Hard-Aware Point-to-Set Deep Metric for Person Re-identification

Rui Yu[1], Zhiyong Dou[1], Song Bai[1], Zhaoxiang Zhang[2], Yongchao Xu[1(✉)], and Xiang Bai[1(✉)]

[1] Huazhong University of Science and Technology, Wuhan, China
{zydou,yongchaoxu,xbai}@hust.edu.cn, yurui.thu@gmail.com,
songbai.site@gmail.com
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
zhaoxiang.zhang@ia.ac.cn

Abstract. Person re-identification (re-ID) is a highly challenging task due to large variations of pose, viewpoint, illumination, and occlusion. Deep metric learning provides a satisfactory solution to person re-ID by training a deep network under supervision of metric loss, *e.g.*, triplet loss. However, the performance of deep metric learning is greatly limited by traditional sampling methods. To solve this problem, we propose a Hard-Aware Point-to-Set (HAP2S) loss with a soft hard-mining scheme. Based on the point-to-set triplet loss framework, the HAP2S loss adaptively assigns greater weights to harder samples. Several advantageous properties are observed when compared with other state-of-the-art loss functions: (1) *Accuracy*: HAP2S loss consistently achieves higher re-ID accuracies than other alternatives on three large-scale benchmark datasets; (2) *Robustness*: HAP2S loss is more robust to outliers than other losses; (3) *Flexibility*: HAP2S loss does not rely on a specific weight function, *i.e.*, different instantiations of HAP2S loss are equally effective. (4) *Generality*: In addition to person re-ID, we apply the proposed method to generic deep metric learning benchmarks including CUB-200-2011 and Cars196, and also achieve state-of-the-art results.

Keywords: Person re-identification · Deep metric learning · Triplet loss

1 Introduction

Person re-identification (re-ID) [47] is a task of searching designated individuals from a large amount of pedestrian images captured by non-overlapping cameras. It attracts extensive attention owing to its significance in video surveillance. The large variations of pose, viewpoint, illumination and occlusion make person re-ID a very challenging problem.

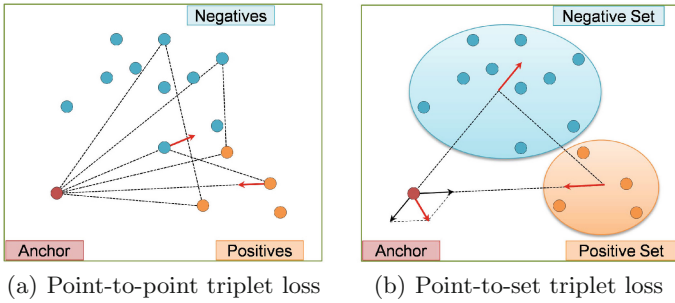(a) Point-to-point triplet loss   (b) Point-to-set triplet loss

Fig. 1. (a) Traditional point-to-point (P2P) triplet loss suffers from the sampling problem. Two of three selected triplets are useless for training. (b) The proposed loss is a point-to-set (P2S) triplet loss.
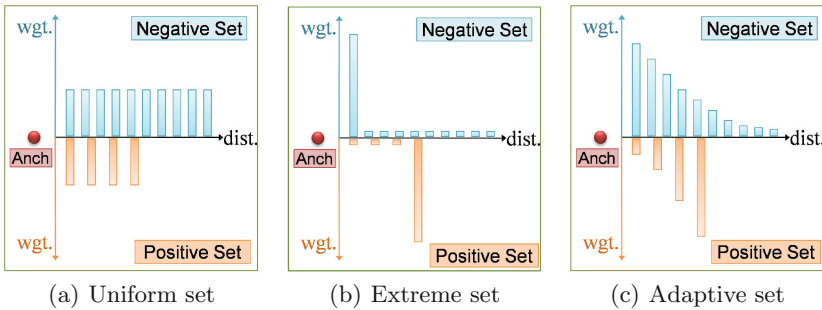

(a) Uniform set   (b) Extreme set   (c) Adaptive set

Fig. 2. Three types of weight distribution for computing the P2S distance. (a) Uniform set. All samples share equal weights. (b) Extreme set. Only the hardest sample is allotted a non-zero weight. (c) Adaptive set. Each sample is allotted a weight according to its "difficulty level". Note that for positive set, the hard samples are the ones far from anchor; while for negative set, the hard samples are near to anchor.

Thanks to the great success of deep learning in computer vision, along with the release of more and more large-scale person re-ID datasets [16,27,46], deep convolutional neural network (CNN) becomes a mainstream person re-ID method for learning discriminative feature representations, which have been proven superior to traditional hand-crafted features.

Since person re-ID can be regarded as a retrieval task in the testing phase, deep metric learning provides an effective methodology for training deep person re-ID models. Accordingly, metric loss function becomes a research hotspot in person re-ID. Several variants of metric loss have been widely applied to person re-ID, including contrastive loss [37], triplet loss [8,10], and quadruplet loss [4].

It is worth noting that the performance of metric loss is significantly influenced by the sampling method. Taking triplet loss as an example, it chooses three samples (anchor, positive, and negative) to constitute a triplet and generally aims to constrain anchor-to-positive distance less than anchor-to-negative

distance. Since a training set with N samples can generate $\mathcal{O}(N^3)$ triplets, it becomes infeasible to use all possible triplets even for a dataset of moderate size (*e.g.*, $N = 10^5$). Meanwhile, in practice, a large proportion of selected triplets have already satisfied the constraint and are useless. As shown in Fig. 1(a), the red, orange and blue points represent the anchor, positives and negatives, respectively. We choose three triplets and connect the points within each triplet to form three triangles. Only the triplet marked with red arrows is effective in training, while the other two triplets already satisfy the constraint and become junk. Thus, it is crucial for metric loss to select hard samples, which are able to violate the constraint and produce gradients with sufficiently large magnitude.

