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.

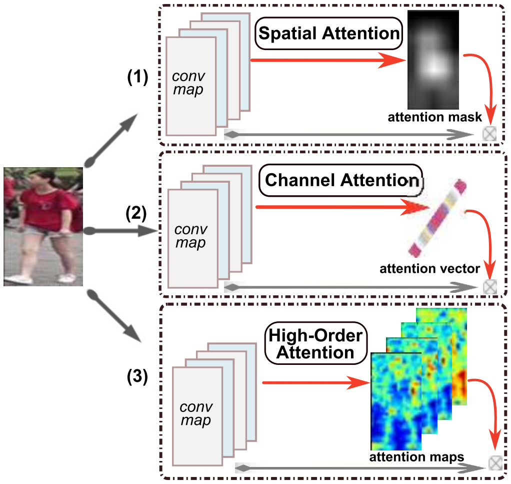

Figure 1. Attention comparison. (1) *Spatial attention* uses softmax-like gated functions to produce a spatial mask. (2) *Channel attention* [19] uses global average pooling and fully connected layers to produce a scale vector. (3) Our *high-order attention* uses high-order polynomial predictor to produce scale maps that contain high-order statistics of convolutional activations.

Affected by the aforementioned factors, the discrimination 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 knowledges 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 at. [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.

second-order statistics via Tensor Sketch [35]. Yin et al. [12] aggregated higher-order statistics by iteratively applying the Tensor Sketch compression to the features. Cai et al. [2] used high-order pooling to aggregate hierarchical convolutional responses. Moreover, the bilinear pooling and high-order pooling methods are also applied in Visual-Question-Answering task, such as [14, 22, 56, 57]. However, different from these above methods which mainly focus on using high-order statistics on top of feature pooling, resulting in high-dimensional feature representations that are not suitable for efficient/fast pedestrian search, we instead intend to enhance the feature discrimination by attention learning. We model high-order attention mechanism to capture the high-order and subtle differences among pedestrians, and to produce the discriminative attention proposals.

Zero-Shot Learning: In ZSL, the model is required to learn from the *seen* classes and then to be capable of utilizing the learned knowledge to distinguish the *unseen* classes. It has been studied in image classification [28, 4], video recognition [13] and image retrieval/clustering [5]. Interestingly, person ReID matches the setting of ZSL well where training identities have no intersection with testing identities, but most the existing ReID works ignore the problem of ZSL. To this end, we propose Mixed High-Order Attention Network (MHN) to explicitly depress the problem of 'biased learning behavior of deep model' [5, 6] caused by ZSL, allowing the learning of all-sided attention information which might be useful for unseen identities, preventing the learning of biased attention information that only benefits to the seen identities.

3. Proposed Approach

In this section, we will first provide the formulation of the general attention mechanism in Sec. 3.1, then detail the proposed *High-Order Attention* (HOA) module in Sec. 3.2, finally show the overall framework of our *Mixed High-Order Attention Network* (MHN) in Sec. 3.3.

3.1. Problem Formulation

Attention acts as a tool to bias the allocation of available resources towards the most informative parts of an input. In convolutional neural network (CNN), it is commonly used to reweight the convolutional response maps so as to highlight the important parts and suppress the uninformative ones, such as *spatial attention* [25, 27] and *channel attention* [19, 27]. We extend these two attention methods to a general case. Specifically, for a convolutional activation output, a 3D tensor $\mathcal{X}$, encoded by CNN and coming from the given input image. We have $\mathcal{X} \in \mathbb{R}^{C \times H \times W}$, where C, H, W indicate the number of channel, height and width, *resp*. As aforementioned, the goal of attention is to reweight the convolutional output, we thus formulate this process as:

$$\mathcal{Y} = \mathcal{A}(\mathcal{X}) \odot \mathcal{X} \tag{1}$$

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 $\mathcal{A}(\mathcal{X}) = rep[M]|^C$ where $M \in \mathbb{R}^{H \times W}$ is a spatial mask and $rep[M]|^C$ means replicate this spatial mask M along channel dimension by C times, Eq. 1 thus is the implementation of *spatial attention*. And if $\mathcal{A}(\mathcal{X}) = rep[V]|^{H,W}$ where $V \in \mathbb{R}^C$ is a scale vector and $rep[V]|^{H,W}$ 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 $\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^C$ denotes a local descriptor at a specific spatial location of $\mathcal{X}$:

$$a(\mathbf{x}) = \sum_{r=1}^{R} \langle \mathbf{w}^r, \otimes_r \mathbf{x} \rangle \tag{2}$$

