Attention in Attention Networks for Person Retrieval

Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Pan Ji, Lars Petersson, and Mehrtash Harandi

Abstract—This paper generalizes the Attention in Attention (AiA) mechanism, in P. Fang et al., 2019 by employing explicit mapping in reproducing kernel Hilbert spaces to generate attention values of the input feature map. The AiA mechanism models the capacity of building inter-dependencies among the local and global features by the interaction of inner and outer attention modules. Besides a vanilla AiA module, termed linear attention with AiA, two non-linear counterparts, namely, second-order polynomial attention and Gaussian attention, are also proposed to utilize the non-linear properties of the input features explicitly, via the second-order polynomial kernel and Gaussian kernel approximation. The deep convolutional neural network, equipped with the proposed AiA blocks, is referred to as Attention in Attention Network (AiA-Net). The AiA-Net learns to extract a discriminative pedestrian representation, which combines complementary person appearance and corresponding part features. Extensive ablation studies verify the effectiveness of the AiA mechanism and the use of non-linear features hidden in the feature map for attention design. Furthermore, our approach outperforms current state-of-the-art by a considerable margin across a number of benchmarks. In addition, state-of-the-art performance is also achieved in the video person retrieval task with the assistance of the proposed AiA blocks.

Index Terms—Attention in attention mechanism, person retrieval, pedestrian representation, convolutional neural network, second-order polynomial kernel, Gaussian kernel

1 Introduction

Person retrieval, also known as person re-identification (re-ID), has attracted an increasing amount of interests in the Computer Vision (CV) community due to its challenging nature and industrial prospects. The task of a person retrieval machine can be characterized as follows: given an image of a specific person, the machine should retrieve all images with the same identity, from a gallery.

There are quite a few factors that can lead to an unreliable person retrieval system, making the re-ID task daunting and challenging. For example, *misalignment* caused by spatial nuances in the person bounding box (e.g., movements of body parts) can negatively affect a re-ID system [2]. That is, the location of the person's body, and its parts, with respect to a reference frame, can be easily displaced due to the change in body orientation, pose, clothing, *etc*. This, in turn, causes mismatching of features during training and testing, leading to inaccurate re-identification. Much effort has been made into studying and addressing these difficulties [3], [4], [5], [6]. However, it still remains an open problem and calls for further study to learn a robust and discriminative representation of the person(s).

In general, solutions to address the misalignment within a person bounding box can be broadly categorized into *human pose*-based, *human attributes*-based as well as *visual attention*-based methods. In recent years, several attempts that rely on human pose estimation have been undertaken to address this in [2], [6], [7]. These algorithms employ additional estimator networks that provide the baseline-network with complementary cues to learn a superior embedding space, thereby outperforming the baseline-network. Other solutions benefit from person attributes [8], [9], [10], that are invariant to variations in human pose, light illumination, background clutter, spatial misalignment, *etc*. Such solutions aim at learning a robust person representation as described by the human attributes. Recently, visual attention-based solutions have received an overwhelming interest in the re-ID task, since it outperforms the pose-based/attribute-based models without the need of any additional pose detector or attribute estimator network.

The attention-based models, inspired by the human visual and attentive sensing processes, aim to localize the discriminative regions within a person bounding box [11], [12], [13]. The inherent attention module (e.g., hard attention [11] or soft attention [14]) is designed to automatically select the informative parts of an image, and is trained in a weakly-supervised manner (i.e., no explicit labeling information is given to identify the areas to attend). In our preliminary study [1], we proposed the Attention in Attention (AiA) mechanism to model the explicit interaction between

- *Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, and Lars Petersson are with the Research School of Electrical, Energy and Material Engineering, Australian National University, Canberra, ACT 2601, Australia, and also with the Data61-CSIRO, Black Mountain Laboratories, Canberra, ACT 2601, Australia. E-mail: {Pengfei.Fang, Jieming.Zhou, Soumava.KumarRoy}@anu.edu.au, lars.petersson@data61.csiro.au.*
- *Pan Ji is with the OPPO US Research Center, Palo Alto, CA 94303 USA. E-mail: peterji1990@gmail.com.*
- *Mehrtash Harandi is with the Department of Electrical and Computer Systems Engineering, Monash University, Clayton, VIC 3800, Australia, and also with the Data61-CSIRO, Melbourne 3800, Australia. E-mail: mehrtash.harandi@monash.edu.*

Manuscript received 25 May 2020; revised 22 Feb. 2021; accepted 5 Apr. 2021. Date of publication 15 Apr. 2021; date of current version 4 Aug. 2022. (Corresponding author: Pengfei Fang.) Recommended for acceptance by L. Wang. Digital Object Identifier no. 10.1109/TPAMI.2021.3073512

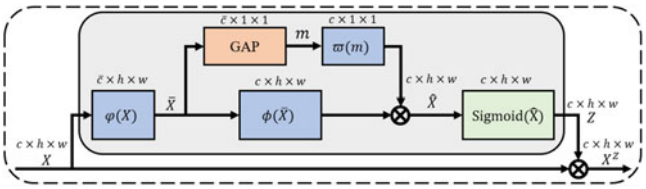


Fig. 1. The structure of Linear attention with AiA. $\varphi(\cdot)$, $\phi(\cdot)$ and $\varpi(\cdot)$ are embedding functions. GAP indicates global average pooling. $\otimes$ indicates element-wise multiplication.

global and local features of the feature map, and used a bilinear mapping [15] that benefits from the second-order statistics to generate the attention values. In this paper, we aim to generalize the AiA mechanism by making use of higher-order statistics, explicitly encoded by non-linear kernel mappings within the AiA framework, to generate the attention map.

Designing non-linear embeddings (e.g., feature space of kernel machines) by making use of the geometry of Reproducing Kernel Hilbert Spaces (RKHS) dates back to the celebrated work of Vapnik [16]. The machinery of RKHS is rich enough to even handle infinite-dimensional representations (through the use of the well-known kernel trick). Also, recent studies show that kernel methods along deep neural networks (DNNs) would help to attain rich models [17], [18], [19], [20]. This inspires us to benefit from the theory of RKHS and its approximations [21], [22] to design attention modules for DNNs. To the best of our knowledge, this is the first attempt where an attention mechanism is implemented from a RKHS perspective.

This paper generalizes the AiA framework by employing explicit non-linear mappings in RKHS to generate attention value(s). The AiA framework consists of an *outer attention* block that encompasses an *inner attention* block such that the inner block is tasked to determine the discriminative regions of the feature map where the outer attention block should focus (See Fig. 1 for a conceptual diagram for AiA structure). Therefore, the AiA block models channel-wise inter-dependencies between the global and local features, while preserving the spatial structural information of the input feature map, in a unified block. Besides a vanilla AiA block, which only exploits linear features of its input feature map, we further propose and develop two non-linear versions of AiA, with each respectively using the second-order polynomial and Gaussian kernels of the feature map along the channels. The intuition behind adopting the features in RKHSs is that such features can benefit from the highly discriminative capacity of high- or infinite-dimensional spaces, thereby helping the attention block to focus on more discriminative areas within the feature maps. Even though functions in RKHS can approximate any function, the operational capacity is limited due to computationally expensive kernel operations on the whole training data [21], [22]. In this paper, we further propose to alleviate these constraints by relying on advanced kernel estimation techniques. More specifically, the second-order polynomial kernel is modeled by a bilinear mapping [15], while the Gaussian kernel is estimated by random Fourier features [22]. By such transformations, learnable parameters are avoided in the non-linear transformation, leading to being optimized easily. We further propose a computationally efficient version of the attention block without the use of the inner attention block. Table 1 summarizes the proposed attention modules.

The *contribution* of our work can be summarized as follows: *(a)* We formulate a generalized Attention in Attention (AiA) mechanism, where the attention map is generated by the interaction between the inner and outer attention modules. This indeed results in modeling inter-dependencies between global and local features of its input feature map, while maintaining the spatial structural information. *(b)* We further develop kernelized versions of the AiA block, namely, *second-order polynomial attention* (SoP-attention) and *Gaussian attention* (Gau-attention), by estimating the second-order polynomial and Gaussian features of the input feature map respectively. Furthermore, we employ advanced kernel estimation techniques to reduce the computational cost of the kernel matrix. *(c)* We propose a novel deep architecture using the AiA block, creating our Attention in Attention Network (AiA-Net), for the task of person retrieval. This AiA-Net extracts complementary person appearance and part features for discriminative person representation learning. *(d)* Extensive experiments performed on large scale standard benchmark datasets including *CUHK03* [23], *Market-1501* [24], *DukeMTMC-reID* [25] and *MSMT17* [26], as well as a small scale benchmark dataset (e.g., *CHUK01* [27]), show that our approach outperforms the current state-of-the-art methods by a considerable margin in terms of mAP and R-1 metrics. Meanwhile, we also conduct extensive ablation studies that verify the superiority of the AiA mechanism and the utility of the non-linear features. *(e)* In the video person retrieval setting, our deep network (e.g., AiA-Net-V) also achieves state-of-the-art results on the popular video benchmark dataset, *MARS* [28]. Additionally, we find an interesting observation that the Gau-attention mechanism is empirically superior to the SoP-attention, in terms of accuracy, computational cost as well as number of parameters.

TABLE 1
Summary of the Proposed Attention Modules

Kernel attention	Kernel formulation:$\mathcal{K}(\boldsymbol{x},\boldsymbol{y})$	AiA formulation	Non-AiA formulation	Kernel approximation
Linear	$\boldsymbol{x}^\top \boldsymbol{y}$	$\boldsymbol{x}^z = \underbrace{\text{Sigmoid}\Big(\phi\big(\varphi(\boldsymbol{x})\big) \otimes \varpi(\boldsymbol{m})\Big)}_{\boldsymbol{z}} \otimes \boldsymbol{x}$	$\boldsymbol{x}^z = \underbrace{\text{Sigmoid}\Big(\phi\big(\varphi(\boldsymbol{x})\big)\Big)}_{\boldsymbol{z}} \otimes \boldsymbol{x}$	Identity mapping
Second-order polynomial	$(\boldsymbol{x}^\top \boldsymbol{y} + c)^2$	$\boldsymbol{x}^z = \underbrace{\text{Sigmoid}\Big(\phi\big(\text{SoP}(\varphi(\boldsymbol{x}))\big) \otimes \varpi(\boldsymbol{m})\Big)}_{\boldsymbol{z}} \otimes \boldsymbol{x}$	$\boldsymbol{x}^z = \underbrace{\text{Sigmoid}\Big(\phi\big(\text{SoP}(\varphi(\boldsymbol{x}))\big)\Big)}_{\boldsymbol{z}} \otimes \boldsymbol{x}$	$\text{SoP}(\cdot)$
Gaussian	$e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}}$	$\boldsymbol{x}^z = \underbrace{\text{Sigmoid}\Big(\phi\big(\text{Gau}(\varphi(\boldsymbol{x}))\big) \otimes \varpi(\boldsymbol{m})\Big)}_{\boldsymbol{z}} \otimes \boldsymbol{x}$	$\boldsymbol{x}^z = \underbrace{\text{Sigmoid}\Big(\phi\big(\text{Gau}(\varphi(\boldsymbol{x}))\big)\Big)}_{\boldsymbol{z}} \otimes \boldsymbol{x}$	$\text{Gau}(\cdot)$

Here, we use feature vectors (e.g., $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^c$) in the attention formulation instead of the tensor shaped feature map, for the purpose of simplicity. $\boldsymbol{z}$ denotes the attention mask, generated by the outer attention, $\boldsymbol{x}$ and $\varpi(\boldsymbol{m})$ denote the associated channel feature and inner attention mask. Refer to Section 3 for more detail.