There are two types of hard sample mining approaches for triplet network: offline mining and online mining. Considering the offline methods are time consuming [1,10], we focus on the online mining manner. Unlike other online methods, we do not design an empirical rule for selecting triplets in a mini-batch. Instead, we propose a novel metric loss, namely Hard-Aware Point-to-Set (HAP2S) loss, which involves an adaptive hard mining mechanism.

First, we generalize the point-to-point (P2P) triplet loss to the point-to-set (P2S) triplet loss, as shown in Fig. 1(b). Given an anchor point, all the positive points and negative points in a mini-batch constitute the positive set and negative set, respectively. The P2S triplet loss constrains the distances from the anchor to the positive/negative set in a similar way as P2P triplet loss.

The key issue of P2S triplet loss is how to define the point-to-set distance. For that, we formulate the anchor-to-set distance by integrating the anchor-to-point distances and consider the contribution (*i.e.*, weight) of each P2P distance to the P2S distance. If all points have the same weight, as the uniform set illustrated in Fig. 2(a), the hard and easy samples are treated equally. The uniform set violates the principle of hard sample mining, and we find it almost impossible to conduct effective training in practice. Conversely, Hermans *et al.* [10] suggest choosing the hardest positive and hardest negative in a mini-batch to form the triplet, discarding the other samples. From the perspective of P2S distance, the method in [10] just assigns zero weight to all points other than the hardest samples (see the extreme set in Fig. 2(b)). However, this solution has two weaknesses: (1) It ignores other hard samples in the set which can also contribute to optimizing the network in the training, and thus may lead to a suboptimal solution. (2) It is easily influenced by a few error-labeled samples, *i.e.*, outliers, because the outliers usually act as the hardest samples and result in an undesired backpropagation.

To overcome these weaknesses, the proposed HAP2S loss introduces a soft hard-mining scheme rather than the traditional "select-or-unselect" manner. The basic idea is the harder samples adaptively gain more weights, as the adaptive set illustrated in Fig. 2(c).

Our main contribution is the HAP2S loss for person re-ID. To demonstrate its effectiveness, we evaluate the performance on three large-scale benchmark datasets including Market-1501 [46], CUHK03 [16], and DukeMTMC-reID [27,49]. Through experiments, HAP2S loss exhibits several advantages over other state-of-the-art deep metrics. (1) *Accuracy*: HAP2S loss consistently yields

higher re-ID accuracies than other alternatives. Simply based on the widely-used ResNet-50 [9] model, HAP2S loss can achieve state-of-the-art performances on the three person re-ID datasets. (2) *Robustness*: Through the experiments training with outliers, HAP2S loss is empirically more robust to outliers than other losses. (3) *Flexibility*: HAP2S loss does not depend on a specific weight function. Two instantiations of HAP2S loss achieve similar re-ID accuracies, implying that the effectiveness of HAP2S loss derives from the essential idea of P2S metric and hard-aware weighting. (4) *Generality*: HAP2S loss is also effective for generic deep metric learning tasks, and yields state-of-the-art results on two popular benchmarks including CUB-200-2011 [38] and Cars196 [13].

2 Related Works

Categorization of deep learning methods for re-ID. Based on testing manner, we categorize deep learning methods for person re-ID into two types: (1) binary classification of image-pair representation; (2) computing distance between single-image representations. The first type is known as verification network [16], which is usually trained under supervision of binary softmax loss, *i.e.*, verification loss. The second type aims to learn a discriminative single-image representation for certain distance metric, *e.g.*, L2-norm. There are two popular categories of loss functions for training single-image representation. One category is multi-class classification loss, *e.g.*, softmax loss (*a.k.a.* cross-entropy loss), which is also called identification loss in person re-ID [47]. The other category is metric loss, *e.g.*, triplet loss [8], which defines a metric among samples to compute the loss. The proposed HAP2S loss belongs to metric loss.

Metric losses for re-ID. A variety of metric losses have been used for training deep re-ID models. Varior *et al.* [37] compute the contrastive loss in a siamese architecture. In [8], the standard triplet loss is applied to person re-ID. Cheng *et al.* [7] add an additional positive-pair constraint to the original triplet loss. Based on the triplet loss, Chen *et al.* [4] propose a quadruplet loss, which further forces the intra-class distance less than the inter-class distance between two other classes. All the metric losses mentioned above are point-to-point losses, the performance of which are greatly influenced by the sampling schemes. Unlike them, our HAP2S loss, as a point-to-set loss, includes all samples in a mini-batch and implicitly applies a soft hard-sampling scheme when computing loss. It is worth mentioning that Zhou *et al.* [52] also propose a point-to-set loss for re-ID. The P2S in [52] is composed of a pairwise term, a triplet term, and a regularizer term. They assign equal weight to each sample in the marginal set for the triplet term, while our HAP2S loss adaptively allocates larger weights to the harder samples and thus significantly outperforms the P2S loss in [52].

Hard sample mining. As previously mentioned, the mining of hard samples plays an essential role in the performance of deep metric learning for person re-ID. Ahmed *et al.* [1] iteratively select the samples that perform worst on current model as the hard negative samples for fine-tuning. Shi *et al.* [30] suggest
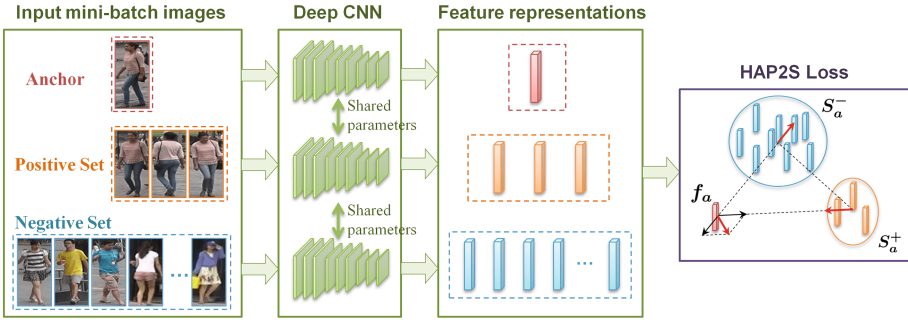
Fig. 3. The pipeline of the deep network using the proposed HAP2S loss

choosing the moderate positive samples by comparing with the hardest negative sample in a mini-batch. Chen *et al.* [4] propose a margin-based online hard negative mining method customized for their quadruplet loss. Hermans *et al.* [10] select the hardest positive and hardest negative of each anchor in a mini-batch to compute the triplet loss. All these methods select hard samples in a "select-or-unselect" manner. In contrast, our HAP2S loss introduces a more robust hard mining strategy by adaptively assigning larger weights to the harder samples.

