Applied Intelligence (2022) 52:4844–4857
https://doi.org/10.1007/s10489-021-02450-y

HOB-net: high-order block network via deep metric learning for person re-identification

Dongyue Chen[1] · **Pengfei Wu**[1] · **Tong Jia**[1] · **Fangbin Xu**[1]

Accepted: 20 April 2021 / Published online: 29 July 2021

Abstract
Learning effective feature representations with deep convolutional neural network (CNN) and metric learning methods to distinguish pedestrians is the key to the success of recent advances in person re-identification (re-ID) tasks. However, the features provided by the common CNN network are not strong enough to distinguish between similar subjects because these common features describe only the scattered local patterns while neglecting the correlation and combinations between them. In fact, the high-order correlations of common features can be significant to the recognition of identities. In this paper, to tackle this problem, we design a flexible high-order block (HOB) module and a scheme of deep metric learning to produce the high-order representations of deep features for the re-identification of pedestrians. Extensive experiments prove the superiority of our proposed HOB module for person re-ID issue. On three large-scale datasets, including Market-1501, DukeMTMC-ReID, and CUHK03-NP, the HOB-net method achieves the competitive results with the state-of-the-arts, particularly in the mAP.

Keywords Person re-identification · CNN · Metric learning · High-order statistics

1 Introduction

Person re-identification is a branch of image retrieval, which aims to identify the same person from multiple detected pedestrian images, typically captured from different cameras without view overlap [1]. In the last few decades, with the development of human recognition and image restoration technology in various real-world scenarios, such as video surveillance, pedestrian detection, and video restoration [2, 3], the algorithm of person re-ID has accomplished quite a lot of important applications in the intelligent security system, individual tracking, and smart mall system, etc. Although the re-ID issue has already achieved dramatic progress with the utilization of CNN, it is still a challenging problem with a few aspects remaining to be bettered due to the complexity of pedestrians in real-world scenarios, e.g. pose changes, environment illumination, and view angle changes, etc. As shown in the Fig. 1.

Affected by the aforementioned factors, the feature representations of pedestrians obtained by the CNN extractor actually cannot represent the input images exactly. Therefore, finding a group of pedestrian representations with good anti-interference, invariance, and distinguish ability has become the key for re-ID issue. For this purpose, many works [4, 5] choose to localize different body parts and align their associated features, while many other works [6, 7] use spatial or channel-based attention selection network to improve feature learning. However, all of the above works are confined by first-order occurrences and just mine simple and coarse information, which cannot be well applied to model person in re-ID cases. Considering the subtle differences among pedestrians caused by complexity in real-world scenarios, a simple representation based on the first-order features is obviously insufficient to capture the interactions of visual parts. As a result, the extracted feature

✉ Tong Jia
1900908@stu.neu.edu.cn

Dongyue Chen
chendongyue@ise.neu.edu.cn

Pengfei Wu
pengfeiwu0919@163.com

Fangbin Xu
xufangbin1996@foxmail.com

[1] College of Information Science and Engineering, Northeastern University, Shenyang, China

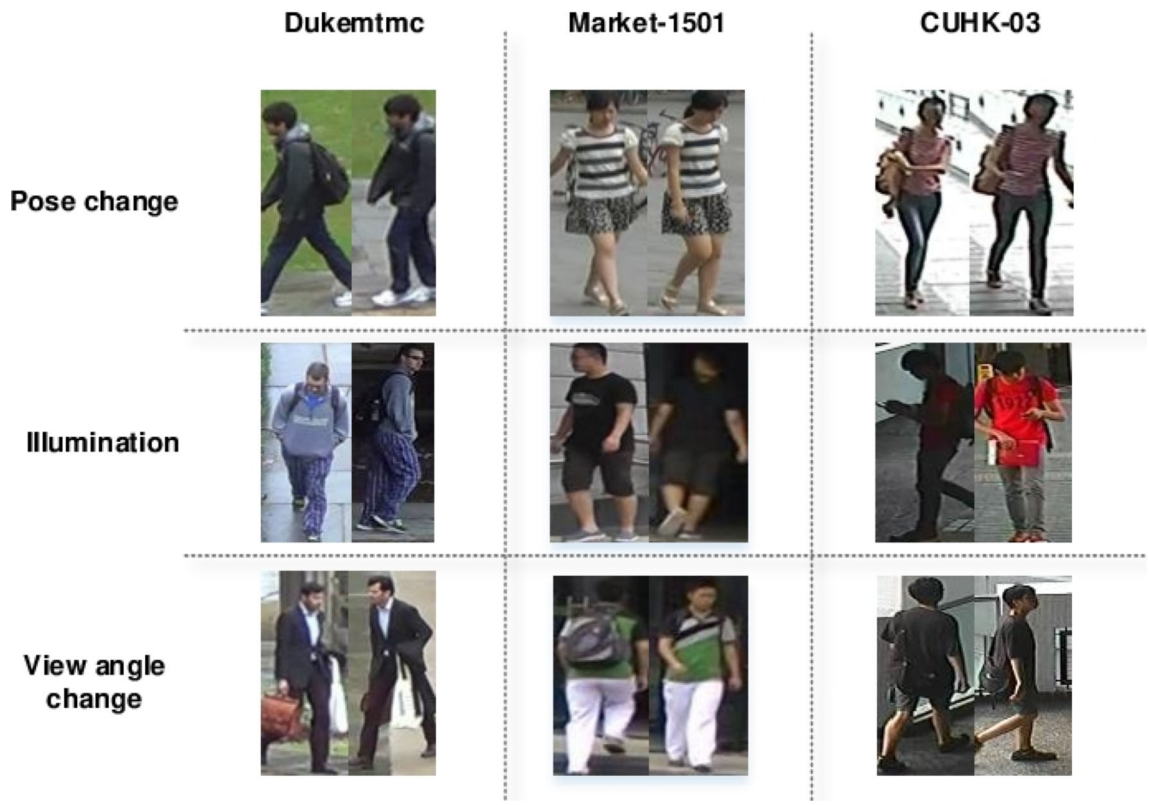

Fig. 1 Typical images from three mainstream datasets: DukeMTMC [8], Market1501 [9], CUHK03 [10]. Each pair of images represent the same person

presentations of pedestrians are not discriminative enough for the target task.

In this paper, we propose a flexible and powerful feature extraction module, referred to as High-Order Block (HOB), to extract high-order statistics from the deep features provided by the backbone network. We dedicate to modeling the deep metric learning mechanism via high-order statistics so as to capture the relationships among training samples and produce refined feature representations for pedestrians. To this end, we design a HOB-based network (HOB-net) to upgrade the quality of features by integrating the high-order statistics information into representations of input images.

Our main contributions of the paper are summarized as follows:

- First of all, based on the comparison between three different layouts of convolutional layers, the High-Order Block (HOB) module with the new architecture is proposed as an embedding part of the backbone network to extract high-order statistics from the deep features.
- Second, a new feature extraction network based on the HOB module in multiple orders is proposed, with which the effectiveness of different combinations between HOB module and loss functions are explored to find out their best scheme for the re-ID task.
- At last, through extensive experiments, we prove the superiority of the proposed HOB-net over a wide range of state-of-the-art re-ID models on three large benchmarks, i.e. DukeMTMC-ReID [8], Market-1501 [9] and CUHK03-NP [10].

