraged to approach the desired discrete values effectively. The distilling is implemented with a L1 loss:

$$\mathcal{L}_d = \|\mathbb{S}^t - \boldsymbol{S}^s\|_1, \tag{2}$$

where $\|\cdot\|$ is the L1-norm of vector, the $\mathbb{S}^t$ and $\boldsymbol{S}^s$ are both vectors concatenated from their own global–local descriptors. In the retrieval process, the concatenated student descriptor $\boldsymbol{S}^s$ is also quantized

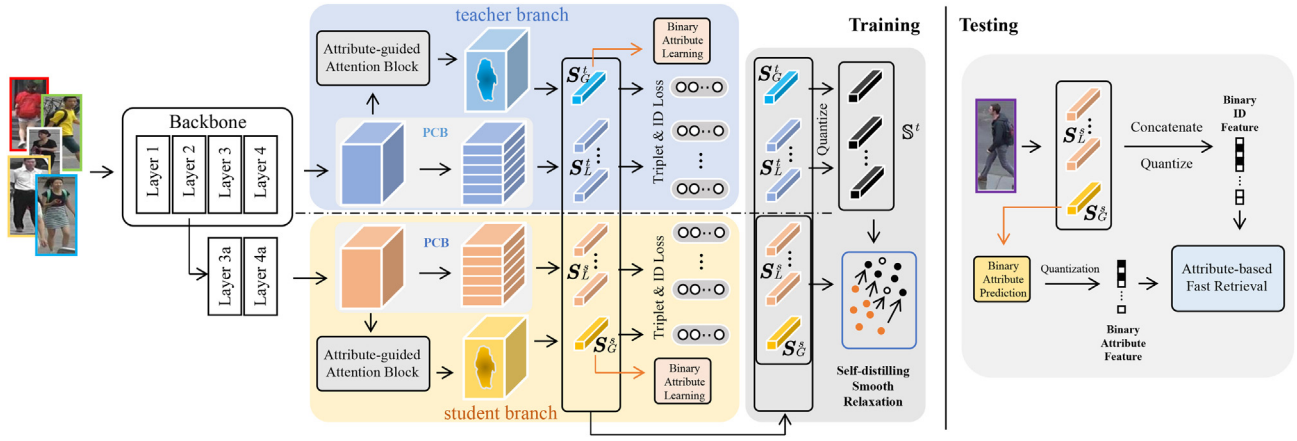


Fig. 1. The general pipeline of proposed framework.

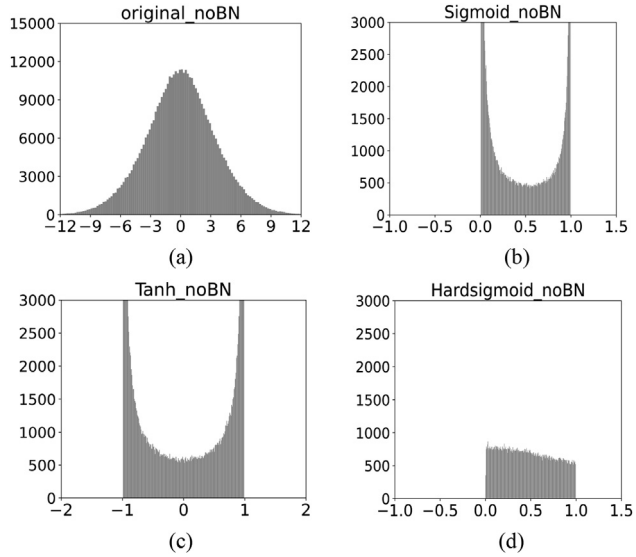


Fig. 2. The distributions of re-ID model output features(without BN). (a) is the distribution of non–hash features on Market-1501 dataset. (b), (c) and (d) are respectively the distribution of features that relaxed by the nonlinear function Sigmoid, Tanh and Hardsigmoid.

in the manner of Eq. (1) and used as ID representation of pedestrian to conduct the Hamming distance ranking.

5. Attribute-based fast retrieval

This section describes the binary classification attribute learning, fast retrieval, and the attribute-guided refinement to global information.

5.1. Binary attribute learning and fast retrieval

In order to make use of attribute prediction to accelerate the retrieval in our hash framework, we replace the classification attribute learning with the binary classification attribute learning. As shown in Fig. 3, take Market-1501 as an example. There are 27 attributes annotated including 26 attributes with two categories (such as gender: male/female, hair length: long/short, etc.), and one attribute with four categories (age: young/teenager/adult/old). Different from the standard classification learning, we use binary classification settings and consider each attribute as an independent binary classification problem. We formulate the prediction of age as four binary classification tasks: young-yes/no, teenager-yes/no, adult-yes/no, and old-yes/No. Then the attribute hair length is modified to longhair-yes/no, etc. As shown in Fig. 3, a Sigmoid activation is imposed after each FC layer of classification branch to fit the 0–1 prediction task. Finally, we have 30 binary classifications for attribute learning. In the training, each attribute's leaning is implemented with a fully-connected layer followed by a Sigmoid function, and the binary cross entropy (BCE) loss. Let $\boldsymbol{D} = \{(x_1, l_1, \boldsymbol{a}_1), \ldots, (x_N, l_N, \boldsymbol{a}_N)\}$ be the pedestrian training set, where x_i, l_i and $\boldsymbol{a}_i$ denote the i-th image, its identity label, and its attributes annotations. N is the total number of images. We can divide $\boldsymbol{D}$ into two parts: $\boldsymbol{D}_L = \{(x_1, l_1), \ldots, (x_N, l_N)\}$ and $\boldsymbol{D}_{Attr} = \{(x_1, \boldsymbol{a}_1), \ldots, (x_N, \boldsymbol{a}_N)\}$, which denote identity labelled set and attribute labelled set (note that $\boldsymbol{D}_{Attr}$ and $\boldsymbol{D}_L$ share the common pedestrian images $\{x_i\}$). Assume that the output of one attribute's FC+Sigmoid layer is $\boldsymbol{z} = [z_0, z_1]$(two categories yes:1 and no:0), the probability of assigning sample x to the attribute class $j \in 0,1$ can be written as:

$$p(j|x) = \frac{\exp(z_j)}{\sum_{n=0}^{1} \exp(z_n)}, \tag{3}$$

for brevity, we omit the correlation between j and x. So, the overall binary cross entropy (BCE) is formulated as:

$$\mathcal{L}_{attr}(\boldsymbol{S}_i, \boldsymbol{a}_i) = -\frac{1}{30} \sum_{\varepsilon=1}^{30} \sum_{j=0}^{1} \log(p(j)) q(j). \tag{4}$$

Where 30 is the total number of attributes. Let y_a be the ground-truth of this attribute label, so that $q(y_a) = 1$ and $q(j) = 0$ for $j \neq y_a$. $\boldsymbol{S}_i$ and $\boldsymbol{a}_i$ are the feature descriptor used for attribute learning and attribute annotations of x_i, respectively. In testing, the proposed attribute-based fast retrieval consists of a fast attribute coarse selection and an ID features-based search. The predicted probabilities of attributes are quantized to binary codes. Assume that the predicted attributes probability set of one image is $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{30}]$, in which each vector $\boldsymbol{z}_v = [z_{v0}, z_{v1}]$ is transformed to a scalar:

$$z'_v = \begin{cases} 1, z_{v0} > z_{v1} \\ -1, z_{v0} \leqslant z_{v1} \end{cases}. \tag{5}$$

The quantized attribute probabilities serve as the attribute binary codes $\boldsymbol{C}_{attr}$ for the first-stage retrieval. The $\boldsymbol{S}^s$ is quantized to ID binary codes $\boldsymbol{C}_{ID}$ by Eq. (1) for the second-stage retrieval. Specifically, for each search we have one target query, let $[\boldsymbol{C}^q_{attr}, \boldsymbol{C}^q_{ID}]$ denotes the hash codes of this query image, and the corresponding