For instance, on the CUHK03 dataset, the mAP/R-1 of AiA-Net with Gau-attention is 77.6/80.6 percent as compared to 77.0/79.4 percent for SoP-attention, while the computational complexity/number of parameters of Gau-attention are three times smaller than that of SoP-attention (e.g., $0.044 \times 10^9/0.58 \times 10^6$ versus $0.117 \times 10^9/1.79 \times 10^6$).

2 Related Work

Our work mainly focuses on person re-identification[1] and the associated attention mechanism. Here, we briefly give an overview of those works. Thereafter, we review the kernel estimation techniques used, namely, the bilinear mapping and kernel approximation.

Person Re-Identification. Early works in the field of person re-ID relied mostly on designing hand-crafted feature representations [29] or learning latent spaces [30]. We refer interested readers to [31] 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 [30], Yi *et al.* proposed a unified framework for feature and similarity learning using Siamese networks [32]. Multi-level similarities are employed in [33] to make more reliable decisions. Xiao *et al.* trained a model across multiple datasets [34] and used domain guided dropouts to mute domain-irrelevant neurons to learn robust features. Structural constraints (e.g., orthogonality, geometry) on the embedding layer [35], [36] have also been shown to learn robust person features and achieve superior results on the person re-ID task. In deep metric learning, some works also concentrate on developing the ranking loss in formulation [37] or mining strategies [12]. Considering the camera distribution, the spatial and temporal signal is further adopted to eliminate the irrelevant, thereby improving the ranking results [38]. Besides the single image presentation, video data also introduces temporal cues to encode a compact and robust video presentation of a pedestrian [39], [40]. In the early work of [40], the clip-level feature is fused by using a simple yet effective temporal pooling technique. A Recurrent Neural Network (RNN) is further employed to leverage the temporal information, and fuse the frame-level features [39]. Temporal attention mechanisms predict the importance of each frame feature and uses weighted sum to fuse them [41]

Attention Mechanism. Attention mechanisms, inspired by the human sensing process, have been studied extensively in Natural Language Processing [42] and Computer Vision [14]. Self-attention is first proposed in [42] and achieves a breakthrough in machine translation, showing its superior performance over the RNN. Thereafter, several visual applications have incorporated this attention module in their formulation, e.g., image classification [43], scene segmentation [44], image captioning [45] as well as video person re-ID [41]. On the other hand, channel attention, such as the Squeeze-and-Excitation block [14], attempts to re-weight each slice of the feature map, thereby emphasizing the informative channel features. More discriminative cues are extracted by incorporating spatial and channel information from the feature map [46].

In person re-ID, the person misalignment [3] and background biases [47] obstruct learning of a robust feature 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) [48], a binary hard attention, was used in [49] to localize the latent body parts of a human. Liu *et al.* [50] proposed a Comparative Attention Network (CAN), which repeatedly localizes discriminative parts and compares different local regions of person pairs. In Harmonious Attention Convolutional Neural Network (HA-CNN) [11], hard region-level attention and soft pixel-level attention are learned in a unified attention block. Wang *et al.* [12] considered both the channel-wise and spatial-wise attention in a Fully Attentional Block (FAB), where the channel information is re-calibrated while the spatial structural information is also preserved. Besides aligning the feature maps, Dual ATtention Matching network (DuATM) [51] also calibrates the features by matching the intra-feature sequence. In [9], the attention learning is driven by person attribute prediction. In the video person re-ID task, attention mechanisms have also been employed in temporal modeling. For example, the attention weights for each frame is generated by temporal convolution [52]. The recent works continue to mine more spatial and temporal information via spatial-temporal attention [53].

Bilinear Mapping. Bilinear mappings and models have been widely considered as a generalization of their linear counterparts. Some prime examples are bilinear classifiers [54], bilinear pooling [55] and bilinear CNNs [15] with applications in visual question answering, fine-grained image recognition, texture classification to name a few. Related to our work, the bilinear pooling [55], is first introduced to model local pairwise feature interactions for fine-grained recognition applications and its representation power is also enhanced by normalizing the higher order statistics [56]. Thereafter, Liu *et al.* [57] proposed a compact form of the bilinear operation to pool a high-dimensional feature representation for the task of person re-ID. In [58], Ustinova *et al.* proposed a patch-based multi-regional bilinear pooling to account for the geometric misalignment problem between the person bounding boxes. Recently, Suh *et al.* [3] used a part-aligned representation to mitigate the misalignment problem by fusing the appearance and part feature maps in a bilinear pooling layer. To avoid a quadratic computational cost, the bilinear features are estimated by a compact representation, e.g., the tensor sketch [55], or the Hadamard product of low-rank bilinear pooling [59].

Kernel Approximation. Feature embedding in RKHS has been commonly used in many machine learning methods, such as, non-linear SVM, kernel PCA and unsupervised learning [60]. Nonetheless, training such kernel machines is N times slower than the vanilla linear machine, where N is the size of the training data [21]. This results in poor scalability of the non-linear kernel based algorithms as the feature learning operates on the kernel matrix, leading to the birth of accelerated kernel machines [21], [22]. One possible attempt is to approximate the high dimensional features by explicit mapping in RKHS, which is scalable linearly to the

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

size of training samples [61]. Maji *et al.* [62] approximated the intersection kernel by sparse projection. Shift-invariant kernels, e.g., Gaussian kernels, Cauchy kernels *etc.*, are estimated by randomly mapping the feature in the Fourier domain of the associated kernel [22]. Approximation to a group of additive homogeneous by spectral analysis is studied by Vedaldi and Zisserman [21], yielding closed-form solutions.

3 Attention in Attention

In this section, we will first describe the Attention in Attention (AiA) framework, which only uses the linear features of the input feature map in the attention block. This vanilla module is termed as *Linear attention with AiA*. Subsequently, its non-linear counterparts will be developed by making use of second-order polynomial and Gaussian kernels in the attention module. Each AiA module will be followed by a discussion of its simplified version (i.e., the attention w/o AiA). All proposed attention blocks are summarized in Table 1.

3.1 Linear Attention

Linear attention (Lin-attention) with AiA refers to the vanilla attention module under the AiA framework as it explicitly uses the linear features over the input feature map. This AiA mechanism models the inter-dependencies between the local and global features, whilst preserving the spatial structure of its input feature map. The architecture of Lin-attention with AiA is shown in Fig. 1.

Let $\boldsymbol{X} \in \mathbb{R}^{c\times h\times w}$ be the input 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) as $\boldsymbol{x}_{ij} \in \mathbb{R}^c, i \in \{1, 2, \ldots, h\},\ j \in \{1, 2, \ldots, w\}$. The embedding function, $\varphi(\cdot)$, first compresses $\boldsymbol{x}$[2] from the original channel dimension c to $\bar{c}$ as follows:

$$\bar{\boldsymbol{x}} = \varphi(\boldsymbol{x}), \tag{1}$$

where $\bar{\boldsymbol{x}} \in \mathbb{R}^{\bar{c}}$ with $\bar{c} = c/r$. The hyper-parameter r is the dimensionality reduction factor and its effect is discussed in Section 5.2.

We note that even though $\bar{\boldsymbol{x}}$ encodes the channel features (i.e., $\boldsymbol{x}$), it doesn't change the spatial location of body parts in the feature map. As a result the misalignment issue within the feature map still persists, which hinders the performance gain by the attention module. To address this shortcoming, we introduce the concept of "Attention in Attention", which aims to adaptively re-weight the channel feature responses by modeling the inter-dependency between the global and local features[3] (See Fig. 1). We first model the global feature of the feature map using a Global Average Pooling (GAP) layer, as follows:

$$\boldsymbol{m} = \frac{1}{hw}\sum_{i=1}^{hw} \bar{\boldsymbol{x}}_i, \tag{2}$$

2. The subscripts have been omitted to avoid cluttering of notations.

3. In this paper, the physical meaning of the "global feature" and "'local feature" indicates the "person's appearance feature" and "part feature".

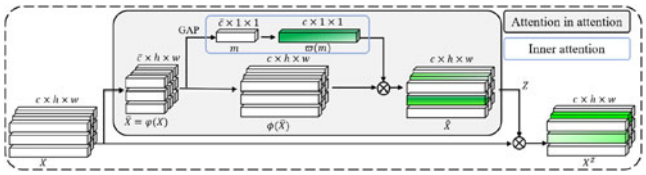


Fig. 2. Details of the Attention in Attention (AiA) mechanism.

where $\boldsymbol{m} \in \mathbb{R}^{\bar{c}}$. The inter-dependency between the embedded global feature $\boldsymbol{m}$ and each embedded local features $\bar{\boldsymbol{x}}$ is calculated as follows:

$$\hat{\boldsymbol{x}} = \varpi(\boldsymbol{m}) \otimes \boldsymbol{\phi}(\bar{\boldsymbol{x}}), \tag{3}$$

where $\otimes$ denotes the standard Hadamard (element-wise) multiplication and $\varpi(\boldsymbol{m}),\ \boldsymbol{\phi}(\bar{\boldsymbol{x}}) \in \mathbb{R}^c$. The embedding functions, $\varpi(\boldsymbol{m})$ and $\boldsymbol{\phi}(\bar{\boldsymbol{x}})$, not only process the channel feature responses, but also recover the dimension of the channel from $\bar{c}$ to c (i.e., the channel size of the input $\boldsymbol{x}$). Refer to Fig. 2 for a detailed pictorial representation of the aforementioned steps. Intuitively, $\varpi(\boldsymbol{m})$ acts as an inner attention and emphasizes the local feature $\boldsymbol{\phi}(\bar{\boldsymbol{x}})$ which are more correlated to the global feature $\varpi(\boldsymbol{m})$ via Eq. (3). In Section 4.4, we give the details of embedding functions (i.e., $\varphi(\cdot)$, $\varpi(\cdot)$ and $\boldsymbol{\phi}(\cdot)$).