2 Related work

2.1 Backbone architecture for re-ID

One of the key issues of re-ID model based on deep learning is the architecture of feature extraction deep network. At present, many advanced architectures of feature networks have been presented, including strip-based, attention-based, and spatial deformation, etc. Stripe-based methods [4, 5, 11, 12] aggregate the salient local features from different body parts and global cues together to improve the representation. In which, [4] split the input feature map horizontally into a fixed number of strips, from which local features are aggregated. Attention-based methods [6, 13–16] enhanced the feature representations with attention mechanism, which guides the feature-extract network to capture

and focus on attentive regions, to handle the imperfect detection of the bounding box and the misalignment of body parts. Besides, spatial deformation methods [17, 18] introduce a spatial transformer network (STN) to align pedestrians or introduce generative adversarial network (GAN) to generate standard posture to alleviate the influence of pose variance. All the three methods above boost features based on new fashions of spatial arrangement while ignoring the representation ability and distinguish ability of the features themselves in nature. Consequently, these methods are limited to distinguish and recognize very similar targets. Hence, finding more detailed, comprehensive, and distinctive features by digging into the structure and learning methods of the deep network is a promising direction in the researches on re-ID issue.

2.2 High-order metric learning

Inspired by some impressive works on fine-grained image classification [19–23], we can obtain the features with stronger capacity of the representation and discrimination, namely high-order metric learning. High-order metric learning is to change the fixed first-order distance measure between feature and loss function into parameterized high-order distance measure. And the parameters of high-order distance measure can be learned in the process of optimization. Furthermore, through gradient back-propagation, the feature representation can get more refined learning based on high-order measurement, so as to improve their ability to distinguish between different targets with the first-order similarity but insufficient high-order similarity. Many recent works [23] about fine-grained visual categorization and large-scale visual recognition tasks have demonstrated that second-order statistics have better performance than descriptors exploiting first-order statistics. However, only using second-order or lower moments information might not be enough when the feature distribution is not Gaussian [21]. Naturally, the higher-order (greater than two) statistics have been explored in many works [19–22]. Among them, [21] utilizes the third-order statistics for person re-ID. [19, 20, 22] exploit higher-order statistics for visual concept detection and fine-grained visual categorization. Although there are some differences between fine-grained classification and re-ID in task requirements, we can transform high-order metric learning into high-order feature learning by adjusting and improving the structure and loss function of the backbone network of the re-ID model. Thus, it is expected to provide high-order features with better discrimination ability for person re-ID task.

2.3 Loss function and learning strategy for re-ID

In addition to the network structure, feature learning still largely depends on loss function and training strategy. The focus includes the mathematical form of the loss function and the corresponding sampling strategy. On the mathematic form, many works proposed the loss function based on distance metrics, such as triplet loss [24], ranked list loss [25], etc. The principle of all loss function design is to better match the basic functional requirements. However, the above loss functions only take into account the distribution of the representations when designing a loss function. Finally, ensemble methods [26–28] have become an increasingly popular way of improving the performances of deep metric learning (DML) architectures. On the learning strategy, many works aim to mine samples that improve robustness, such as hard triplet mining [29], margin sample mining [30], etc. Since deep metric learning methods are sensitive to the samples of pairs, selecting suitable samples to train the model by mining strategy is shown to be effective. Furthermore, to learn high-order features through deep networks, it is necessary to select and design a suitable loss function and sampling strategy according to the characteristics of high-order metric learning.

The above works provide a theoretical basis on the backbone architecture design, high-order statistics computation, and metric learning strategy. In which, the works in Section 2.1 focus our research on the backbone architecture. And works in Section 2.2 prompt us to exploit higher-order statistics for re-ID task. The loss function and learning strategy in Section 2.3 make it possible to learn high-order person representations through high-order metric learning.

3 Proposed approach

In this section, we will first describe the pipeline of the proposed method in Section 3.1, then detail the proposed High-order Block module in Section 3.2, finally show the design of loss function for the overall framework in Section 3.3.

3.1 Network architecture

As described in Fig. 2, a person image patch is fed into a plain CNN to extract a deep feature map of size $h \times w \times d$, where h and w are the height and width of the feature map respectively, and d is the number of feature channels. Following standard first-order DML practices, in the training stage, these features are

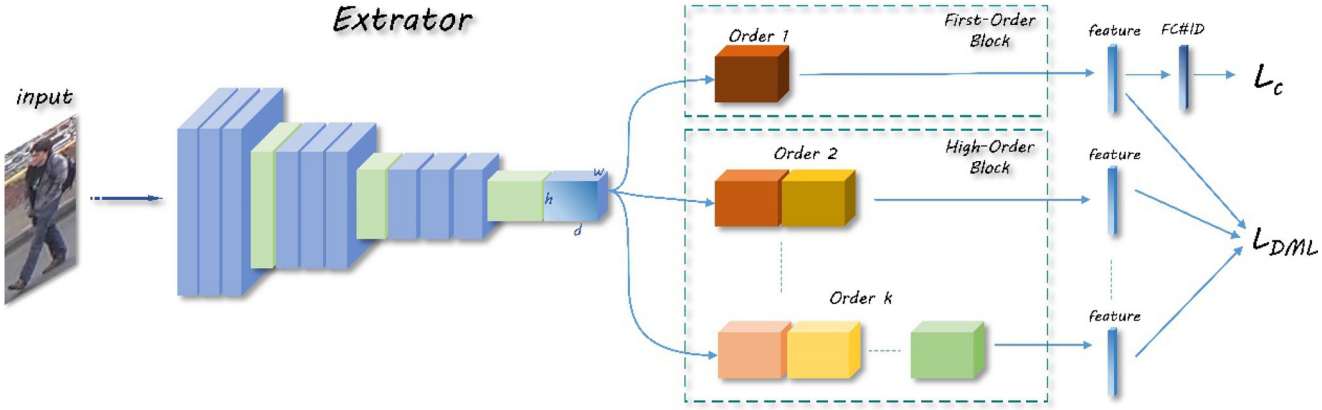

Fig. 2 Illustration on the architecture of the proposed HOB-net. The deep convolutional neural network extracts 3D($h \times w \times d$) tensor features from input person images. Through the CNN-based feature extractor, both the First-order block (on the top) and the High-order block (on the bottom) are introduced and combined to compute two loss functions: the classification loss L_C and the deep metric learning loss L_{DML}

aggregated using Global Average Pooling (GAP) to build a representation which is projected into an embedding space through the first-order block directly to optimize both the classification loss and the similarity loss. Within this pipeline, however, these representations without high-order metric learning are relatively coarse and are unable to capture the complex interactions among different parts, resulting in less discriminative in the deep features and scattering distribution in the deep embedding space. To this end, we dedicate to modeling with high-order statistics.

In our proposed HOB-net, we directly modulate the embedding feature space by minimizing the high-order distance between samples with the same person-ID while maximizing the high-order distance between samples with different IDs. By "high-order distance", we mean a distance-like metric that can be approximated by calculating some high-order moments. Combining all these high-order metrics together, a DML-based loss function can be computed and combined with the classification loss to guide the learning of all the features representations through the proposed feature network in an end-to-end fashion. During training, as shown in Fig. 2, the deep metric learning loss is applied on both the first-order branch and the high-order branch, and the classification loss is applied only on the first-order branch. During testing, the features from both branches are concatenated into a long plain vector to describe the testing image patch of the person.