3   Proposed Method

3.1   Overview

The objective of deep metric learning is to learn a deep neural network that maps an image $\boldsymbol{x}$ to a corresponding feature representation $f_{\boldsymbol{\Theta}}(\boldsymbol{x})$, which is suited to a predefined metric. The parameters of the network, *i.e.*, the weights and biases, are included in $\boldsymbol{\Theta}$. As for person re-ID, we can extract the features of probe and gallery images through a well-trained deep model and then compute the distances between the features to obtain a ranking list. The role of metric loss is to provide a discriminative metric for supervising the network training.

The proposed pipeline is depicted in Fig. 3. We adopt a pre-trained CNN model as the backbone network, which transfers each pedestrian image into an intermediate feature embedding. The backbone network employed in this work is ResNet-50 [9] model, which consists of five down-sampling blocks and one global average pooling layer. The backbone network is followed by two fully connected (FC) layers with 1024 and 128 neurons, respectively. After backbone and two FC layers, the network extracts the output features to compute HAP2S loss in a mini-batch. The Euclidean distance is employed as the point-to-point (P2P) metric. Details about the proposed method is given in the following.

3.2   Revisit Triplet Loss

Triplet loss [28,40] is one of the most representative metric losses. Given a training mini-batch $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N_s}$ with labels $\{y_i\}_{i=1}^{N_s}$, we select a triplet $\{\boldsymbol{x}_a, \boldsymbol{x}_p, \boldsymbol{x}_n\}$

where the *anchor* $\boldsymbol{x}_a$ and the *positive* $\boldsymbol{x}_p$ are two images from the same person, while the *negative* $\boldsymbol{x}_n$ is an image from another person. The corresponding features are $f_{\boldsymbol{\Theta}}(\boldsymbol{x}_a)$, $f_{\boldsymbol{\Theta}}(\boldsymbol{x}_p)$, and $f_{\boldsymbol{\Theta}}(\boldsymbol{x}_n)$. To simplify the notation, we use $\boldsymbol{f}_a$ to replace $f_{\boldsymbol{\Theta}}(\boldsymbol{x}_a)$, and so forth. Despite of several variants [7,25], the most common expression of triplet loss is as follows

$$\mathcal{L}_{trp} = \frac{1}{N_t} \sum_{\substack{y_p = y_a \\ y_n \neq y_a}} \left[d(\boldsymbol{f}_a, \boldsymbol{f}_p) - d(\boldsymbol{f}_a, \boldsymbol{f}_n) + m \right]_+, \tag{1}$$

where $[\cdot]_+ = \max(\cdot, 0)$, N_t represents the number of all possible triplets in the mini-batch, and d is a predefined distance metric. It can be seen from Eq. (1) that triplet loss aims to force the distance between an intra-class pair less than an inter-class pair by at least a margin m. While training a CNN with triplet loss, many of the possible triplets would easily satisfy the constraint

$$d(\boldsymbol{f}_a, \boldsymbol{f}_p) + m < d(\boldsymbol{f}_a, \boldsymbol{f}_n). \tag{2}$$

It makes the selected triplet equal to 0, *i.e.*, useless for training. Thus, the hard sample mining is critical to triplet loss. Hermans *et al.* [10] present a variant of triplet loss with a simple yet powerful hard-mining scheme, defined as

$$\mathcal{L}_{trpBH} = \frac{1}{N_s} \sum_{a=1}^{N_s} \left[\max_{y_p = y_a} d(\boldsymbol{f}_a, \boldsymbol{f}_p) - \min_{y_n \neq y_a} d(\boldsymbol{f}_a, \boldsymbol{f}_n) + m \right]_+, \tag{3}$$

where the hardest positive and hardest negative for each anchor in a mini-batch (Batch Hard) are selected to constitute a triplet. Based on this variant, they reported state-of-the-art results on two large-scale datasets.

3.3 Hard-Aware P2S Loss

Triplet loss is a type of P2P loss, since it only includes distances between points. Though interesting results can be obtained with the P2P triplet loss using a simple hard-mining scheme [10], as discussed in Sect. 1, such a simple hard-mining scheme may lead to two problems: (1) it excludes the contributions of other hard samples in the gradient descent training; (2) it is vulnerable to the outliers which usually serve as the hardest samples, causing undesired backpropagation. These two problems reveal that the simple solution of hard mining is not robust enough.

In this work, we generalize the P2P triplet loss to point-to-set (P2S) triplet loss. Given an anchor with label y_a, let $\boldsymbol{S}_a^+ = \{\boldsymbol{f}_p | y_p = y_a\}$ denote the positive set which contains all positive points in the mini-batch and similarly $\boldsymbol{S}_a^- = \{\boldsymbol{f}_n | y_n \neq y_a\}$ be the negative set. The P2S triplet loss is defined as

$$\mathcal{L}_{P2S} = \frac{1}{N_s} \sum_{a=1}^{N_s} \left[D\left(\boldsymbol{f}_a, \boldsymbol{S}_a^+\right) - D\left(\boldsymbol{f}_a, \boldsymbol{S}_a^-\right) + m \right]_+, \tag{4}$$

where D represents the P2S distance. The P2S triplet loss is a more generic form, which can be transferred to the P2P triplet loss in Eq. (3) if the P2S distance is defined as

$$
\begin{cases}
D\left(\boldsymbol{f}_a, \boldsymbol{S}_a^+\right) = \max_{\boldsymbol{f}_i \in \boldsymbol{S}_a^+} d\left(\boldsymbol{f}_a, \boldsymbol{f}_i\right) \\
D\left(\boldsymbol{f}_a, \boldsymbol{S}_a^-\right) = \min_{\boldsymbol{f}_j \in \boldsymbol{S}_a^-} d\left(\boldsymbol{f}_a, \boldsymbol{f}_j\right)
\end{cases} . \tag{5}
$$

From the perspective of P2S loss, the triplet loss in Eq. (3) only selects the hardest sample to represent the whole set.