The final attention mask of input $\boldsymbol{x}$ is obtained by bounding $\hat{\boldsymbol{x}}$. In this paper, we use $\mathrm{Sigmoid}(\cdot)$ for this purpose (i.e., $\boldsymbol{z} = \mathrm{Sigmoid}(\hat{\boldsymbol{x}})$). This resulting vector will act as an outer attention map, and emphasize/attenuate the significant/insignificant elements of its input feature vector $\boldsymbol{x}$ at the same spatial position as shown below:

$$\boldsymbol{x}^z = \boldsymbol{z} \otimes \boldsymbol{x}. \tag{4}$$

Remark 1. The operations described by Eqs. (2) and (3) resemble the Squeeze-and-Excitation (SE) Networks [14]. 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 the channel attention as an inner attention module to perform significance weighting of the attention-dependent feature map (e.g., $\boldsymbol{\phi}(\bar{\boldsymbol{x}})$) in AiA and produce the output feature map (e.g., $\hat{\boldsymbol{X}}$). Subsequently, our AiA module will further process $\hat{\boldsymbol{X}}$ to generate the final attention map (e.g., $\boldsymbol{Z}$). In Figs. 1 and 3, we further illustrate the difference between the SE block and the proposed AiA block. Mathematically, for a given feature maps $\boldsymbol{X} \in \mathbb{R}^{c\times h\times w}$ as input, the output of SE block is given by

$$\boldsymbol{X}^z = \mathrm{Sigmoid}\Big(\sigma\big(\xi(\mathrm{GAP}(\mathbf{X}))\big)\Big) \otimes \boldsymbol{X}, \tag{5}$$

where GAP indicates Global Average Pooling and $\xi(\cdot)$, $\sigma(\cdot)$ are the gating functions. In contrast, the output of our proposed AiA block is formulated as

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

By comparing Eqs. (5) and (6), one can observe that if $\varphi(\cdot)$ is the identity mapping, $\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{I}$, and $\varpi(\cdot) = \sigma\big(\xi(\cdot)\big)$, then our AiA block realizes the SE block. In other words, SE block in a special case in our AiA framework. It is noted $\boldsymbol{I} \in \mathbb{R}^{c\times h\times w}$ represents identity tensor here. Since

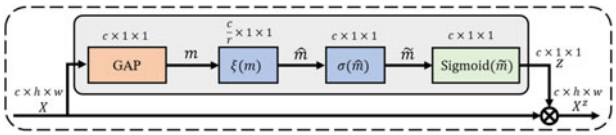


Fig. 3. The structure of Squeeze and Excitation block.

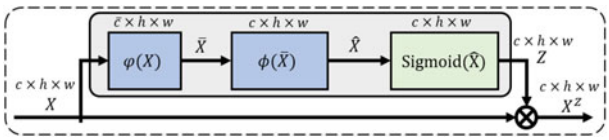


Fig. 4. The structure of Linear attention without AiA.

our AiA block also encodes local features (e.g., $\phi(\bar{X})$), we believe our attention maintains the spatial structural information of the input feature map (e.g., X), which essentially improves the performance of the attention block (Refer the study in Section 5.2).

3.1.1 Linear Attention Without AiA

In case the number of parameters in the AiA module becomes a concern, one can resort to a simplified version which we denote as *Lin-attention without AiA* (See Fig. 4). This simplification reduces the number of parameters in the Lin-attention block while still obtaining competitive performance with respect to the current algorithms for person re-ID tasks. (Refer to Section 5.2.2 for a comparison against various benchmarks). Formally, we have

$$\boldsymbol{x}^z = \mathrm{Sigmoid}\Big(\phi\big(\varphi(\boldsymbol{x})\big)\Big) \otimes \boldsymbol{x}. \tag{7}$$

In the Lin-attention module, the attention map is generated based on the linear property of the input feature map. To boost its discriminative capacity, we estimate second-order polynomial and Gaussian kernels to extract non-linear features from the input feature map so as to generate the attention map (or values). The two attention modules are called *second-order polynomial attention* and *Gaussian attention*, respectively (Refer to Fig. 5 for a more detailed description).

3.2 Second-Order Polynomial Attention

In the second-order polynomial attention (SoP-attention), we make use of the concept of polynomial kernels within

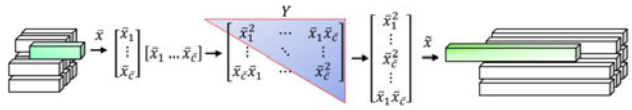


Fig. 6. Processing of bilinear pooling and second order feature rearrangement, denoted by $\mathrm{SoP}(\cdot)$. In this operation, we sample the elements in the upper triangle of Y and vectorize those elements to a new feature vector $\tilde{x}$.

AiA. The architecture of SoP-attention is shown in Fig. 5a. In SoP-attention, we first obtain

$$\begin{aligned} \boldsymbol{Y} &= \varphi(\boldsymbol{x})\varphi(\boldsymbol{x})^\top = \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top \\ &= \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_{\bar{c}} \end{bmatrix} [\bar{x}_1 \dots \bar{x}_{\bar{c}}] \\ &= \begin{bmatrix} \bar{x}_1^2 & \dots & \bar{x}_1\bar{x}_{\bar{c}} \\ \vdots & \ddots & \vdots \\ \bar{x}_{\bar{c}}\bar{x}_1 & \dots & \bar{x}_{\bar{c}}^2 \end{bmatrix}. \end{aligned} \tag{8}$$

Since $\boldsymbol{Y}$ is a symmetric matrix, we only consider its upper triangular elements in the subsequent processing. This simple step reduces the feature dimensionality from $\bar{c}^2$ to $\bar{c}\cdot(\bar{c}+1)/2$, thereby resulting in faster and efficient processing in the subsequent modules (Refer to Fig. 6). Specifically, we perform

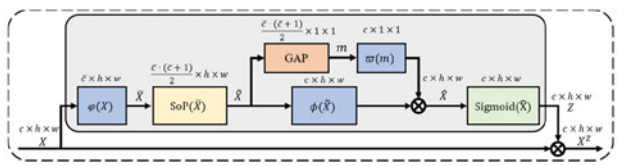
$$\tilde{\boldsymbol{x}} = \mathrm{Vec}\big(\mathrm{UTri}(\boldsymbol{Y})\big), \tag{9}$$

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

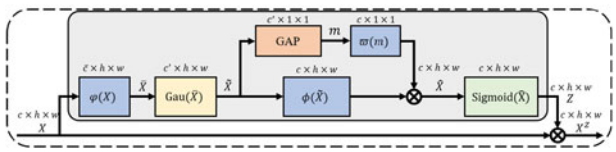
Given the second order features (e.g., $\tilde{\boldsymbol{x}}$) and following the similar aforementioned steps from Eqs. (2) to (4), we propose

$$\mathrm{Sigmoid}\Big(\varpi(\boldsymbol{m}) \otimes \phi(\tilde{\boldsymbol{x}})\Big), \tag{10}$$

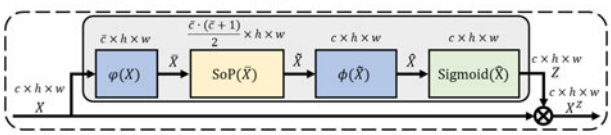
as the attention map for $\boldsymbol{x}$, where $\boldsymbol{m} = \frac{1}{hw}(\sum_{i=1}^{hw} \tilde{\boldsymbol{x}}_i)$. It is worth mentioning that $\boldsymbol{m}$ contains the *second order statistical* information (i.e., the vectorized version of the empirical auto-correlation matrix of $\bar{X}$) of the input to AiA.



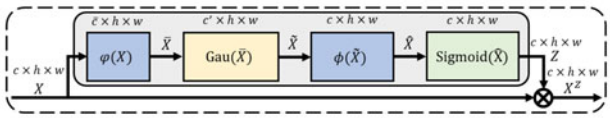
(a) Second-order polynomial attention with AiA



(b) Gaussian attention with AiA



(c) Second-order polynomial attention without AiA



(d) Gaussian attention without AiA

Fig. 5. The structure of AiA modules employing non-linear features in the feature map. (a): Second-order polynomial attention with AiA, (b): Gaussian attention with AiA, (c): Second-order polynomial attention without AiA, (d): Gaussian attention without AiA. $\mathrm{SoP}(\cdot)$ indicates the bilinear pooling and second order feature rearrangement function. $\mathrm{Gau}(\cdot)$ indicates the random Fourier feature mapping function.

Remark 2. The inner product between two vectors is widely used as a means of similarity matching. As an insight on the properties of SoP-attention, consider the inner product between $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$, (i.e., the output of $\mathrm{SoP}(\cdot)$ function)

$$\begin{aligned}\tilde{\boldsymbol{x}}_i^\top \tilde{\boldsymbol{x}}_j &= \mathrm{SoP}(\bar{\boldsymbol{x}}_i)^\top \mathrm{SoP}(\bar{\boldsymbol{x}}_j) \\ &= \mathrm{Vec}\big(\mathrm{UTri}(\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top)\big)^\top \mathrm{Vec}\big(\mathrm{UTri}(\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^\top)\big) \\ &= \sum_u (\bar{x}_{iu}\cdot\bar{x}_{ju})^2 + \sum_u\sum_{s\neq u}(\bar{x}_{iu}\bar{x}_{is}\cdot\bar{x}_{ju}\bar{x}_{js}). \end{aligned} \tag{11}$$

Here, $\bar{x}_{iu}$ is the uth element in vector $\bar{\boldsymbol{x}}_i$. This shows that with second order pooling, one can introduce higher order statistics (e.g., second term in Eq. (11)) into making decisions. This, as we will see empirically, boosts the accuracy of the model substantially.

SoP-attention also has its simplified counterpart, shown in Fig. 5c. This formulation approximately halves the number of parameters of the SoP-attention block, while still benefiting from second order information (using bilinear mapping). Here, the attended feature map is calculated as

$$\boldsymbol{x}^z = \mathrm{Sigmoid}\Big(\boldsymbol{\phi}\big(\mathrm{SoP}(\varphi(\boldsymbol{x}))\big)\Big)\otimes\boldsymbol{x}. \tag{12}$$

3.3 Gaussian Attention

The SoP-attention module requires a large set of parameters if its input feature map is high-dimensional. To address this difficulty, we propose the Gaussian attention or Gau-attention for short (Refer to Fig. 5b for a conceptual diagram). The Gau-attention makes use of the theory of random Fourier features to approximate the infinite dimensional feature space of a Gaussian kernel. This, as will be shown shortly, drastically reduces the number of parameters of the model and required FLOPs (See Table 8 in Section 5).