3.2 High-order computation

Suppose that a person image I is passed by a plain CNN, and the corresponding 3D feature tensor of the output convolutional layer is denoted by $x \in R^{h\times w\times d}$. Most of the existing person re-ID models turn the feature tensor $\mathbf{x}$ into a vector $\varphi(\mathbf{x}) \in \Re^d$ using global average pooling and then transform it into a new vector $v \in \Re^l$ through a fully connected (FC) layer. Technically, it is expected that the representation vector v should focus on the local features which keep invariant within the same person while distinguishing between different persons. Going further along with this approach, we consider that the high-order relationships between local features could be a better representation than the first-order features. Therefore, we design a group of high-order mapping modules $f_k(\mathbf{x})$ to produce multiple feature vectors in different orders.

In practice, the dimension of the high-order features of an input image is extremely high and the corresponding calculation cost is too large to implement in real applications. Fortunately, for the person re-ID task, it only needs to calculate the similarity of the high-order feature vectors between a query image Q and the image S to be matched. If it can be proved mathematically that the inner product of the high-order feature vectors between Q and S can be approximated by the inner product of two relatively lower-dimensional vectors, we will only need to design a network to extract the lower-dimensional features as image representation. In fact, the RM algorithm [31] provides an effective solution to this problem, which relies on a set of random projectors to approximate the inner product between the kth-order moments of two basic feature vectors $\mathbf{x}, \mathbf{y} \in \Re^l$, which is denoted as $\langle \mathbf{x}, \mathbf{y}\rangle^k$.

$$\begin{aligned}\langle \mathbf{x}, \mathbf{y}\rangle^k &= \left\langle \underbrace{\mathbf{x}\otimes\cdots\otimes\mathbf{x}}_{k\ times}, \underbrace{\mathbf{y}\otimes\cdots\otimes\mathbf{y}}_{k\ times} \right\rangle \\ &= E_{w_1,\ldots,w_k\sim p_w}[\Phi_k(\mathbf{x})\Phi_k(\mathbf{y})] \end{aligned} \quad (1)$$

where $\otimes$ is the Kronecker product, $E_{w_1,\ldots,w_k\sim p_w}$ is the expectation over the random projectors $w_1, \ldots, w_k \in \Re^l$,

whose elements follow the uniform distribution p_w in the interval of $[-1, +1]$, and $\Phi_k(\mathbf{x}) \in \Re$ is defined as follows:

$$\Phi_k(\mathbf{x}) = \prod_i^k \langle w_i, \mathbf{x} \rangle \tag{2}$$

According to the works in [20], the kth-order inner product $\langle \mathbf{x}, \mathbf{y} \rangle^k$ can be further approximated to the inner product between two lower-dimensional vectors $\psi_k(\mathbf{x}), \psi_k(\mathbf{y}) \in \Re^s, s \ll l^k$

$$\langle \mathbf{x}, \mathbf{y} \rangle^k \approx \frac{1}{s} \langle \psi_k(\mathbf{x}), \psi_k(\mathbf{y}) \rangle \tag{3}$$

Using Tensor Tucker Decomposition [32], $\psi_k(\mathbf{x})$ can be written as:

$$\psi_k(\mathbf{x}) = (W_{k,1}^T \mathbf{x}) \odot (W_{k,2}^T \mathbf{x}) \odot \cdots \odot (W_{k,k}^T \mathbf{x}) \tag{4}$$

in which, $W_{k,i} \in \Re^{l \times s}, i = 1, 2, \ldots, k$ are a series of random matrices sampled independently, $\odot$ represents the Hadamard (element-wise) product. In this case, the feature x in (4) is exactly a 3D tensor in the size of $l = h \times w \times d$, so the calculation of $W_{k,i}^T x, i = 1, 2, \ldots, k$ can be evaluated by feeding the 3D tensor $x \in \Re^{h \times w \times d}$ into multiple individual convolutional layers respectively.

Actually, two typical architectures have been used in relative works to evaluate $\psi_k(\mathbf{x})$, as shown in Fig. 3b and c. The former applies a cascade architecture that evaluates the kth-order component based on the $k-1$ order gradually [20]:

$$\psi_k(\mathbf{x}) = \psi_{k-1}(\mathbf{x}) \odot (W_{k,k}^T \mathbf{x}) \tag{5}$$

Equation (5) can be regarded as a particular case of (4) under the hypothesis that $W_{p,k} = W_{q,k}, \forall q, p \geq k$. For the duplicate architecture given by Fig. 3c, it is supposed that [22]:

$$\psi_k(\mathbf{x}) = \underbrace{(W_{k,k}^T \mathbf{x}) \odot (W_{k,k}^T \mathbf{x}) \odot \cdots \odot (W_{k,k}^T \mathbf{x})}_{k \ times} \tag{6}$$

That means all the random matrices $W_{k,i}^T x, i = 1, 2, \ldots, k$ for the kth-order moment are supposed to be the same, namely $W_{k,k}^T x, \forall i \leq k$. The above two architectures can reduce the number of trainable parameters and save the computational cost. However, it also lowers the ability of the network to approximate the high-order moments. Considering the challenge of pedestrian Re-ID task, we still employ the original result of Tensor Tucker Decomposition [32] according to (4). In other words, each convolutional layer corresponding to the parameter matrix is learned independently in the fashion of end-to-end from the training data, and then the output tensors of all the convolutional layers at the same order are multiplied element-wise, as shown as Fig. 3a.

In comparison, suppose that the maximal order is K, both the cascade and duplicate architectures need K parameter matrices while our model uses $K(K+1)/2$ matrices. That means our model introduces more trainable weights to improve the performance of the network. Although more parameters will increase computation cost, to avoid over fitting in practical problems, the maximal order K is usually not large. Experimental results have revealed that the performance of the network starts to saturate at $K = 6$, which means that the ratio of the trainable parameters between our model and the other two models is about 15:6. The additional calculation cost is completely acceptable.

At last, the Hadamard product of all the convolutional layers' outputs in the kth-order block is transformed to a feature vector through the modules of GAP and a FC layer, as illustrated in Fig. 4. The final feature vector of the kth-order block is denoted as $\mathbf{f}_k(\mathbf{x}), k = 1, 2, \ldots, K$.

