Bilinear Attention Networks for Person Retrieval

Pengfei Fang[1,2], Jieming Zhou[1], Soumava Kumar Roy[1,2], Lars Petersson[1,2], Mehrtash Harandi[2,3]
[1]The Australian National University, [2]DATA61-CSIRO, Australia, [3]Monash University
{Pengfei.Fang, u5761794, Soumava.KumarRoy}@anu.edu.au
Lars.Petersson@data61.csiro.au, mehrtash.harandi@monash.edu

Abstract

This paper investigates a novel ***Bilinear attention*** *(Bi-attention) block, which discovers and uses second order statistical information in an input feature map, for the purpose of person retrieval. The Bi-attention block uses bilinear pooling to model the local pairwise feature interactions along each channel, while preserving the spatial structural information. We propose an* ***Attention in Attention*** *(AiA) mechanism to build inter-dependency among the second order local and global features with the intent to make better use of, or pay more attention to, such higher order statistical relationships. The proposed network, equipped with the proposed Bi-attention is referred to as* ***Bilinear ATtention network*** *(BAT-net). Our approach outperforms current state-of-the-art by a considerable margin across the standard benchmark datasets (*e.g.*, CUHK03, Market-1501, DukeMTMC-reID and MSMT17).*

1. Introduction

Person retrieval[1], also known as person re-identification (re-ID), has attracted an increasing amount of attention in the Computer Vision (CV) community due to its significant industrial potential as well as academic importance in terms of creating highly discriminative feature representations, with one of the earliest works being [6]. In short, the task of a person retrieval machine can be characterised as follows: given an image of a specific person, the machine should retrieve all images, from a gallery, that contain a person with the same identity (ID).

This is a challenging task, and one of the main issues causing an unreliable person retrieval system is that of *misalignment*. That is, the location of the person's body, and its parts, with respect to a reference frame, can easily change due to body shape, pose, clothing *etc*. This, in turn, causes feature mismatches during training and testing, leading to inaccurate re-identification. Much effort has gone into studying and addressing these issues [20, 36, 38, 32, 23, 41, 17, 22, 35]; however, it still remains a dominant problem and calls for further study.

Some attempts [34, 29] developed over the years to address this problem rely on human pose estimation. These estimator networks supplement the baseline-network with additional cues to learn a superior embedding space, thereby resulting in increased accuracy over the baseline-network. Other solutions benefit from person attributes [16, 35], which are invariant across pose, illumination, misalignment *etc*. However, person attribute learning also requires training a network on an additional person attribute dataset or labelling attributes within existing person re-ID datasets.

Recently, several solutions have been inspired by the human visual sensing process using *visual attention* mechanisms [20, 43, 33, 22], to focus on the discriminative regions within a person bounding box. The inherent attention module is designed to automatically select the meaningful parts of an image, and is trained in a weakly-supervised manner (*i.e.*, no explicit labelling information is given to identify the areas to attend). However, current attention models tend to only utilize first order information, such as the pattern itself in the feature map, ignoring *higher order statistical information* that may be hidden in the feature map.

Bilinear mappings and models have been widely adopted as a generalization of their linear counterparts. Some prime examples are bilinear classifiers [25], bilinear pooling [5] and bilinear CNNs [21] with applications in visual question answering, fine-grained image recognition, texture classification to name but a few. To the best of our knowledge, attention mechanisms equipped with bi-linear models have not been developed or studied before despite their intriguing properties.

The **contribution** of this paper can be summarized as follows: **(a)** We formulate a novel Bilinear attention (Bi-attention) block with an *Attention in Attention* (AiA) mechanism. The AiA mechanism can be understood as having an attention module inside another, with the *inner* one de-

[1]In the remainder of this paper, we will use the terms "person retrieval", "person re-identification" and "person re-ID" interchangeably.

termining where to focus for the *outer* attention module. As such, the Bi-attention with an AiA block utilizes second order statistical information and builds inter-dependency among second order local and global features, channel-wise, in a unified block, while preserving the spatial structure information of the input feature map. **(b)** We propose a novel deep architecture using the Bi-attention block, creating our Bilinear ATtention network (BAT-net), for the task of person retrieval. To the best of our knowledge, this is the first time a bilinear attention mechanism for representation learning has been developed. **(c)** Extensive experiments performed on the standard benchmark datasets including **CUHK03** [18], **Market-1501** [52], **DukeMTMC-reID** [26] and **MSMT17** [46], show that our approach outperforms the current state-of-the-art methods by a considerable margin.

2. Related Work

Person Re-identification. Early works in the area of person re-ID relied mostly on hand-crafted feature representations [6] and learning latent spaces [49]. We refer interested readers to [8] for more details regarding traditional methods. Convolutional Neural Networks (CNN) are currently the method of choice for representation learning, delivering state-of-the-art results in person re-ID. In [49], Yi *et al*. proposed a unified framework for feature and similarity learning using Siamese networks [27]. Multi-level similarities are employed in [45] to make more reliable decisions. Having robustness in mind, Xiao *et al*. trained a model across multiple datasets [48] and used domain guided dropouts to mute domain-irrelevant neurons. Structures, such as orthogonality constraints [37] and geometry constraints [1], have also shown to help achieving better, or more robust, decisions in person re-identification.

Attention Mechanism. Recently, attention mechanisms, inspired by the human sensing process, have been studied extensively in Natural Language Processing [42] and Computer Vision [20]. In person re-ID, the person misalignment [36] and background biases [40] hinder learning a robust representation. Visual attention mechanisms aim at emphasizing informative regions for identification, while depreciating harmful ones (*e.g*., background and occluded regions).