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

Considering $\mathbf{w}^r$ with large r will introduce excessive parameters and incur the problem of overfitting, we suppose that when $r > 1$, $\mathbf{w}^r$ can be approximated by D^r rank-1 tensors by Tensor Decomposition [23], i.e. $\mathbf{w}^r = \sum_{d=1}^{D^r} \alpha^{r,d} \mathbf{u}_1^{r,d} \otimes \cdots \otimes \mathbf{u}_r^{r,d}$ when $r > 1$, where $\mathbf{u}_1^{r,d} \in \mathbb{R}^C, \ldots, \mathbf{u}_r^{r,d} \in \mathbb{R}^C$ are vectors, $\otimes$ 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:

$$\begin{aligned} a(\mathbf{x}) &= \langle \mathbf{w}^1, \mathbf{x} \rangle + \sum_{r=2}^{R} \langle \sum_{d=1}^{D^r} \alpha^{r,d} \mathbf{u}_1^{r,d} \otimes \cdots \otimes \mathbf{u}_r^{r,d}, \otimes_r \mathbf{x} \rangle \\ &= \langle \mathbf{w}^1, \mathbf{x} \rangle + \sum_{r=2}^{R} \sum_{d=1}^{D^r} \alpha^{r,d} \prod_{s=1}^{r} \langle \mathbf{u}_s^{r,d}, \mathbf{x} \rangle \\ &= \langle \mathbf{w}^1, \mathbf{x} \rangle + \sum_{r=2}^{R} \langle \boldsymbol{\alpha}^r, \mathbf{z}^r \rangle \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha}^r = [\alpha^{r,1}, \cdots, \alpha^{r,D^r}]^T$ is the weight vector, $\mathbf{z}^r = [z^{r,1}, \cdots, z^{r,D^r}]^T$ with $z^{r,d} = \prod_{s=1}^{r} \langle \mathbf{u}_s^{r,d}, \mathbf{x} \rangle$. For later convenience, Eq. 3 can also be written as:

$$a(\mathbf{x}) = \mathbf{1}^T(\mathbf{w}^1 \odot \mathbf{x}) + \sum_{r=2}^{R} \mathbf{1}^T(\boldsymbol{\alpha}^r \odot \mathbf{z}^r) \qquad (4)$$

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

$$\mathbf{a}(\mathbf{x}) = \mathbf{P}^{1^T}(\mathbf{w}^1 \odot \mathbf{x}) + \sum_{r=2}^{R} \mathbf{P}^{r^T}(\boldsymbol{\alpha}^r \odot \mathbf{z}^r) \qquad (5)$$

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

$$\mathbf{a}(\mathbf{x}) = \widehat{\mathbf{w}}^{1^T}\mathbf{x} + \sum_{r=2}^{R} \widehat{\boldsymbol{\alpha}}^{r^T}\mathbf{z}^r \qquad (6)$$

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

$$\mathbf{a}(\mathbf{x}) = \widehat{\boldsymbol{\alpha}}^{1^T}(\widehat{\mathbf{v}}^T\mathbf{x}) + \sum_{r=2}^{R} \widehat{\boldsymbol{\alpha}}^{r^T}\mathbf{z}^r = \sum_{r=1}^{R} \widehat{\boldsymbol{\alpha}}^{r^T}\mathbf{z}^r \qquad (7)$$

where $\mathbf{z}^1 = \widehat{\mathbf{v}}^T\mathbf{x}$, and when $r > 1$, $\mathbf{z}^r$ is the same as in Eq. 3. $\widehat{\boldsymbol{\alpha}}^r \in \mathbb{R}^{D^r \times C}$ are the trainable parameters.