3.3 Loss functions

With the proposed architecture of the HOB-net, we compute a DML-based loss function on each of the high-order moments using the empirical estimator, such that similar

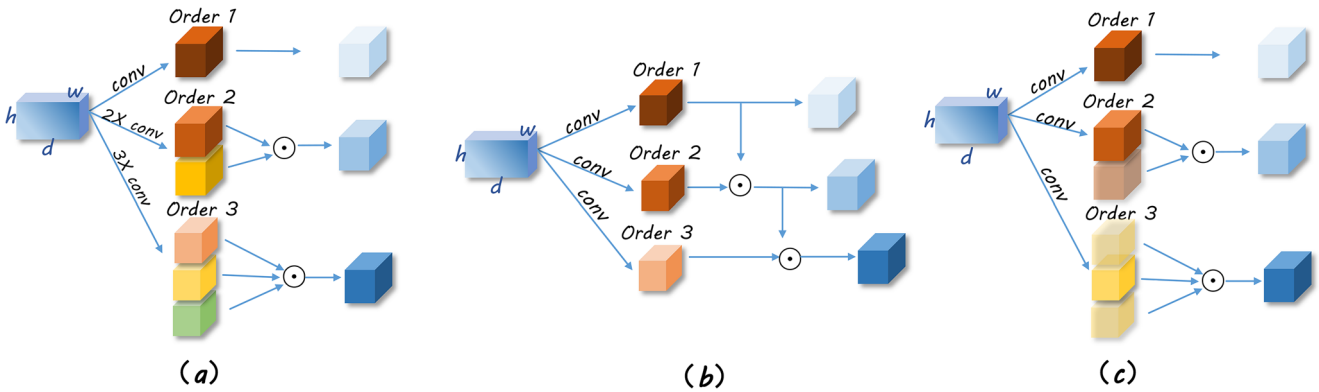

Fig. 3 Illustration of three architectures for higher-order (with maximal order $K = 3$) moment approximation. **a** our proposed architecture, **b** cascade architecture [20], **c** duplicate architecture [19]

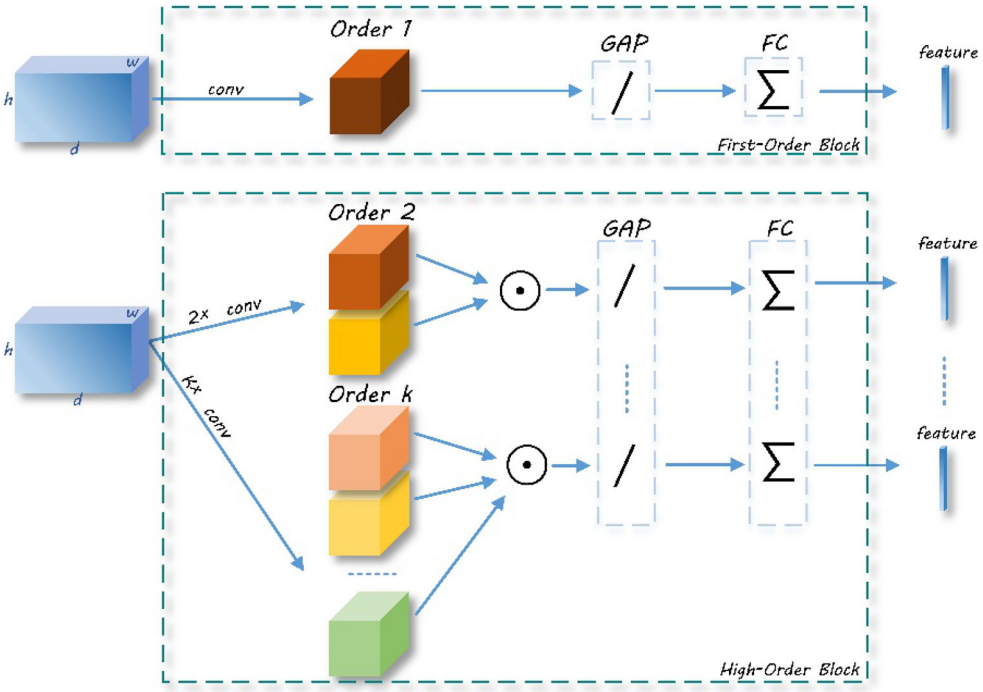

Fig. 4 Illustration of high-order block (HOB) modules

(respectively dissimilar) images have similar (respectively dissimilar) higher-order moments:

$$\begin{aligned} L_{DML}(I, J) &= \sum_{k=1}^{K} L_k \left(E_{x\sim I}[\mathbf{f}_k(\mathbf{x}), \mathbf{f}_k(\mathbf{y})]\right) \\ &= \sum_{k=1}^{K} L_k \left(\frac{1}{|I|}\sum_{x_i}\mathbf{f}_k(\mathbf{x}_i), \frac{1}{|J|}\sum_{x_j}\mathbf{f}_k(\mathbf{x}_j)\right) \end{aligned} \tag{7}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the sets of deep features exacted from person images I and J. As describe in Section 3.1, in our practical application for the classification and DML, we utilize the softmax loss and the batch-hard triplet loss proposed in [29] respectively. We randomly sample P identities and N instances for each identity in each mini-batch to meet the requirement of the batch-hard triplet loss. Typically, the loss function

$$\begin{aligned} L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{N} \Bigg(\alpha + \max_{\substack{p=1,\ldots,N;\\ p\neq a}} D\left(f(x_a^{(i)}), f(x_p^{(i)})\right) \\ - \min_{\substack{n=1,\ldots,N;\\ j=1,\ldots,P;\\ n\neq a, j\neq i}} D\left(f(x_a^{(i)}), f(x_n^{(j)})\right) \Bigg)_+ \end{aligned} \tag{8}$$

where $f(x_a^{(i)}), f(x_p^{(i)}), f(x_n^{(j)})$ are the features extracted from the anchor, positive, and negative samples respectively. $D(\mathbf{x}, \mathbf{y})$ computes the Euclidean distance between vectors $\mathbf{x}$ and $\mathbf{y}$, and α is the margin hyper parameter. In addition to batch-hard triplet loss, we employ softmax cross entropy loss for the discriminative learning of the model as well, which can be formulated as follows:

$$L_{softmax} = -\sum_{i=1}^{P}\sum_{a=1}^{N} log\frac{e^{\phi_{i,a,y_{a,i}}}}{\sum_{k=1}^{C} e^{\phi_{i,a,k}}} \tag{9}$$

where $y_{a,i}$ is the ground truth identity of the a-th sample of the i-th identity in the mini-batch, $\phi_{i,a,k}$ denotes the k-th element of the corresponding feature, and C is total number of person identities. Then the overall loss function for optimization is the combination of softmax loss and batch-hard triplet loss. Specifically, according to the following experimental analysis, we employ batch-hard triplet loss for all orders but softmax cross entropy loss only for the first order. Therefore, the overall loss function for the HOB-net is formulated as follows:

$$L_{HOB-net} = L_{softmax} + \beta \cdot L_{triplet} \tag{10}$$

where $\beta \in [0, 1]$ is the scale factor of $L_{triplet}$. By minimizing the overall loss function given by (10), the proposed HOB-net can be trained well to provide the multi-order features which are used to measure the similarity between the query images and the images to be matched.

4 Experiments

4.1 Datasets

Our approach is evaluated on three popular used re-ID datasets: Market-1501, DukeMTMC-ReID and CUHK03-NP.

Market-1501 is a large-scale person re-ID dataset collected from six cameras, which contains 32668 annotated images of 1501 identities. For evaluation, there are 12936 training images of 751 identities and 19732 testing images of 750 identities. Gallery and query sets have 19,732 and 3,368 images respectively with another 750 identities.

DukeMTMC-ReID is a subset of DukeMTMC, which is a multi-target, multi-camera pedestrian tracking dataset. It includes 36411 bounding boxes of 1404 identities. We divide the dataset into 16522 images of 702 identities for training and 17661 images of the other 702 identities for testing. The 2208 query images are picked from 17661 gallery images set.