The spatial transformer network (STN) [13], a binary hard attention, was used in [17] to localize the latent body parts of a human. Liu *et al*. proposed a Comparative Attention Network (CAN), which repeatedly localizes discriminative parts and compares different local regions of person pairs [22]. In Harmonious Attention Convolutional Neural Network (HA-CNN) [20], hard region-level attention and soft pixel-level attention are learned in a unified attention block. In [43], Wang *et al*. considered both the channel-wise and spatial-wise attention in a Fully Attentional Block (FAB), where the channel information is re-calibrated and the spatial structure information is also preserved.

Bilinear Pooling. Bilinear pooling [5, 51], is first introduced to model local pairwise feature interactions for fine-grained recognition problems and its representation power is also enhanced by normalizing the higher order statistics [21, 15]. Thereafter, Liu *et al*. utilized a compact form of the bilinear operation to pool a high-dimensional feature representation for the task of person re-ID [23]. In [41], Ustinova *et al*. proposed a patch-based multi-regional bilinear pooling to account for the geometric misalignment problem between person bounding boxes. Recently, Suh *et al*. used a part-aligned representation to reduce the misalignment problem by fusing the appearance and part feature maps in a bilinear pooling layer [36].

3. Bilinear Attention

In this section, we will first detail the Bilinear attention block and employ it in a novel Attention in Attention mechanism. A simplified version of Bilinear attention is then introduced, reducing the learnable parameters by almost half.

3.1. Bilinear Attention with AiA

Bilinear attention (Bi-attention) with the Attention in Attention (AiA) mechanism, captures the second order statistical information along each channel of the feature map and builds inter-dependency among second order local and global features in a unified cell, while preserving the spatial structure of the input feature map. The architecture of Bi-attention with AiA is depicted in Fig. 1.

Let $\boldsymbol{X} \in \mathbb{R}^{c\times h\times w}$ be a feature map, where c, h and w stand for the number of channels, height and width, respectively. We denote the local feature at spatial location (i, j) with $\boldsymbol{x}_{ij} \in \mathbb{R}^c, i \in \{1, 2, \ldots, h\},\ j \in \{1, 2, \ldots, w\}$. The bilinear pooling of a vector $\varphi(\boldsymbol{x})$, an embedding of $\boldsymbol{x}$ (subscript is omitted for simplicity), is obtained as (see Fig. 2)

$$\begin{aligned}\boldsymbol{Y} &= \varphi(\boldsymbol{x})\varphi(\boldsymbol{x})^T = \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T \\ &= \begin{bmatrix} \bar{x}_1^2 & \ldots & \bar{x}_1\bar{x}_{\bar{c}} \\ \vdots & \ddots & \vdots \\ \bar{x}_{\bar{c}}\bar{x}_1 & \ldots & \bar{x}_{\bar{c}}^2 \end{bmatrix},\end{aligned} \tag{1}$$

where $\varphi(\boldsymbol{x}) \in \mathbb{R}^{\bar{c}}$, $\bar{c} = c/r$, and $\boldsymbol{Y} \in \mathbb{R}^{\bar{c}\times\bar{c}}$. The hyper-parameter r is a dimensionality reduction factor and its effect is discussed in § 5.4. Having efficiency in mind, and since $\boldsymbol{Y}$ is a symmetric matrix, we only consider its upper triangular elements in the subsequent processing. This helps reducing the feature dimensionality from $\bar{c}^2$ to $\bar{c}\cdot(\bar{c}+1)/2$ (see Fig. 2). Formally,

$$\tilde{\boldsymbol{x}} = \mathrm{Vec}\big(\mathrm{UTri}(\boldsymbol{Y})\big), \tag{2}$$

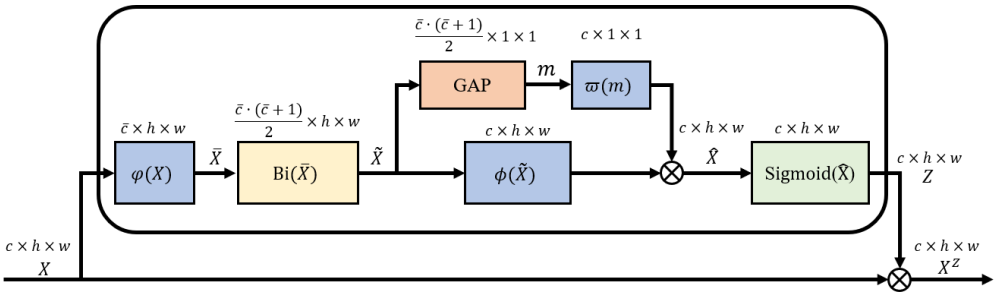

Figure 1. The structure of Bilinear attention with the Attention in Attention mechanism. $\varphi(\cdot)$, $\phi(\cdot)$ and $\varpi(\cdot)$ are embedding functions. $\mathrm{Bi}(\cdot)$ indicates the bilinear pooling and second order feature rearrangement function. GAP operates global average pooling. $\otimes$ indicates element-wise multiplication.

where $\mathrm{Vec}(\cdot)$ and $\mathrm{UTri}(\cdot)$ indicate vectorization and the operator that extracts the upper triangular elements of a matrix, respectively. We abstract the bilinear pooling and feature rearrangement with: $\mathrm{Bi}(\bar{\boldsymbol{x}}) = \mathrm{Vec}\big(\mathrm{UTri}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T)\big)$.

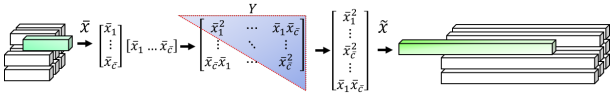

Figure 2. Processing of bilinear pooling and second order feature rearrangement, denoted by $\mathrm{Bi}(\cdot)$. In this operation, we sample the elements in the upper triangle of $\boldsymbol{Y}$ and vectorize those elements to a new feature vector $\tilde{\boldsymbol{x}}$.