To solve the problems of P2P triplet loss described above, we present a hard-aware P2S (HAP2S) loss with an adaptive hard mining scheme. The HAP2S loss has the same form as Eq. (4). The key of HAP2S loss is to assign different weights to the points in each set by computing the P2S distance as

$$
\begin{cases}
D\left(\boldsymbol{f}_a, \boldsymbol{S}_a^+\right) = \dfrac{\sum\limits_{\boldsymbol{f}_i \in \boldsymbol{S}_a^+} d(\boldsymbol{f}_a, \boldsymbol{f}_i) w_i^+}{\sum\limits_{\boldsymbol{f}_i \in \boldsymbol{S}_a^+} w_i^+} \\[2em]
D\left(\boldsymbol{f}_a, \boldsymbol{S}_a^-\right) = \dfrac{\sum\limits_{\boldsymbol{f}_j \in \boldsymbol{S}_a^-} d(\boldsymbol{f}_a, \boldsymbol{f}_j) w_j^-}{\sum\limits_{\boldsymbol{f}_j \in \boldsymbol{S}_a^-} w_j^-}
\end{cases} , \tag{6}
$$

where w_i^+ and w_j^- represent the weights of the elements $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ in the positive and negative set respectively. As discussed in Sect. 1, an effective hard mining strategy should assign higher weights to the hard samples in a set. Considering the metric loss, the "difficulty level" of a sample lies in the distance from the anchor to it. Accordingly, for the positive set, the remote points to the anchor are the hard ones and deserve higher weights. On the contrary, for the negative set, the nearest point to the anchor is the hardest. To this end, we introduce two weighting schemes to the proposed HAP2S loss.

(i) Exponential weighting. The first weighting scheme is exponential weighting. The weights of the elements in each set are defined as

$$
\begin{cases}
w_i^+ = \exp\left(\dfrac{d(\boldsymbol{f}_a, \boldsymbol{f}_i)}{\sigma}\right) & \text{if } \boldsymbol{f}_i \in \boldsymbol{S}_a^+ \\
w_j^- = \exp\left(-\dfrac{d(\boldsymbol{f}_a, \boldsymbol{f}_j)}{\sigma}\right) & \text{if } \boldsymbol{f}_j \in \boldsymbol{S}_a^-
\end{cases} , \tag{7}
$$

where $\sigma > 0$ is a coefficient for adjusting the weight distribution. In this way, the weight of each sample exponentially adapts to its "difficulty level". The complete formula of HAP2S loss $\mathcal{L}_{HAP2S}$ is composed of Eqs. (4), (6) and (7).

(ii) Polynomial weighting. Instead of exponential weighting, we can define an alternate HAP2S loss by assigning weights to the elements in each set via a univariate polynomial function with real coefficients, as

$$
\begin{cases}
w_i^+ = \left(d\left(\boldsymbol{f}_a, \boldsymbol{f}_i\right) + 1\right)^{\alpha} & \text{if } \boldsymbol{f}_i \in \boldsymbol{S}_a^+ \\
w_j^- = \left(d\left(\boldsymbol{f}_a, \boldsymbol{f}_j\right) + 1\right)^{-2\alpha} & \text{if } \boldsymbol{f}_j \in \boldsymbol{S}_a^-
\end{cases} , \tag{8}
$$

where $\alpha > 0$ is also a coefficient for adjusting the weight distribution. The weighting scheme of Eq. (8) is similar to that of Eq. (7) by assigning greater weights to harder samples. This instantiation of $\mathcal{L}_{HAP2S}$ consists of Eqs. (4), (6) and (8). In order distinguish the two instantiations, we denote the former one as **HAP2S_E** and the latter one as **HAP2S_P**.

To demonstrate why HAP2S loss outperforms other alternatives, we analyze the gradient to show how the loss optimizes the network parameters. Due to space limitation, the detailed analyses are given in the supplementary material.

3.4   Multi-loss Training

Since various losses establish different optimization objectives, joint supervision of different losses usually helps to train better deep re-ID models. For example, McLaughlin *et al.* [23] adopt softmax loss and contrastive loss to train a recurrent neural network; Chen *et al.* [5] optimize a multi-task deep network jointly by verification loss, triplet loss, and contrastive loss.

We notice that the metric loss (*e.g.*, HAP2S loss) does not fully utilize the annotations provided by the training set. It only verifies the labels of two samples, but ignores the specific class ID. In contrast, the classification loss (*e.g.*, softmax loss) exactly uses the multi-class labels as the supervision information. Based on this observation, we can combine the proposed HAP2S loss with softmax loss to further boost re-ID performance. The details of multi-loss training (including network, algorithm, and experiments) are given in the supplementary material.

4   Experiments

4.1   Datasets and Evaluation Protocols

We evaluate the proposed method on three challenging large-scale benchmark datasets including Market-1501 [46], CUHK03 [16], and DukeMTMC-reID [27, 49].

Market-1501. This dataset comprises 32,668 labeled images of 1,501 identities captured by six cameras. Within the dataset, 12,936 images of 751 identities are used for training, while the rest are used for testing. Among the testing data, fixed 3,368 images constitute the probe set. The testing set also contains 2,793 distractor images, which makes this dataset very challenging.

CUHK03. This dataset consists of 14,096 images of 1,467 identities captured by six cameras. Each identity shows in two camera views and has 4.8 images on average in one view. The dataset provides two types of data. One is a set of manually labeled bounding boxes of pedestrians, while the other set contains automatically detected bounding boxes by the DPM detector. We conduct experiments on both "labeled" and "detected" datasets.