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

$$A(\mathbf{x}) = sigmoid(\mathbf{a}(\mathbf{x})) = sigmoid(\sum_{r=1}^{R} \widehat{\boldsymbol{\alpha}}^{r^T}\mathbf{z}^r) \qquad (8)$$

where $A(\mathbf{x}) \in \mathbb{R}^C$ and the value of each element in $A(\mathbf{x})$ 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:

$$A(\mathbf{x}) = sigmoid(\sum_{r=1}^{R} \widehat{\boldsymbol{\alpha}}^{r^T}\sigma(\mathbf{z}^r)) \qquad (9)$$

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

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

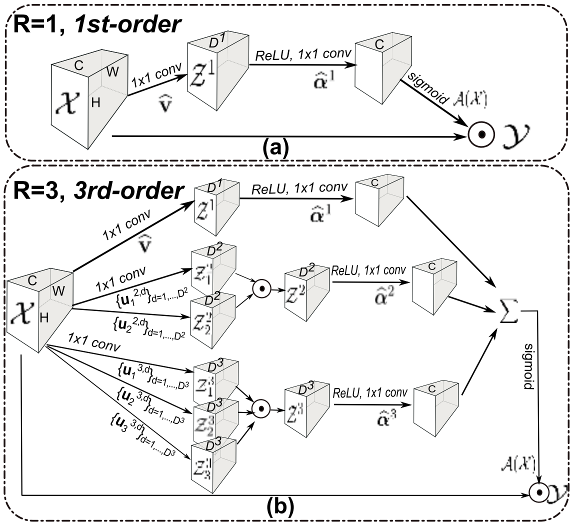

Figure 2. Illustration of ***High-Order Attention*** (HOA) modules.

Implementation: Since the learnable parameters are shared among spatial locations, all operations in $\mathcal{A}(\mathcal{X})$ can be implemented by Convolution. As illustrated in Fig. 2.(a), when $R = 1$, matrixes $\{\widehat{\mathbf{v}}, \widehat{\boldsymbol{\alpha}}^1\}$ are implemented by 1x1 convolution layers with D^1 and C output channels, *resp*. When $R > 1, r > 1$, we first employ $\{\mathbf{u}_s^{r,d}\}_{d=1,\cdots,D^r}$ as a set of D^r 1x1 convolutional filters on $\mathcal{X}$ so as to produce a set of feature maps $\mathcal{Z}_s^r$ with channels D^r, then feature maps $\{\mathcal{Z}_s^r\}_{s=1,\cdots,r}$ are combined by element-wise product to obtain $\mathcal{Z}^r = \mathcal{Z}_1^r \odot \cdots \odot \mathcal{Z}_r^r$, where $\mathcal{Z}^r = \{\mathbf{z}^r\}$, and $\widehat{\boldsymbol{\alpha}}^r$ 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

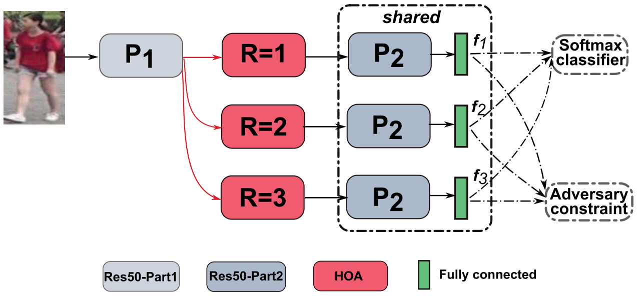

Figure 3. Illustration of ***Mixed High-Order Attention Network*** (MHN). Our MHN is model-agnostic, it can be applied in both IDE [63] and PCB [43] architectures, here for clarity, we take ResNet50 [18] based IDE for example. The adversary constraint is used to regularize the order of HOA modules.