We note that though $\tilde{\boldsymbol{x}}$ contains second order information of $\bar{\boldsymbol{x}}$, it is sensitive to spatial misalignment. To address this shortcoming, we introduce the concept of the Attention in Attention (AiA) mechanism (see Fig. 3). The idea is to adaptively re-weight the second order feature responses by modelling the inter-dependencies between the second order global and local features (see Fig. 1). We model the second order global feature by

$$\boldsymbol{m} = \frac{1}{hw}\sum_{i=1}^{hw}\tilde{\boldsymbol{x}}_i. \tag{3}$$

This formulation contains the second order *statistical* information (*i.e.*, the vectorized version of the empirical auto-correlation matrix of $\bar{\boldsymbol{X}}$) of the input of AiA.

The inter-dependency between embedded second order global feature $\boldsymbol{m}$ and each embedded second order local features $\tilde{\boldsymbol{x}}$ is:

$$\hat{\boldsymbol{x}} = \varpi(\boldsymbol{m}) \otimes \phi(\tilde{\boldsymbol{x}}), \tag{4}$$

where $\otimes$ denotes element-wise multiplication and $\varpi(\boldsymbol{m})$, $\phi(\tilde{\boldsymbol{x}}) \in \mathbb{R}^c$. The embedding functions, $\varpi(\boldsymbol{m})$ and $\phi(\tilde{\boldsymbol{x}})$, do not merely re-weight the second order feature responses, but also reduce the dimension of the second order feature from $\bar{c}\cdot(\bar{c}+1)/2$ to c (*i.e.*, the channel size of the input $\boldsymbol{x}$). In Fig. 3, we further detail the aforementioned steps. Intuitively, $\varpi(\boldsymbol{m})$ acts as an inner attention and local

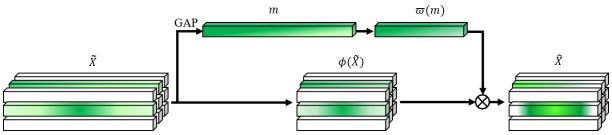

Figure 3. Insight of Attention in Attention mechanism. The *inner* attention module in AiA produces channel-wise attention values of its input feature maps, (*e.g.*, $\tilde{\boldsymbol{X}}$), conceptually weighting or calibrating them for future processing.

features $\phi(\tilde{\boldsymbol{x}})$ that are more correlated to the global feature $\varpi(\boldsymbol{m})$, are emphasized by Eq. (4).

Finally, the attention mask of input $\boldsymbol{x}$ is obtained by normalizing $\hat{\boldsymbol{x}}$. In this paper, we use $\mathrm{Sigmoid}(\cdot)$ as a normalization function (*i.e.*, $\boldsymbol{z} = \mathrm{Sigmoid}(\hat{\boldsymbol{x}})$). This normalized vector will act as a channel mask, and emphasize the significant elements of its input feature vector $\boldsymbol{x}$ at the same spatial position, by element-wise multiplication as:

$$\boldsymbol{x}^z = \boldsymbol{z} \otimes \boldsymbol{x}. \tag{5}$$

Remark 1 *The operations, described by Eq. (3) and (4), resemble the Squeeze-and-Excitation (SE) Networks [10]. However, there is an essential difference: The SE Network first squeezes the information in each channel to a scalar which is then used to scale all the elements of a channel uniformly. In contrast, we use channel attention as the inner attention module to weight the significance of attention-dependent feature maps (*e.g., $\tilde{\boldsymbol{X}}$*) in AiA.*

3.2. Bilinear Attention without AiA

In case the number of parameters in AiA becomes a concern, one can resort to the simplified version called *Bi-attention without AiA* (see Fig. 4). This simplification approximately halves the number of parameters of the Bi-attention block while still keeping competitive performance with regards to the person re-ID task. (See § 5 for a comparison against various benchmarks). Formally, we have

$$\boldsymbol{x}^z = \mathrm{Sigmoid}\Big(\phi\big(\mathrm{Bi}(\varphi(\boldsymbol{x}))\big)\Big) \otimes \boldsymbol{x}. \tag{6}$$

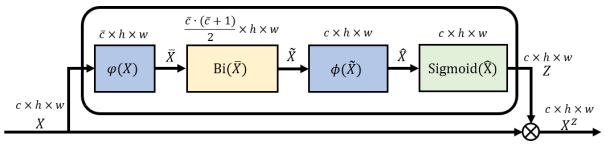

Figure 4. The structure of Bilinear attention without AiA.

Remark 2 *The structure of Bi-attention without AiA is similar to the Fully Attentional Block (FAB) [43] in the sense that both types of attention keep spatial structure information of the feature map, but with a fundamental difference that our module exploits the second order information along the channel, while FAB considers only the vanilla, first order, channel pattern.*

It is worth mentioning that both the proposed Bi-attention modules can be seamlessly placed in any existing convolutional neural network to enhance the representation learning similar to what most existing attention blocks do. In section 5, we will show the effectiveness of both Bi-attention modules in a person re-ID application.

4. Bilinear Attention Networks for Person Retrieval

In this section, the architecture of the proposed network, Bilinear ATtention network (BAT-net), will be detailed, starting from the problem formulation of our application.

4.1. Problem Formulation

Let $\boldsymbol{p}_i \in \mathbb{R}^{C\times H\times W}$ denote an input image, where C, H, and W are the number of channels, height and width, respectively. Each image $\boldsymbol{p}_i$ is labeled by its identity, denoted by $y_i \in \{1,\dots,k\}$. Thus, the training set with N images, can be described as $\{\boldsymbol{p}_i, y_i\}_{i=1}^N$. The person retrieval system, $\mathcal{F}(\boldsymbol{p},\theta)$, parameterized by θ, aims at encoding an image $\boldsymbol{p}$ to an embedding space, such that the intra-person variations are minimized while the inter-person variations are maximized. In this work, the embedding space is the concatenation of the person-appearance embedding space, *i.e.* $f_a = \mathcal{F}_a(\boldsymbol{p},\theta_a)$, and the person-part embedding space, *i.e.* $f_p = \mathcal{F}_p(\boldsymbol{p},\theta_p)$, satisfying that $\mathcal{F}(\boldsymbol{p},\theta) = [f_a^T,\ f_p^T]^T$.

