Mixed High-Order Attention Network for Person Re-Identification

Binghui Chen, Weihong Deng,* Jiani Hu
Beijing University of Posts and Telecommunications
chenbinghui@bupt.edu.cn, whdeng@bupt.edu.cn, jnhu@bupt.edu.cn

Abstract

Attention has become more attractive in person re-identification (ReID) as it is capable of biasing the allocation of available resources towards the most informative parts of an input signal. However, state-of-the-art works concentrate only on coarse or first-order attention design, e.g. spatial and channels attention, while rarely exploring higher-order attention mechanism. We take a step towards addressing this problem. In this paper, we first propose the High-Order Attention (HOA) module to model and utilize the complex and high-order statistics information in attention mechanism, so as to capture the subtle differences among pedestrians and to produce the discriminative attention proposals. Then, rethinking person ReID as a zero-shot learning problem, we propose the Mixed High-Order Attention Network (MHN) to further enhance the discrimination and richness of attention knowledge in an explicit manner. Extensive experiments have been conducted to validate the superiority of our MHN for person ReID over a wide variety of state-of-the-art methods on three large-scale datasets, including Market-1501, DukeMTMC-ReID and CUHK03-NP. Code is available at http://www.bhchen.cn/.

1. Introduction

Since the quest for algorithms that enable cognitive abilities is an important part of machine learning, person re-identification (ReID) has become more attractive, where the model is requested to be capable of correctly matching images of pedestrians across videos captured from different cameras. This task has drawn increasing attention in many computer vision applications, such as surveillance [49], activity analysis [31, 32] and people tracking [55, 44]. It is also challenging because the images of pedestrians are captured from disjoint views, the lighting-conditions/person-poses differ across cameras, and occlusions are frequent in real-world scenarios.

Affected by the aforementioned factors, the discrimina-

dtion of feature representations of pedestrian images actually is not good enough. In order to obtain discriminative feature representations, many research works [30, 27, 25, 52, 21, 58, 45] resort to attention mechanism so as to highlight the informative parts (e.g. spatial locations) of convolutional responses and suppress the noisy patterns (e.g. background). Specifically, spatial attention [25, 27, 52] is a form of visual attention that involves directing attention to a location in space, it allows CNN to selectively process visual information of an area within the visual field. While, in spatial attention, the processing strategy of spatial masking is coarse and has no intrinsic effect on modulating the fine-grained channel-knowledge. Recently, channel attention [10, 19, 27] is proposed to adaptively recalibrates channel-wise convolutional responses by explicitly modelling interdependencies among channels. And the combination of spatial and channel attention has also been successfully applied in person ReID [27]. However, we emphasize that these commonly used attention methods (i.e. spatial and channel attention) are either coarse or first-order, being confined to mining only simple and coarse information, in person ReID cases, they are insufficiently rich to capture the

*Corresponding author
complex/high-order interactions of visual parts and the subtle differences among pedestrians caused by various viewpoints/person poses, as a result, the produced attention maps are neither discriminative or detailed. To this end, we dedicate to modeling the attention mechanism via high-order statistics of convolutional activations so as to capture more complex and high-order relationships among parts and to produce powerful attention proposals.

Moreover, we rethink the problem of person ReID as a zero-shot learning (ZSL) task where there is no intersection of pedestrian identities between training and testing sets. Zero-shot learning has large gap with conventional full-shot learning (e.g. classification on CIFAR [8, 7], Imagenet [38]), and in zero-shot settings, the phenomenon of ‘partial/biased learning behavior of deep model’ [5] largely affects the embedding performance, i.e. the deep model will only focus on the biased visual knowledge that only benefit to the seen identities and ignore the other helpful ones that might be useful for identifying the unseen identities. In other words, deep models easily learn to focus on surface statistical regularities rather than more general abstract concepts. However, many ReID works ignore this intrinsic problem of zero-shot learning. To this end, proposing detail-preserving attention framework remains important.