CUHK03-NP is a new training-testing dataset, following the new protocol. It splits the CUHK03 into two subsets which contain labeled (by human) and detected (by a person detector) person images. The detected set includes 7365 training images from two cameras, 1,400 query images and 5,332 gallery images. The labeled set contains 7,368 training images, 1,400 query and 5,328 gallery images respectively. And according to the new protocol, their training and testing sets are split into 767 and 700 identities.

4.2 Experimental setting

Implementation details As described in Section 3.3, The proposed HOB network is constructed based on the architecture of ResNet50 [33]. We train both the baseline model and our model according to the strategy presented in [34]. Specifically, we keep the aspect ratio of all images and resize them to 288×144. Two data augmentation methods, random cropping and random horizontal flipping, are employed during training. To meet the requirement of hard-batch triplet loss, each mini-batch is sampled with randomly selected $P = 16$ identities and randomly sampled $N = 8$ images for each identity from the training set, so that the mini-batch size is 128. The Warmup learning rate is applied to bootstrap the network. The initial learning rate is set to 0.00035, which increases linearly to 0.0035 in 10 epochs and decreases by 0.1 at the 40th epoch and the 70th epoch respectively. We use the Adam solver to optimize the parameters for a total of 120 epochs. Our model is implemented on the Pytorch platform and trained with one NVIDIA 2080Ti GPU. All our experiments on different datasets follow the same settings as above.

Evaluation metrics In the test phase, to take advantage of the high-order moments, the final feature vectors are concatenated to form the final representation. Then, the metrics of cumulative matching characteristic (CMC) and mean average precision (mAP) are used to evaluate the overall performance of our model. For the sake of fairness, the re-ranking tricks are not adopted.

4.3 Comparison with the state-of-the-art methods

We present the superiority of our method by comparing our results with the state-of-the-arts in Tables 1 and 2, in which all methods have been divided into different types, including mask-guided [1, 35–37], stripe-based [4, 5, 11, 12], attention-based [6, 13–16], GAN-based [38], and Global feature-based [33, 39–46]. And for a fair comparison, we select the ResNet50-based backbone models, and the top two results are shown in red and blue respectively. From these two tables, it is noticeable that our HOB methods can significantly improve the performances over the baseline methods (e.g. comparing with baseline, HOB-6 achieve 4.5%/5.1% improvements of R-1/mAP on Market-1501 and 8.9%/11.4% improvements of R-1/mAP on DukeMTMC-ReID). Moreover, compared with other recent methods of different types, our method achieves state-of-the-art performance on both Market-1501 and DukeMTMC-ReID datasets. For CUHK03-NP dataset, our method achieves competitive results with the state-of-the-arts, particularly in the mAP. The above results prove the superiority of our HOB method.

4.4 Component analysis

Analysis on the maximal order In order to find the proper maximal order of the HOB module for person re-ID task, we conduct comparisons between HOB networks with the same backbone and loss functions but different maximal orders. The results on CUHK03, DukeMTMC-ReID, and Market-1501 datasets are shown in Fig. 5. From the results, it is noticeable that all the HOB modules in different orders can significantly improve the performances over the baseline methods. Concretely, comparing the HOB-2 with the baseline, there are 3.9%/6.2% on Market-1501, 7.1%/4.8% on DukeMTMC-ReID and 3.0%/4.4% on CUHK03-L improvements of R-1/mAP respectively. Moreover, it can be observed that the higher-order HOB modules can boost the capability of feature representations and achieve better performances. Specifically, e.g. on CUHK03-L, the scores of R-1/mAP enhance by 5.3%/2.1% from 64.1%/59.0 to 69.4%/66.5% when the order of HOB module increases from 2 to 6. The similar improvements can be found on the other two datasets. The experimental results show that employing higher-order HOB modules benefits

Table 1 Comparison of the proposed method with the art on Market-1501 and DukeMTMC-ReID

Type	Method	Reference	Market1501(%)		DukeMTMC(%)	
			R-1	mAP	R-1	mAP
Mask-guided	SPReID [35]	CVPR2018	92.5	81.3	84.4	71.0
	BDB [1]	ICCV2019	94.2	84.3	86.8	72.1
	MMP+LTL [36]	IEEE Access2020	93.4	80.5	83.4	67.2
Stripe-based	PABR [4]	ECCV2018	91.7	79.6	84.4	69.3
	PCB+RPP [5]	ECCV2018	93.8	81.6	83.3	69.2
	CAMA [11]	CVPR2019	94.7	84.5	85.8	72.9
Attention-based	HA-CNN [6]	CVPR2018	91.2	75.7	80.5	63.8
	Mancs [13]	ECCV2018	93.1	82.3	84.9	71.8
	CASN+PCB [14]	CVPR2019	94.4	82.8	87.7	73.7
	MHN-6(IDE) [16]	ICCV2019	93.6	83.6	87.5	75.2
	DUNet [15]	APIN2020	91.6	75.9	–	–
GAN-based	Camstyle [38]	TIP2018	89.5	71.5	78.3	57.6
	PN-GAN	ECCV2018	89.4	72.6	73.6	53.2
Global feature	IDE [33]		89.0	73.9	80.1	64.2
	MLFN [39]	CVPR2018	90.0	74.3	81.0	62.8
	DaRe [40]	CVPR2018	89.0	76.0	80.2	64.5
	IANet [41]	CVPR2019	94.4	83.1	87.1	73.4
	APR [42]	PR2019	87.0	66.9	73.9	55.6
	DHA-Net [43]	TIP2019	91.3	75.9	81.3	64.1
	expAT [44]	TIP2020	94.7	86.6	87.6	77.1
	IRN+ARN [45]	APIN2020	92.8	79.5	–	–
Ours	Baseline		90.2	79.2	79.5	65.8
	Ours		94.7	86.3	88.2	77.2

The best/second results are shown in red/blue respectively

to the ability of recognizing the person identities. Besides, the CMC and mAP results of HOB-net over all the three benchmarks demonstrate the effectiveness of our method. As a complementary account, when the further increase in the order of HOB module, e.g. 7 or 8, the performance improvements are few. Even with the higher order, the performance may start to decline. In this case, the number of network parameters increase sharply, which will increase the burden of model in training phase. Therefore, we select 6 as the proper maximal order of our proposed HOB module.