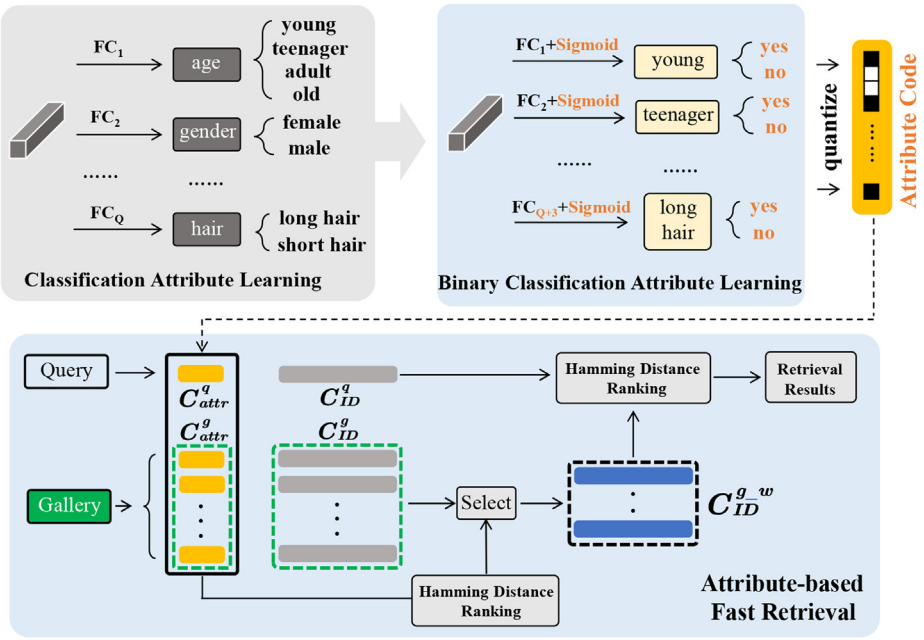


Fig. 3. The diagram of classification attribute learning, proposed binary classification attribute learning, and the attribute-based fast retrieval (AFR).

codes of the all gallery image are denoted as $[\boldsymbol{C}^g_{attr}, \boldsymbol{C}^g_{ID}]$. First, we choose the attribute hash codes $\boldsymbol{C}^q_{attr}$ and $\boldsymbol{C}^g_{attr}$ to calculate SIM_{attr} for the first-step similarity measurement:

$$SIM_{attr} = dist_{ham}\boldsymbol{C}^q_{attr}, \boldsymbol{C}^g_{attr}. \tag{6}$$

Where $dist_{ham}$ is the Hamming distance matrix of vectors. We rank the result and select the first w candidates (the closest w gallery images to the query). Then, the selected gallery's ID features $\boldsymbol{C}^{g\_w}_{ID}$ are used to do the second-stage search with query ID features $\boldsymbol{C}^q_{ID}$ by SIM_{ID}:

$$SIM_{ID} = dist_{ham}\boldsymbol{C}^q_{ID}, \boldsymbol{C}^{g\_w}_{ID}. \tag{7}$$

Then the ranking and evaluation of rank-k (k can be 1, 5, 10) and mAP [47] are conducted.

5.2. Attribute-guided attention block

Since attribute information (see in attribute annotations) such as clothes color and length of upper/lower body locate in different parts of the human body, it is difficult to learn attributes using horizontally divided local features. So, in our framework, we conduct attribute learning only on global descriptors. However, the attribute and ID learning focus on different knowledge granularity, simply combing them could impair the discrimination of ID features. Therefore, we fuse attribute and ID recognition tasks not only on loss level, but also on feature level. Inspired by the success in attention mechanism [48,11,45], we propose to leverage the spatial properties and high-level semantic cues contained by attribute, to help filter out background interference and enhance the feature representation.

Specifically, the attention mechanism module includes spatial attribute attention block (SAAB) and channel attribute attention block (CAAB). By exploring the long-range dependencies between pixels or channels, these two blocks are expected to learn to capture local part structural information and highlight meaningful channels during training. In SAAB and CAAB, as shown in Fig. 4, to reduce the computation costs, we first do 1×1 convolution to reduce the channels of input feature F_{In} and get two sets of feature maps: $F^{S1}, F^{S2}, F^{S3} \in \mathbb{R}^{C/r \times H \times W}$ for SAAB, and $F^{C1}, F^{C2}, F^{C3} \in \mathbb{R}^{C/r \times H \times W}$ for CAAB, and reshape them to $\mathbb{R}^{C/r \times X}$ $(X = H \times W)$. Where r is a reduction ratio to reduce the channel dimension, C, H, and W respectively represent the number of channels, height, and width of the feature map. Then we perform matrix multiplication between F^{S1} and transposed F^{S2}, and normalize the result to obtain the spatial semantic relation map $\mathcal{R}^S \in \mathbb{R}^{X \times X}$. Simultaneously, F^{C2} and the transposed F^{C1} are multiplied to obtain the channel relation map $\mathcal{R}^C \in \mathbb{R}^{C/r \times C/r}$. Specifically, the semantic similarity between any two pixels is calculated as:

$$\mathcal{R}^S_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_{p=1}^{X} \exp\left(f_i^T f_p\right)}, \tag{8}$$

where $f^i, f^j \in \mathbb{R}^C$ denote the features in the i_{th} position of F^{S1} and j_{th} position of $\left(F^{S2}\right)^T$. Accordingly, the semantic similarity between any two channels is calculated as:

$$\mathcal{R}^C_{nm} = \frac{\exp\left(f_n^T f_m\right)}{\sum_{l=1}^{C/r} \exp\left(f_n^T f_l\right)}, \tag{9}$$

where $f^n, f^m \in \mathbb{R}^C$ denote the features in the n_{th} and m_{th} channel of F^C. Then, two relation maps are multiplied with F^{S3} and F^{C3} to obtain the aggregated attention feature. After reshaping and a dimensionality raising 1×1 convolution, we get the outputs of two attention blocks F^S_{Out} and F^C_{Out}. Finally, they are added together and serve as a 3-D mask to dot with F_{In}:

$$F^{ref} = \left(F^S_{Out} + F^C_{Out}\right)F_{In} + F_{In}, \tag{10}$$

where F^{ref} is the final refined global feature whose discrimination is enhanced by spatial-channel attribute attention.

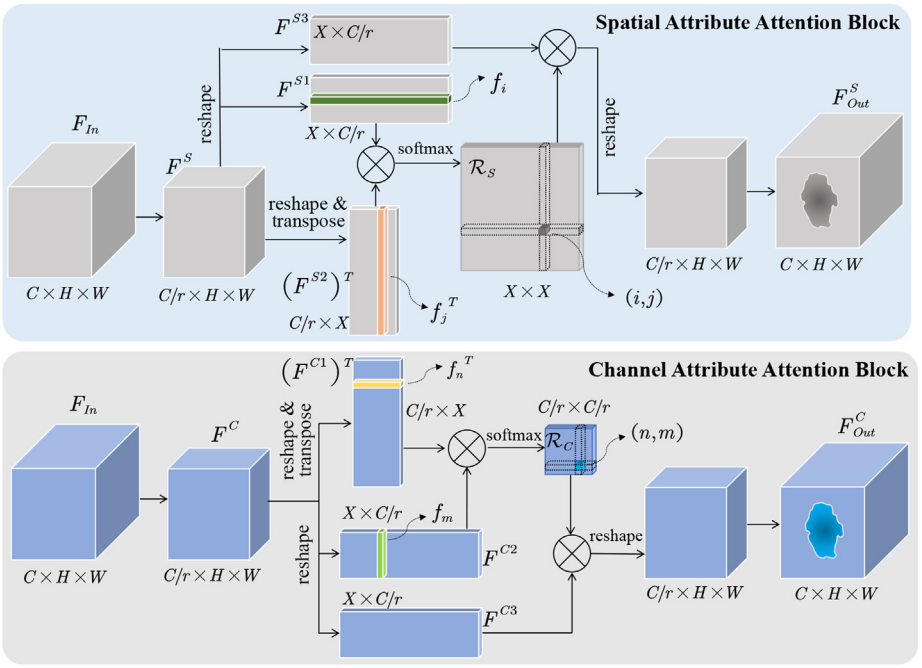


Fig. 4. The diagram of spatial attribute attention block and channel attribute attention block.