4.2. Overview

The BAT-net has two feature extractors, namely, a person-appearance feature extractor (denoted by $\mathcal{F}_a$) and a person part-feature extractor (denoted by $\mathcal{F}_p$). The overall architecture of the BAT-net is shown in Fig. 5. The person overall appearance is encoded by the appearance feature extractor; while the part feature extractor aims to encode the different parts of the person.

The appearance feature extractor consists of 4 convolutional blocks. After the second convolutional block, which learns mid-level features, a Bi-attention is added to capture the second order statistical information of the feature map and highlight its discriminative regions. This bilinearly attended feature map encourages the following layers to learn a holistic representation of the person.

Recent studies in person re-identification suggest that independent modeling of part regions can enhance the precision of the overall system [36, 38, 20]. We also equip the BAT-net with such part-based learning ability. More specifically, we use a simple sub-network as a part feature extractor, which aims to learn distinct and discriminative parts in the input image. The bilinearly attended feature map $\boldsymbol{X}^z \in \mathbb{R}^{c\times h\times w}$ is divided into T non-overlapping regions $\boldsymbol{X}_t^z$ s.t. $\boldsymbol{X}_t^z \in \mathbb{R}^{c\times \frac{h}{T}\times w}, t = 1,\dots,T$. Each of the non-overlapped regions is resized to $c \times h \times w$ by bilinear interpolation and fed to the t-th stream of the part feature extractor network; which generates the part-feature embeddings.

Remark 3 *Our part feature extractor network is different from the current part-based solutions [36, 20, 38]. For example, in [36], the part feature is extracted via a pose estimation network called OpenPose [2]. In [20], the part regions are sampled through a hard attention network. In [38], the parts are split in the final feature map. By contrast, and in addition to the structural differences, each part model in the BAT-net works independently from the others in the sense that no weight-sharing between part-models is envisaged. This, in turn, can increase the diversity of the learnt parts leading to a more generalized discriminative embedding space for retrieval purposes.*

4.3. Multi-Task Training

Multi-Task Training (MTT) has shown to be effective in modern person re-ID solutions. As the name suggests, MTT formulates the overall learning procedure as a combination of several sub-tasks; each having its own importance in the overall learning mechanism. [50] uses cross-entropy loss for the classification task and triplet loss for the ranking task. [43] combines triplet loss, focal loss and cross-entropy loss to train a state-of-the-art model. Following [50], we train our network for the tasks of ranking and classification.

Ranking Task. We use triplet loss for the ranking task. In a mini batch, $\{\boldsymbol{p}_i\}_{i=1}^{N_m}$, one possible triplet can be denoted as $\{\boldsymbol{p}_i, \boldsymbol{p}_i^+, \boldsymbol{p}_i^-\}$ such that the anchor $\boldsymbol{p}_i$ shares the same identity with the positive sample $\boldsymbol{p}_i^+$ and the negative sample $\boldsymbol{p}_i^-$ belongs to a different identity. In the embedding space $\mathcal{F}(\cdot)$, the triplet loss is formulated as follows:

$$\mathcal{J}_{\text{rank}} = \frac{1}{N_{\text{tri}}}\sum_{i=1}^{N_{\text{tri}}}\left[d_i^+ - d_i^- + m\right]_+, \tag{7}$$

where $[\cdot]_+ = \max(\cdot, 0)$, N_{tri} indicates the number of triplets in one batch, m is a margin. $d_i^+ = \|(\mathcal{F}(\boldsymbol{p}_i) -$

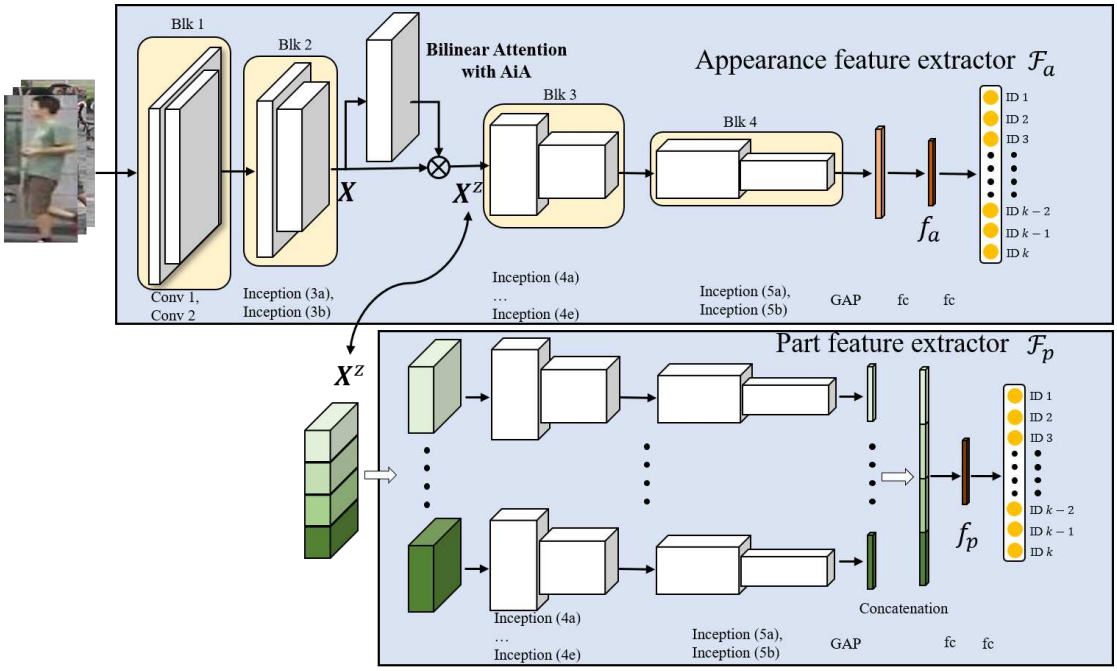

Figure 5. The architecture of Bilinear ATtention network (BAT-net). BAT-net has two feature extractor, *e.g.*, person appearance feature extractor and person part feature extractor. The feature map, $\boldsymbol{X}^z$, which is imposed by bilinear attention net, is feed to following stages of convolutional layers for appearance and part feature embedding.