In this paper, we first propose High-Order Attention (HOA) module, a novel and powerful attention mechanism, to model the complex and high-order relationships among visual parts via high-order polynomial predictor, such that the subtle differences among pedestrian can be captured and discriminative attention results can be produced. Then, rethinking person ReID as a zero-shot problem, we propose Mixed High-Order Attention Network (MHN) to prevent the problem of ‘biased learning behavior of deep model’ [5] and to encourage the richness of attention information. It is mainly achieved by employing multiple HOA modules with different orders to model diverse high-order statistics, such that all-sided attention knowledge can be preserved and thus the unseen pedestrian identity can be successfully recognized. Additionally, we introduce the adversarial learning constraint for MHN to further prevent the order collapse problem during training \(^1\), so as to explicitly enhance the discrimination of MHN. Our contributions can be summarized as follows:

- The High-Order Attention (HOA) module is proposed to capture and use high-order attention distributions. To our knowledge, it is the first work to propose and apply high-order attention module in Person-ReID.
- We rethink ReID as zero-shot learning task and propose the Mixed High-Order Attention Network (MHN) to efficiently utilize multiple HOA modules, so as to enhance the richness of attention by explicitly suppressing the ‘biased learning behavior of deep model’. And adversary learning constraint is introduced to further prevent the problem of order collapse.

- MHN is a generally applicable and model-agnostic framework, it can be easily applied in the popular baseline architectures, such as IDE [63] and PCB [43].
- Extensive experiments demonstrate the superiority of the proposed MHN over a wide range of state-of-the-art ReID models on three large benchmarks, i.e. Market-1501 [61], DukeMTMC-ReID [37, 65] and CUHK03-NP [26, 66].

2. Related work

**Person ReID & Attention Mechanism:** Person ReID intends to correctly match images of pedestrians across videos captured from different cameras, it has been widely studied, such as ranking by pairwise constraints [34, 48], metric learning [54, 51], deep embedding learning [63, 43], re-ranking [62, 16] and attributes learning [40, 60]. Recently, attention methods [10, 53, 19, 46] in deep community are more attractive, in this paper, we focus on improving the performance of ReID via attention strategy.

Attention serves as a tool to bias the allocation of available resources towards the most informative parts of an input. Li et al. [24] propose a part-aligning CNN network for locating latent regions (i.e. hard attention) and then extract and exploit these regional features for ReID. Zhao et at. [59] employ the Spatial Transformer Network [20] as the hard attention model for finding discriminative image parts. Except hard attention methods, soft attention strategies are also proposed to enhance the performance of ReID. For example, Li et al. [25] use multiple spatial attention modules (by softmax function) to extract features at different spatial locations. Xu et al. [52] propose to mask the convolutional maps via pose-guided attention module. Li et al. [27] employ both the softmax-based spatial attention module and channel-wise attention module [19] to enhance the convolutional response maps. However, spatial attention and channel attention are coarse and first-order respectively, and are not capable of modeling the complex and high-order relationships among parts, resulting in loss of fine-grained information. Thus, to capture detailed and complex information, we propose High-Order Attention (HOA) module.

**High-order statistics:** It has been widely studied in traditional machine learning due to its powerful representation ability. And recently, the progresses of challenging fine-grained visual categorization task demonstrates integration of high-order pooling representations with deep CNNs can bring promising improvements. For example, Lin et al. [29] proposed bilinear pooling to aggregate the pairwise feature interactions. Gao et al. [15] proposed to approximate the

---

\(^1\) Although, the proposed high-order attention module has ability to capture complex and high-order statistics, but suffering from ‘biased learning behavior of deep model’, in zero-shot settings, the high-order module might collapse to lower-order module.
can be reformulated as:

\[
\mathcal{Y} = \mathcal{A}(\mathcal{X}) \odot \mathcal{X}
\]  

(1)

where \(\mathcal{A}(\mathcal{X}) \in \mathbb{R}^{C \times H \times W}\) is the attention proposal output by a certain attention module, \(\odot\) is the Hadamard Product (element-wise product). As \(\mathcal{A}(\mathcal{X})\) serves as a reweighting term, the value of each element of \(\mathcal{A}(\mathcal{X})\) should be in the interval \([0, 1]\). Based on the above general formulation of attention, \(\mathcal{A}(\mathcal{X})\) can take many different forms. For example, if \(\mathcal{A}(\mathcal{X}) = \text{rep}[M]^{C}\) where \(M \in \mathbb{R}^{H \times W}\) is a spatial mask and \(\text{rep}[M]^{C}\) means replicate this spatial mask \(M\) along channel dimension by \(C\) times, Eq. 1 thus is the implementation of spatial attention. And if \(\mathcal{A}(\mathcal{X}) = \text{rep}[V]^{H,W}\) where \(V \in \mathbb{R}^{C}\) is a scale vector and \(\text{rep}[V]^{H,W}\) means replicate this scale vector along height and width dimensions by \(H\) and \(W\) times resp, Eq. 1 thus is the implementation of channel attention.