DukeMTMC-reID. This dataset contains 36,411 images of 1,812 identities, which are manually cropped from multi-camera tracking dataset DukeMTMC [27]. The images are captured by eight cameras, and 1,404 identities

Table 1. Comparison with other losses based on pre-trained ResNet-50 model. Note that test-phase data augmentations are *not* applied throughout the experiments.

Loss	Market1501		CUHK03 (labeled)			CUHK03 (detected)			DukeMTMC	
	mAP	r = 1	r = 1	r = 5	r = 10	r = 1	r = 5	r = 10	mAP	r = 1
Softmax [47]	58.16	79.25	72.81	95.09	97.89	71.61	92.20	95.64	49.33	71.01
Triplet [8]	53.40	70.84	78.11	97.17	98.15	75.71	94.55	97.07	48.16	66.02
Improved Triplet[7]	55.10	71.56	80.56	97.69	98.48	76.10	95.43	97.36	49.63	66.92
P2S [52]	54.30	70.99	77.82	97.67	98.40	75.94	96.01	97.64	49.74	66.88
OIM [41]	60.48	81.26	77.88	94.86	97.54	75.09	93.41	95.84	51.32	71.99
Quadruplet [4]	64.88	81.47	85.17	97.64	98.52	84.13	97.70	98.55	54.29	73.47
Hard Triplet [10]	67.22	82.13	87.65	98.46	99.37	86.34	97.28	98.65	57.08	74.37
HAP2S_E	**69.76**	84.20	90.22	98.76	99.37	88.13	97.67	98.69	59.58	**76.08**
HAP2S_P	69.43	**84.59**	**90.39**	**99.54**	**99.90**	**88.90**	**98.44**	**99.09**	**60.64**	75.94

appear in more than one camera. Following [49], the 1,404 identities are divided into two halves, with 702 identities for training and the others for testing.

Evaluation protocols. We adopt the standard evaluation protocols. For Market-1501 and DukeMTMC-reID, since the data are divided fixedly, we directly evaluate the cumulative matching characteristics (CMC) and mean average precision (mAP), and report the average results of two independent trials. For CUHK03, 100 identities are selected for testing and the rest are used for training. The CUHK03 provides 20 different train/test splits, so we report the average CMC on the 20 trails. All experiments are by default under the single query setting. We also report multiple query evaluation results for Market-1501.

4.2   Implementation Details

We implement the deep model on PyTorch framework. The network is trained with Adam [12] under supervision of HAP2S loss. The learning rate is 4×10^{-4} at the first 100 epochs and gradually decay to 4×10^{-7} at the 150th epoch.

All images are first resized to 256×128. Standard random crop and horizontal flipping are adopted for data augmentation in training phase. Following [8,10], we select a fixed number of images for each person to form a mini-batch. In our experiments, eight images from each of 32 persons are randomly chosen as a 256-size mini-batch. We set the parameters of HAP2S loss with margin $m = 2.5$, and weight coefficient $\sigma = 0.5$ (*resp.* $\alpha = 10$) for HAP2S_E (*resp.* HAP2S_P). It costs less than an hour to train the model on two GTX TITAN X 12GB GPUs.

In testing phase, we do not apply any data augmentations (*e.g.*, five crops and flips [10]) on account of efficiency. We extract the intermediate feature given by the backbone network for each image and produce the ranking results according to Euclidean metric. Note that the results of multi-loss training (Sect. 3.4) are not reported below, but given in the supplementary material.

4.3 Comparison with Other Losses

We compare HAP2S loss with state-of-the-art losses reported in recent person re-ID works, including softmax loss [47], triplet loss [8], improved triplet loss [7], P2S loss [52], OIM loss [41], quadruplet loss [4], and hard triplet loss [10]. We separately evaluate the performance of each loss with the pre-trained ResNet-50 model on the three datasets. To be fair in comparison, we apply the same mini-batch configuration and tune the parameters to optimum for each loss. The hard mining approaches presented by original papers are also reproduced.
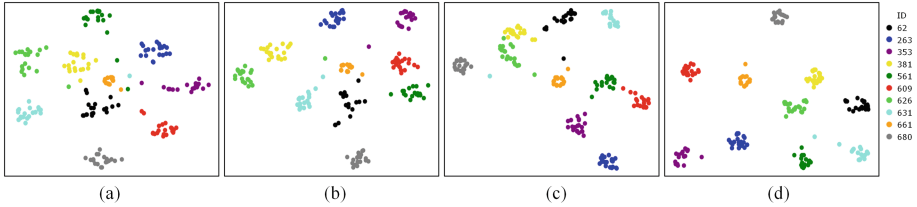

Fig. 4. Visualization of deeply-learned features by (a) softmax loss [47], (b) triplet loss [8], (c) hard triplet loss [10], (d) HAP2S loss using exponential weighting. We randomly select 10 identities from the testing set of Market-1501. The points with different colors denote features from different identities. (Best viewed in color)
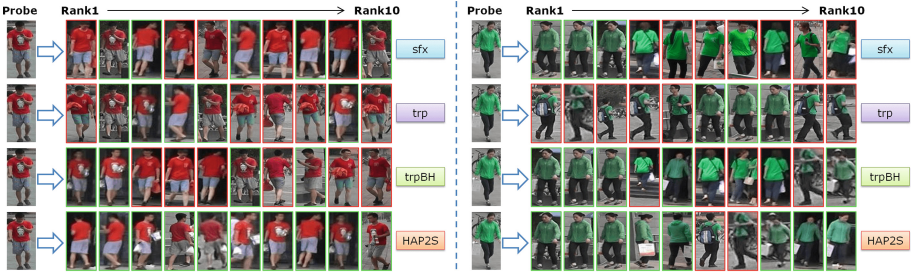

Fig. 5. Examples of some retrieval results on Market-1501. Each row contains the top-10 rank images retrieved by the corresponding method (*sfx* - softmax loss [47]; *trp* - triplet loss [8]; *trpBH* - hard triplet loss [10]). The correct and false matches are enclosed in green and red boxes, respectively. (Best viewed in color)