$\mathcal{F}(\boldsymbol{p}_i^+)\|_2$, and $d_i^- = \|\mathcal{F}(\boldsymbol{p}_i) - \mathcal{F}(\boldsymbol{p}_i^-)\|_2$. In the triplet selection, for each anchor, we mine one hard positive and 5 hard negatives, coming up with 5 triplets. This is to avoid collapsing to local minima in the early stages of optimization.

Classification Task. The triplet loss does not fully take into account the identity specific (intra-person) information and only encodes relative similarity (inter-person) information within a particular triplet. Thus, we augment the triplet loss with the cross-entropy based classification loss $\mathcal{J}_{\text{cls}}$ to encode the class specific information.

4.4. Implementation Details

Network Architecture. We implemented our BAT-net model in the PyTorch [24] deep learning framework. The backbone network is the first version of GoogLeNet [39], pretrained on ImageNet [28] with Batch Normalization [12]. The spatial size of the input image is fixed to 256×128. In the appearance feature extractor, the size of the feature after global average pooling (GAP) is 1024, which is followed by the 512-dimensional person appearance embedding layer f_a. Another fc layer is connected to predict person identity using the person appearance embedding. In the part feature extractor, we follow the work in [20], and fix $T = 4$ across all experiments. The output features of each of the T streams are concatenated, and is passed through a 512-dimensional part embedding f_p. A fc layer is further connected to predict person identity using the person part embedding. In the testing stage, f_a and f_p are concatenated to give the final person representation f, such that $f = [f_a^T, f_a^T]^T \in \mathbb{R}^{1024}$.

In the Bi-attention block, the embedding functions $\varphi(\cdot)$, $\phi(\cdot)$ and $\varpi(\cdot)$ are 1×1 convolutional layers with a following batch normalization layer and a nonlinear layer. Here, the nonlinear layer uses the $\text{ReLU}(\cdot)$ function. In $\varphi(\cdot)$, the dimensionality reduction factor, r, is set to 8 for the CUHK03 [18] dataset, and to 4 for the other datasets. The details of datasets will be presented in § 5.1.

Network Training. We use the Adam [14] optimizer with default momentum values (0.9, 0.999) for (β_1 and β_2). The weight decay is set to 0.0001. The learning rate is initialized to 1×10^{-3} for CUHK03 [18] and 5×10^{-4} for Market-1501 [52], DukeMTMC-reID [26] and MSMT17 [46]. We train the network for 300 epochs. The learning rate is decayed by a factor of 0.1 at 150, 200, 250 epochs respectively for all the datasets. In the multi-task training, we pose the ranking task and classification task in both appearance and part feature extractors separately; this is inspired by [38] where supervision on each respective feature extractor is vital for learning discriminative features. Our training images are randomly flipped in the horizontal direction, followed by random erasing [54]. Here, the random erasing is used to provide the momentum to jump out of local optima, inspired by [11]; thus, we apply this data augmentation after 50 epochs. No such augmentation is used during the testing phase. We report the performance

of the trained network at the last epoch. Moreover, it is worth noting that we do not apply re-ranking to boost the ranking result in the testing phase.

5. Experiment

5.1. Datasets

In this section, we evaluate our proposed algorithm across four standard benchmark datasets, *i.e.*, **CUHK03** [18], **Market-1501** [52], **DukeMTMC-reID** [26] and **MSMT17** [46].

CUHK03 This dataset consists of $13,164$ person images of $1,467$ identities, captured by 6 cameras. Each person is observed by two disjoint camera views. CUHK03 offers both hand-labeled and DPM-detected [4] bounding boxes, and we evaluate our model on both sets. We adopt the new training/testing protocol proposed in [53]. In this protocol, the training set contains 767 identities and the testing set contains the remaining 700 identities.

Market-1501 This dataset consists of $32,668$ person images of $1,501$ identities observed under a maximum of 6 different camera views. The dataset is split into $12,936$ training images of 751 identities and $19,732$ testing images of the remaining 750 identities, and both training and testing images are detected using a DPM [4].

DukeMTMC-reID This dataset is collected with 8 different cameras and was originally proposed for video-based person tracking and re-identification. It has $1,404$ identities and includes $16,522$ training images of 702 identities, $2,228$ query images of 702 identities and $17,661$ gallery images.

MSMT17 This is the largest person re-ID dataset, consisting of $126,441$ person images, detected by Faster R-CNN [7], and $4,101$ identities. This dataset is collected with 15 cameras and covers 4 days with different weather conditions over a month. The training set consists of $32,621$ images belonging to $1,041$ identities, whereas the test set contains $93,820$ images of the remaining $3,060$ identities. The test set is further randomly split into $11,659$ query images and the remaining $82,161$ are used as gallery images.

5.2. Evaluation Protocol

We use both mean average precision (mAP) and cumulative matching characteristic (CMC) to evaluate the model performance. The CMC curve measures the correct matching rate for a given query image against the gallery images at various rank, whereas the mAP measures the probability of all correct matches in the gallery images for a given query image, measuring the overall ranking performance.

5.3. Comparison with State-of-the-Art Methods

To show the effectiveness of the attention block with higher order information and the AiA mechanism, Bilinear ATtention network *with* AiA and *without* AiA are tested across the four datasets.