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 *layer2*[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 the 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, $\mathbf{a}(\mathbf{x})$ 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 $\mathbf{x}$, 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 Rs 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:

$$\max_{HOA|_{R=1}^{R=k}} \min_{F}(L_{adv}) = \max_{HOA|_{R=1}^{R=k}} \min_{F}\Big(\sum_{j,j',j\neq j'}^{k} \|F(f_j) - F(f_{j'})\|_2^2\Big) \tag{10}$$

where $HOA|_{R=1}^{R=k}$ 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_j is the feature representation vector learned from the corresponding HOA module with $R = j$. In Eq. 10, the adversary network F tries to minimize the discrepancies among features f_j 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:

$$\min(L_{ide}) + \lambda(\max_{HOA|_{R=1}^{R=k}} \min_{F}(L_{adv})) \tag{11}$$

where L_{ide} indicates the identity loss based on Softmax classifier, λ 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 5,332 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_j is 256 and the two FC layers in F have 128, 128 neurons *resp*, we set all $D^r|_{r=1}^{R}$ 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

		Market-1501 (%)			
Methods	Ref	R-1	R-5	R-10	mAP
BoW+kissme [61]	ICCV15	44.4	63.9	72.2	20.8
SVDNet [42]	ICCV17	82.3	-	-	62.1
DaRe(De)+RE [50]	CVPR18	89.0	-	-	76.0
MLFN [3]	CVPR18	90.0	-	-	74.3
KPM [39]	CVPR18	90.1	96.7	97.9	75.3
HA-CNN [27]	CVPR18	91.2	-	-	75.7
DNN-CRF [9]	CVPR18	93.5	97.7	-	81.6
PABR [41]	ECCV18	91.7	96.9	98.1	79.6
PCB+RPP [43]	ECCV18	93.8	97.5	98.5	81.6
Mancs [47]	ECCV18	93.1	-	-	82.3
CASN+PCB [64]	CVPR19	94.4	-	-	82.8
IDE* [63]		89.0	95.7	97.3	73.9
MHN-6 (IDE)		*93.6*	*97.7*	*98.6*	*83.6*
PCB* [43]		93.1	97.5	98.5	78.6
MHN-6 (PCB)		***95.1***	***98.1***	***98.9***	***85.0***

Table 1. Results comparisons over Market-1501 [61] under Single-Query settings. * indicates the re-implementation by our code. The best/second results are shown in red/blue, *resp*.

		DukeMTMC-ReID (%)			
Methods	Ref	R-1	R-5	R-10	mAP
BoW+kissme [61]	ICCV15	25.1	-	-	12.2
SVDNet [42]	ICCV17	76.7	-	-	56.8
DaRe(De)+RE [50]	CVPR18	80.2	-	-	64.5
MLFN [3]	CVPR18	81.0	-	-	62.8
KPM [39]	CVPR18	80.3	89.5	91.9	63.2
HA-CNN [27]	CVPR18	80.5	-	-	63.8
DNN-CRF [9]	CVPR18	84.9	92.3	-	69.5
PABR [41]	ECCV18	84.4	92.2	93.8	69.3
PCB+RPP [43]	ECCV18	83.3	-	-	69.2
Mancs [47]	ECCV18	84.9	-	-	71.8
CASN+PCB [64]	CVPR19	87.7	-	-	73.7
IDE* [63]		80.1	90.7	93.5	64.2
MHN-6 (IDE)		*87.5*	*93.8*	*95.6*	*75.2*
PCB* [43]		83.9	91.8	94.4	69.7
MHN-6 (PCB)		***89.1***	***94.6***	***96.2***	***77.2***

Table 2. Results comparisons over DuckMTMC-ReID [37, 65]. * indicates the re-implementation by our code. The best/second results are shown in red/blue, *resp*.