HAP2S loss consistently outperforms other alternatives. The experimental results are presented in Table 1. As can be seen, the identification losses (softmax [47] and OIM [41]) achieve favorable re-ID accuracies, while the metric losses (*e.g.*, quadruplet [4] and hard triplet [10]) can realize higher performances with certain hard mining strategies. The proposed HAP2S loss using either exponential weighting or polynomial weighting outperforms all other competitors

on the three datasets. On Market-1501 and DukeMTMC-reID, the performance gaps between HAP2S and other losses are consistently more than +2.2% in mAP. As for CUHK03, HAP2S loss also outperforms other alternatives with a noticeable improvement in rank-1 accuracy on both labeled and detected datasets. It is worth noting that the results of HAP2S_P are on par with that of HAP2S_E. Thus, HAP2S loss does not depend on a specific weight function. More generally, it is expected that any weighting functions with similar properties as Eq. (7) or (8) would produce other effective instantiations of HAP2S loss.

Table 2. Comparison with state-of-the-art methods. (Best results are highlighted)

Method	Market-1501				CUHK03						DukeMTMC	
	Single query		Multi query		Labeled			Detected				
	mAP	r = 1	mAP	r = 1	r = 1	r = 5	r = 10	r = 1	r = 5	r = 10	mAP	r = 1
OL-MANS [51]	-	60.67	-	74.00	54.7	86.5	93.9	45.0	76.0	83.5	-	-
DNS [43]	35.68	61.02	46.03	71.56	62.6	90.1	94.8	54.7	84.8	94.8	-	-
Gated SCNN[36]	39.55	65.88	48.45	76.04	-	-	-	68.1	88.1	94.6	-	-
P2S [52]	44.27	70.72	55.73	85.78	-	-	-	-	-	-	-	-
MTDnet [5]	-	-	-	-	74.7	96.0	97.5	-	-	-	-	-
CRAFT-MFA[6]	42.30	68.70	50.30	77.00	-	-	-	84.3	97.1	98.3	-	-
ACRN [29]	62.60	83.61	-	-	-	-	-	62.6	89.7	94.7	51.96	72.58
CADL [19]	47.11	73.84	55.58	80.85	-	-	-	-	-	-	-	-
MSCAN [15]	57.53	80.31	66.70	86.79	74.2	94.3	97.5	68.0	91.0	95.4	-	-
Quadruplet [4]	-	-	-	-	75.5	95.2	99.2	-	-	-	-	-
PAN [48]	63.35	82.81	71.72	88.18	-	-	-	-	-	-	51.51	71.59
Re-ranking [50]	63.63	77.11	-	-	61.6	-	-	58.5	-	-	-	-
APR [20]	64.67	84.29	-	-	-	-	-	-	-	-	51.88	70.69
SSM [2]	68.80	82.21	76.18	88.18	76.6	94.6	98.0	72.7	92.4	96.1	-	-
MuDeep [26]	-	-	-	-	76.9	96.1	98.4	75.6	94.4	97.5	-	-
SVDNet [35]	62.10	82.30	-	-	-	-	-	81.8	95.2	97.2	56.80	76.70
JLML [17]	65.50	85.10	74.50	89.70	83.2	98.0	99.4	80.6	96.9	98.7	-	-
Part-Aligned[45]	63.40	81.00	-	-	85.4	97.6	99.4	81.6	97.3	98.4	-	-
Spindle [44]	-	76.90	-	-	88.5	97.8	98.6	-	-	-	-	-
PDC [34]	63.41	84.14	-	-	88.7	98.6	99.2	78.3	94.8	97.2	-	-
LSRO [49]	66.07	83.97	76.10	88.42	-	-	-	84.6	97.6	98.9	47.13	67.68
MLFN [3]	74.30	90.00	82.40	92.30	-	-	-	82.8	-	-	62.80	**81.00**
HA-CNN [18]	**75.70**	**91.20**	**82.80**	**93.80**	-	-	-	-	-	-	**63.80**	80.50
HAP2S_E	69.76	84.20	76.50	88.69	90.2	98.8	99.4	88.1	97.7	98.7	59.58	76.08
HAP2S_P	69.43	84.59	76.75	90.20	**90.4**	**99.5**	**99.9**	**88.9**	**98.4**	**99.1**	60.64	75.94

Visualization analysis. The t-SNE [21] tool is adopted to visualize the feature embeddings learned by the losses. We randomly choose 10 identities and 20 images for each identity from the testing set of Market-1051. The visualization results of the features are plotted in Fig. 4. As it shows, HAP2S loss achieves larger inter-class variances and smaller intra-class variances than other losses. In addition, we show several example re-ID results on Market-1501 in Fig. 5. HAP2S loss can find more correct matches than other losses in top ranks.

In sum, based on the quantitative and qualitative results, the superiority of HAP2S loss is not only proved by the better retrieval performance (see Table 1), but also by the superior clustering quality (see Fig. 4).

4.4   Comparison with State-of-the-Arts

We compare in Table 2 the proposed method with the state-of-the-art methods on the three datasets described in Sect. 4.1.

Comparison on Market-1501. For the sake of testing efficiency, we apply neither test-phase augmentation nor post-ranking. Despite this, the proposed method still outperforms most of state-of-the-arts under both single query (SQ) and multiple query (MQ) settings. When compared with a recently reported multi-loss model JLML [17], the proposed method achieves an improvement of about +4% mAP for SQ. It is also worth mentioning that the performance of the proposed method can be further boosted by re-ranking tools. For example, when applying the re-ranking approach [50], the proposed method can achieve 79.91% mAP and 85.72% rank-1 accuracy for SQ.
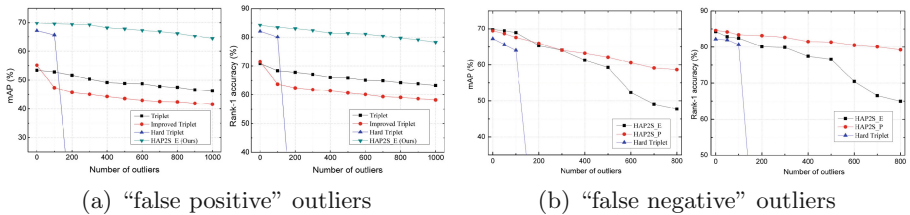

(a) "false positive" outliers   (b) "false negative" outliers

Fig. 6. Re-ID results on Market-1501 with different numbers of outliers in training set.