CUHK03 We evaluated our model on both the *labeled* and *detected* person bounding boxes of CUHK03. Table 1 clearly shows that our model improves over the current state-of-the-art in both settings significantly. In particular, when compared against the current state-of-the-art Mancs, we observe that BAT-net w/o AiA outperforms it by a significant margin: *i.e.* 8.1% in mAP and 5.2% in Rank-1 accuracy on the manually labeled set and by 8.2% in mAP and 5.9% in Rank-1 accuracy on the detected set. This significant improvement shows that the use of such second order information increases the discriminative ability in representation through attention by itself. Incorporation of the AiA mechanism leads to a further improvement against Mancs: *i.e.* 4.1% in mAP and 4.4% in Rank-1 accuracy on the manually labeled set and by 4.5% in mAP and 4.8% in Rank-1 accuracy on the detected set. This validates the need of our design choices in BAT-net along with the importance of the AiA mechanism to obtain superior discriminative embeddings for person-retrieval.

Table 1. Evaluation on the CUHK03 [18] dataset in both labeled and detected bounding box. 1^{st} / 2^{nd} best in red / blue.

	@ Labeled		@ Detected	
Model	mAP	R-1	mAP	R-1
SVDNet [37]	-	-	37.3	41.5
HA-CNN [20]	41.0	44.4	38.6	41.7
AOS [11]	-	-	47.1	43.4
MLFN [3]	49.2	54.7	47.8	52.8
MGCAM [33]	50.2	50.1	46.9	46.7
DaRe [45]	52.2	56.4	50.1	54.3
PCB+RPP [38]	57.5	63.7	-	-
Mancs [43]	63.9	69.0	60.5	65.5
BAT-net w/o AiA	72.0	74.2	68.7	71.4
BAT-net w/ AiA	76.1	78.6	73.2	76.2

Market-1501 We further evaluate our proposed BAT-net against the recent state-of-the-art methods on Market-1501 in the single query setting. The results are shown in Table 2. Like before, BAT-net w/o AiA outperforms Mancs by 3.2% on mAP and 1.0% on Rank-1 accuracy respectively. Similarly, with addition of the AiA module, we observe a further improvement of 5.1% / 2.0% in terms of mAP and Rank-1 accuracy over Mancs. Moreover, when compared against PBR, which uses bilinear pooling for part alignment, both BAT-net w/o AiA and BAT-net w/ AiA outperform it by 9.5% / 3.9% and 11.4% / 4.9% respectively in-terms of mAP / Rank-1 measures.

DukeMTMC-reID The evaluation of our proposed algorithm for DukeMTMC-reID is shown in Table 3. Compared to Mancs, BAT-net w/o AiA improves the their evaluated results by 4.0% on mAP and 1.2% on Rank-1 accu-

Table 2. Evaluation on the Market-1501 [52] dataset under single query setting. 1st / 2nd best in red / blue.

Model	mAP	R-1	R-5	R-10
MSCAN [17]	57.5	80.3	-	-
SVDNet [37]	62.1	82.3	92.3	95.2
PDC [34]	63.4	84.1	92.7	94.9
JLML [19]	65.5	85.1	-	-
DaRe [45]	69.9	86.0	-	-
AOS [11]	70.4	86.5	-	-
Glad [47]	73.9	89.9	-	-
MGCAM [33]	74.3	83.8	-	-
MLFN [3]	74.3	90.0	-	-
DKPM [31]	75.3	90.1	96.7	97.9
HA-CNN [20]	75.7	91.2	-	-
PBR [36]	76.0	90.2	96.1	97.4
DuATM [32]	76.6	91.4	97.1	-
PCB+RPP [38]	81.6	93.8	97.5	98.5
Mancs [43]	82.3	93.1	-	-
SGGNN [30]	82.8	92.3	96.1	97.4
BAT-net w/o AiA	85.5	94.1	98.2	99.1
BAT-net w/ AiA	87.4	95.1	98.2	98.9

racy. Equipped with AiA, BAT-net outperforms Mancs by 5.5% on mAP and 2.8% on Rank-1 accuracy. Moreover, BAT-net w/o AiA and BAT-net w/ AiA improves the mAP / Rank-1 over PBR by 11.6% / 4.0%, and 13.1% / 5.6%, respectively.

Table 3. Evaluation on the DukeMTMC-reID [26] dataset under single query setting. 1st / 2nd best in red / blue.

Model	mAP	R-1	R-5	R-10
DaRe [45]	56.3	74.5	-	-
SVDNet [37]	56.8	76.7	86.4	89.9
AOS [11]	62.1	79.2	-	-
MLFN [3]	62.8	81.0	-	-
DKPM [31]	63.2	80.3	89.5	91.9
HA-CNN [20]	63.8	80.5	-	-
DuATM [32]	64.6	81.8	90.2	-
PBR [36]	64.2	82.1	-	-
SGGNN [30]	68.2	81.1	88.4	91.2
PCB+RPP [38]	69.2	83.3	-	-
Mancs [43]	71.8	84.9	-	-
BAT-net w/o AiA	75.8	86.1	93.9	95.6
BAT-net w/ AiA	77.3	87.7	94.7	96.3

MSMT17 Table 4 shows the result of our proposed network without/with AiA mechanism on the new challenging MSMT17 dataset. As observed, both of our proposed networks outperform the baseline algorithms by a tangible margin. More specifically, BAT-net w/o AiA and w/ AiA outperforms the next-best algorithm, *i.e.* Glad, by 16.4% / 12.7% and 22.8% / 18.1% with regards to mAP and Rank-1 accuracy, respectively.

Table 4. Evaluation on the MSMT17 [46] dataset under single query setting. "*" indicates the results of the algorithms as reported in [46] . 1st / 2nd best in red / blue.