Analysis on the combined loss function Since loss function has a dramatic impact on the final performance of the re-ID models, comparisons between different combinations of loss functions are conducted to find out the best solution. More specifically, with the fixed maximal order, and Softmax and Triplet loss as the candidates of the loss functions, the corresponding CMC and mAP results on DukeMTMC-ReID and Market-1501 datasets have shown in Table 3, from which it can be seen clearly that the combination of "Softmax +Triplet" for the first-order and "Triplet" for the high-orders achieves the best performances. Specifically, the way of using only Softmax loss or Triplet loss reaches 90.6% or 93.0% of R-1 respectively on Market-1501 dataset. The combinations of both the Softmax loss and Triplet loss can significantly improve the performances, up to over 93.4% of R-1. Moreover, there is also a gap between different combinations. As shown in Table 3, the combination of Softmax loss with the first-order and Triplet loss with the high-order modules can achieve 94.3% of R-1 on Market-1501 dataset. And when assigning Triplet loss to the first-order, it is up to 94.7%. However, the score is down to 94.1% of R-1 when the high-order modules are allotted with Softmax loss. A similar degradation of performance can be also observed on the DukeMTMC-ReID dataset, which indicates the alliance between the

Table 2 Comparison of the proposed method with the art on CUHK03-NP. The best/second results are shown in red/blue respectively

Type	Method	Reference	CUHK03-NP(%)			
			Labeled		Detected	
			R-1	mAP	R-1	mAP
Mask-guided	MGCAM [37]	CVPR2018	50.1	50.2	46.7	46.9
	MMP+LTL+T [36]	IEEE Access2020	66.3	59.9	62.3	56.7
Stripe-based	PCB [5]	ECCV2018	–	–	61.3	54.2
	PCB+RPP [5]	ECCV2018	–	–	63.7	57.5
	CAMA [11]	CVPR2019	–	–	66.6	64.2
	M-DFFNet+SA [12]	MTA2020	62.4	66.7	62.7	56.0
Attention -based	HA-CNN [6]	CVPR2018	44.4	41.0	41.7	38.6
	Mancs [13]	ECCV2018	69.0	63.9	65.5	60.5
	MHN-6(IDE) [16]	ICCV2019	69.7	65.1	67.0	61.2
	CASN+PCB [14]	CVPR2019	73.7	68.0	71.5	64.4
	DUNet [15]	APIN2020	54.6	52.2	51.6	49.9
Global feature	IDE [33]		52.9	48.5	50.4	46.3
	MLFN [39]	CVPR2018	54.7	49.2	52.8	47.8
	DaRe [40]	CVPR2018	66.1	61.6	63.3	59.0
	NSL+SML [46]	IEEE Access2020	67.9	65.5	64.2	62.7
Ours	Baseline		61.1	59.0	60.6	58.9
	Ours		70.2	67.5	69.2	66.8

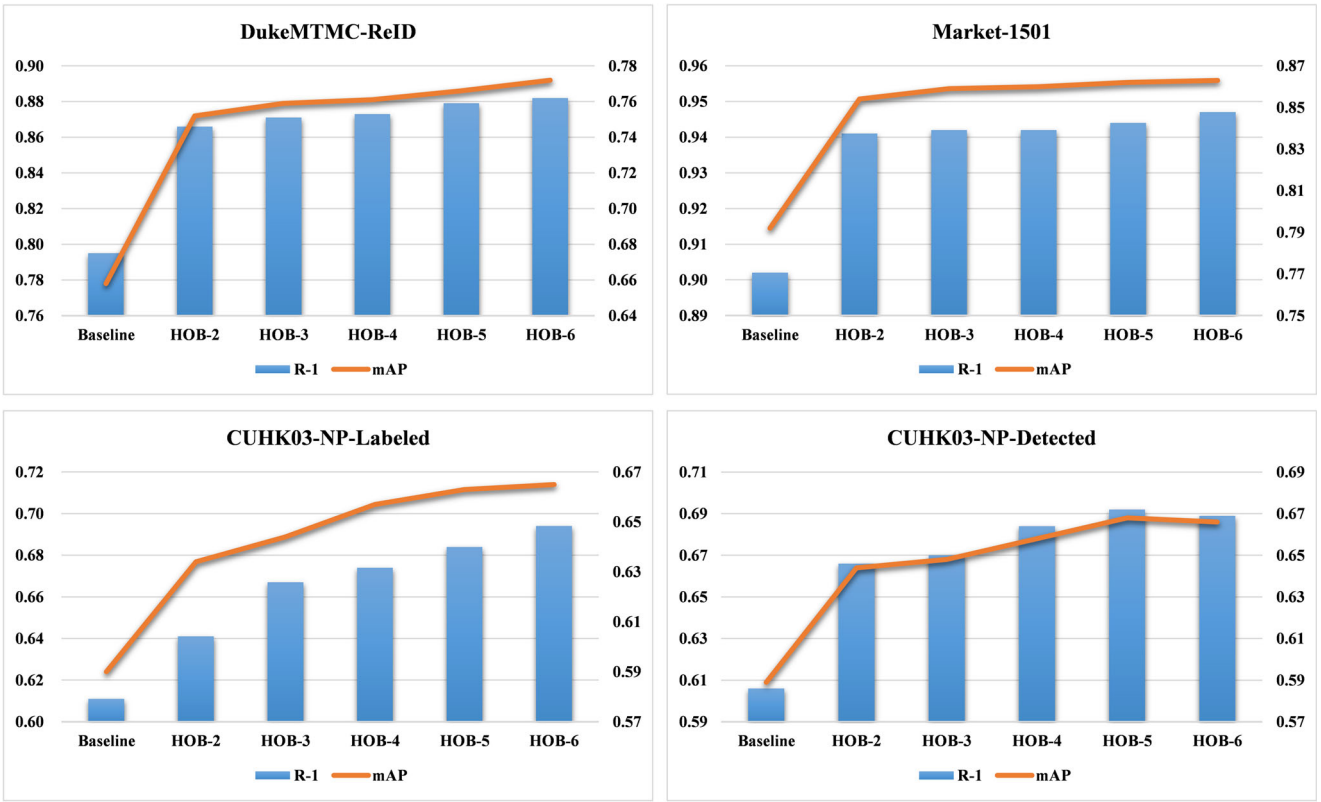

Fig. 5 Comparison of the effect of different number of order on Market-1501, DukeMTMC-ReID and CUHK03-NP

Table 3 Effect of the combination mode about loss function

Loss function		Market-1501		DukeMTMC-ReID	
First-order	Higher-order	R-1	mAP	R-1	mAP
Triplet	None	90.6	79.3	81.0	67.8
Softmax	None	93.0	82.0	85.0	70.0
Softmax	Triplet	94.3	85.7	87.7	76.0
Softmax	Softmax+Triplet	93.4	82.5	85.8	70.5
Softmax+Triplet	Triplet	94.7	86.3	88.2	77.2
Softmax+Triplet	Softmax	93.4	82.5	85.8	70.5
Softmax+Triplet	Softmax+Triplet	94.1	85.4	87.6	77.0

The best/second results are shown in red/blue respectively

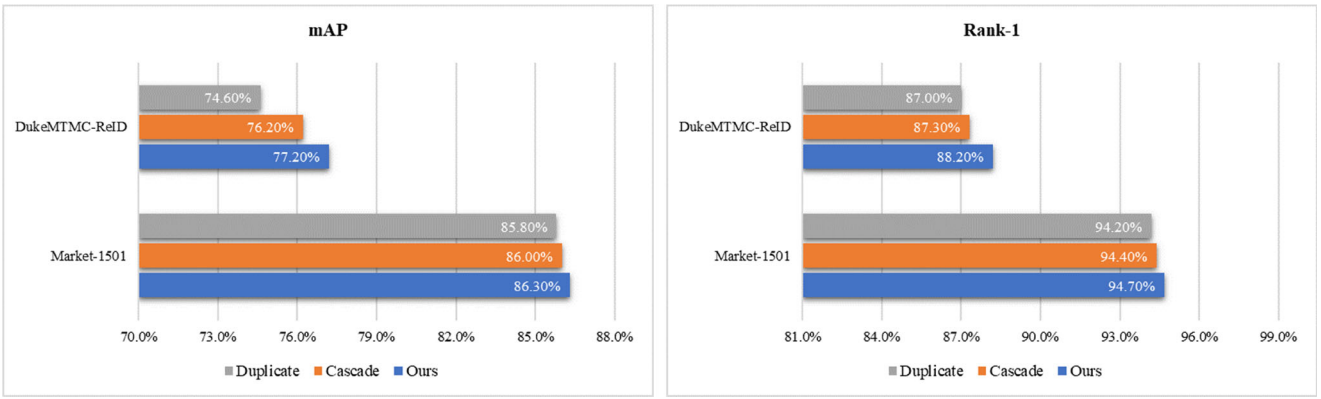

Fig. 6 Comparison of the different architecture of HOB module on Market-1501 and DukeMTMC-ReID