However, in spatial attention or channel attention, \(\mathcal{A}(\mathcal{X})\) is coarse and unable to capture the high-order and complex interactions among parts, resulting in less discriminative attention proposals and failing in capturing the subtle differences among pedestrians. To this end, we dedicate to modeling \(\mathcal{A}(\mathcal{X})\) with high-order statistics.

### 3.2. High-Order Attention Module

To model the complex and high-order interactions within attention, we first define a linear polynomial predictor on top of the high-order statistics of \(x\), where \(x \in \mathbb{R}^{C}\) denotes a local descriptor at a specific spatial location of \(\mathcal{X}\):

\[
a(x) = \sum_{r=1}^{R} \left\langle w^r, \odot, x \right\rangle
\]

(2)

where \(\left\langle \cdot, \cdot \right\rangle\) indicates the inner product of two same-sized tensors, \(R\) is the number of order, \(\odot, x\) is the \(r\)-th order outer-product of \(x\) that comprises all the degree-\(r\) monomials in \(x\), and \(w^r\) is the \(r\)-th order tensor to be learned that contains the weights of degree-\(r\) variable combinations in \(x\).

Considering \(w^r\) with large \(r\) will introduce excessive parameters and incur the problem of overfitting, we suppose that when \(r > 1\), \(w^r\) can be approximated by \(D^r\) rank-1 tensors by Tensor Decomposition [23], i.e. \(w^r = \sum_{d=1}^{D^r} \alpha^{r,d} u^{r,d}_1 \odot \cdots \odot u^{r,d}_r\) when \(r > 1\), where \(u^{r,d}_1, \ldots, u^{r,d}_r \in \mathbb{R}^{C}\) are vectors, \(\odot\) is the outer-product, \(\alpha^{r,d}\) is the weight for \(d\)-th rank-1 tensor. Then according to the tensor algebra, Eq. 2 can be reformulated as:

\[
a(x) = \langle w^1, x \rangle + \sum_{r=2}^{R} \sum_{d=1}^{D^r} \alpha^{r,d} u^{r,d}_1 \odot \cdots \odot u^{r,d}_r, \odot, x \rangle
\]

\[
= \langle w^1, x \rangle + \sum_{r=2}^{R} \sum_{d=1}^{D^r} \alpha^{r,d} \prod_{s=1}^{r} (u^{r,d}_s, x)\rangle
\]

\[
= \langle w^1, x \rangle + \sum_{r=2}^{R} \langle \alpha^r, z^r \rangle
\]

(3)
where $\alpha' = [\alpha_r^1, \cdots, \alpha_r^{D'}]^T$ is the weight vector, $z' = [z_r^1, \cdots, z_r^{D'}]^T$ with $z_r^{d} = \prod_{r=1}^R (u_{r,d}^r, x)$. For later convenience, Eq. 3 can also be written as:

$$a(x) = 1^T (w^1 \odot x) + \sum_{r=2}^R 1^T (\alpha_r \odot z_r')$$

where $\odot$ is Hadamard Product and $1^T$ is a row vector of ones. Then, to obtain a vector-like predictor $a(x) \in \mathbb{R}^C$, Eq. 4 is generalized by introducing the auxiliary matrices $P^r$:

$$a(x) = P^{1T} (w^1 \odot x) + \sum_{r=2}^R P^{rT} (\alpha_r \odot z_r')$$

where $P^1, w^1, \alpha'$ are parameters to be learned, for implementation convenience, we can integrate $\{P^1, w^1\}$ into a new single matrix $\tilde{w}^1 \in \mathbb{R}^{C \times C}$ according to matrix algebra, and $\{P^r, \alpha_r\}$ into $\tilde{\alpha}_r \in \mathbb{R}^{D \times C}$ (simple proof is in Supplementary file). Then Eq. 5 can be expressed as:

$$a(x) = \tilde{w}^{1T} x + \sum_{r=2}^R \tilde{\alpha}_r^T z_r'$$

The above equation contains two terms, for clarity, we intend to formulate it into a more general case. Suppose $\tilde{w}^1$ can be approximated by the multiplication of two matrices $\tilde{v} \in \mathbb{R}^{C \times D}$ and $\tilde{\alpha} \in \mathbb{R}^{D \times C}$, i.e. $\tilde{w}^1 = \tilde{v} \tilde{\alpha}^1$. then Eq. 6 can be reformulated as:

$$a(x) = \tilde{\alpha}^1 T (\tilde{v}^T x) + \sum_{r=2}^R \tilde{\alpha}_r^T z_r = \sum_{r=1}^R \tilde{\alpha}_r^T z_r'$$

where $z' = \tilde{v}^T x$, and when $r > 1$, $z_r'$ is the same as in Eq. 3. $\tilde{\alpha}_r \in \mathbb{R}^{D \times C}$ are the trainable parameters.

In Eq.7, $a(x)$ is capable of modeling and using the high-order statistics of the local descriptor $x$, thus, we can obtain the high-order vector-like attention ‘map’ by performing Sigmoid function on Eq. 7:

$$A(x) = \text{sigmoid}(a(x)) = \text{sigmoid}(\sum_{r=1}^R \tilde{\alpha}_r^T z_r')$$

where $A(x) \in \mathbb{R}^C$ and the value of each element in $A(x)$ is in the interval $[0, 1]$.

Nonlinearity: Moreover, in order to further improve the representation capacity of this high-order attention ‘map’, inspired by the common design of CNN, we provide a variation of Eq.8 by introducing nonlinearity as follows:

$$A(x) = \text{sigmoid}(\sum_{r=1}^R \tilde{\alpha}_r^T \sigma(z_r'))$$

where $\sigma$ denotes an arbitrary non-linear activation function, here, we use ReLU [33] function. $A(x)$ in Eq.9 is finally employed as the required high-order attention ‘map’ for the corresponding local descriptor $x$. The experimental comparisons between Eq.8 and Eq.9 are in Sec. 4.

Full module: As aforementioned, $A(x)$ is defined on a local descriptor $x$, to obtain $A(\mathcal{X})$ which is defined on 3D tensor $\mathcal{X}$, we generalize Eq.9. Specifically, we share the learnable weights in $A(x)$ among different spatial locations of $\mathcal{X}$ and let $A(\mathcal{X}) = \{A(x_{(1,1)}), \cdots, A(x_{(H,W)})\}$, where $x_{(h,w)}$ indicates a local descriptor at spatial location point $(h, w)$ of $\mathcal{X}$. Employing this attention map $A(\mathcal{X})$ in CNN has two benefits. (1) Sharing weights among different spatial locations will not incur excessive parameters. (2) $A(\mathcal{X})$ can be easily implemented by 1x1 convolution layers. After obtaining the high-order attention map $A(\mathcal{X})$, our High-Order Attention (HOA) module can be formulated in the same way as Eq. 1, i.e. $\mathcal{Y} = A(\mathcal{X}) \odot \mathcal{X}$.

Implementation: Since the learnable parameters are shared among spatial locations, all operations in $A(\mathcal{X})$ can be implemented by Convolution. As illustrated in Fig. 2.(a), when $R = 1$, matrices $\{\tilde{v}, \tilde{\alpha}^1\}$ are implemented by 1x1 convolution layers with $D$ and $C$ output channels, resp. When $R > 1, r > 1$, we first employ $\{u_{r,d}^r\}_{d=1,\cdots, D'}$ as a set of $D'$ 1x1 convolutional filters on $\mathcal{X}$ so as to produce a set of feature maps $Z_s^r$ with channels $D'$, then feature maps $\{Z_s^r\}_{s=1,\cdots, r}$ are combined by element-wise product to obtain $Z^r = Z_1^r \odot \cdots \odot Z_r^r$, where $Z^r = \{z_r^r\}$, and $\tilde{\alpha}_r'$ can also be implemented by 1x1 convolution layer. A toy example of HOA when $R = 3$ is illustrated in Fig.2.(b).