		CUHK03-NP (%)			
Methods	Ref	Labeled		Detected	
		R-1	mAP	R-1	mAP
BoW+XQDA [61]	ICCV15	7.9	7.3	6.4	6.4
SVDNet [42]	ICCV17	-	-	41.5	37.3
DaRe(De)+RE [50]	CVPR18	66.1	61.6	63.3	59.0
MLFN [3]	CVPR18	54.7	49.2	52.8	47.8
HA-CNN [27]	CVPR18	44.4	41.0	41.7	38.6
PCB+RPP [43]	ECCV18	-	-	63.7	57.5
Mancs [47]	ECCV18	69.0	63.9	65.5	60.5
CASN+PCB [64]	CVPR19	73.7	68.0	71.5	64.4
IDE* [63]		52.9	48.5	50.4	46.3
MHN-6 (IDE)		*69.7*	*65.1*	*67.0*	*61.2*
PCB* [43]		61.9	56.8	60.6	54.4
MHN-6 (PCB)		***77.2***	***72.4***	***71.7***	***65.4***

Table 3. Results comparisons over CUHK03-NP [26, 66]. * indicates the re-implementation by our code. The best/second results are shown in red/blue, *resp*.

Methods	CUHK03-NP [26, 66]				DukeMTMC-ReID [37, 65]				Market-1501 [61]			
	Labeled		Detected									
	R-1	mAP	R-1	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
IDE* [63]	52.9	48.5	50.4	46.3	80.1	90.7	93.5	64.2	89.0	95.7	97.3	73.9
IDE*+era	61.4	55.71	56.9	51.3	83.6	92.1	94.3	67.4	90.3	96.5	97.6	75.9
MHN-2 (IDE)	65.9	59.1	60.9	54.8	84.5	92.6	94.7	68.9	90.6	96.1	97.6	76.1
MHN-4 (IDE)	67.4	60.3	62.7	55.8	86.3	93.1	95.6	72.4	91.8	97.6	98.5	80.1
MHN-6 (IDE)	**69.7**	**65.1**	**67.0**	**61.2**	**87.5**	**93.8**	**95.6**	**75.2**	**93.6**	**97.7**	**98.6**	**83.6**
PCB* [43]	61.9	56.8	60.6	54.4	83.9	91.8	94.4	69.7	93.1	97.5	98.5	78.6
PCB*+era	57.4	52.5	54.3	49.9	83.4	91.5	94.3	68.2	91.9	97.4	98.4	76.8
MHN-2 (PCB)	71.2	66.3	67.9	61.9	86.9	93.3	95.3	73.5	94.0	97.8	98.5	82.5
MHN-4 (PCB)	75.1	70.6	71.6	**66.1**	88.7	94.4	95.9	76.8	94.5	98.0	98.6	84.2
MHN-6 (PCB)	**77.2**	**72.4**	**71.7**	65.4	**89.1**	**94.6**	**96.2**	**77.2**	**95.1**	**98.1**	**98.9**	**85.0**

Table 4. Effect (%) of attention modules. * indicates the re-implementation and 'era' means random erasing.

$R = \{1, \cdots, k\}$ *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.

Methods	DukeMTMC-ReID		Market-1501	
	R-1	mAP	R-1	mAP
IDE* [63]	80.1	64.2	89.0	73.9
MHN-6 (IDE) w/o L_{adv}	85.5	70.8	91.8	80.0
MHN-6 (IDE)	**87.5**	**75.2**	**93.6**	**83.6**
PCB* [43]	83.9	69.7	93.1	78.6
MHN-6 (PCB) w/o L_{adv}	87.7	75.4	93.9	83.2
MHN-6 (PCB)	**89.1**	**77.1**	**95.1**	**85.0**

Table 5. Effect (%) of adversary constraint. * indicates the re-implementation by our code.

Methods	DukeMTMC-ReID		Market-1501	
	R-1	mAP	R-1	mAP
MHN-6 (IDE) w/o *nonli*	87.1	74.9	93.3	83.1
MHN-6 (IDE)	**87.5**	**75.2**	**93.6**	**83.6**
MHN-6 (PCB) w/o *nonli*	88.7	76.8	95.0	84.5
MHN-6 (PCB)	**89.1**	**77.1**	**95.1**	**85.0**

Table 6. Effect (%) of nonlinearity.