Given the embedded feature $\bar{\boldsymbol{x}} = \varphi(\boldsymbol{x}) \in \mathbb{R}^{\bar{c}\times h\times w}$, the function $\mathrm{Gau}(\bar{\boldsymbol{x}})$ estimates the Gaussian kernel along each channel, such that

$$\mathcal{K}(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{x}}_j) = e^{-\frac{\|\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}}_j\|^2}{2\sigma^2}} \approx \kappa(\bar{\boldsymbol{x}}_i)^\top\kappa(\bar{\boldsymbol{x}}_j), \tag{13}$$

where $\kappa(\cdot)$ is a randomized embedding. The form of $\kappa(\cdot)$ for a Gaussian kernel [22] is shown below as

$$\kappa(\bar{\boldsymbol{x}}) = \sqrt{\tfrac{1}{c'}}\begin{bmatrix}\cos(\boldsymbol{\omega}_1^\top\bar{\boldsymbol{x}})\\ \vdots \\ \cos(\boldsymbol{\omega}_{c'}^\top\bar{\boldsymbol{x}})\\ \sin(\boldsymbol{\omega}_1^\top\bar{\boldsymbol{x}})\\ \vdots\\ \sin(\boldsymbol{\omega}_{c'}^\top\bar{\boldsymbol{x}})\end{bmatrix} \in \mathbb{R}^{2c'}, \tag{14}$$

where the weights (i.e., $\boldsymbol{\omega}_i,\ i=1,\ldots,c'$) are drawn from the scaled Fourier transformation of a Gaussian kernel. That is, we sample from

$$p(\boldsymbol{\omega}) = (2\pi)^{-\bar{c}/2}\exp\left(-\frac{\|\boldsymbol{\omega}\|^2}{2}\right) = \frac{1}{2\pi}\int e^{-\mathrm{j}\boldsymbol{\omega}^\top\boldsymbol{\delta}}e^{-\frac{\|\boldsymbol{\delta}\|^2}{2\sigma^2}}d\boldsymbol{\delta}. \tag{15}$$

The above processing is summarized as the $\mathrm{Gau}(\cdot)$ function, with $\tilde{\boldsymbol{x}} = \mathrm{Gau}(\bar{\boldsymbol{x}})$. Given the estimated random features (i.e., $\tilde{\boldsymbol{x}}$ or $\kappa(\bar{\boldsymbol{x}})$), Gau-attention generates the attention map and attends to the input feature map, following Eqs. (10) and (4) respectively.

Remark 3. Here, we provide a brief analysis how the $\mathrm{Gau}(\cdot)$ function equips the input feature $\bar{\boldsymbol{x}}$ with the discriminative power of a Gaussian kernel. Given any two random feature vectors, $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$, their similarity matching is calculated as follows:

$$\begin{aligned}&\mathbb{E}\big[\tilde{\boldsymbol{x}}_i^\top\tilde{\boldsymbol{x}}_j\big] = \mathbb{E}\big[\kappa(\bar{\boldsymbol{x}}_i)^\top\kappa(\bar{\boldsymbol{x}}_j)\big] \\ &= \frac{1}{c'}\mathbb{E}\Big[\sum_k\big(\cos(\boldsymbol{\omega}_k^\top\bar{\boldsymbol{x}}_i)\cos(\boldsymbol{\omega}_k^\top\bar{\boldsymbol{x}}_j)+\sin(\boldsymbol{\omega}_k^\top\bar{\boldsymbol{x}}_i)\sin(\boldsymbol{\omega}_k^\top\bar{\boldsymbol{x}}_j)\big)\Big] \\ &= \mathbb{E}\big[\cos(\boldsymbol{\omega}^\top(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}}_j))\big] = \int_{\mathbb{R}^{\bar{c}}} p(\boldsymbol{\omega})e^{\mathrm{j}\boldsymbol{\omega}^\top(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}}_j)}d\boldsymbol{\omega} \\ &= \mathcal{K}(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{x}}_j),\end{aligned} \tag{16}$$

where the last equality follows from the Bochner theorem [22]. In Section 5, we also empirically verify the superior performance of Gaussian attention, which not only saves parameter numbers and computational overhead significantly, but also outperforms the other two in the majority of the experiments.

The simplified version of Gau-attention is shown in Fig. 5d and is denoted as Gaussian attention without AiA, and its formulation is shown as follows:

$$\boldsymbol{x}^z = \mathrm{Sigmoid}\Big(\boldsymbol{\phi}\big(\mathrm{Gau}(\varphi(\boldsymbol{x}))\big)\Big)\otimes\boldsymbol{x}. \tag{17}$$

Remark 4. Similar to the Fully Attentional Block (FAB) [12], both SoP-attention and Gau-attention without AiA modules maintain the spatial structural information of the input feature map. However, unlike FAB that considers only the first order channel information, the aforementioned attention blocks additionally exploit the non-linear channel information in the second-order polynomial and Gaussian kernel spaces, so as to learn a superior discriminative embedding space for the re-ID task.

It is worth mentioning that the proposed 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 the proposed attention modules in the person re-ID application.

4 Attention in Attention Network

In this section, we will first provide an overview of the problem formulation. Subsequently, it will be followed by a detailed description of the architecture of the proposed deep convolutional network, the Attention in Attention Network (AiA-Net).

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 represent the number of channels and its height and

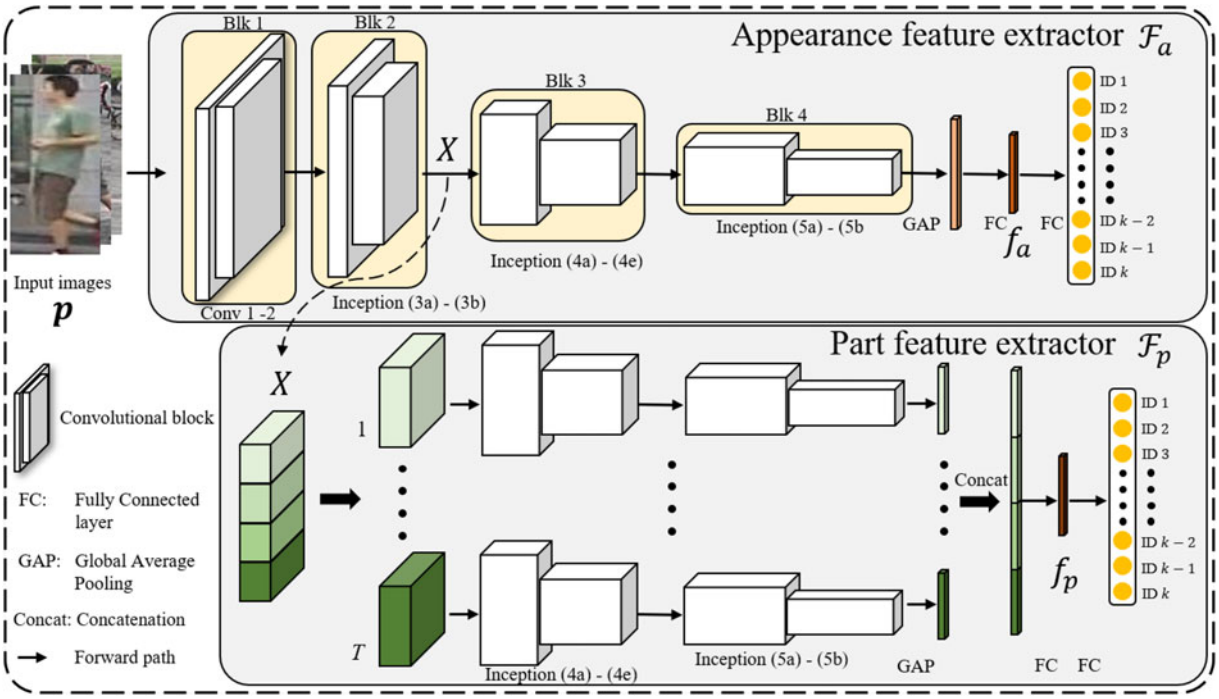


Fig. 7. The deep architecture of the proposed feature extractor. AiA-Net has two feature extractors, e.g., the person appearance feature extractor (i.e., $\mathcal{F}_a$) and the part feature extractor (i.e., $\mathcal{F}_p$). f_a and f_p are concatenated to give the final person representation as $f = [f_a^\top, f_a^\top]^\top$.

width, respectively. Each image $\boldsymbol{p}_i$ is labeled by its identity, denoted as $y_i \in \{1, \ldots, k\}$, where k represents the total number of distinct identities of the training data. Thus, the training set with N_{train} images, can be represented as $\{\boldsymbol{p}_i, y_i\}_{i=1}^{N_{\text{train}}}$. 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 our work, the final embedding space is obtained by concatenating the person-appearance embedding space, i.e. $f_a = \mathcal{F}_a(\boldsymbol{p}, \theta_a)$, and the person-part embedding space, i.e. $f_p = \mathcal{F}_p(\boldsymbol{p}, \theta_p)$, such that $\mathcal{F}(\boldsymbol{p}, \theta) = [f_a^\top,\ f_p^\top]^\top$.

4.2 Overview

The AiA-Net has two feature extractors, namely, (*1*) a person-appearance feature extractor (denoted by $\mathcal{F}_a$) and (*2*) a person part-feature extractor (denoted by $\mathcal{F}_p$). The overall architecture of the AiA-Net is shown in Fig. 7. The person holistic appearance is encoded by the appearance feature extractor; while the part feature extractor aims at encoding the different parts of the person.

The appearance feature extractor consists of 4 convolutional blocks. After each convolutional block, an AiA block is added to align the feature map and highlight its discriminative regions. The attended feature map encourages the network to learn a holistic representation (i.e., f_a in Fig. 7) of the person.

Recent studies of the person re-identification task suggest that an independent modeling of the part regions can enhance the precision of the overall system [3], [4], [11]. We also equip the AiA-Net with a parts-based learning ability. More specifically, we use a simple sub-network as a part feature extractor, which aims at learning distinct and discriminative parts in the input image. In the part feature extractor, the aligned feature map $\boldsymbol{X} \in \mathbb{R}^{c \times h \times w}$ is divided into T non-overlapping regions $\boldsymbol{X}_t$ s.t. $\boldsymbol{X}_t \in \mathbb{R}^{c \times \frac{h}{T} \times w}, t = 1, \ldots, T$. Each of the non-overlapped regions is resized to $c \times h \times w$ by bilinear interpolation and fed to the tth stream of the part feature extractor network; which generates the part-feature embedding. Then, T part features are concatenated to represent the final person part representation (i.e., f_p in Fig. 7).

Remark 5. Our part feature extractor network is different from the current part-based solutions [3], [4], [11], [63], [64]. For example, in [3], the part feature is extracted via a pose estimation network called OpenPose [65]. Zhao *et al.* uses an implicitly defined part detector to align the part features [64]. In [11], the parts are sampled through a hard attention network. In [4], [63], the parts are split evenly in the final feature map. In addition to the structural differences, each part model within the AiA-Net works independently from the others as no weights are shared between them. This, in turn, leads to an increased diversity of the learned parts, thereby learning 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. Yu *et al.* uses cross-entropy loss for the classification task and triplet loss for the ranking task [66]. Mancs combines the triplet loss, focal loss and cross-entropy loss and learns a superior embedding space for person re-iD against the baseline algorithms [12]. Recent works in [67], [68] also show person re-ID can benefit from

various regularization, e.g., L2 regularization, angular regularization *etc*. Following the protocol prescribed in [66], we train our network for the tasks of ranking and classification. The ablation study of MTT is reported in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3073512.