Model	mAP	R-1	R-5	R-10
GoogLeNet* [39]	23.0	47.6	65.0	71.8
PDC* [34]	29.7	58.0	73.6	79.4
Glad* [47]	34.0	61.4	76.8	81.6
GoogLeNet (Ours)	39.3	65.8	80.5	85.3
+ Bi-attention w/AiA	43.1	69.5	82.7	87.2
BAT-net w/o AiA	50.4	74.1	86.4	89.7
BAT-net w/ AiA	56.8	79.5	89.1	91.1

5.4. Ablation Study

We further perform extra experiments to verify the effectiveness of our proposed Bi-attention with AiA on Market-1501 [52] under the single query setting and CUHK03 [18] with the detected bounding boxes.

Effect of Bilinear Attention. We first evaluate the effect of bilinear attention on the feature extractors, and the results are shown in Table 5. The results on both the datasets convince us that: **(1)** The Bi-attention brings retrieval gain in person appearance feature extractor. **(2)** The retrieval accuracy increases when the appearance feature extractor is equipped with the part feature extractor. **(3)** Further addition of Bi-attention continues to improve the overall performance of the network as a whole. This shows that our design is effective in exploiting the complementary information between feature extractors and the attention model.

Table 5. Effect of Bi-attention on the Market-1501 [52] and CUHK03 [18] datasets.

		Market @ SQ		CUHK03 @ D	
Model		mAP	R-1	mAP	R-1
(i)	$\mathcal{F}_a$	80.7	91.6	64.5	67.1
(ii)	+ Bi-attention w/ AiA	83.4	93.2	67.4	70.6
(iii)	$\mathcal{F}_a + \mathcal{F}_p$	85.1	93.8	67.8	71.1
(iv)	BAT-net w/ AiA	**87.4**	**95.1**	**73.2**	**76.2**

Effect of the Position of Bilinear Attention. Table 6 shows the effect of Bi-attention when added to different positions of the baseline GoogLeNet network. p_1, p_2, p_3 and p_4 indicate the position of the output of Blk 1, Blk 2, Blk 3 and Blk 4 along the appearance feature extractor respectively (Refer to Fig. 5). Table 6 shows that: **(1)** Using Bi-attention in the early stages, *i.e.* p_1, p_2, is superior to inserting it in the later stages *i.e.* p_3, p_4. A similar observation is also made in [44], where the non-local network enhances the performance of ResNet [9] in its early stages. (**2**) Moreover, the performance of adding Bi-attention in p_2 surpasses the performance compared to when it is added in

p_1. One reasonable explanation is that, the feature maps in p_2 have richer channel information than that in p_1, while it still maintains spatial structure information, thereby enabling the network to emphasize more on the statistical information. **(3)** In the CUHK03 dataset, which has a smaller training set, the performance of person retrieval degrades when Bi-attention is inserted in p_4. This is observed as the embedding layer of the Bi-attention module overfits on the training set due to the high dimensionality of the feature maps in p_4.

Table 6. Effect of the position of Bi-attention on the Market-1501 [52] and CUHK03 [18] datasets.

		Market @ SQ		CUHK03 @ D	
Model		mAP	R-1	mAP	R-1
(i)	w/o attention	85.1	93.8	67.8	71.1
(ii)	p_1	86.8	94.4	71.4	73.2
(iii)	p_2	**87.4**	**95.1**	**73.2**	**76.2**
(iv)	p_3	85.5	93.9	69.8	72.9
(v)	p_4	85.3	94.0	68.9	70.8

Effect of the Dimensionality Reduction Factor r. The reduction factor r in the embedding function $\varphi(\cdot)$ is an important hyperparameter that affects the information pooled by the bilinear operation. The results and comparisons shown in Table 7 reveal that: **(1)** The performance does not improve monotonically with a decreased factor. The main interpretation is that the parameter size will increase exponentially by decreasing the factor, which leads to the overfitting of embedding functions (*e.g.*, $\varpi(\cdot)$, $\phi(\cdot)$) in the training set. **(2)** We observe that while $r = 4$ gives the best results in the Market dataset, the best value of r is observed to be 8 when the network is trained on CUHK03. One possible explanation is that the network trained on Market is less prone to overfitting due to its larger training set in comparison to CUHK03.

Table 7. Effect of the Dimensionality Reduction Factor r for embedding function $\varphi(\cdot)$ on the Market-1501 [52] and CUHK03 [18] datasets.

		Market @ SQ		CUHK03 @ D	
Model		mAP	R-1	mAP	R-1
(i)	w/o attention	85.1	93.8	67.8	71.1
(ii)	$r = 2$	87.1	94.9	72.3	75.4
(iii)	$r = 4$	**87.4**	**95.1**	72.6	74.9
(iv)	$r = 8$	87.2	94.5	**73.2**	**76.2**
(v)	$r = 16$	86.9	94.4	72.5	75.6
(vi)	$r = 32$	86.9	94.1	72.1	74.8

Visualisation of Bilinear Attention. We visualise the Bi-attention for person images in both the Market-1501 dataset in Fig. 6(a) and CUHK03 detected-set in Fig. 6(b).

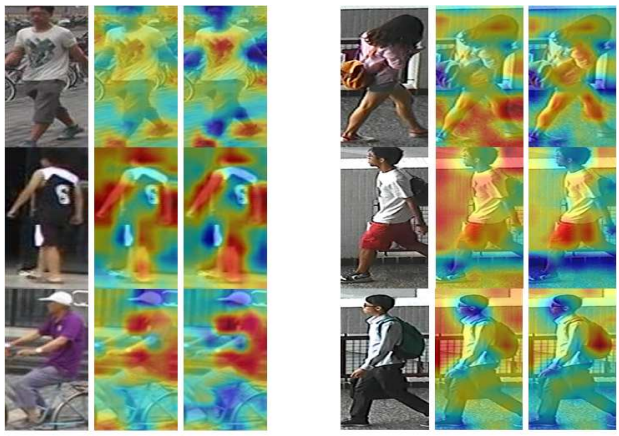

Figure 6. Visualisation of our Bi-attention in person images, sampled from the Market dataset (a) and the CUHK03 dataset (b). In each dataset, from left to right, (1) the input person image, (2) the input feature map to Bi-attention and (3) the masked feature map. In the heat map, the response increases from blue to red. Best viewed in color.