6. Loss function

In our proposed framework, the learning task includes ID recognition, attribute recognition, self-distilling smooth relaxation, and metric learning. For metric learning, we adopt the hard example mining triplet learning [10]. Specifically, in one batch, we randomly select $\mathcal{P}$ persons and pick $\mathcal{K}$ images of each person, i.e. total $\mathcal{PK}$ images. Our goal is to make the distance between features of the same ID smaller than the distance between features of different IDs. Given a training image x_i^l whose ID is l, its feature descriptor is $\boldsymbol{S}_i^l$, then descriptors $\boldsymbol{S}_f^l$ for all $f \neq i$ are regarded as positive examples $\left\{\boldsymbol{S}_i^{(pos)}\right\}$, and for all $\boldsymbol{S}_j^e$ that $e \neq l$ are negative examples denoted as $\left\{\boldsymbol{S}_i^{(neg)}\right\}$. For each $\boldsymbol{S}_i^l$ in this batch, we find its hardest positive and negative example. The hard example mining triplet loss is given by:

$$\mathcal{L}_{Tri} = \sum_i^{\mathcal{PK}} \left(\tau + \max_{k=1,\ldots,\mathcal{K}} Eu\left(\boldsymbol{S}_i, \boldsymbol{S}_i^{(pos)}\right) - \min_{k=1,\ldots,\mathcal{K}} Eu\left(\boldsymbol{S}_i, \boldsymbol{S}_i^{(neg)}\right) \right), \tag{11}$$

where $Eu(\bullet)$ means Euclidean Distance calculation, and τ is margin enforced between positive and negative examples. Simultaneously, $\boldsymbol{S}_i^l$ is imposed with an identity classification loss. Assume that the output of FC in ID classifiers is $\boldsymbol{v} = [v_1, \ldots, v_l]$. The predicted probability of each ID label n is calculated as:

$$p(n|x) = \frac{\exp(v_n)}{\sum_{\varphi=1}^{l} \exp(v_\varphi)}. \tag{12}$$

The cross entropy loss of ID classification is formulated as:

$$\mathcal{L}_{ID}(\boldsymbol{S}_i, l) = -\sum_{n=1}^{l} \log(p(n)) q(n). \tag{13}$$

Let y_l be the ground-truth ID label, so that $q(y_l) = 1$ and $q(n) = 0$ for all $n \neq y_l$. In this case, minimizing the cross entropy is equivalent to maximizing the possibility of being classified to the ground-truth category. Thus, the final loss is:

$$\mathcal{L} = \frac{1}{14}\sum_{\phi=1}^{14}\mathcal{L}_{Tri} + \frac{1}{14}\sum_{\phi=1}^{14}\mathcal{L}_{ID} + \frac{\alpha}{7}\sum_{\omega=1}^{7}\mathcal{L}_d + \frac{\beta}{2}\sum_{\mu=1}^{2}\mathcal{L}_{attr} \tag{14}$$

where parameters α and β balance the contribution of self-distilling regularization and attribute learning. Following the setting in [29], the local parts number is set to 6. It should be noted that as shown in Fig. 1, the ID learning and triplet learning are both conducted on each vector in the student branch and teacher branch. And as described in Section 5.2, the attribute learning is only implemented on global vectors of two branches. Thus, the denominator 14 denotes the number of all vectors in two branches in Fig. 1, the denominator 7 denotes the number of pairs made of one student vector and one teacher vector. And the denominator 2 means the number of vectors used to do the binary attribute learning.

7. Experiments

In Section 7.1, we introduce the datasets, protocol of evaluation, and the implementation details. In Section 7.2, we compare our method with the state-of-the-art. In Section 7.3, we evaluate our method on large-scale benchmark. And in Section 7.4, the effectiveness of self-distilling smooth relaxation is described. The ablation study of the attribute-based fast retrieval is discussed in Section 7.5. Finally, we study the generality of our framework in Section 7.6.

7.1. Experimental settings

7.1.1. Dataset

We use four authoritative benchmarks to evaluate our method, including Market-1501 [47], DukeMTMC-reID [49], CUHK03-NP [51], Market-1501+500k [47]. Market-1501 has 12,936 training images with 751 different identities. Gallery and query sets have 15,913 and 3,368 images, respectively, with another 750 identities. DukeMTMC-reID includes 16,522 training images of 702 identities, 2,228 query and 17,661 gallery images of another 702 identities.

CUHK03-NP [51] is a new training–testing split protocol for CUHK03 and consists of 1,467 pedestrians, it provides 14,096 manually labelled images and 14,097 DPM-detected images. In this protocol, 767 identities are used for training and the remaining for testing. The gallery sizes are 5,328 and 5,332 for labelled and detected images. Market-1501+500k provides an additional 500,000 distractors recorded at another time, these distractors are composed of background images and a large number of irrelevant pedestrians.

Annotations of the CUHK03 attributes: Lin et al. [18] labelled 27 and 23 pedestrian attributes for the Market-1501 and DukeMTMC-reID datasets, respectively. For a more comprehensive evaluation of our method, following the attribute labelling standard of [18], we manually annotate 27 attributes for each person in CUHK03. These attributes are gender (male, female), hair length (long, short), sleeve length (long, short), length of lower-body clothing (long, short), carrying backpack (yes, no), carrying handbag (yes, no), carrying other types of bag (yes, no),10 colours of upper-body clothing (black, white, red, purple, yellow, grey, blue, green, pink, brown) and 10 colours of lower-body clothing (black, white, red, purple, yellow, grey, blue, green, pink, brown). This annotation will be released later.

7.1.2. Evaluation protocols

For the evaluation of person re-ID accuracy, we use the cumulative matching characteristic (CMC) and the mean average precision (mAP). The evaluation packages are provided by [47,49]. For each query, the average precision (AP) is obtained from its precision-recall curve, and mAP is the mean value of average precisions across all queries. For Market-1501 in comparison with the state-of-the-art, both single query and multiple query settings are considered, all reported results in this paper are compared to the state-of-the-art without re-ranking.

7.1.3. Implementation details

The input images are resized to 384×128 after random flipping and erasing. Stochastic gradient descent is applied with a momentum of 0.9. We set the batch size to 64 and the model is trained for 350 epochs. The base model is pre-trained ResNet50 [9]. In the preceding 20 epochs, the learning rate gradually increases from 0.001 to 0.01 with a warm-up trick [9]. Besides, the learning rates of all pre-trained layers are the same as the base learning rate. The channel reduction ratio r in AAB is set to 8. The GPU we utilize is NVIDIA GTX 1080Ti, the CPU is an Intel Xeon CPU E5-2609 (1.70 GHz), and the RAM is 32 GB. We use PyTorch 1.0.0 to establish our whole framework, training our framework on Market-1501 consumes for nearly 145 min. In addition, When the performance results we show do not include AFR, as shown in Table 1,2,3,4, 6, 7, 10, our attribute codes do not participate in retrieval and all codes are ID codes. In this case, when the total bit-length is set 2048, 512, and 128 in comparison with the state-of-the-art, each local descriptor is set 256-bits, 64-bits, and 16-bits respectively, and the global descriptor is set 512-bits, 128-bits, and 32-bits respectively. We also construct a baseline for comparison. In this baseline, we ***remove*** all of our proposed methods (SSR, AAB and AFR) from our framework. Then, our whole framework can be denoted as "baseline(SSR)+AAB+AFR". The "baseline(Sigmoid)" denotes the baseline with a relaxation of Sigmoid, and the "baseline(SSR)" means the baseline with a relaxation of our SSR. When we use a Sigmoid (which is the most commonly used in the existing hashing researches) as the relaxation for comparison, batch normalization (BN) layer is removed to obtain its best performance according to the results in Section 7.4.

Based on the ablation studies on Market-1501 in Sections 7.4 and 7.5, in comparison with state-of-the-art in Sections 7.2 and 7.3, the super-parameters in our method are set $\alpha = 0.3$, $\beta = 0.3$. The best w in AFR varies on different datasets, which is stated in the discussion corresponding to each table. Additionally, to implement the variable-controlled approach, in each ablation study section corresponding to one hyper-parameter, other hyper-parameters set the values which lead to the best performance.

7.2. Comparisons with the state-of-the-art

We first compare our framework with 11 hashing state-of-the-art methods, including CNNH [36], NINH [15], DSRH [46], StructDH [35], DCH [32], DRSCH [43], PDH [52], HashNet [2], CSBT [3], ABC [22], and CPDH [16]. Among them, CNNH [36] decomposes the learning process into a stage of learning approximate hash codes; NINH [15] is a deep method, which adopts a triplet loss to preserve relative similarities; DSRH [46] aims to preserve multi-level semantic similarity between multi-label images; DRSCH [43] is a deep framework, which is optimized by triplet ranking loss; PDH [52] is also a framework in which batches of triplet samples are employed as the input of deep hashing architecture; StructDH [35] is a structured deep hashing network embedded with hash function learning; DCH [32], CSBT [3], ABC [22], and CPDH [16] are all deep CNN networks which introduce binary appropriation layers in the residual network. Since the recent CSBT [3], ABC [22], and CPDH [16] all adopt ResNet50 as their backbone. For fairness, we report the performance of our method using ResNet50 as the backbone on three datasets.

For a more intuitive comparison, in this part, we only show the performance enhancement of SSR and AAB, the speed up results of AFR are reported in Sections 7.3 and 7.5.2. The re-ID performance comparison with hashing state-of-the-art under different bit-lengths are shown in Table 1 (on Market-1501), Table 2 (on DukeMTMC-reID), and Table 3 (CUHK03-NP). We can see that on the one hand, the proposed method outperforms the previous studies by a large margin on three datasets, even compared with the latest works based on the same deep neural network ResNet50 as ours, such as CSBT, ABC, and CPDH. On the other hand, compared with the baseline(Sigmoid), our specially designed hashing relaxation method SSR outperforms by +(9.5%/14.9%) in rank-1/mAP on Market-1501 with the longest 2048 bits in single query setting. On DukeMTMC-reID, this enhancement is +(16.5%/25.2%). And on CUHK03-NP, it is +(11.2%/ 12.5%) on CUHK03-NP-Labelled and +(11.9%/11.5%) on CUHK03-NP-Detected. The accuracy enhancement of AAB seems weak. However, according to the results in Section 7.5.1, if we remove AAB, there would be a performance deterioration when introducing attribute learning task. AAB actually turns the effects of attribute learning from negative to positive.