Comparison on CUHK03. The proposed method yields the best rank-1 accuracies on both labeled and detected datasets. For labeled dataset, the pose-driven deep convolutional (PDC) model [34] reports previous best performance of 88.7% in rank-1 accuracy. The proposed method outperforms PDC [34] on the labeled dataset by about +1.5% rank-1 accuracy, and increases the performance gap with PDC by +9.8% rank-1 accuracy on the detected dataset. Therefore, the proposed method exhibits greater robustness than the body-part-based method in the automatically detected scenario, which is closer to practical applications.

Comparison on DukeMTMC-reID. The state-of-the-art performances are achieved by MLFN [3] and HA-CNN [18]. In [3], a novel multi-level factorization net is proposed to learn latent discriminative factors without manual annotation. In [18], a well-designed CNN can learn discriminative features by soft pixel attention and hard regional attention. The loss function in [3] and [18] is softmax loss. By contrast, we only use pre-trained ResNet-50 network. It can be expected that HAP2S loss would outperform current state-of-the-arts by using better backbone networks such as HA-CNN [18].

4.5   Robustness to Outliers

In order to assess the robustness of different losses to outliers, we conduct experiments with outliers. Specifically, we randomly select a certain number of images

Table 3. Re-ID results on Market-1501 with different σ (HAP2S_E) or α (HAP2S_P)

σ	$\sigma\to 0$[10]	$\sigma=0.25$	$\sigma=0.5$	$\sigma=0.75$	$\sigma=1$	$\sigma=1.25$	$\sigma=1.5$	$\sigma=2$	$\sigma=4$	$\sigma=8$	$\sigma\to\infty$
mAP	67.22	68.86	**69.76**	69.25	68.21	67.88	67.23	66.57	63.68	62.30	1.36
r = 1	82.13	83.28	**84.20**	83.76	82.48	82.27	81.09	79.87	78.15	76.22	3.30
α	$\alpha=0$	$\alpha=4$	$\alpha=5$	$\alpha=6$	$\alpha=7$	$\alpha=8$	$\alpha=9$	$\alpha=10$	$\alpha=11$	$\alpha=12$	$\alpha\to\infty$[10]
mAP	3.11	59.07	64.75	67.46	68.48	69.29	69.39	**69.43**	69.20	68.94	67.22
r = 1	7.27	76.07	80.26	82.93	83.76	84.03	84.31	**84.59**	83.64	83.21	82.13

from CUHK03 and add them to the training set of Market-1501 as outliers. Each outlier is randomly labeled with an ID number from the 751 identities of Market-1501 training set. Then, we use the new training set to train a deep network and conduct the standard evaluation on the testing set of Market-1501.

Several variants of triplet loss are evaluated in this experimental setting. The re-ID performances with varying number of outliers are depicted in Fig. 6(a). The proposed HAP2S loss is least affected by outliers. Even when 1,000 outliers are present, HAP2S_E loss still achieves a high performance of 64.48% mAP. Traditional triplet loss [8] is more robust to outliers than the improved triplet loss [7]. In line with previous analyses, hard triplet loss [10] is very sensitive to outliers. When the number of outliers is larger than 200, the model training with hard triplet loss is prone to collapse due to multiple outliers in a mini-batch.

Besides the "false positive" outliers, we also report the re-ID performances with "false negative" outliers. In particular, we randomly assign false ID labels to varying number of images in original Market-1501 training set. The evaluation results depicted in Fig. 6(b) also verifies the robustness of HAP2S loss.

4.6   Parameter Analysis

For either the exponential weighting in Eq. (7) or the polynomial weighting in Eq. (8), a coefficient (σ or α) is introduced to adjust the weight distribution. Here, we empirically analyze the impact of the coefficient. According to Eqs. (7) and (8), HAP2S loss is equivalent to hard triplet loss [10] when $\sigma\to 0$ or $\alpha\to\infty$. Conversely, HAP2S loss turns into uniform weighting when $\sigma\to\infty$ or $\alpha=0$.

We evaluate the re-ID performance of HAP2S loss on Market-1501 dataset under different parameters. As seen in Table 3, HAP2S loss can yield both high accuracy and training stability when $\sigma \leq 1.5$ or $\alpha \geq 6$. The uniform weighting ($\sigma\to\infty$ or $\alpha=0$) tends to produce no loss and fail in training.

4.7   Generic Deep Metric Learning

Datasets and evaluation metrics. The proposed HAP2S deep metric has demonstrated its superiority on person re-ID task in previous experiments. To further assess the effectiveness of HAP2S loss, we evaluate it on two popular deep metric learning benchmarks: CUB-200-2011 [38] and Cars196 [13]. We follow the same training/testing split and standard protocol described in [31,33]. The

Table 4. Comparison with state-of-the-art methods of generic deep metric learning

Method	CUB-200-2011					Cars196				
	NMI	R@1	R@2	R@4	R@8	NMI	R@1	R@2	R@4	R@8
Lifted Struct [33]	56.50	43.57	56.55	68.59	79.63	56.88	52.98	65.70	76.01	84.27
N-pairs [31]	57.24	45.37	58.41	69.51	79.49	57.79	53.90	66.76	77.75	86.35
Clustering [32]	59.23	48.18	61.44	71.83	81.92	59.04	58.11	70.64	80.27	87.81
Proxy NCA [24]	59.53	49.21	61.90	67.90	72.40	**64.90**	73.22	82.42	86.36	88.68
Smart Mining [14]	59.90	49.78	62.34	74.05	83.31	59.50	64.65	76.20	84.23	90.19
HDC [42]	-	53.60	65.70	77.00	85.60	-	73.70	83.20	89.50	93.80
Angular Loss [39]	61.10	54.70	66.30	76.00	83.90	63.20	71.40	81.40	87.50	92.10
HAP2S_E	**63.42**	**56.08**	**68.43**	**79.20**	**86.93**	63.05	**74.09**	**83.45**	**89.85**	**94.13**

CUB-200-2011 dataset [38] consists of 11,788 bird images of 200 species, where the first 100 species (5,864 images) are used for training and the others for testing. The *Cars196* dataset [13] contains 16,185 car images of 196 classes. We use the first 98 classes (8,054 images) for training and the rest for testing. The performances are measured by two standard metrics. The normalized mutual information (NMI) [22] is used for evaluating the clustering quality, while the Recall@K metric [11] serves to measure the retrieval performance.