Fig. 6 shows that: **(1)** The attention mask filters out the non-informative background clutter in person images, **(2)** The attention mask further emphasizes on the discriminative parts of a person bounding box, which reduces the prevalent misalignment problem in the retrieval task.

6. Conclusion

In this paper, we propose a novel Bilinear attention (Bi-attention) block for person retrieval. The Bi-attention block uses bilinear pooling to model the local pairwise feature interactions along each channel, while preserving the spatial structural information. Then an Attention in Attention (AiA) mechanism is proposed to build inter-dependency among second order local and global features with the intent to make better use of, or pay more attention to, such higher order statistical relationships. We also introduce a simplified version called Bi-attention without AiA, which approximately halves the number of parameters of the Bi-attention block, while still keeping competitive performance in visual tasks. We incorporated the aforementioned two Bi-attention blocks in our network, BAT-net, and showed that state-of-the-art performances could be achieved by benefiting from higher order attention in representation learning. This includes extensive evaluations on four standard person re-ID benchmarks along with the required ablation studies to understand the effect of the Bi-attention block.

Future works include analyzing the AiA for addressing other visual tasks and developing other forms of attention mechanisms by exploiting higher-order information.

References

[1] Song Bai, Xiang Bai, and Qi Tian. Scalable Person Re-identification on Supervised Smoothed Manifold. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[3] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-Level Factorisation Net for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 7

[4] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 6

[5] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact Bilinear Pooling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2

[6] Niloofar Gheissari, Thomas B. Sebastian, Peter H. Tu, Jens Rittscher, and Richard Hartley. Person Reidentification Using Spatiotemporal Appearance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006. 1, 2

[7] Ross Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, 2015. 6

[8] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person Re-Identification*. Springer, January 2014. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[11] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially Occluded Samples for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 6, 7

[12] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015. 5

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015. 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] Piotr Koniusz, Hongguang Zhang, and Fatih Porikli. A Deeper Look at Power Normalizations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5774–5783, 2018. 2

[16] Ryan Layne, Timothy Hospedales, and Shaogang Gong. Person Re-identification by Attributes. In *23rd British Machine Vision Conference (BMVC)*, September 2012. 1

[17] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2017. 1, 2, 7

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 5, 6, 7, 8

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Person Re-identification by Deep Joint Learning of Multi-Loss Classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, 2017. 7

[20] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4, 5, 6, 7

[21] Tsung-Yu Lin and Subhransu Maji. Improved Bilinear Pooling with CNNs. In *British Machine Vision Conference (BMVC)*, 2017. 1, 2

[22] Hao Liu, Jiashi Feng, Jianguo Jiang, and Shuicheng Yan. End-to-End Comparative Attention Networks for Person Re-identification, 2016. arXiv:1606.04404 [cs.CV]. 1, 2

[23] Jian Liu, Zhen Yang, Zhang Tao, and Xiong Huilin. MULTI-PART COMPACT BILINEAR CNN FOR PERSON RE-IDENTIFICATION. In *2017 IEEE International Conference on Image Processing (ICIP)*, February 2017. 1, 2

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *31st Conference on Neural Information Processing Systems (NIPS)*, December 2017. 5

[25] Hamed Pirsiavash, Deva Ramanan, and Charless C. Fowlkes. Bilinear classifiers for visual recognition. In *Advances in Neural Information Processing Systems 22*, pages 1482–1490. Curran Associates, Inc., 2009. 1

[26] 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 workshop on Benchmarking Multi-Target Tracking*, 2016. 2, 5, 6, 7

[27] Soumava Kumar Roy, Mehrtash Harandi, Richard Nock, and Richard Hartley. Siamese networks: The Tale of Two Manifolds. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[28] 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. 5

[29] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood

Re-Ranking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[30] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person Re-identification with Deep Similarity-Guided Graph Neural Network. In *European Conference on Computer Vision (ECCV)*, 2018. 7

[31] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-End Deep Kronecker-Product Matching for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[32] Jianlou Si, Hongguang Zhang, Chunguang Li, Jason Kuen, Xiangfei Kong, Kot. C. Alex, and Wang Gang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 7

[33] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-Guided Contrastive Attention Model for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 6, 7

[34] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven Deep Convolutional Model for Person Re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3960–3969, 2017. 1, 7

[35] 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 (ECCV)*, September 2016. 1

[36] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-Aligned Bilinear Representations for Person Re-Identification. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 4, 7

[37] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. SVDNet for Pedestrian Retrieval. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3820–3828. IEEE, 2017. 2, 6, 7

[38] 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 *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 4, 5, 6, 7

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1–9, 2015. 5, 7

[40] Maoqing Tian, Shuai Yi, Li Hongsheng, Li Shihua, Xueshen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating Background-bias for Robust Person Re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[41] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. Multi-region Bilinear Convolutional Neural Networks for Person Re-Identification, 2015. arXiv:1512.05300 [cs.CV]. 1, 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. 2

[43] 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. 1, 2, 4, 6, 7

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[45] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 6, 7

[46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person Trasfer GAN to Bridge Domain Gap for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6, 7

[47] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-Local-Alignment Descriptor for Pedestrian Retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, 2017. 7

[48] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[49] Dong Yi, Zhen Lei, and Stan Z Li. Deep Metric Learning for Person Re-identification. In *2014 22nd International Conference on Pattern Recognition (ICPR)*, December 2014. 2

[50] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *European Conference on Computer Vision (ECCV)*, September 2018. 4

[51] Hongguang Zhang and Piotr Koniusz. Power Normalizing Second-order Similarity Network for Few-shot Learning. In *Winter Conference on Applications of Computer Vision (WACV)*, 2019. 2

[52] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 5, 6, 7, 8

[53] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking Person Re-identification with k-reciprocal Encoding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017. 6

[54] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 5