Effect of Adversary Constraint: From Tab. 5, when comparing {MHN-6 (IDE) w/o L_{adv}} with {IDE} and comparing {MHN-6 (PCB) w/o L_{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

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

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

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 [3] [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.

Methods	Market-1501	
	R-1	mAP
①:$P_1=\{conv1\sim layer1\}$,$P_2=\{layer2\sim GAP\}$	92.2	81.8
②:$P_1=\{conv1\sim layer2\}$,$P_2=\{layer3\sim GAP\}$	**93.6**	**83.6**
③:$P_1=\{conv1\sim layer3\}$,$P_2=\{layer4\sim GAP\}$	92.7	82.1

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.

Models	PN (million)	Depth	R-1 (on Market)
IDE [63]	24.2	50	89.0%
SENet50 [19]	27.4	50	90.0%
MHN-2 (IDE)	24.4	50	90.6%
MHN-4 (IDE)	25.2	50	91.8%
MHN-6 (IDE)	26.8	50	93.6%

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

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 '*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.

[3] We fine-tune the pre-trained SENet50 released at https://github.com/moskomule/senet.pytorch.

References

[1] https://github.com/layumi/Person_reID_baseline_pytorch. 6

[2] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 511–520, 2017. 3

[3] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018. 6, 7

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016. 3

[5] Binghui Chen and Weihong Deng. Energy confused adversarial metric learning for zero-shot image retrieval and clustering. In *AAAI Conference on Artificial Intelligence*, 2019. 2, 3, 5

[6] Binghui Chen and Weihong Deng. Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2750–2759, 2019. 3

[7] Binghui Chen, Weihong Deng, and Junping Du. Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[8] Binghui Chen, Weihong Deng, and Haifeng Shen. Virtual class enhanced discriminative embedding learning. In *Advances in Neural Information Processing Systems*, pages 1946–1956, 2018. 2

[9] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2018. 6, 7

[10] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5659–5667, 2017. 1, 2

[11] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3828–3836, 2015. 4

[12] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2930, 2017. 3

[13] Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1857–1860. ACM, 2013. 3

[14] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016. 3

[15] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016. 2

[16] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1305–1313, 2015. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1, 2, 3, 4, 5, 8

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2

[21] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018. 1

[22] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 3

[23] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 3

[24] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017. 2

[25] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018. 1, 2, 3, 8

[26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on*

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

[27] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018. 1, 2, 3, 4, 6, 7, 8

[28] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7463–7471, 2018. 3

[29] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 2

[30] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017. 1

[31] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE, 2009. 1

[32] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010. 1

[33] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 4

[34] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015. 2

[35] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013. 3

[36] Pytorch. https://pytorch.org/. 5, 6

[37] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 2, 6, 7

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[39] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6886–6895, 2018. 6, 7

[40] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European conference on computer vision*, pages 475–491. Springer, 2016. 2

[41] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018. 6, 7

[42] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3800–3808, 2017. 6, 7

[43] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 2, 5, 6, 7

[44] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017. 1

[45] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016. 1

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2

[47] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *The European Conference on Computer Vision (ECCV)*, September 2018. 6, 7

[48] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014. 2

[49] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013. 1

[50] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8042–8051, 2018. 6, 7

[51] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014. 2

[52] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018. 1, 2

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

Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2

[54] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014. 2

[55] Shoou-I Yu, Yi Yang, and Alexander Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3714–3720, 2013. 1

[56] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 3

[57] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, (99):1–13, 2018. 3

[58] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 1

[59] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017. 2

[60] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151, 2014. 2

[61] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 2, 6, 7

[62] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1741–1750, 2015. 2

[63] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2, 5, 6, 7, 8

[64] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019. 6, 7, 8

[65] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017. 2, 6, 7

[66] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 2, 6, 7

[67] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 6