Remark: The proposed HOA module can be easily implemented by the commonly used operations, such as 1x1 convolution and element-wise product/addition. Equipped by the powerful high-order predictor, the attention proposals could be more discriminative and is capable of modeling the complex and high-order relationships among parts. Moreover, the channel attention module in [19, 27] is called to be first-order because (1) GAP layer only collects first-order statistics while neglecting richer higher-order ones, suffering from limited representation ability [11] (2) fully-connected layers can be regarded as 1x1 convolution layers and thus the two cascaded fully-connected layers used in
channel-attention [19] are equivalent to our HOA module when $R = 1$ (regardless of the spatial dimensions and see Fig. 2(a)). In summary, the full channel attention module can only collect and utilize the first-order information, being insufficiently rich to capture the complex interactions and to produce the discriminative attention maps. And if without using GAP, the channel-attention module can be regarded as a special case of our HOA with $R = 1$, further demonstrating it indeed is first-order.

### 3.3. Mixed High-Order Attention Network

Considering that Person ReID essentially pertains to zero-shot learning (ZSL), where there is no intersection between training identities and testing identities, we should explicitly suppress the problem of ‘biased learning behavior of deep model’ caused by zero-shot settings [5]. Specifically, in ZSL, the deep model easily learn to focus on surface statistical regularities rather than more general abstract concepts, in other words, deep model will selectively learn the biased knowledge that are only useful to distinguish the seen identities, while ignore the knowledge that might be useful for unseen ones. Therefore, to correctly recognize the unseen identities, we propose Mixed High-Order Attention Network (MHN) to utilize multiple HOA modules with different orders such that the diverse and complementary high-order information can be explicitly used, encouraging the richness of the learned features and preventing the learning of partial/biased visual information.

For a toy example as shown in Fig. 3, the proposed MHN is constituted by several different HOA modules such that the diverse statistics of visual knowledge could be modeled and used. In particular, ResNet50 is first decomposed into two parts, i.e. $P_1$ (from conv1 to layer3\(^2\)) and $P_2$ (from layer3 to GAP). $P_1$ is used to encode the given image from raw pixel space to mid-level feature space. $P_2$ is used to encode the attentional information to the high-level feature space where data can be classified. HOA modules with different orders (e.g. \{ $R = 1, 2, 3$ \}) are placed between $P_1$ and $P_2$ so as to produce the diverse high-order attention maps and intensify the richness within learned knowledge. Worthy of mention is that our MHN won’t introduce excessive parameters since $P_2$ modules share the same weights across different attention streams.

However, simply employing multiple HOA modules with different orders won’t lead the best performance of MHN, since one HOA module with higher order might collapse to a relatively lower order module due to ‘the partial/biased learning behavior of deep model’. Specifically, from Eq. 7, one can observe that for a $k$-th order HOA module, $a(x)$ also contains the $l$-th order sub-term (where $l < k$). In theory, HOA module with $R = k$ can capture and use the $k$-th order statistics of local descriptor $x$, but in actual, especially in zero-shot learning settings, due to the fact that the deep model will selectively learn surface statistical regularities that are the easiest ones to distinguish the seen classes [5], the $k$-th order attention module might collapse to a lower-order counterpart as lower-order statistics are common and are much easier to collect than higher-order statistics. Therefore, these HOA modules with different $R$s actually collapse to some similar lower-order counterparts, and the wanted diverse higher-order attention information are not captured. To this end, inspired by GAN [17], we introduce the adversary constraint for regularizing the order of HOA to be different, as shown in Fig. 3, it can be formulated as:

$$
\max_{HOA_{R=1}^{k}} \min_{F} \left( \sum_{l=1}^{k} \| (F(f_{l}) - F(f_{l}')) \|^{2} \right)
$$

where $HOA_{R=1}^{k}$ means there are $k$ HOA modules (from first-order to $k$-th order) in MHN, $F$ indicates the encoding function parameterized by the adversary network which contains two fully-connected layers, $f_{l}$ is the feature representation vector learned from the corresponding HOA module with $R = l$. In Eq. 10, the adversary network $F$ tries to minimize the discrepancies among features $f_{l}$ while HOA modules try to maximize these discrepancies. After obtaining the Nash Equilibrium, the orders of HOA modules will be different with each other, since during the optimization of Eq.10, $P_2$ shares across streams and the only differentiating parts in MHN are HOA modules, when maximizing the feature discrepancies, the only solution is to make the HOA modules have different orders and produce diverse attention knowledge. In other words, only diverse HOA modules will make $L_{adv}$ large. Thus the problem of order collapse can be suppressed.