Ranking Task. We use the well studied triplet loss for the ranking task. In a mini-batch, $\{\boldsymbol{p}_i\}_{i=1}^{N_{\text{batch}}}$, a 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:

$$\mathcal{J}_{\text{rank}} = \frac{1}{N_{\text{tri}}} \sum_{i=1}^{N_{\text{tri}}} \left[d_i^+ - d_i^- + \tau \right]_+, \tag{18}$$

where $[\cdot]_+ = \max(\cdot, 0)$, N_{tri} indicates the number of triplets within one batch, τ is a margin. $d_i^+ = \|\mathcal{F}(\boldsymbol{p}_i) - \mathcal{F}(\boldsymbol{p}_i^+)\|_2$, and $d_i^- = \|\mathcal{F}(\boldsymbol{p}_i) - \mathcal{F}(\boldsymbol{p}_i^-)\|_2$. In the triplet mining, for each anchor, we mine one hard positive and 5 hard negatives, thus obtaining 5 triplets per anchor sample. This mining strategy is to avoid collapsing to local minima in the early stages of optimization [69].

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

$$\mathcal{J}_{\text{cls}} = \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} -\log\big(p(y_i|\mathcal{F}(p_i))\big), \tag{19}$$

where $p(y_i|\mathcal{F}(p_i))$ is the predicted probability that p_i belongs to identity y_i, and N_{batch} is the number of samples in one mini-batch.

4.4 Implementation Details

Network Architecture. We implemented our AiA-Net model in the PyTorch [70] deep learning framework. The backbone network is the GoogLeNet-V1 [71], pretrained on ImageNet [72] with Batch Normalization [73]. 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 fully connected (FC) layer is connected to predict the person identity using the person appearance embedding. In the part feature extractor, we follow the work in [11], 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 the person identity using the person part embedding. During testing, f_a and f_p are concatenated to give the final person representation f, where $f = [f_a^\top, f_a^\top]^\top \in \mathbb{R}^{1024}$. The study of the choice of T and size of Dim is reported in the supplementary material, available online.

In the AiA block, the embedding functions $\varphi(\cdot)$, $\phi(\cdot)$ and $\varpi(\cdot)$ are 1×1 convolutional layers, followed by a 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 [23] and CUHK01 [27] datasets, and to 4 for the other datasets. The dimension of the random feature (i.e. c') in Eq. (14) is set to 960 for DukeMTMC-reID dataset and to 480 for the other datasets. The details of the datasets will be presented in Section 5.1.

Network Training. We use the Adam [74] optimizer with the default momentum values of (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 [23] and CUHK01 [27], and 5×10^{-4} for Market-1501 [24], DukeMTMC-reID [25] and MSMT17 [26]. The size of the mini-batch (i.e., N_{batch} in Eq. (19)) is set to 64 for all experiments. 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 the appearance and part feature extractors separately; this is inspired by [4] where supervision on each respective feature extractor is vital for learning discriminative features. In the triplet loss, we set the margin (i.e., τ in Eq. (18)) to 1.5 for the CUHK03 and CUHK01 datasets and 1 for the other datasets. We randomly apply horizontal flip to the input images. Similar to [75], we also apply random erasing [76] after 50 epochs of training in order to avoid any local optima within the embedding space. No such augmentations are used during the testing phase. We report the performance of the network after training it for 250 epochs. Moreover, it is worth noting that we do not apply any re-ranking algorithms to boost the ranking result in the testing phase.

5 Experiment

5.1 Datasets

In this section, we evaluate our proposed algorithm across four large scale datasets, i.e., *CUHK03* [23], *Market-1501* [24], *DukeMTMC-reID* [25] and *MSMT17* [26], as well as one small scale dataset, i.e. *CUHK01* [27]. In the supplementary, available online, we will show samples from the aforementioned datasets.

CUHK03. This dataset consists of 13,164 person images of 1,467 identities, captured by 6 non-overlapping cameras. Each person is observed by two disjoint camera views. CUHK03 offers both hand-labeled and deformable part model (DPM)-detected [77] bounding boxes, and we evaluate our model on both sets. In the CUHK03 dataset, there are two training/testing protocols. In the vanilla training protocol, the training set contains 1,367 identities, while the remaining 100 identities constitute the test set. However in the new protocol [78], the training set contains 767 identities and the testing set contains the remaining 700 identities. In this paper, we adopt both the protocols to verify the effectiveness of the proposed attention blocks.

Market-1501. Market-1501 is one of the most popular re-ID dataset which consists of 32,668 person images of 1,501 identities observed under a maximum of 6 different cameras. The dataset is split into 12,936 training images of 751 identities and 19,732 testing images of the remaining 750

identities. Both the training and testing images are detected using a DPM [77]. In this dataset, we use both the single query and the multi query setting to evaluate our algorithm.

DukeMTMC-reID. This dataset is collected using 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. In this dataset, the person bounding boxes are manually labeled.

MSMT17. This is the largest person re-ID dataset, consisting of 126,441 person images from 4,101 different identities, which are detected using Faster R-CNN [79]. This dataset is collected with using 15 different cameras. 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.

CHUK01. This is a small scale person re-ID dataset, which contains 3,884 images of 971 identities. The person images are captured by two cameras with each person having two images in each camera view. The person bounding boxes are labeled manually. We adopt the 485/486 training protocol to evaluate our network.

We use both the 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 ranks, whereas the mAP measures the probability of all correct matches in the gallery images for a given query image, thereby measuring the overall ranking performance.

5.2 Ablation Study

We first perform experiments to verify the effectiveness of our proposed AiA mechanism and its variants on CUHK03, Market-1501, DukeMTMC-reID and MSMT17 under the single query setting (i.e., SQ). For the CUHK03 dataset, we use the most difficult setting, i.e., the new protocol with detected bounding boxes (i.e., ND).

5.2.1 Effect of the Proposed Feature Extractor

In the field of person retrieval, ResNet-50 [80] and GoogLeNet [71] are the most commonly used backbones [3], [12], [81]. Since we also want the network to own the capacity of learning part features, the part feature extractor is further developed. We compare the performance of the ResNet-50 and GoogLeNet, with each equipped with the part feature extractor. As suggested in Table 2, we could observe that: *(1)* the retrieval accuracy increases when the GoogLeNet is equipped with the part feature extractor, thereby showing that our design is indeed effective in exploiting the complementary information between the two feature extractors. *(2)* GoogLeNet + part feature extractor is superior to the ResNet-50 counterpart in both the performance and the network size. Hence, Hence, we use the ImageNet pre-trained GoogLeNet against the ResNet-50 in our experiments. In the rest of the paper, the GoogLeNet and part feature extractor are represented by $\mathcal{F}_a$ and $\mathcal{F}_p$, respectively. The results under ResNet-50 are reported in the supplementary material, available online.

TABLE 2
Result of Various Backbone Networks on the CUHK03 and Market-1501 Datasets

Model	CUHK03 @ ND		Market @ SQ		PNs
	mAP	R-1	mAP	R-1	($\times 10^6$)
GoogLeNet	64.5	67.1	80.7	91.6	9.45
+ Part feature extractor	**67.8**	**71.1**	**85.1**	93.8	30.16
ResNet-50	64.0	67.6	85.0	**94.5**	25.61
+ Part feature extractor	65.3	68.2	84.1	94.2	46.84

PNs: parameter numbers. We use the bold *to indicate the best result in each category.*

TABLE 3
Effect of the Attention in Attention Nechanism on the CUHK03 and Market-1501 Datasets

Model	CUHK03 @ ND		Market-1501 @ SQ		PNs	Inf-time
	mAP	R-1	mAP	R-1	($\times 10^6$)	(ms)
$\mathcal{F}_a$	64.5	67.1	80.7	91.6	9.45	3.2
Lin-attention w/o AiA	64.8	67.9	80.9	92.4	0.12	3.2
Lin-attention w/ AiA	**66.8**	**70.4**	**82.5**	**92.7**	0.18	3.4
SE block [14]	65.2	68.3	81.2	92.4	0.12	3.4
NL block [43]	65.6	68.9	81.4	92.0	0.23	3.8
$\mathcal{F}_a + \mathcal{F}_p$	67.8	71.1	85.1	93.8	30.16	8.4
Lin-attention w/o AiA	68.5	71.4	85.3	94.0	0.12	8.4
Lin-attention w/ AiA	**72.2**	**75.1**	**87.1**	**94.7**	0.18	8.9
SE block [14]	70.7	73.0	86.3	94.4	0.12	8.6
NL block [43]	70.9	72.9	86.8	94.5	0.23	9.2

PNs: parameter numbers; Inf-time: inference time. We use the bold *to indicate best result in each category.*

5.2.2 Effect of the Attention in Attention Mechanism

We then evaluate the effectiveness of the proposed AiA mechanism and use the Linear attention for this study on the CUHK03 and Market-1501 datasets. In this study, we compare the Lin-attention without AiA and with AiA employed in the two feature extractors, e.g., $\mathcal{F}_a$ and $\mathcal{F}_a + \mathcal{F}_p$. The attention block is added after the second convolutional block (i.e., Blk 2 in Fig. 7). In the attention block, we use the identical dimensionality reduction factor, i.e., $r = 4$. The results are listed in Table 3. The table shows that: Addition of Lin-attention with and without AiA leads to an increase in the retrieval accuracy across either of the feature extractors, with the former outperforming the latter in terms of mAP and R-1 values respectively. This indeed verifies the design intuition of the AiA mechanism. Further, we replace the Lin-attention block by other popular attention blocks, e.g., the Squeeze-and-Excitation (SE) block [14] and the Non-local (NL) block [43], in the same position of the feature extractor (i.e., $\mathcal{F}_a$ and $\mathcal{F}_a + \mathcal{F}_p$). We set the dimensionality reduction factor as 4 in both SE and NL blocks. In this study, we also compare the parameter numbers and inference time of attention networks. As suggested in Table 3, our attention outperforms the other two significantly without bringing any additional heavy computational cost,[4] thereby verifying the effectiveness of our proposed AiA mechanism.