As we introduce attribute annotations into our system, we present the comparisons with person re-ID methods which also leverage attribute information for fairness. Researches APR [18], DHANet [34], APDR [17], AFFNet [25], and AANet [30] are listed as competitors. Since we are the first to annotate the attribute for CUHK03, there are only comparisons on Market-1501 and DukeMTMC-reID. As shown in Table 4, our method outperforms the best competitor AANet by +(0.9%/3.5%) in rank-1/mAP on Market-1501 and +(2.0%/ 6.0%) in rank-1/mAP.

7.3. Robustness in the wild

To validate the effectiveness of our method under practical conditions, we also study our performance on a large dataset Market-1501+500K,which provides an additional 500, 000 distractors.

Experimental results on Market-1501+500K are shown in Table 5. The state-of-the-art results on Market-1501+500K are listed for comparison and they all employ ResNet50 as their backbone. Since the publicly available results of state-of-the-art on this dataset are all non-hashing methods and use floating-point fea-

Table 1
The comparison with the hashing state-of-the-art methods on the Market-1501 dataset under different bit-lengths, Rank-1 accuracy (%), mAP (%), and search time are shown.

Methods	Bits	Single-query		Multi-query	
		R1	mAP	R1	mAP
CNNH [36]		–		16.4	34.5
NINH [15]		–		37.7	37.8
DSRH [46]		–		34.3	42.3
StructDH [35]]		–		48.0	48.2
CPDH [16]	128	**83.1**	**67.2**	–	
baseline(Sigmoid)		78.3	63.1	79.9	66.0
baseline(SSR)		88.4	71.2	90.0	78.1
baseline(SSR) +AAB		**89.1**	**71.7**	**90.6**	**78.6**
PDH [52]		47.9	26.1	56.8	31.7
CSBT [3]		42.9	20.3	–	
CPDH [16]		**88.4**	**74.9**	–	
baseline(Sigmoid)	512	84.1	70.0	85.7	71.5
baseline(SSR)		93.1	83.1	93.9	86.8
baseline(SSR) +AAB		**93.6**	**83.6**	**94.2**	**87.6**
ABC [22]		81.4	64.7	–	
CPDH [16]		**89.5**	**77.1**	–	
baseline(Sigmoid)	2048	85.0	70.7	85.2	77.7
baseline(SSR)		94.5	85.6	96.2	89.7
baseline(SSR) +AAB		**94.8**	**86.0**	**96.6**	**90.2**

Table 2
The comparison with the hashing state-of-the-art methods on the DukeMTMC-reID dataset under different bit-lengths, Rank-1 accuracy (%), mAP (%) and search time are shown.

Methods	Bits	DukeMTMC-reID	
		R1	mAP
ABC [22]		60.3	–
CPDH [16]		**75.5**	**56.9**
baseline(Sigmoid)	128	71.1	50.2
baseline(SSR)		79.0	61.1
baseline(SSR)+AAB		**79.5**	**61.7**
DRSCH [43]		19.3	13.6
CSBT [3]		47.2	33.1
HashNet [2]		40.8	28.6
DCH [32]		57.4	37.3
ABC [22]	512	65.5	–
CPDH [16]		**80.6**	**65.3**
baseline(Sigmoid)		76.6	61.0
baseline(SSR)		87.0	74.6
baseline(SSR)+AAB		**87.5**	**75.0**
ABC [22]		82.5	61.2
CPDH [16]		**81.6**	**66.4**
baseline(Sigmoid)	2048	71.4	53.0
baseline(SSR)		87.9	78.0
baseline(SSR)+AAB		**88.4**	**78.6**

tures, in order to ensure the fairness and comprehensiveness of the comparison, we have implemented the hashing version of a non-hashing method MGN [31]. In this hashing-MGN implementation, we use Sigmoid or Hardtanh as relaxation to obtain binary features.

In addition to retrieval accuracy, the comparisons of computations and storage efficiency are provided. Note that in the application of person re-ID, a target pedestrian image is often used to compare with the stored gallery to find his or her other images. Thus, following this setting, for each dataset, we first evaluate the time for searching one query's similar images in the whole gallery, then report the memory for storing all the gallery features of the dataset. Compared with the real-valued floating feature which requires 32-bit for storing the value per dimension, binary hashing code requires only 1-bit for each dimension. For fairness, we reproduce the searching process of these state-of-the-art according to the feature dimension their paper reported.

In Table 5, we can see that the baseline(SSR) still achieves remarkable improvement over the baseline(Sigmoid)(from 81.9%/66.5% to 89.8%/76.6%in rank-1/mAP). Then, AAB gains +0.6%/0.5% in rank-1/mAP. Besides, our re-ID accuracy surpasses the state-of-the-art. Thereinto, although the search time of baseline(SSR)+AAB (521.1 ms) is longer than the low-bit competitor TriNet (252.7 ms), under AFR (the w is set 1e+05 here), our search time is reduced to 107.2 ms, with an advantage of +14.5% in rank-1 and +24.7% in mAP over TriNet. And compared with hashing-MGN, baseline(SSR)+AAB+AFR has 30-bits more attribute codes, but AFR can save 79.5% of retrieval time while its performance is +0.1%/2.9% higher than the strong competitor MGN (Hardtanh).

7.4. Ablation study on self-distilling smooth relaxation

To find how hyper-parameter α in Eq. (14) and the learning method influence the final results of SSR, we first set experiments with different α. Then, two types of distilling method end-to-end and pretrained are implemented for comparison. In the end-to-end mode, the model is trained in the way shown in Fig. 1, the teacher model dynamically evolves as training proceeds, and the informative sources distilled from teacher model also real-time updates. While in the pretrained mode, the distillation process includes two stages, the teacher model is trained in the first stage. In the second stage, the student model distills knowledge from the static and pretrained teacher model.

Besides, batch normalization (BN) is an approach which could mitigate the problem of internal covariate shift, where parameter initialization and changes in the distribution of the inputs of each layer affect the learning rate of the neural network [12]. In [24], BN was proved to be particularly effective on promoting person re-ID accuracy. In our experiments, we find that the features relaxed by the SSR could get more benefit from BN than the features relaxed by the traditional activation function (which can be seen in Table 6). Thus, although the effect of BN is not our contribution, we also present the results with/without BN.

The experimental results on Market-1501 are shown in Fig. 5. We see that the upper performance limits of two types of distillation method are the same, which indicates the informative representation learning relies more on optimization target loss

Table 3
The comparison of our method with the hashing state-of-the-art methods on the CUHK03-NP-Labelled, UHK03-NP-Detected dataset under different bit-lengths, Rank-1 accuracy (%), mAP (%) and search time are shown.

Methods	Bits	Labelled		Detected	
		R1	mAP	R1	mAP
CPDH [16]		–		**56.6**	**52.0**
baseline(Sigmoid)		55.6	51.2	52.2	48.3
baseline(SSR)	128	60.7	54.4	57.0	56.5
baseline(SSR) +AAB		**61.2**	**54.9**	**57.4**	**56.9**
DRSCH [43]		25.4	–	–	
CSBT [3]		55.5	–	–	
DCH [32]		44.4	41.3	–	
CPDH [16]	512	–		**63.2**	**58.7**
baseline(Sigmoid)		59.4	56.3	56.2	53.8
baseline(SSR)		72.7	69.3	69.3	65.0
baseline(SSR) +AAB		**73.1**	**69.8**	**69.6**	**65.4**
CPDH [16]		–		**66.4**	**61.9**
baseline(Sigmoid)		63.2	58.6	60.1	55.0
baseline(SSR)	2048	74.3	71.0	71.9	66.2
baseline(SSR) +AAB		**74.8**	**71.6**	**72.5**	**66.9**

Table 4
The comparison of our method with the state-of-the-art methods using attribute on the Market-1501 and DukeMTMC-reID dataset, Rank-1 accuracy (%) and mAP (%) are shown

Methods	Market-1501		DukeMTMC-reID	
	R1	mAP	R1	mAP
APR [18]	87.0	66.9	73.9	55.6
DHANet [34]	91.3	76.0	81.3	64.1
APDR [17]	93.1	80.1	84.3	69.7
AFFNet [25]	93.9	81.7	84.6	70.7
AANet [30]	93.9	82.5	86.4	72.6
baseline(SSR)+AAB	**94.8**	**86.0**	**88.4**	**78.6**

function than on distilling way. However, two-stage distillation needs a longer training time: in our experiments on Market-1501, the training time of using pretrained distillation is 2.2 times longer than end-to-end. So, we recommend using the end-to-end SSR. For the influence of α, take the end-to-end curves as an example, we observe that under the setting without BN, when α changes from 0 to 0.7, the rank-1 accuracy and mAP gradually increase from 92.6%/79.9% to 93.9%/83.5%. It indicates the importance of self-distillation in the re-ID task. When α increases to 1, the rank-1 accuracy and mAP of the model decrease to 93.5%/81.4%, which indicates that an over-increased α could break the balance of multi-task learning and compromise the feature representation. Similarly, under the setting with BN, rank-1 and mAP curves also rise first, then fall, and the best re-ID performance is obtained when $\alpha = 0.3$. Therefore, we use $\alpha = 0.3$ when BN is adopted, and $\alpha = 0.7$ when BN layer is removed.