Then, the overall objective function of MHN is as:

$$
\min(L_{ide}) + \lambda \left( \max_{HOA_{R=1}^{k}} \min_{F} L_{adv} \right)
$$

where $L_{ide}$ indicates the identity loss based on Softmax classifier, $\lambda$ is the coefficient.

**Remark**: From Eq.11, one can observe that we regularize the order/diversity of HOA modules by imposing constraint on the encoded feature vectors, instead of directly on

\(^2\)Named in pytorch [36] manner.
the high-order attention maps, since these attention maps come from the complex high-order statistics and the definition of the order difference of HOA modules in the attention space is too hard to be artificially made. Thus, the order constraint is imposed on the feature vectors. Moreover, it seems that using Hinge loss based constraint instead of the adversary strategy to maximize the feature discrepancies is also feasible. However, we want to emphasize that in Hinge loss based function there is another margin-controller ‘m’ which needs extra tuning, and the discrepancies between features that coming from different HOA modules will be heterogeneous, thus to determine the optimal margin ‘m’, many redundant experiments must be executed. To this end, we employ the adversary constraint so as to allow the automatic learning of the optimal discrepancies.

By preventing the problem of order collapse, the HOA modules are explicitly regularized to model the wanted high-order attention distributions and thus can produce the discriminative and diverse attention maps which could be benefit for recognizing the unseen identities.

4. Experiments

Datasets: We use three popular benchmark datasets based on zero-shot learning (ZSL) settings, i.e. Market-1501 [61], DukeMTMC-ReID [37, 65] and CUHK03-NP [26, 66]. Market-1501 have 12,936 training images with 751 different identities. Gallery and query sets have 19,732 and 3,368 images respectively with another 750 identities. DukeMTMC-ReID includes 16,522 training images of 702 identities, 2,228 query and 17,661 gallery images of another 702 identities. CUHK03-NP is a new training-testing split protocol for CUHK03, it contains two subsets which provide labeled and detected (from a person detector) person images. The detected CUHK03 set includes 7,365 training images, 1,400 query images and 5,328 gallery images. The labeled set contains 7,368 training, 1,400 query and 5,328 gallery images respectively. The new protocol splits the training and testing sets into 767 and 700 identities.

Implementation: The proposed MHN is applied on both ResNet50-based IDE [63] and PCB [43] architectures. For both architectures, we adopt the SGD optimizer with a momentum factor of 0.9, set the start learning rate to be 0.01 for backbone CNN and ten times learning rate for the new added layers, and a total of 70 epochs with the learning rate decreased by a factor of 10 each 20 epochs. The dimension of feature $f_i$ is 256 and the two FC layers in $F$ have 128, 128 neurons respectively, we set all $D_i^{(l)} | r=1$ to be 64. For IDE, the images are resized to 288x144. For PCB, the images are resized to 336x168. We set the batch size to 32 in all experiments and use one 1080Ti GPU. MHN is implemented by Pytorch [36] and modified from the public code [1], random erasing [67] is also applied. Notation: We use ‘MHN-k’ to denote that in MHN there are k HOA modules with orders
\( R = \{1, \cdots, k\} \) resp, and ‘MHN-\( k \) (IDE/PCB)’ to denote using IDE/PCB architectures, resp.

**Evaluation:** In testing, the feature representations \( f_j, j \in \{1, \cdots, k\} \) are concatenated after L2 normalization. Then, the metrics of cumulative matching characteristic (CMC) and mean Average Precision (mAP) are used for evaluation. **No re-ranking tricks are adopted.**

### 4.1. Comparison with State-of-the-Art Methods

In order to highlight the significance of the proposed MHN for person ReID task, we compare it with some recent remarkable works, including methods of alignment [39, 41, 64, 43], deep supervision [50], architectures [63, 43], attention [27, 64, 47] and others [42, 9, 3], over the popular used benchmarks Market-1501, DukeMTMC-ReID and CUHK03-NP. For fair comparison, we re-implement the baseline models, i.e. ResNet50-based IDE and PCB, with the same training configurations as ours. MHN is then applied over both IDE and PCB architectures. The comparison results are listed in Tab. 1, Tab. 2 and Tab. 3. From these tables, one can observe that by explicitly intensify the discrimination and diversity within the deep embedding via high-order attention modules, our MHN-6 can significantly improve the performances over both the baseline methods IDE and PCB (e.g. comparing with PCB, MHN-6 (PCB) has 2\%/6.4\% improvements of R-1/mAP on Market and 5.2\%/7.5\% improvements of R-1/mAP on DukeMTMC), demonstrating the effectiveness of our high-order attention idea. And our MHN-6 (PCB) achieves the new SOTA performances on all these three benchmarks, showing the superiority of our method.