To further verify the superiority of the AiA block, we compare the learned attention between Lin-attention and its

4. In the inference time, we calculate the averaging inference time per image on NVIDIA GeForce RTX TITAN V

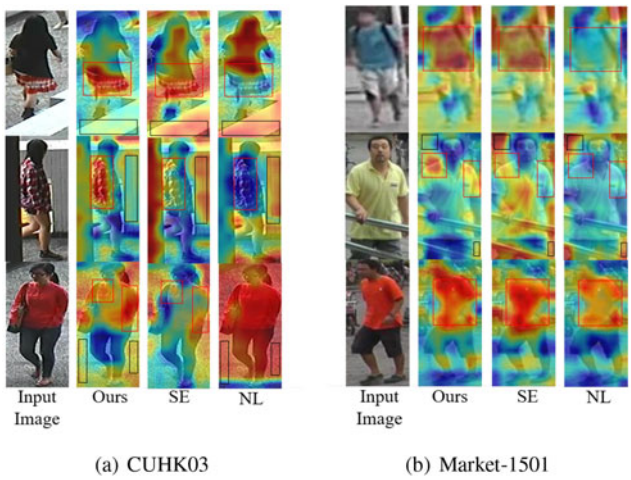


Fig. 8. Comparison of the learned attention on CUHK03 (a) and Market-1501 (b) datasets. In each dataset, we compare the the feature map from Lin-attention and its alternatives (e.g., SE block and NL block). In the heat map, the response increases from blue to red. Best viewed in color.

alternatives (e.g., SE block and NL block) in Fig. 8. We sample the person images in CUHK03 and Market-1501 datasets. Fig. 8 shows that our AiA block either highlights the informative foreground (denoted by red rectangles) or filters the non-informative background areas (denoted by black rectangles), thereby clearly demonstrating the benefits of the AiA mechanism.

5.2.3 Effect of Employing Non-Linear Features in Attention

Then, we study the effect of using non-linear features for attention design on the baseline network $\mathcal{F}_a + \mathcal{F}_p$. In this study, we first evaluate that the AiA framework benefits from the manual non-linear features in RKHSs (i.e., SoP-attention w/AiA and Gau-attention w/AiA). We also verify that the manual non-linear features are superior to the learned non-linear features. We has two settings of learned non-linear feature: one is naive nonlinear activations and another one is a stack of nonlinear activations. They are denoted by non-linear attention V1 and non-linear attention V2, respectively. Note that both the two versions of the attention block are incorporated into the AiA framework.

The results on CUHK03 and Market-1501 datasets are shown in Table 4. It is observed that the non-linear features, modeled by bilinear mapping and random Fourier features, has superior performance compared to their linear counterpart, thereby highlighting the importance of using non-linear features to locate the highly discriminative regions in the input feature map. In addition, we also observe that AiA-Net with Gau-attention has superior performance over the other two attention variants across both the datasets, which reveals that Gau-attention can learn more complicated non-linear functions than the other two attention blocks. Table 4 also reveals that both the versions of learned non-linearity in AiA achieve similar performance to the Lin-attention with AiA, while the manual non-linear features improves the performance over its linear counterpart, showing the advantage of manually designed non-linearity. It might be that the manual ones enjoy high discrimination power in RKHSs, and are easier to optimize, as compared to the learned non-linear features. In the supplementary material, available online, we also study the effect of the interactive terms in Eq. (8) of SoP-attention block.

TABLE 4
Effect of the Learned Non-Linearity in Attention Mechanism on the CUHK03 and Market-1501 Datasets

Model	CUHK03 @ ND		Market-1501 @ SQ		PNs
	mAP	R-1	mAP	R-1	($\times 10^6$)
$\mathcal{F}_a + \mathcal{F}_p$	67.8	71.1	85.1	93.8	30.16
Lin-attention w/ AiA	72.2	75.1	87.1	94.7	0.18
SoP-attention w/ AiA	73.2	76.2	87.4	95.1	1.79
Gau-attention w/ AiA	74.0	76.8	87.5	95.2	0.58
Non-linear attention V1	72.7	75.0	87.0	94.6	0.69
Non-linear attention V2	72.4	74.9	86.9	95.0	0.60

PNs: parameter numbers.

5.2.4 Effect of the Dimensionality Reduction Factor

In the section, we study the effect of the reduction factor r in the embedding function $\varphi(\cdot)$ on CUHK03 and Market-1501 datasets. All the experiments for this study are conducted using the SoP-attention with AiA, as r is an important hyperparameter that directly affects the information pooled by the bilinear operation. The results and their comparisons, as shown in Table 5, reveal that: *(1)* even though r is an important parameter, which influences the size of the attention model (i.e., the learnable parameters within $\varpi(\cdot)$, $\phi(\cdot)$), our network has a weak dependency on r as changes in r lead to minuscule changes in the performance of our network across all datasets. *(2)* We further observe that while $r = 4$ obtains the best results in the large datasets (i.e., Market-1501, DukeMTMC-reID and MSMT17), the best value of r is observed to be 8 when the network is trained on CUHK03. One plausible explanation is that the network trained on the large datasets is less prone to over-fitting due to its larger training set in comparison to CUHK03. The study on DukeMTMC-reID and MSMT17 datasets is reported in the supplementary material, available online.

5.2.5 Effect of the Dimensionality in Random Features

In $\mathrm{Gau}(\cdot)$, we approximate the channel features in the Gaussian kernel space via a random Fourier mapping. Therefore, we study the result of varying the dimensionality of the random feature (i.e., c') in this section. Here, we have set r to 4 in the embedding function $\varphi(\cdot)$. The results are

TABLE 5
Effect of the Dimensionality Reduction Factor r in the Embedding Function $\varphi(\cdot)$ on the CUHK03 and Market-1501 Datasets

Model	CUHK03 @ ND		Market-1501 @ SQ	
	mAP	R-1	mAP	R-1
$\mathcal{F}_a + \mathcal{F}_p$	67.8	71.1	85.1	93.8
$r = 2$	72.3	75.4	87.1	94.9
$r = 4$	72.6	74.9	**87.4**	**95.1**
$r = 8$	**73.2**	**76.2**	87.2	94.5
$r = 16$	72.5	75.6	86.9	94.4
$r = 32$	72.1	74.8	86.9	94.1

We use the bold *to indicate best the result in each category.*

TABLE 6
Effect of the Dimensionality c' in Random Features on the CUHK03 and Market-1501 Datasets

Model	CUHK03 @ ND		Market-1501 @ SQ	
	mAP	R-1	mAP	R-1
$\mathcal{F}_a + \mathcal{F}_p$	67.8	71.1	85.1	93.8
$c' = 120$	72.7	75.3	86.7	94.7
$c' = 240$	73.1	75.9	87.1	94.8
$c' = 480$	**74.0**	**76.8**	**87.5**	**95.2**
$c' = 960$	72.9	75.6	87.3	95.0

We use the bold *to indicate the best result in each category.*

shown in Table 6. One can observe that: along any dimension value (i.e., c'), the random Fourier feature helps to improve retrieval performance of the network. In addition, the network attains the best performance when $c' = 480$ for both datasets. Further, there is a negligible change in the performance of our proposed network with changes in c', thus clearly demonstrating the weak dependency of AiA-Net on c'. The study on DukeMTMC-reID and MSMT17 datasets is reported in the supplementary material, available online, and it can be observed that the network attains the best performance when $c' = 960$ for DukeMTMC-reID dataset and $c' = 480$ for MSMT17 dataset.

5.2.6 Effect of the Position of the Attention Block

Table 7 shows the effect of adding the Lin-attention with the AiA block to different positions along the baseline network on the CUHK03 and Market-1501 datasets. 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. 7). Table 7 shows that: *(1)* using Lin-attention in the early stages, i.e. p_1, p_2, is superior to using it in the later stages i.e. p_3, p_4. A similar observation is also made in [43], where the non-local block enhances the performance of ResNet [80] in its early stages. (2) Moreover, the performance of adding Lin-attention in p_2 surpasses the performance compared to when it is added in p_1. One reasonable explanation is that the feature maps at p_2 consist of richer channel, as well as spatial, structural information in comparison to the feature maps at p_1, thereby enabling the network to emphasize more on the discriminative areas of the images. *(3)* In the CUHK03 dataset, which has a smaller training set, the performance of person retrieval degrades when Lin-attention is inserted at p_4. This is observed as the embedding layer of the Lin-attention module overfits on the training set due to the high dimensionality of the feature map at p_4. *(4)* It is also observed that the network with multiple attention blocks can further bring performance gain. In the rest of the paper, AiA-Nets indicate plugging multiple attention blocks along with the baseline network (i.e., $\mathcal{F}_a + \mathcal{F}_p$).

TABLE 7
Effect of the Position of the AiA Block on the CUHK03 and Market-1501 Datasets

Model	CUHK03 @ ND		Market-1501 @ SQ	
	mAP	R-1	mAP	R-1
$\mathcal{F}_a + \mathcal{F}_p$	67.8	71.1	85.1	93.8
p_1	71.2	73.1	86.5	94.1
p_2	72.2	75.1	87.1	94.7
p_3	69.1	72.4	85.6	93.9
p_4	68.6	70.9	85.1	93.8
p_1 - p_4	**72.8**	**75.8**	**87.2**	**95.0**

Here, we use Lin-attention in AiA-Net. We use the bold *to indicate the best result in each category.*

TABLE 8
Computational Complexity and Module Size of Proposed Attention Modules

	Lin-attention	SoP-attention	Gau-attention	$\mathcal{F}_a + \mathcal{F}_p$
Hyper Parameter	$r = 4$	$r = 8$	$r = 4$, $c' = 480$	-
FLOPs ($\times 10^9$)	0.015	0.117	0.044	2.82
PNs ($\times 10^6$)	0.18	1.79	0.58	30.16

FLOPs: the number of floating-point operations; PNs: number of parameters.

5.2.7 Computational Complexity and Model Size

In Section 5.2.3, we have studied the effect of non-linearity within the AiA module. In this part, we study the block properties (i.e., computational complexity and module size) of each of the AiA blocks and the baseline network (i.e., $\mathcal{F}_a + \mathcal{F}_p$). The computational complexity and model size are measured by the number of floating-point operations (FLOPs) and parameter numbers (PNs) respectively. This study is performed on the CUHK03 dataset and the results are shown in Table 8, along with the parameter settings of each attention block. The size of the input feature map to the attention block and input image to baseline network are set to $480 \times 16 \times 8$ and $3 \times 256 \times 128$ respectively. Table 8 depicts that: *(1)* compared against the baseline network, the computational complexity and model size of the attention blocks are insignificant, indicating that the performance gain significantly relies on the attention mechanism, rather than increasing the number of parameters. *(2)* Lin-attention and Gau-attention are light weight attention blocks, which can be used in other resource-constrained applications. *(3)* Taking into account the results obtained in Table 4, it is clearly observed that Gau-attention is superior to the SoP-attention as it results in a large performance gain (See Table 4), while using significantly fewer number parameters than the SoP-attention (i.e., only $1/3$ of the number of parameters of SoP-attention). This clearly indicates the hidden potential of the use of non-linear features in the Gaussian kernel space in attention design.

5.3 Comparison With State-of-the-Art Methods

To show the superiority of the proposed deep architecture, we compare the performance of AiA-Nets with the current state-of-the-art methods across five datasets.