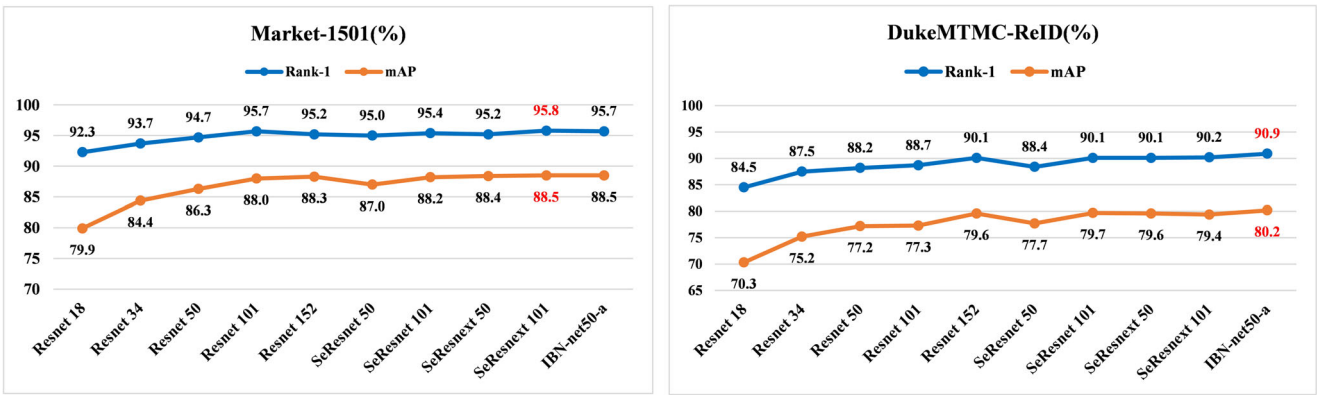

Fig. 7 The performance of different backbone network of our model on Market-1501 and DukeMTMC-ReID. The best results are shown in red

Fig. 8 Example results of six query images on Market-1501. Each row shows the top-10 retrieved images of Baseline and HOB-net. Person surrounded by green box denotes the same person as the query image, the red one is different

Fig. 9 Example results of six query images on DukeMTMC-ReID. Each row shows the top-10 retrieved images of Baseline and HOB-net.Person surrounded by green box denotes the same person as the query image, the red one is different

high-order modules and deep metric learning loss functions (e.g. triplet loss) benefits to the enhancement of the final performance.

Analysis on architecture of HOB module As illustrated in Fig. 3, there are three different architectures that can be used to construct the HOB module. Therefore, we carried out some experiments to validate the superiority of our proposal (shown in Fig. 3a). The CMC and mAP results on DukeMTMC-ReID and Market-1501 datasets are shown in Fig. 6, where our proposal achieves the best results, reaching 94.7%/86.3% and 88.2%/77.2% of R-1/mAP on the two datasets respectively. The cascade architecture comes second, reaching 94.4%/86.0% and 87.3%/76.2% of R-1/mAP. And the duplicate architecture obtains the result of 94.2%/85.8% and 87.0%/74.6% of R-1/mAP. Comparing with the other two methods, cascade architecture is the most concise architecture, which limits the number of trainable weights. On the contrary, our module provides the maximal number of trainable weights among the three ones due to the complicated architecture, which enhances the nonlinear expressive ability of the whole network model and completes more complex feature extraction for re-ID task. Consequently, our proposed HOB module achieves the best performance on both two datasets.

Analysis on backbone network and effectiveness of HOB In order to confirm the backbone network of our model, we tried many popular CNN-based feature networks by fixing the maximal order and the loss function. From the results shown in Fig. 7, it can be observed that the proposed HOB module is universally valid for different backbone networks. For the Market-1501 dataset, SeResnext 101 achieves the best, reaching 95.8%/88.5% of R-1/mAP. Also, for the DukeMTMC-ReID dataset, using IBN-net50 as the backbone shows the highest scores that 90.9%/88.2% of R-1/mAP, which is the best performance ever on DukeMTMC-ReID dataset even compared with other state-of-the-art results. Furthermore, several person re-ID examples on Market-1501 and DukeMTMC-ReID produced by Baseline and HOB-net are shown in Figs. 8 and 9, where six query images with different pedestrians and camera views are selected respectively. From the retrieval results, it can be observed that our proposed HOB-net performs more superior than the baseline model, which effectively ranks more true person at the top of the ranking list, even with large variations in the gallery including pose change, illumination, view angle change, and so on.

Model size and time/memory complexity We present the model size of the proposed HOB-net and time/memory complexity in the training phase in Table 4. From the results in the table, it can be observed that the number of parameters, training time, and maximum memory of our HOB-net increase with the order. While comparing with the baseline, the growth of training time and the maximum memory of each order HOB-net is reasonable. In terms of performance, our model has been greatly improved, showing that the HOB module is indeed flexible and efficient.

Table 4 Model size and time/memory complexity comparisons

Models	NP(million)	Time(minutes)	Memory(MB)	mAP(%)
baseline	24.95	82	6441	65.8
HOB-2	29.08	108	6703	75.2
HOB-4	37.14	134	7005	76.1
HOB-6	49.39	164	7639	77.2

NP means the number of parameters. Time and Memory denote training time and maximum memory respectively. And the results of training time, max memory and mAP are based on DukeMTMC-ReID

5 Conclusion

In this paper, we propose a flexible High-order Block (HOB) module which boosts the feature representations by introducing the high-order statistics and the corresponding high-order deep metric learning. And we study the influence of the different architectures of HOB and different maximal orders on the final performance. With taking the advantage of the HOB module, we propose the High-order Block Network (HOB-net) to improves the distinguishing ability of the deep features. Furthermore, we explore the training schemes of different combinations between HOB module and loss functions. As shown in the experimental results, the proposed HOB-net achieves very competitive performances on the three popular benchmarks, which validates the motivations and the conclusions that the proposed model can effectively improve the reliability and distinguishing ability of the features. In addition, the proposed HOB-net provides a strong baseline for the higher-order framework of re-ID feature network, and it can be also applied in real applications to enhance the recognition precision.

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grant U1613214 and in part supported by the Fundamental Research Funds for the Central Universities of China under Grant N170402008.

Data Availability All the experimental data in this paper are published data sets.

Code Availability The code is available on https://github.com/NothingToSay99/HOB-net.

Declarations

Ethics approval Compliance with ethical standards.

Conflict of Interests The authors declare that they have no conflict of interests.