We then follow the best performance setting of α on Market-1501, and validate the effectiveness of SSR on other three datasets. In this set of experiments, except for using the traditional activation function as relaxation, we add the results of using DSH [20]. DSH is a competitive relaxation method designed for hashing image retrieval, which proposes a regularizer to fit the shape of the feature distribution to the distribution of hash codes. We implement this relaxation and applied it in our person re-ID system for comparison. In addition, we also report the performance of teacher descriptor (SSR) and student descriptor (SSR) in SSR. We use Sigmoid, HardSigmoid, Tanh, Hardtanh activation functions, and DSH regularizer as relaxation, and compare with our method. The performance results are shown in Table 6. According to the results, we conclude as follows:

First, whether it is with or without BN, DSH [20] performs better as relaxation than traditional activation functions, which is in line with the results reported in [20]. And our student descriptor can always achieve higher performance than DSH. This is because, compared with the activation function, although DSH can make the distribution of features more effectively fit the true distribution of binary codes, the design concept of DSH and activation function is similar, and both impose restrictions on the distribution of features. While SSR introduces less hard regularization (such as restricting the region that features distribute). Instead, it learns informative semantic knowledge from self-distillation, thus achieves stronger feature representation.

Second, in the scene without BN, DSH surpasses our teacher descriptor with a slight advantage, but in the case of BN, DSH failed to do so. This shows that in person re-ID, the hard relaxations (such

Table 5
The comparison of our method with the state-of-the-art methods on the Market-1501+500k dataset, Rank-1 accuracy (%) and mAP (%) are shown

Methods	Feature's Dim./Type	R1	mAP	Storage Memory	Search Time
2Stream [50]	4096/float	68.3	45.3	7812.5 MB	8084.8 ms
APR [18]	2048/float	75.4	49.8	3906.3 MB	4042.4 ms
TriNet [10]	128/float	74.7	53.6	244.1 MB	252.7 ms
MGN(Sigmoid)	2048/binary	82.3	66.6	122.1 MB	521.1 ms
MGN(Hardtanh)	2048/binary	**89.0**	**75.4**	122.1 MB	521.1 ms
baseline(Sigmoid)	2048/binary	81.9	66.5	122.1 MB	521.1 ms
baseline(SSR)	2048/binary	89.8	76.6	122.1 MB	521.1 ms
baseline(SSR)+AAB	2048/binary	90.4	77.1	122.1 MB	521.1 ms
baseline(SSR)+AAB+AFR	2078/binary	**89.1**	**78.3**	123.7 MB	107.2 ms

Table 6
Performance of using relaxation of SSR, Sigmoid, Hardsigmoid,Tanh, Hardtanh, and DSH, with/without bn on the Market-1501, DukeMTMC-reID, CUHK03-NP, Rank-1 accuracy (%), and mAP (%) are shown

Relaxation	with BN	Market-1501		DukeMTMC -reID		CUHK03-NP-Labelled		CUHK03-NP-Detected	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
Sigmoid	×	85.6	71.2	72.2	53.3	66.6	60.7	60.6	55.4
HardSigmoid	×	86.4	72.8	72.3	53.5	66.8	60.8	60.8	55.5
Tanh	×	93.0	82.4	85.0	72.9	72.0	66.8	69.2	64.3
Hardtanh	×	93.2	82.5	85.2	73.0	72.2	66.9	69.3	64.4
DSH	×	93.5	83.1	85.6	74.1	72.6	67.8	69.5	65.2
SSR	×	93.3	82.9	85.3	73.9	72.4	67.6	69.3	65.0
SSR	×	**93.9**	**83.5**	**86.0**	**74.4**	**72.9**	**68.2**	**69.9**	**65.5**
Sigmoid	✓	78.4	60.5	66.5	48.9	53.7	49.1	51.5	45.8
HardSigmoid	✓	78.5	60.8	66.7	49.0	53.8	49.4	51.7	45.9
Tanh	✓	93.8	83.6	86.5	74.1	73.5	68.5	71.2	63.4
Hardtanh	✓	93.9	83.8	86.7	74.5	73.7	68.7	71.3	63.6
DSH	✓	94.0	85.3	87.6	78.0	74.1	70.9	71.7	66.1
SSR	✓	94.2	85.4	87.8	78.1	74.2	71.1	71.9	66.3
SSR	✓	**94.8**	**86.0**	**88.4**	**78.6**	**74.8**	**71.6**	**72.5**	**66.9**

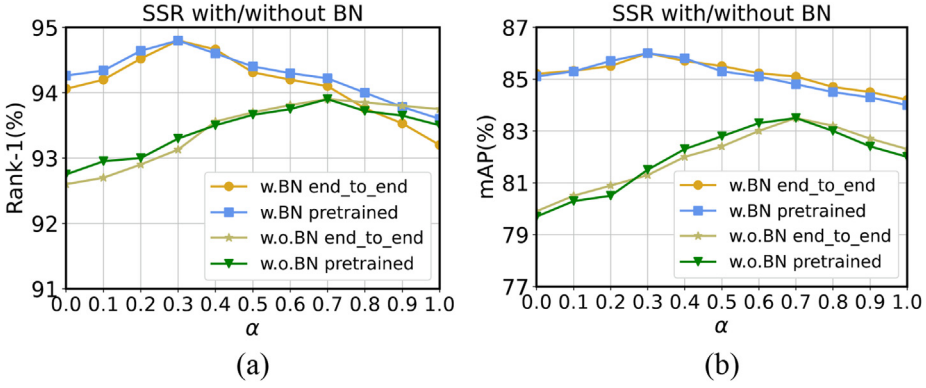


Fig. 5. Rank-1 and mAP curves of the end-to-end and pretrained distilling method as function of *alpha* on Market-1501.

as activation functions or DSH) that greatly change the feature distribution can hardly coexist with BN. This may be because, according to the description in [12], BN will process the features into a form close to the Gaussian distribution, which will be destroyed by these hard relaxations, thus affects the robustness of feature representation. Therefore, we recommend using our SSR as the relaxation in the application of hashing person re-ID, especially when using BN at the same time.

7.5. Ablation study on attribute-based fast retrieval

In this section, we give a detailed study over different settings of the proposed AFR. First, we study the effectiveness of attribute-guided attention block, and then we analyze the experimental results of fast retrieval.

7.5.1. Effectiveness of attribute-guided attention block

In the study of AAB, we first discuss the determination of the hyper-parameter β in Eq. (14), which controls the strength of attribute learning. We observe the influence of β from two aspects: the attribute prediction precision and re-ID accuracy. We follow the attribute accuracy evaluation method of [18], and use the mean accuracy of 30 attributes in our system on Market-1501. The experimental results are shown in Fig. 6.

We can see that, in Fig. 6(a), the prediction accuracies of the attribute learning without AAB and attribute learning with AAB both increase as the values of β becomes larger when $\beta < 0.3$ (when β =0, there is no attribute learning thus the accuracy is also 0), within this range, more weights the attribute learning loss has, more benefits the attribute prediction will gain. The best performances of two models are both achieved at β = 0.3 and equal to 93.8%. This implies that the AAB doesn't promote the attribute learning ability. In Fig. 6(b), we find that in the setting of person re-ID without AAB, the re-ID accuracy drops along with the increasing of β, which means simply adding attribute loss to re-ID task can compromise the feature representation. While in the setting person re-ID with AAB, by leveraging the spatial properties and high-level semantic cues contained by the attribute, re-ID accuracy gradually increases from 94.5% to 94.8% when β changes from 0 to 0.3. When $\beta > 0.3$, performance drops with a larger β. This indicates that in an appropriate weight of attribute loss, with AAB adopted, the attribute learning can refine the ID features discrimination.