CUHK03. In the CUHK03 dataset, we evaluate our network under all data settings, that is, both labeled and detected data for the two training set protocols. Tables 9 and 10 show the results for both training protocols. We observe that our methods outperform the current state-of-the-art results in vanilla setting and achieve competitive results in the new setting. In the vanilla training set protocol (Refer to Table 9), our AiA-Net with Gau-attention improves over the state-of-the-art result by 0.9/7.6 percent on mAP for labeled and detected sets, respectively. With

TABLE 9
Evaluation on the CUHK03-Vanilla Dataset in Both Labeled and Detected Bounding Box

Model	@ Labeled		@ Detected	
	mAP	R-1	mAP	R-1
DKPM [82]	89.2	91.1	-	-
IANet [83]	-	92.4	-	90.1
MVP Loss [84]	-	93.7	-	91.8
SGGNN [85]	94.3	95.3	-	-
MuDeep [13]	-	95.8	-	93.7
AiA-Net w/ Lin-attention	94.8	96.1	91.5	93.6
AiA-Net w/ SoP-attention	**95.2**	**96.8**	92.1	94.0
AiA-Net w/ Gau-attention	94.9	96.6	**92.4**	**94.1**

We use the bold *to indicate the best result in each category.*

TABLE 10
Evaluation on the CUHK03-New Dataset in Both Labeled and Detected Bounding Box

Model	@ Labeled		@ Detected	
	mAP	R-1	mAP	R-1
HPM [86]	-	-	57.5	63.9
Mancs [12]	63.9	69.0	60.5	65.5
OSNet [87]	-	-	67.8	72.3
Auto-ReID [88]	73.0	77.9	69.3	73.3
RGA [89]	77.4	**81.1**	74.5	**79.6**
AiA-Net w/ Lin-attention	76.4	79.1	72.8	75.8
AiA-Net w/ SoP-attention	77.0	79.4	74.2	76.9
AiA-Net w/ Gau-attention	**77.6**	80.6	**74.8**	77.8

We use the bold *to indicate the best result in each category.*

respect to the R-1 value, our network beats the current state-of-the-art result by 1.0/0.4 percent across the labeled and detected sets. In the new training set protocol (Refer to Table 10), our AiA-Net improves the present state-of-the-art mAP value by 0.2/0.3 percent and achieves competitive results on R-1 value. This validates the utility of our design choices in AiA-Net along with the importance of the various attention modules to obtain a superior discriminative embedding for person-retrieval.

Market-1501. We further evaluate our proposed AiA-Net against the recent state-of-the-art methods on the Market-1501 in both the single query and multi query settings. The results are shown in Table 11. In the single query setting, our method (e.g., AiA-Net w/ Gau-attention) achieves very competitive results over the RGA and ABD-Net. Moreover, our AiA-Nets with Lin-attention, SoP-attention and Gau-attention outperform the present state-of-the-art Mancs by 3.7, 4.0 and 4.3 percent on mAP, and by 0.4, 0.7 and 1.2 percent on R-1, respectively in the multi query setting.

DukeMTMC-reID. The evaluation of our proposed algorithm on DukeMTMC-reID is shown in Table 11. It is obvious that our AiA-Nets obtain a competitive performance with respect to mAP and R-1 value. The AiA-Net with Gau-attention improves over DG-Net by 3.1 percent on mAP and 1.6 percent on Rank-1 accuracy. As for ABD-Net, AiA-Net with Gau-attention has competitive performance on the R-1 value (88.8 versus 89.0 percent), while achieving the same performance on mAP value. It is worth mentioning that ABD-Net uses larger image sizes, which demands more computation resources.

MSMT17. Table 11 shows the result of our proposed network on the challenging MSMT17 dataset. As observed, our proposed networks outperform RGA by 1.3 percent on mAP value and. However, the present state-of-the-art method (i.e., ABD-Net) beats our network considerably.

CUHK01. Besides learning a discriminative feature representation on large scale datasets, we also compare the performance of the AiA-Nets against the state-of-the art algorithms in the CUHK01 benchmark dataset, thereby demonstrating the generalization ability of our proposed networks in learning discriminative representations on a small scale dataset. Table 12 compares our AiA-Net with current state-of-the-art methods. We observe that each of the AiA-Nets outperform the existing state-of-the-art

TABLE 11
Evaluation on the Market-1501, DukeMTMC-reID, and MSMT17 Datasets

Model	Market-1501 @ SQ				Market-1501 @ MQ				DukeMTMC-reID @ SQ				MSMT17 @ SQ			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
MSCAN [49]	57.5	80.3	-	-	66.7	86.8	-	-	-	-	-	-	-	-	-	-
SVDNet [35]	62.1	82.3	92.3	95.2	-	-	-	-	56.8	76.7	86.4	89.9	-	-	-	-
PDC [2]	63.4	84.1	92.7	94.9	-	-	-	-	-	-	-	-	-	-	-	-
DaRe [33]	69.9	86.0	-	-					56.3	74.5	-	-	-	-	-	-
AOS [76]	70.4	86.5	-	-	88.6	92.5	-	-	62.1	79.2	-	-	-	-	-	-
MLFN [91]	74.3	90.0	-	-	82.4	92.3	-	-	62.8	81.0	-	-	-	-	-	-
DKPM [83]	75.3	90.1	96.7	97.9	-	-	-	-	63.2	80.3	89.5	91.9	-	-	-	-
HA-CNN [11]	75.7	91.2	-	-	82.8	93.8	-	-	63.8	80.5	-	-	-	-	-	-
PBR [3]	79.6	91.7	96.9	98.1	85.2	94.0	98.0	98.8	64.2	82.1	-	-	-	-	-	-
DuATM [51]	76.6	91.4	97.1	-	-	-	-	-	64.6	81.8	90.2	-	-	-	-	-
PCB+RPP [4]	81.6	93.8	97.5	98.5	-	-	-	-	69.2	83.3	-	-	-	-	-	-
Mancs [12]	82.3	93.1	-	-	87.5	95.4	-	-	71.8	84.9	-	-	-	-	-	-
SGGNN [86]	82.8	92.3	96.1	97.4	-	-	-	-	68.2	81.1	88.4	91.2	-	-	-	-
HPM [87]	82.7	94.2	97.5	98.5	-	-	-	-	74.3	86.6	-	-	-	-	-	-
IANet [84]	83.1	94.4	-	-	-	-	-	-	73.4	87.1	-	-	46.8	75.5	85.5	88.7
AANet [9]	83.4	93.9	-	98.5	-	-	-	-	74.3	87.6	-	-	-	-	-	-
OSNet [88]	84.9	94.8	-	-	-	-	-	-	73.5	88.6	-	-	52.9	78.7	-	-
DG-Net [82]	86.0	94.8	-	-	-	-	-	-	74.8	86.6	-	-	52.3	77.2	87.4	90.5
ABD-Net [92]	88.3	95.6	-	-	-	-	-	-	**78.6**	**89.0**	-	-	**60.8**	**82.3**	-	-
RGA [90]	**88.4**	**96.1**	-	-	-	-	-	-	-	-	-	-	57.5	80.3	-	-
AiA-Net w/ Lin-attention	87.2	95.0	97.4	98.5	91.2	95.8	98.6	99.2	77.3	88.0	94.6	96.0	56.2	78.2	88.1	90.6
AiA-Net w/ SoP-attention	87.4	95.3	98.5	99.2	91.5	96.1	98.9	99.3	77.5	88.2	94.8	96.4	57.6	79.6	89.3	91.4
AiA-Net w/ Gau-attention	87.9	95.6	**98.5**	**99.1**	**91.8**	**96.6**	**99.0**	**99.6**	**78.6**	88.8	**94.9**	**96.7**	58.8	80.0	**89.7**	**92.0**

In the Market-1501 dataset, we apply both single query and multi query to evaluate the model. We use the bold *to indicate the best result in each category.*

TABLE 12
Evaluation on the CUHK01 Dataset

Model	R-1	R-5	R-10	R-20
DGD [34]	66.6	-	-	-
Zhao *et al.* [64]	75.0	93.5	95.7	97.7
Spindle Net [92]	79.9	94.4	97.1	98.6
PBR [3]	80.7	94.4	97.3	98.6
Baseline ($\mathcal{F}_a + \mathcal{F}_p$)	82.0	94.4	97.7	99.0
AiA-Net w/ Lin-attention	82.8	94.7	97.7	99.0
AiA-Net w/ SoP-attention	83.5	**95.6**	97.9	**99.3**
AiA-Net w/ Gau-attention	**83.9**	95.5	**98.0**	**99.3**

We use the bold *to indicate best the result in each category.*

approach (i.e., PBR) by a large margin. In particular, our three AiA-Nets with Lin-/Sop-/Gau-attention improve the state-of-the-art accuracy by 2.1, 2.8 and 3.2 percent on R-1. It is also noted that PBR is pre-trained on the CHUK03 dataset and further fine-tuned on the CUHK01 dataset to avoid over-fitting, while our network is solely trained on the CUHK01 dataset. This indeed shows that our network is able to generalize well while trained on a small dataset from scratch without the need of any such pre-training step.

5.4 Experiments on Video Person Retrieval

In this section, we further evaluate our AiA modules on the video person retrieval setting on the MARS [28] benchmark dataset. The network architecture and training details are given in the supplementary material, available online. Table 13 compares the result of our approach against the current state-of-the-art algorithms. It clearly shows that our AiA-Net-V improves the state-of-the-art mAP value by 1.0 percent and the R-1 value by 0.2 percent. It should be noted that AiA-Net-V only considers the spatial information in each frame to calculate the attention values and unlike [41], [52], [53], it doesn't take into account the modeling of the temporal attention to fuse the frame features. This improvement clearly shows that our AiA-Net-V makes better use of spatial structure information and attends to the informative areas in each frame.

5.5 Visualization of the Attention in Attention Module

We visualize the heat maps of the input (i.e., X) and output (i.e., X^z) of the Gau-attention block for person images in both the CUHK03 detected-set in Fig. 9a and Market-1501 dataset in Fig. 9b. In each dataset, from left to right, (1): the input person image, (2): the input feature map to attention block, and (3): the masked feature map from attention block. In (2), we use black rectangles to bound the non-informative background clutters in images, which will be filtered by attention block. In (3), we use red rectangles to bound the discriminative parts of the person body parts, which are further emphasized by attention blocks. This visualization indeed reveals the proposed AiA can focus on the discriminative areas of person images, thereby aligning the feature maps.

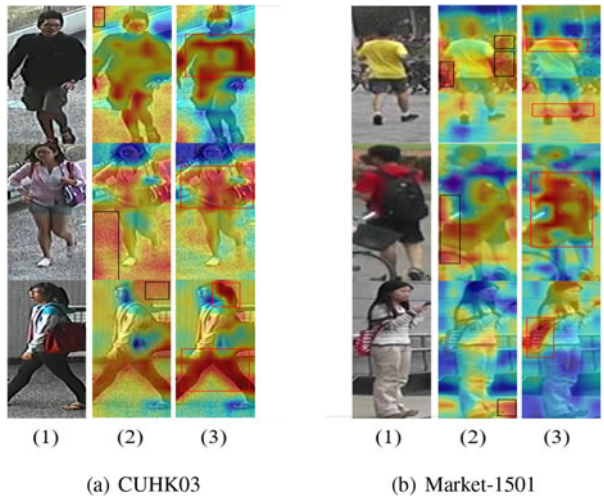


Fig. 9. Visualization of the attention mechanism in person images, sampled from the CUHK03 dataset (a) and the Market dataset (b). In each dataset, from left to right, (1) the input person image, (2) the input feature map to attention and (3) the masked feature map. The heat maps are generated in AiA-Net with Gau-attention. In the heat map, the response increases from blue to red. Best viewed in color.