Consent for Publication The authors declare that they consent to publication.

References

1. Dai Z, Chen M, Gu X, Zhu S, Tan P (2019) Batch DropBlock network for person re-identification and beyond. In: Proceedings of the IEEE international conference on computer vision, pp 3691–3701
2. Yu X, Ye X, Gao Q (2020) Infrared handprint image restoration algorithm based on apoptotic mechanism. IEEE Access 8:47334–47343
3. Khan MA, Javed K, Khan SA, Saba T, Habib U, Khan JA, Abbasi A (2020) Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimedia Tools Appl :1–27
4. Suh Y, Wang J, Tang S, Mei T, Mu Lee K (2018) Part-aligned bilinear representations for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 402–419
5. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV), pp 480–496
6. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
7. Tay CP, Roy S, Yap KH (2019) Aanet: Attribute attention network for person re-identifications. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7134–7143
8. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision, pp 17–35
9. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
10. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 152–159
11. Yang W, Huang H, Zhang Z, Chen X, Huang K, Zhang S (2019) Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1389–1398
12. Xiang S, Fu Y, Chen H, Ran W, Liu T (2020) Multi-level feature learning with attention for person re-identification. Multimed Tools Appl 79(43):32079–32093
13. Wang C, Zhang Q, Huang C, Liu W, Wang X (2018) Mancs: a multi-task attentional network with curriculum sampling for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 365–381
14. Zheng M, Karanam S, Wu Z, Radke RJ (2019) Re-identification with consistent attentive siamese networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5735–5744
15. Li R, Zhang B, Teng Z, Fan J (2020) A divide-and-unite deep network for person re-identification. Appl Intell :1–13
16. Chen B, Deng W, Hu J (2019) Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE International conference on computer vision, pp 371–381
17. Chen D, Chen P, Yu X, Cao M, Jia T (2019) Deeply-learned spatial alignment for person re-identification. IEEE Access 7:143684–143692
18. Qian X, Fu Y, Xiang T, Wang W, Qiu J, Wu Y, Xue X (2018) Pose-normalized image generation for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 650–667
19. Cai S, Zuo W, Zhang L (2017) Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: Proceedings of the IEEE international conference on computer vision, pp 511–520
20. Jacob P, Picard D, Histace A, Klein E (2019) Metric learning with horde: high-order regularizer for deep embeddings. In: Proceedings of the IEEE international conference on computer vision, pp 6539–6548
21. Gou M, Camps O, Sznaier M (2017) Mom: Mean of moments feature for person re-identification. In: Proceedings of the IEEE international conference on computer vision workshops. pp 1294–1303
22. Koniusz P, Yan F, Gosselin PH, Mikolajczyk K (2016) Higher-order occurrence pooling for bags-of-words: visual concept detection. IEEE Trans Pattern Anal Mach Intell 39(2):313–326
23. Li P, Xie J, Wang Q, Zuo W (2017) Is second-order information helpful for large-scale visual recognition? In: Proceedings of the IEEE international conference on computer vision, pp 2070–2078
24. Cheng D, Gong Y, Zhou S, Wang J, Zheng N (2016) Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1335–1344
25. Wang X, Hua Y, Kodirov E, Hu G, Garnier R, Robertson NM (2019) Ranked list loss for deep metric learning. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 5207–5216
26. Kim W, Goyal B, Chawla K, Lee J, Kwon K (2018) Attention-based ensemble for deep metric learning. In: Proceedings of the European conference on computer vision (ECCV), pp 736–751
27. Xuan H, Souvenir R, Pless R (2018) Deep randomized ensembles for metric learning. In: Proceedings of the European conference on computer vision (ECCV), pp 723–734
28. Zhong Z, Zheng L, Li S, Yang Y (2018) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European conference on computer vision (ECCV), pp 172–188
29. Hermans A, Beyer L, Leibe B (2017)
30. Xiao Q, Luo H, Zhang C (2017) Margin sample mining loss: a deep learning based method for person re-identification. arXiv:1710.00478
31. Kar P, Karnick H (2012) Random feature maps for dot product kernels. In: Artificial intelligence and statistics, pp 583–591
32. Kolda TG, Bader BW (2009) Tensor decompositions and applications. SIAM Rev 51(3):455–500
33. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: Past, present and future. arXiv:1610.02984
34. Luo H, Gu Y, Liao X, Lai S, Jiang W (2019) Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 0–0

35. Kalayeh MM, Basaran E, Gökmen M., Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1062–1071
36. Zheng F, Cai T, Wang Y, Deng C, Chen Z, Zhu H (2020) A mask-pooling model with local-level triplet loss for person re-identification. IEEE Access 8:138191–138202
37. Song C, Huang Y, Ouyang W, Wang L (2018) Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1179–1188
38. Zhong Z, Zheng L, Zheng Z, Li S, Yang Y (2018) Camstyle: a novel data augmentation method for person re-identification. IEEE Trans Image Process 28(3):1176–1190
39. Chang X, Hospedales TM, Xiang T (2018) Multi-level factorisation net for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2109–2118
40. Wang Y, Wang L, You Y, Zou X, Chen V, Li S, Weinberger KQ (2018) Resource aware person re-identification across multiple resolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8042–8051
41. Hou R, Ma B, Chang H, Gu X, Shan S, Chen X (2019) Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9317–9326
42. Lin Y, Zheng L, Zheng Z, Wu Y, Hu Z, Yan C, Yang Y (2019) Improving person re-identification by attribute and identity learning. Pattern Recogn 95:151–161
43. Wang Z, Jiang J, Wu Y, Ye M, Bai X, Satoh SI (2019) Learning sparse and identity-preserved hidden attributes for person re-identification. IEEE Trans Image Process 29(1):2013–2025
44. Ye H, Liu H, Meng F, Li X (2020) Bi-directional exponential angular triplet loss for rgb-infrared person re-identification. IEEE Trans Image Process (TIP) 2020:3045261
45. Yin J, Fan Z, Chen S, Wang Y (2020) In-depth exploration of attribute information for person re-identification. Appl Intell 50(11):3607–3622
46. Yang Z, Liu T, Liu J, Wang L, Zhao S (2020) A novel soft margin loss function for deep discriminative embedding learning. IEEE Access 8:202785–202794

Dongyue Chen received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Fudan University, Shanghai, China, in 2002 and 2007, respectively. From 2014 to 2015, he was an International Visiting Scholar with the Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. He is currently a Professor with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His current research interests include biologically motivated visual modeling, computer vision, and image processing.

Pengfei Wu received the bachelor's degree from the College of Information Science and Engineering, Northeastern University, Shenyang, China, in 2015. He is currently pursuing the master's degree with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His current research interests include computer vision and deep learning.

Tong Jia received the Ph.D. degree from the College of Information Science and Engineering, Northeastern University, Shenyang, China, in 2008, where he is currently a Professor. His current research interests include stereoscopic and omni-directional vision for mobile intelligent robot, computer vision, structured light system, and biomedical image processing and analysis.

Fangbin Xu received the bachelor's degree from the School of Control Engineering, Northeastern University, Shenyang, China, in 2019. He is currently pursuing the master's degree with the College of Information Science and Engineering, North- eastern University, Shenyang, China. His current research in terests include computer vision, deep learning and person reidentication.