Comparison with state-of-the-arts. We adopt the same network architecture as for person Re-ID task (Fig. 3) for training and testing. As the proposed HAP2S_E and HAP2S_P loss performs similarly, for simplicity, we only compare HAP2S_E with state-of-the-art methods, as shown in Table 4. The proposed HAP2S loss outperforms the other methods by about $+2\%$ in terms of both NMI and Recall@K scores on the CUB-200-2011 dataset. For the Cars196 dataset, HAP2S loss achieves the best Recall@K scores in all ranks, while reaching comparable NMI with the state-of-the-art methods.

5   Conclusions

The selection of training samples is a crucial problem in deep metric learning for person re-ID. In this paper, we propose a novel loss function, Hard-Aware Point-to-Set (HAP2S) loss, to solve the sampling problem in a robust manner. Unlike traditional solutions, HAP2S loss does not focus on how to select samples, but to distribute different weights to the samples. Based on the P2S triplet loss framework, HAP2S loss adaptively assigns greater weights to harder samples. We conduct extensive experiments on three large-scale person re-ID benchmarks. Benefiting from the soft hard-mining scheme, HAP2S loss achieves state-of-the-art re-ID accuracies on the three datasets. Besides, HAP2S loss performs more robust than other alternatives when some outliers are present in the training set. Moreover, HAP2S loss is also able to yield state-of-the-art performances on generic deep metric learning benchmarks. In this work, we mainly target on loss function and adopt the off-the-shelf ResNet-50 network. It can be expected

that the performance of the proposed HAP2S loss would be further boosted by bespoke re-ID networks [3,18].

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR, pp. 3908–3916 (2015)
2. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. In: CVPR, pp. 2530–2539 (2017)
3. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: CVPR, pp. 2109–2118 (2018)
4. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR, pp. 403–412 (2017)
5. Chen, W., Chen, X., Zhang, J., Huang, K.: A multi-task deep network for person re-identification. In: AAAI, pp. 3988–3994 (2017)
6. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE TPAMI **40**(2), 392–408 (2018)
7. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: CVPR, pp. 1335–1344 (2016)
8. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognit. **48**(10), 2993–3003 (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
10. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification (2017). arXiv:1703.07737
11. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE TPAMI **33**(1), 117–128 (2011)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
13. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV Workshop, pp. 554–561 (2013)
14. Kumar, V.B., Harwood, B., Carneiro, G., Reid, I., Drummond, T.: Smart mining for deep metric learning. In: ICCV, pp. 2821–2829 (2017)
15. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR, pp. 384–393 (2017)
16. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR, pp. 152–159 (2014)
17. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: IJCAI, pp. 2194–2200 (2017)
18. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR, pp. 2285–2294 (2018)

19. Lin, J., Ren, L., Lu, J., Feng, J., Zhou, J.: Consistent-aware deep learning for person re-identification in a camera network. In: CVPR, pp. 5771–5780 (2017)
20. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning (2017). arXiv:1703.07220
21. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR **9**, 2579–2605 (2008)
22. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
23. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: CVPR, pp. 1325–1334 (2016)
24. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: ICCV, pp. 360–368 (2017)
25. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR, pp. 4004–4012 (2016)
26. Qian, X., Fu, Y., Jiang, Y.G., Xiang, T., Xue, X.: Multi-scale deep learning architectures for person re-identification. In: ICCV, pp. 5399–5408 (2017)
27. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshop, pp. 17–35 (2016)
28. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: CVPR, pp. 815–823 (2015)
29. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: CVPR Workshop, pp. 1435–1443 (2017)
30. Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., Li, S.Z.: Embedding deep metric for person re-identification: a study against large variations. In: ECCV, pp. 732–748 (2016)
31. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: NIPS, pp. 1857–1865 (2016)
32. Song, H.O., Jegelka, S., Rathod, V., Murphy, K.: Deep metric learning via facility location. In: CVPR, pp. 5382–5390 (2017)
33. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR, pp. 4004–4012 (2016)
34. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV, pp. 3960–3969 (2017)
35. Sun, Y., Zheng, L., Deng, W., Wang, S.: SVDNet for pedestrian retrieval. In: ICCV, pp. 3800–3808 (2017)
36. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: ECCV, pp. 791–808 (2016)
37. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: ECCV, pp. 135–153 (2016)
38. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
39. Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: ICCV, pp. 2593–2601 (2017)
40. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. JMLR **10**, 207–244 (2009)
41. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR, pp. 3415–3424 (2017)
42. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: ICCV, pp. 814–823 (2017)

43. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR, pp. 1239–1248 (2016)
44. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR, pp. 1077–1085 (2017)
45. Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. In: ICCV, pp. 3219–3228 (2017)
46. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV, pp. 1116–1124 (2015)
47. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future (2016). arXiv:1610.02984
48. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification (2017). arXiv:1707.00408
49. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV, pp. 3754–3762 (2017)
50. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR, pp. 1318–1327 (2017)
51. Zhou, J., Yu, P., Tang, W., Wu, Y.: Efficient online local metric adaptation via negative samples for person re-identification. In: ICCV, pp. 2420–2428 (2017)
52. Zhou, S., Wang, J., Wang, J., Gong, Y., Zheng, N.: Point to set similarity based deep feature learning for person re-identification. In: CVPR, pp. 3741–3750 (2017)