### 4.2. Component Analysis

**Effect of MHN:** We conduct quantitative comparisons on MHN as in Tab. 4. From this table, one can observe that the proposed MHN can significantly improve the performances of person ReID task over both IDE and PCB baseline architectures. Specifically, comparing MHN-2(IDE/PCB) with IDE/PCB, we can see that using higher-order attention information indeed encourage the discrimination of the learned embedding. Moreover, the performances will further increase with the number of HOA modules, e.g. on CUHK03-NP Labeled dataset, applying MHN on PCB, when increasing the number of HOA modules from 2 to 6 the performance of R-1 will be increased from 71.2\% to 77.2\%, the same phenomenon can be observed in other datasets and architecture. This phenomenon also shows that employing multiple HOA modules is benefit for modeling diverse and discriminative information for recognizing the unseen identities, and MHN-6 outperforms all the baseline models by a large margin over all the three benchmarks, demonstrating the effectiveness of our method. However, when further increase the number of HOA modules, e.g. \( k = 8 \), the performance improvements are few, thus we don’t report it here.

**Effect of Adversary Constraint:** From Tab. 5, when comparing \{MHN-6 (IDE) w/o \( L_{\text{adv}} \)\} with \{IDE\} and comparing \{MHN-6 (PCB) w/o \( L_{\text{adv}} \)\} with \{PCB\}, one can observe that on both DukeMTMC and Market datasets the performances of R-1 and mAP can be improved by simply employing multiple HOA modules without any regularizing constraint, showing that using higher-order attention information will indeed increase the discrimination of the learned knowledge in ZSL settings. However, as mentioned in Sec. 3.3, the task of person ReID pertains to zero-shot settings, the problem of ‘partial/biased learning behavior...
of deep model will incur the problem of order collapse of HOA modules, i.e. the deep model will partially model the easy and lower-order information regardless the theoretical capacity of HOA module. Therefore, we introduce the adversary constraint to explicitly prevent the problem of order collapse. After equipping with \( L_{adv} \), MHN-6 (IDE/PCB) can further improve the performances over both the benchmarks, demonstrating the effectiveness of \( L_{adv} \) and implying that explicitly learning diverse high-order attention information is essential for recognizing the unseen identities.

**Effect of Nonlinearity:** The nonlinearity comparisons are listed in Tab. 6, from this table, one can observe that by adding nonlinearity into the high-order attention modules, the performances can be further improved.

**Comparison to other attention methods:** To demonstrate the effectiveness of our idea of high-order attention, we compare with some other attention methods as in Tab. 7. Specifically, our MHN-6 (IDE) outperforms both the spatial and channel attention methods, i.e. HA-CNN [27] and SENet50 [19], showing the superiority of high-order attention model to these coarse/first-order attention methods. Moreover, although {SpaAtt+Q} [25] employs multiple diverse attention modules like MHN to enhance the richness of attention information, the used attention method is spatial attention which is coarse and insufficiently rich to capture the complex and high-order interactions of parts, failing in producing more discriminative attention proposals and thus performing worse than MHN-6 (IDE). {CASN+IDE} [64] regularizes the attention maps of the paired images belonging to the same identity to be similar and indeed improves the results, but it still performs worse than MHN-6 (IDE) since the consistence constraint for attention maps is only based on the the coarse spatial attention maps.

In summary, because of the ability of modeling and using complex and high-order information, the proposed MHN can significantly surpass all the listed coarse/first-order attention methods as shown in Tab. 7.