We then set $\beta = 0.3$ and validate the effectiveness of AAB in four datasets Market-1501, DukeMTMC-reID, CUHK03-NP-Labelled, and CUHK03-NP-Detected. The experimental results are shown in Table 7, we set four groups of experiments to observe the influence of attribute learning and the AAB. First, in the comparison of mode a with b, we can see that without attribute information, AAB can help to enhance the feature representation only +0.1%/0.1% in rank-1/mAP on average. Second, in the comparison of mode a with c, a clear re-ID accuracy degradation can be observed when introducing attribute learning without AAB adopted, which proves that the heteroscedasticity [33] of attribute and ID learning can cause re-ID deterioration. In the comparison of mode c with d, we see that the AAB brings significant re-ID promotion, on Market-1501,

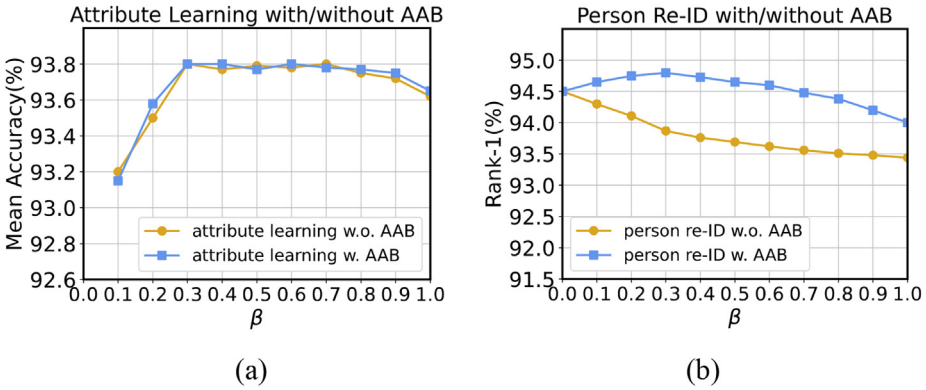


Fig. 6. The attribute prediction accuracy and person re-ID accuracy curves of the proposed method as function of β on Market-1501.

the enhancement is +(0.7%/0.9%) in rank-1/mAP, on DukeMTMC-reID, CUHK03-NP-Labelled, and CUHK03-NP-Detected, the enhancement are +(0.8%/1.0%), +(0.9%/1.1%), and +(1.0%/ 1.2%). It should be noted that these performance gains include two parts, the promotion from ID information-based attention and attribute information-based attention. But as shown in the comparison of modes a and b, the performance promotion from ID information-based attention is small, which verifies the effectiveness of AAB when attribute learning is introduced into person re-ID.

7.5.2. Effectiveness of fast retrieval

In this section, we follow the hyper-parameter setting of attribute learning determined in last part, and report the experimental results of AFR. The line charts of numerical results on Market-1501, DukeMTMC-reID, CUHK03-NP-Labelled, and CUHK03-NP-Detected are shown in Fig. 7. The gallery sizes of four datasets are 15,912, 17,661, 5,328, and 5,332. Apparently, when w is set the same as the gallery size, the accuracy will be equal to not employing AFR (i.e., baseline(SSR)+AAB). For the convenience of comparison, we show the performance curves from w = 50 to w = 15,000 on Market-1501 and DukeMTMC-reID, and from w = 50 to w = 5000 for CUHK03-NP-Detected and CUHK03-NP-Labelled.

Two points can be summarized from Fig. 7. First, on all datasets, rank-1 accuracies always increase with a larger w, until w is close to the number of gallery pictures, the rank-1 performance approaches the baseline(SSR) +AAB settings (which are reported in Tables 1–3). Second, the mAP curves first increase with a larger w and then gradually decline. We think this mAP advantage AFR gains over the baseline(SSR) +AAB could come from the attribute description, which helps roughly filter examples in the first stage of faster-retrieval. Specifically in our experiments, on Market-1501, DukeMTMC-reID, CUHK03-NP-Labelled and CUHK03-NP-Detected, the best mAPs are respectively achieved under the settings of w = 1,111, w = 2,297, w = 1,500, w = 1,500, where performances are 93.7%/87.2%, 87.0%/80.6%, 71.6%/68.2%, and 74.2%/72.8% in rank-1/mAP, respectively.

In order to provide the person re-ID community with more thoughtful recommendation for the use of AFR, we also study the recommended value of w. For each dataset, although Rank-1 will increase as w increases, mAP reaches its peak before w increases to gallery size. Therefore, based on the consideration of finding the optimal mAP for w, we think that it may be possible to find some rules about w through the re-ID accuracy of using only attribute codes to search. We did the following experiment: For each dataset, we only use attribute codes to do re-ID and observe the performance of rank-w. As shown in Fig. 7, the dashed lines show the performance obtained by using only attribute codes to search on the different datasets. In the end, we got rank-1111 on Market-1501, rank-2297 DukeMTMC-reID, rank-1500 on CUHK03-NP-Labelled, and rank-1500 on CUHK03-NP-Detected. They were 97.1%, 96.9%, 96.8% and 96.8%, respectively. This shows that although the optimal value of w for mAP is different on different datasets, we can still use only the attribute codes to retrieve and observe the performance, so as to quickly find the recommended value of w on each dataset. Based on the research on four datasets, we recommend that when using AFR on a dataset, first use only attribute codes to search, and then find a rank-w approximately equal to 97%. This w is likely to achieve the optimal mAP.

The rank-1 performance would surely increase along with the w increasing, but it is worth noting out that a smaller w can lead to lower computational complexity, as well as a shorter search time. Therefore, to validate that the APR genuinely ***make sense***, we propose an evaluation method:

Take Market-1501 dataset as an example, the gallery has 19,732 pictures, the best performance of our framework needs a bit-length of 2,048. To decrease the search time, there are two methods: reducing the bits number of features, for example, setting a lower bit-length such as 128, 256 or 512; another is adopting the AFR, by setting a different w, the approach could save time in various degrees. Both these two methods can lead to a decline in re-ID performance. We observe when they take equal time consuming, whether the re-ID accuracy employs the AFR outperforms the strategy of reducing bit-length.

The numerical results are reported in Table 8 for different conditions with bit limitation on each dataset. We find the w that makes the AFR match the search time of the method lower dimension (i.e., reducing the bit-length to save search time) under different bit limitations. As is shown, in all cases of low-bit setting, our strategy outperforms the method lower bit-length by a considerable margin. Specifically, for example, when we match the search time of AFR to the 128-bit condition, the mAP advantages of AFR over reducing bit-length gained on Market-1501, DukeMTMC-reID, CUHK03-NP-Labelled and CUHK03-NP-Detected are +14.4%, +15.1%, +8.3%, +5.7%, respectively. Although faster-retrieval cannot save storage memory, experimental results show that this strategy can preserve an excellent re-ID performance while greatly decreasing time-consumption. The value of w on the left side of Table 8 is as follows: Take Market-1501 as an example, which has 15912 images in its gallery set. Suppose that the time required to calculate the Hamming distance of a pair of point-to-point (that is, an XOR operation) is T. When we want to make the calculation time of the Hamming distance of AFR (w=W) equal to the 128-bits of the non-AFR model, we can formulate: 128*15912*T=30*15912*T +2048*W*T. Then we get W≈761, which is the ideal value of w. And since AFR has an extra sorting compared to no-AFR, in the real measurement, the value of w should be smaller than 761. Thus, we

Table 7
Performance of our method with/without AAB, and with/without attribute learning on the Market-1501, DukeMTMC-reID, CUHK03-np, Rank-1 accuracy (%), and mAP (%) are shown

Mode	with attribute learning	with AAB	Market-1501		DukeMTMC -reID		CUHK03- NP-labelled		CUHK03- NP-detected	
			R1	mAP	R1	mAP	R1	mAP	R1	mAP
a	×	×	94.5	85.6	87.9	78.0	74.3	71.0	71.9	66.2
b	×	✓	94.6	85.7	88.0	78.1	74.4	71.2	72.0	66.3
c	✓	×	94.1	85.1	87.6	77.6	73.9	70.5	71.5	65.7
d	✓	✓	94.8	86.0	88.4	78.6	74.8	71.6	72.5	66.9