**Ablation study on the configurations of \( P_1 \) & \( P_2 \):** As mentioned in Sec. 3.3, the HOA modules are placed between \( P_1 \) and \( P_2 \), to investigate the effect of the placed position of HOA modules, we conduct experiments as in Tab. 8. One can observe that placing HOA modules after

<table>
<thead>
<tr>
<th>Methods</th>
<th>DukeMTMC-ReID R-1 mAP</th>
<th>Market-1501 R-1 mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDE [63]</td>
<td>80.1 64.2</td>
<td>89.0 73.9</td>
</tr>
<tr>
<td>SENet50 [19]</td>
<td>81.2 64.8</td>
<td>90.0 75.6</td>
</tr>
<tr>
<td>HA-CNN [27]</td>
<td>80.5 63.8</td>
<td>91.2 75.7</td>
</tr>
<tr>
<td>SpaAtt+Q [25]</td>
<td>84.7 69.6</td>
<td>91.6 77.4</td>
</tr>
<tr>
<td>CASN+IDE [64]</td>
<td>84.5 67.0</td>
<td>92.0 78.0</td>
</tr>
<tr>
<td>MHN-6 (IDE)</td>
<td>87.5 75.2</td>
<td>93.6 83.6</td>
</tr>
</tbody>
</table>

Table 7. Comparison to other attention methods (%). * indicates our reproducing.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Market-1501 R-1 mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_1 = {\text{conv1} \sim \text{layer1}}, P_2 = {\text{layer2} \sim \text{GAP}} ) &amp; 92.2 81.8</td>
<td></td>
</tr>
<tr>
<td>( P_1 = {\text{conv1} \sim \text{layer2}}, P_2 = {\text{layer3} \sim \text{GAP}} ) &amp; 93.6 83.6</td>
<td></td>
</tr>
<tr>
<td>( P_1 = {\text{conv1} \sim \text{layer3}}, P_2 = {\text{layer4} \sim \text{GAP}} ) &amp; 92.7 82.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Ablation study on the configurations of \( P_1 \) and \( P_2 \). All the layer names are shown in Pytorch manner. Here, for convenience we conduct experiments with MHN-6 (IDE) and test three configurations, i.e. \( 1, 2 \) and \( 3 \).

<table>
<thead>
<tr>
<th>Models</th>
<th>PN (million)</th>
<th>Depth</th>
<th>R-1 (on Market)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDE [63]</td>
<td>24.2</td>
<td>50</td>
<td>89.0%</td>
</tr>
<tr>
<td>SENet50 [19]</td>
<td>27.4</td>
<td>50</td>
<td>90.0%</td>
</tr>
<tr>
<td>MHN-2 (IDE)</td>
<td>24.4</td>
<td>50</td>
<td>90.6%</td>
</tr>
<tr>
<td>MHN-4 (IDE)</td>
<td>25.2</td>
<td>50</td>
<td>91.8%</td>
</tr>
<tr>
<td>MHN-6 (IDE)</td>
<td>26.8</td>
<td>50</td>
<td>93.6%</td>
</tr>
</tbody>
</table>

Table 9. Model size comparisons. PN means Parameter Number.

[layer2] (i.e. using the configuration 2) performs the best since when placing it at the relatively lower layer (i.e. using the configuration 1) the knowledge input to HOA module is more relevant to the low-level texture information and contains much noise, while placing it at relatively higher layer (i.e. using the configuration 3), some useful knowledge for recognizing the unseen identities might be already lost during the forward propagation of information as a result of partial/biased learning behavior. To this end, we use the configuration 2 for both IDE and PCB architectures throughout the experiments.

**Model size:** We compare the model size as in Tab. 9, from this table, one can observe that the parameter number of our MHN increases with the order. While comparing with SENet50 [19], the total parameter number of each MHN is not so much, and in terms of the performance, each MHN can outperform SENet50, showing that our MHN is indeed ‘light and sweet’.

5. **Conclusion**

In this paper, we first propose the High-Order Attention (HOA) module so as to increase the discrimination of attention proposals by modeling and using the complex and high-order statistics of parts. Then, considering the fact that the person-ReID task pertains to zero-shot learning where the deep model will easily learn the biased knowledge, we propose the Mixed High-Order Attention Network (MHN) to utilize the HOA modules at different orders, preventing the learning of partial/biased visual information that only benefit to the seen identities. The adversary constraint is further introduced to prevent the problem of order collapse of HOA module. And Extensive experiments have been conducted over three popular benchmarks to validate the necessity and effectiveness of our method.

**Acknowledgments:** This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61871052, 61573068, 61471048, and BUPT Excellent Ph.D. Students Foundation CX2019307.
References


Computer Vision and Pattern Recognition, pages 152–159, 2014. 2, 6, 7


[53] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua