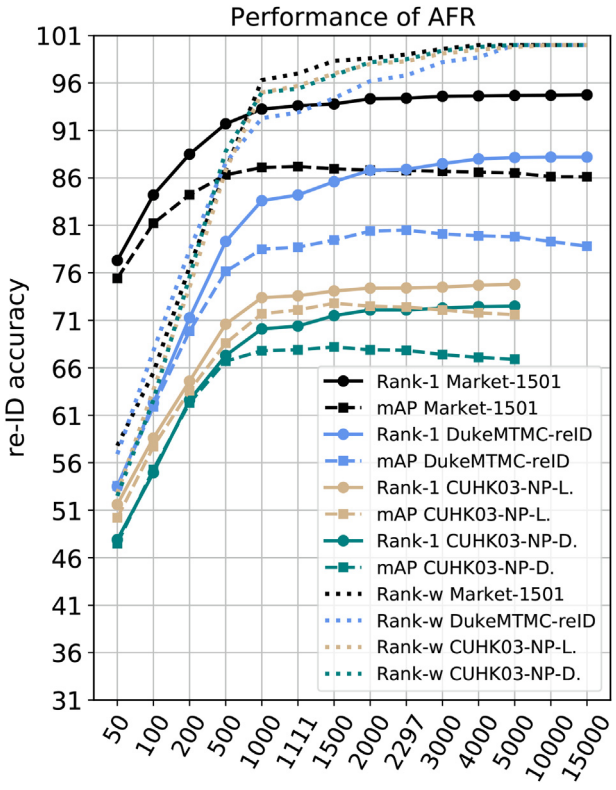


Fig. 7. Performances of AFR strategy under different w (returning different numbers of candidates of first-step retrieval for second-step searching) on the Market-1501, DukeMTMC-reID, CUHK03-NP-Detected (CUHK03-NP-D.) and CUHK03-NP-Labelled (CUHK03-NP-L.) datasets.

will use 761 as the reference value and gradually reduce the value of w to finally obtain the final value of w (=553) taken in the AFR, to fit the search time same to the different number of bits without AFR.

7.6. Study of the generality

In this section, we study the generality of our method, including two subsections, the generality of AFR and the generality of using stronger baseline.

7.6.1. Generality of AFR

In addition to being applied to hashing person re-ID, AFR can also speed up the general (non-hashing) person re-ID method. We use two non-hashing state-of-the-art person re-ID methods MGN [31] and ABD-Net [5] to verify the effectiveness of AFR. Unlike the MGN in Section 7.3, the non-hashing methods here do not use the nonlinear function to generate binary codes, but use the original floating features of the algorithm to perform the second-stage retrieval of AFR. The results on the large dataset Market-1501+500k can best demonstrate the benefits of AFR. As shown in Table 9, based on the floating features of MGN [31] and ABD-Net [5], we add 30-bit binary attribute codes to implement AFR. Using the w value method introduced in Section 7.5.2, we set w as 1e+05 and let the retrieval rank-w of the attribute codes in the two methods equal 96.8% and 97.0%. We can see that AFR significantly reduces the search time at the cost of slight reduction in Rank-1, and obtains 1.3% and 1.2% increases in mAP on MGN [31] and ABD-Net [5], respectively. This illustrates the generality of AFR in non-hashing methods.

7.6.2. Generality of using stronger baseline

In our framework, in order to reduce the complexity of the network structure and the amount of calculation as much as possible, we use a simple PCB structure as the base model. In this part, we verify the generality of SSR and AAB by observing whether they can still bring performance gains when using a stronger baseline. In the literature [38], the authors designed a powerful baseline AGW [38] by analyzing the advantages of the existing person re-ID methods, and achieved state-of-the-art both in single- and multi-modality re-ID tasks. The main innovation of AGW [38] lies in the fusion of non-local attention, generalized-mean pooling,

Table 8
Performance comparisons of AFR and reducing bit-length under the same real search time on the Market-1501, DukeMTMC-reID, CUHK03-NP, Rank-1(%) and mAP(%) are shown.

Datasets	with the AFR			Search	with lower bit-length		
	Strategy	R-1	mAP	Time	Bit-length	R-1	mAP
Market-1501	w = 553	91.8	86.1	1.04 ms	128-bits	89.2	71.7
	w = 1067	93.3	86.9	2.07 ms	256-bits	92.5	79.3
	w = 3105	94.2	86.5	4.15 ms	512-bits	93.6	83.6
Duke MTMC-reID	w = 654	80.1	76.8	1.15 ms	128-bits	79.5	61.7
	w = 1352	84.6	78.6	2.30 ms	256-bits	83.9	69.1
	w = 3352	87.8	80.1	4.60 ms	512-bits	87.5	75.0
CUHK03-NP-Labelled	w = 265	65.7	63.2	0.35 ms	128-bits	61.2	54.9
	w = 510	70.9	67.8	0.69 ms	256-bits	68.8	62.0
	w = 951	73.5	71.7	1.39 ms	512-bits	73.1	69.8
CUHK03-NP-Detected	w = 265	63.1	62.6	0.35 ms	128-bits	57.4	56.9
	w = 510	67.5	66.7	0.69 ms	256-bits	67.2	59.4
	w = 951	70.1	67.7	1.39 ms	512-bits	69.6	65.4

Table 9
The re-ID performance of MGN, MGN+AFR, ABD-Net, and ABD-Net +AFR on the Market-1501+500k dataset, Rank-1 accuracy (%) and mAP (%) are shown

Methods	Feature's Dim./Type	R1	mAP	Storage Memory	Search Time
MGN [31]	2048/float	91.1	78.2	3906.3 MB	4042.4 ms
MGN [31]+AFR	2048/float+30/binary	**90.5**	**79.5**	3907.9 MB	811.6 ms
ABD-Net [5]	2048/float	91.0	79.7	3906.3 MB	4042.4 ms
ABD-Net [5]+AFR	2048/float+30/binary	**90.5**	**80.9**	3907.9 MB	811.6 ms

Table 10
Performance of using the AGW as baseline on the Market-1501, DukeMTMC-reID, CUHK03-NP, Rank-1 accuracy (%), and mAP (%) are shown

Method	Market-1501		DukeMTMC -reID		CUHK03-NP-Labelled		CUHK03-NP-Detected	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
baseline(Sigmoid)	85.0	70.7	71.4	53.0	63.2	58.6	60.1	55.0
AGW(Sigmoid)	85.3	71.7	72.1	53.5	65.4	60.7	62.9	57.8
baseline(SSR)	94.5	85.6	87.9	78.0	74.3	71.0	71.9	66.2
AGW(SSR)	94.6	86.5	88.4	78.3	75.9	73.0	74.4	68.3
baseline(SSR)+AAB	94.8	86.0	88.4	78.6	74.8	71.6	72.5	66.9
AGW(SSR)+AAB	94.9	86.9	88.8	78.8	76.3	73.5	74.9	68.9

and weighted regularization triplet loss. Table 10 shows the performance of SSR and AAB when AGW is the baseline.

We can see that on the one hand, when our baseline is replaced with AGW, the performances on the three datasets have been improved. Among them, on the CUHK03-NP-Labelled and CUHK03-NP-Detected, in the SSR experiment, the performance gains obtained by AGW are +1.6%/2.0% and +2.5%/2.1% respectively in Rank-1/mAP. In the SSR+AAB experiment, the gains are +1.5%/1.9% and +2.4%/2.0%. On the other hand, we can also see that when AGW is used as the baseline, the performance gains obtained by SSR and AAB are similar to when using our baseline. This confirms that the proposed SSR and AAB have good generality when using a different baseline.

8. Conclusion

In this paper, we focus on both the accuracy and efficiency of deep hashing person re-ID. First, we propose an attribute-based fast retrieval (AFR) for hashing re-ID, AFR leverages the attribute prediction of the model which is trained in a binary classification manner tailor-made for hashing. The attribute information is also used to refine the global feature representation by the attribute-guided attention block (AAB). Then, to fully exploit deep representation to generate the hash codes, we propose a binary code learning method self-distilling smooth relaxation (SSR). Four authoritative public benchmarks (Market-1501, Market-1501 +500K, CUHK03, and Duke-MTMC-reID). The experimental results indicate that with the SSR and AAB, we surpass all the state-of-the-art hashing methods, and compared with reducing the number of feature bits, the AFR strategy is more effective to save search time.

CRediT authorship contribution statement

Hanyang Jin: Writing - original draft, Writing - review & editing, Methodology, Software. ***Shenqi Lai:*** Writing - original draft, Methodology, Software. ***Guoshuai Zhao:*** Writing - review & editing, Formal analysis. ***Xueming Qian:*** Validation, Supervision, Writing - original draft, Writing - review & editing.

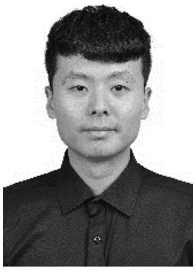
Declaration of Competing Interest

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.

References

[1] S. Bak, P. Carr, One-shot metric learning for person re-identification, CVPR (2017) 1571–1580.
[2] Z. Cao, M. Long, J. Wang, P.S. Yu, Hashnet: Deep learning to hash by continuation, ICCV (2017) 5609–5618.
[3] J. Chen, Y. Wang, J. Qin, L. Liu, L. Shao, Fast person re-identification via cross-camera semantic binary transformation, CVPR (2017) 5330–5339.
[4] K. Chen, Y. Chen, C. Han, N. Sang, C. Gao, Hard sample mining makes person re-identification more efficient and accurate, Neurocomputing 382 (2019) 259–267.
[5] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, Z. Wang, Abd-net: Attentive but diverse person re-identification, ICCV (2019) 8350–8360.
[6] Y. Chen, S. Duffner, A. Stoian, J. Dufour, A. Baskurt, Deep and low-level feature based attribute learning for person re-identification, Image Vis. Comput. 79 (2018) 25–34.
[7] Y. Chen, X. Lu, Deep discrete hashing with pairwise correlation learning, Neurocomputing 385 (2020) 111–121.
[8] H. Feng, N. Wang, J. Tang, J. Chen, F. Chen, Multi-granularity feature learning network for deep hashing, Neurocomputing 423 (2021) 274–283.
[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CVPR (2016) 770–778.
[10] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, CoRR (2017), URL: http://arxiv.org/abs/1703.07737.
[11] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-and-aggregation network for person re-identification, CVPR (2019) 9317–9326.
[12] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, ICML (2015) 448–456.
[13] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, CVPR (2018) 1062–1071.
[14] S. Khamis, C. Kuo, V.K. Singh, V.D. Shet, L.S. Davis, Joint learning for attribute-consistent person re-identification, in: ECCV, Springer, 2014, pp. 134–146.
[15] H. Lai, Y. Pan, Y. Liu, S. Yan, Simultaneous feature learning and hash coding with deep neural networks, CVPR (2015) 3270–3278.
[16] D. Li, Y. Gong, D. Cheng, W. Shi, X. Tao, X. Chang, Consistency-preserving deep hashing for fast person re-identification, Pattern Recognit. 94 (2019) 207–217.
[17] S. Li, H. Yu, R. Hu, Attributes-aided part detection and refinement for person re-identification, Pattern Recognit. 97 (2020).
[18] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognit. 95 (2019) 151–161.
[19] H. Ling, Z. Wang, P. Li, Y. Shi, J. Chen, F. Zou, Improving person re-identification by multi-task learning, Neurocomputing 347 (2019) 109–118.
[20] H. Liu, R. Wang, S. Shan, X. Chen, Deep supervised hashing for fast image retrieval, Int. J. Comput. Vis. 127 (2019) 1217–1234.

[21] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–126.
[22] Z. Liu, J. Qin, A. Li, Y. Wang, L.V. Gool, Adversarial binary coding for efficient person re-identification, ICME (2019) 700–705.
[23] Z. Liu, D. Wang, H. Lu, Stepwise metric promotion for unsupervised video person re-identification, ICCV (2017) 2448–2457.
[24] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multim. 22 (2020) 2597–2609.
[25] J. Luo, Y. Liu, C. Gao, N. Sang, Learning what and where from attributes to improve person re-identification, ICIP (2019) 165–169.
[26] T. Matsukawa, E. Suzuki, Person re-identification using CNN features learned from combination of attributes, ICPR (2016) 2428–2433.
[27] X. Shi, X. Qian, Exploring spatial and channel contribution for object based image retrieval, Knowl. Based Syst. (2019) 186.
[28] C. Su, F. Yang, S. Zhang, Q. Tian, L.S. Davis, W. Gao, Multi-task learning with low rank attribute embedding for person re-identification, ICCV (2015) 3739–3747.
[29] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline), ECCV (2018) 501–518.
[30] C. Tay, S. Roy, K. Yap, Aanet: Attribute attention network for person re-identifications, CVPR (2019) 7134–7143.
[31] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, ACM Multimedia (2018) 274–282.
[32] J. Wang, Y. Li, X. Zhang, Z. Miao, G. Tao, Deep classification hashing for person re-identification, ICGIP (2018) 226–229.
[33] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, CVPR (2018) 2275–2284.
[34] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, S. Satoh, Learning sparse and identity-preserved hidden attributes for person re-identification, IEEE Trans. Image Process. 29 (2020) 2013–2025.
[35] L. Wu, Y. Wang, Z. Ge, Q. Hu, X. Li, Structured deep hashing with convolutional neural networks for fast person re-identification, Comput. Vis. Image Underst. 167 (2018) 63–73.
[36] R. Xia, Y. Pan, H. Lai, C. Liu, S. Yan, Supervised hashing for image retrieval via image representation learning, AAAI (2014) 2156–2162.
[37] M. Ye, J. Shen, D.J. Crandall, L. Shao, J. Luo, Dynamic dual-attentive aggregation learning for visible-infrared person re-identification, ECCV (2020) 229–247.
[38] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C., 2021. Deep learning for person re-identification: A survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell, 1–1 doi: 10.1109/TPAMI.2021.3054775..
[39] M. Ye, J. Shen, X. Zhang, P.C. Yuen, S.F. Chang, Augmentation invariant and instance spreading feature for softmax embedding, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1.
[40] M. Ye, P.C. Yuen, Purifynet: A robust person re-identification model with noisy labels, IEEE Trans. Inf. Forensics Secur. 15 (2020) 2655–2666.
[41] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, ICPR (2014) 34–39.
[42] H. Yu, A. Wu, W. Zheng, Cross-view asymmetric metric learning for unsupervised person re-identification, ICCV (2017) 994–1002.
[43] R. Zhang, L. Lin, R. Zhang, W. Zuo, L. Zhang, Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification, IEEE Trans. Image Process. 24 (2015) 4766–4779.
[44] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Densely semantically aligned person re-identification, CVPR 2019 (2019) 667–676.
[45] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: CVPR, IEEE, 2020, pp. 3183–3192.
[46] F. Zhao, Y. Huang, L. Wang, T. Tan, Deep semantic ranking based hashing for multi-label image retrieval, CVPR (2015) 1556–1564.
[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, ICCV (2015) 1116–1124.
[48] M. Zheng, S. Karanam, Z. Wu, R.J. Radke, Re-identification with consistent attentive siamese networks, CVPR (2019) 5735–5744.
[49] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, ICCV (2017) 3774–3782.
[50] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned CNN embedding for person reidentification, ACM Trans. Multim. Comput. Commun. Appl. 14 (2018) 13:1–13:20.
[51] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, CVPR (2017) 3652–3661.
[52] F. Zhu, X. Kong, L. Zheng, H. Fu, Q. Tian, Part-based deep hashing for large-scale person re-identification, IEEE Trans. Image Process. 26 (2017) 4806–4817.
[53] L. Wang, X. Qian, Y. Zhang, J. Shen, X. Cao, Enhancing Sketch-Based Image Retrieval by CNN Semantic Re-ranking., IEEE Trans. Cybern. 50 (2020) 3330–3342.
[54] L. Wang, X. Qian, X. Zhang, X. Hou, Sketch-Based Image Retrieval With Multi-Clustering Re-Ranking., IEEE Trans. Circuits Syst. Video Technol. 30 (2020) 4929–4943.
[55] C. Kang, L. Zhu, X. Qian, J. Han, M. Wang, Y. Tang, Geometry and Topology Preserving Hashing for SIFT Feature., IEEE Trans. Multim. 21 (2019) 1563–1576.
[56] H. Zhai, S. Lai, H. Jin, X. Qian, T. Mei, Deep Transfer Hashing for Image Retrieval., IEEE Trans. Circuits Syst. Video Technol. 31 (2021) 742–753.
[57] X. Li, S. Lai, X. Qian, DBCFace: Towards Pure Convolution Neural Network Face Detection., IEEE Trans. Circuits Syst. Video Technol. (2021), https://doi.org/10.1109/TCSVT.2021.3082635.



Hanyang Jin received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2014, and M.S. degree from Guilin University Of Electronic Technology, Guilin, China, in 2017. He is currently pursing the Ph.D. degree with Xi'an Jiaotong University. His research interests are computer vision and multimedia retrieval.

Shenqi Lai received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2014, where he received the M.S. degree from the School of Software Engineering. His research interests are multimedia retrieval, neural network acceleration and computational aesthetics

Guoshuai Zhao, received the B.S. degree from Heilongjiang University, Harbin, China, in 2012, the M.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2015, and the Ph.D. degree at Xi'an Jiaotong University, Xi'an, China in 2019. He is with the Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, and the SMILES LAB of Xi'an Jiaotong University. He is mainly engaged in the research of social media big data analysis and recommender systems.

Xueming Qian (M'10) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, Xi'an, China, in 2008. He was a Visiting Scholar with Microsoft Research Asia, Beijing, China, from 2010 to 2011. He was previously an Assistant Professor at Xi'an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory, Xi'an Jiaotong University. His rsearch interests include social media big data mining and search.