TABLE 13
Evaluation on the MARS Dataset in Video Person Retrieval Setting

Model	mAP	R-1	R-5	R-10
PBR [3]	72.2	83.0	92.8	95.0
Zhao *et al.* [10]	78.2	87.0	95.4	-
GLTR [41]	78.4	87.0	95.7	-
COSAM [93]	79.9	84.9	95.5	-
STA [53]	80.8	86.3	95.7	-
Baseline	77.3	83.1	94.2	96.0
AiA-Net-V w/ Lin-attention	81.3	86.4	94.7	96.7
AiA-Net-V w/ SoP-attention	**81.8**	86.7	95.4	97.0
AiA-Net-V w/ Gau-attention	81.7	**87.2**	**95.6**	**97.2**

We use the bold *to indicate the best result in each category.*

5.6 Discussion

Statistical Significance of the Proposed Method. In Sections 5.2 and 5.3, a thorough study has been studied to verify the superiority of proposed attention blocks. We further study their statistical significance using t-test. We adopt the AiA-Net w/Gau-attention and CUHK03 (See Table 9) in this study, and we obtain the p-value of 0.0026/0.0033, meaning that our results are significant ($p < 0.05$ is significant). Thus we believe that our AiA-Net is superior to the MuDeep [13]. We also plug the Gau-attention with AiA to the ResNet-50 backbone, The results of our AiA read 96.1/93.9 as compared to 95.8/93.7 of MuDeep, again showing the superiority of the AiA block. In this study, the p-values are 0.0003/0.0041, still showing that the results are significant.

Analysis of the "Attention in Attention" Mechanism and "Single Attention" Mechanism. In Table 3, we compared AiA against a simplified version, which still benefits from the use of an attention block without the use of any inner attention module (Figs. 4 versus 1 in Section 3). Empirically, we observe that by incorporating the inner attention module, improved results can be obtained in both baseline architectures(i.e., $\mathcal{F}_a$ and $\mathcal{F}_a + \mathcal{F}_p$). To further verify this, we replace our AiA with the current state-of-the-art attention modules, namely the Squeeze-and-Excitation block [14] and

Fig. 10. Some failure cases on person re-ID datasets. In each ranking list, to the left is the query person and to the right is the corresponding ranked list in the gallery set. The correct and false matches are enclosed in green and red boxes. Best viewed in color.

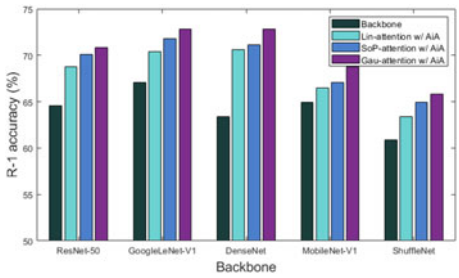


Fig. 11. Evaluation for attention blocks on different backbone networks on the CUHK03 dataset.

the Non-local block [43], and evaluate the resulting structure on the CUHK03 and Market-1501 datasets. The results on Table 3 and Fig. 8 clearly show the superiority of AiA over both the Squeeze-and-Excitation and Non-local blocks, even though only the linear kernel is used in this study.

Analysis of Failure Cases. In this section, we show some ranking lists of the failure cases (i.e., the identity mismatch of R-1 retrieved images for certain query images) obtained by AiA-Net with Gau-attention across the person re-ID datasets. Fig. 10 shows that the AiA-Net may be affected by persons with similar distractors, such as similar clothing and stature (e.g., the first and second ranking lists). Further, for the DukeMTMC-reID dataset, our network is also affected (e.g., the third and forth ranking lists) by occlusions (i.e., bike, car). Nonetheless, taking a closer look at those failure cases highlighted with red rectangles, they are in fact perceptually very similar to its respective query image (i.e., color of clothes, body orientation *etc.*). Having said that, these observations motivate us to further develop more robust person re-ID algorithms so as to differentiate such subtle changes successfully.

Generalization of Attention Blocks. To verify the generalization of proposed attention blocks, we employ other backbones to evaluate the effectiveness of AiA blocks, including ResNet-50 [80], GoogLeNet-V1 [71], DenseNet [94], MobileNet [95] as well as ShuffleNet [96]. This study is conducted on CUHK03 dataset. Fig. 11 reveals that our AiA blocks can consistently bring performance gain across various backbones, clearly showing the generalization and superiority of the AiA block. The study on other datasets is reported in the supplementary material, available online. We also conduct experiments on large-scale image classification to verify the generalization to other tasks in the supplementary material, available online.

6 Conclusion

In this paper, we generalize the Attention in Attention (AiA) mechanism for the person retrieval task. This AiA mechanism uses an inner attention, which encodes the global features of the input feature map, to re-weight the feature map. Thereafter, this feature map is further processed by an outer attention, to generate a well focused attention map. Besides the linear version of AiA, we propose and develop non-linear versions of AiA, where the features are approximated using the second-order polynomial and Gaussian kernel spaces respectively. We further propose simplified versions of the aforementioned attention blocks which exclude the inner attention (i.e. without AiA). With regards to the person retrieval task, we also propose an efficient feature extractor, which encodes both person appearance and part features. We incorporate the aforementioned AiA blocks in our network, termed AiA-Net, and empirically show that state-of-the-art performances can be achieved by incorporating the AiA modules in representation learning. This includes extensive evaluations on five standard person re-ID benchmarks along with the required ablation studies to understand the effect of various AiA blocks. Furthermore, our AiA-Net-V also achieves state-of-the-art result on the video person retrieval task, showing the generalization to video data.

Future works involve analyzing the AiA for addressing other visual tasks and developing additional forms of attention mechanisms by exploiting complementary non-linear information.

References

[1] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8030–8039.

[2] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3960–3969.

[3] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–437.

[4] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–518.

[5] Y. Chen, X. Zhu, W. Zheng, and J. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[6] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, "Pose-guided representation learning for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 16, 2019, doi: 10.1109/TPAMI.2019.2929036.

[7] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 420–429.

[8] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 475–491.

[9] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7134–7143.

[10] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4913–4922.

[11] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2285–2294.

[12] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 384–400.

[13] X. Qian, Y. Fu, T. Xiang, Y. Jiang, and X. Xue, "Leader-based multi-scale attention deep architecture for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 371–385, Feb. 2020.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[15] T.-Y. Lin, A. R. Chowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer, 2000.

[17] J. Xu, J.-F. Ton, H. Kim, A. R. Kosiorek, and Y. W. Teh, "MetaFun: Meta-learning with iterative functional updates," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 10617–10627.

[18] B. Peng *et al.*, "Correlation congruence for knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5007–5016.

[19] S. Jayasumana, S. Ramalingam, and S. Kumar, "Kernelized classification in deep networks," 2020, *arXiv:2012.09607*.

[20] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2921–2930.

[21] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, Mar. 2012.

[22] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2008, pp. 1177–1184.

[23] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 152–159.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[25] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. Workshop Benchmarking Multi-Target Tracking*, 2016, pp. 17–35.

[26] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 79–88.

[27] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 31–44.

[28] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[29] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1528–1535.

[30] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 34–39.

[31] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*. Berlin, Germany: Springer, 2014.

[32] S. K. Roy, M. Harandi, R. Nock, and R. Hartley, "Siamese networks: The tale of two manifolds," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3046–3055.

[33] Y. Wang *et al.*, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8042–8051.

[34] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1249–1258.

[35] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3800–3808.

[36] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2530–2539.

[37] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 403–412.

[38] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8933–8940.

[39] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 701–716.

[40] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1325–1334.

[41] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representation for video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3958–3967.

[42] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[43] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7794–7803.

[44] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[45] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4634–4643.

[46] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[47] M. Tian *et al.*, "Eliminating background-bias for robust person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5794–5803.

[48] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[49] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 384–393.

[50] H. Liu, J. Feng, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[51] J. Si *et al.*, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5363–5372.

[52] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person reid," 2018, *arXiv: 1805.02104*.

[53] Y. Fu, X. Wang, Y. Wei, and T. Huang, "STA: Spatial-temporal attention for large-scale video-based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 8287–8294.

[54] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Bilinear classifiers for visual recognition," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1482–1490.

[55] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 317–326.

[56] T.-Y. Lin and S. Maji, "Improved bilinear pooling with CNNs," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[57] J. Liu, Z. Yang, Z. Tao, and X. Huilin, "Multi-part compact bilinear CNN for person re-identification," in *Proc. Int. Conf. Image Process.*, 2017, pp. 2309–2313.

[58] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," 2015, *arXiv:1512.05300*.

[59] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. Int. Conf. Learn. Representations*, 2017.

[60] T. Hofmann, B. Scholkopf, and A. J. Smola, "Kernel methods in machine learning,",*The Annals of Statistics*. Cambridge, U. K.: Cambridge Univ. Press, 2008.

[61] T. Joachims, "Training linear SVMs in linear time," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 217–226.

[62] S. Maji and A. C. Berg, "Max-margin additive classifiers for detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 40–47.

[63] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 902–917, Mar. 2021.

[64] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3219–3228.

[65] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[66] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 196–212.

[67] X. Ni, L. Fang, and H. Huttunen, "Adaptive L2 regularization in person re-identification," in *Proc. Int. Conf. Pattern Recognit.*, 2020.

[68] Z. Zhu *et al.*, "Viewpoint-aware loss with angular regularization for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 13114–13121.

[69] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1440–1448.
[70] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshop*, 2017, pp. 1–4.
[71] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[72] O. Russakovsky *et al.* "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, 2015, pp. 211–252.
[73] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[75] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5098–5107.
[76] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv: 1708.04896*.
[77] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[78] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1318–1327.
[79] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1318–1327, 2015.
[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[81] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2138–2147.
[82] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6886–6895.
[83] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9317–9326.
[84] H. Sun, Z. Chen, S. Yan, and L. Xu, "MVP matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6736–6746.
[85] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conf. Comput. Vis*, 2018, pp. 508–526.
[86] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8295–8302.
[87] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3702–3712.
[88] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-ReID: Searching for a part-aware ConvNet for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3750–3759.
[89] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3186–3195.
[90] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2109–2118.
[91] T. Chen *et al.*, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8351–8361.
[92] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1077–1085.
[93] A. Subramaniam, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 562–572.
[94] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[95] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[96] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 122–138.

Pengfei Fang received the BE degree in automation from Hangzhou Dianzi University, in 2014, and the ME degree in mechatronics from the Australian National University (ANU), in 2017. He is currently working toward the PhD degree from the ANU and Data61-CSIRO. His research interests include computer vision, machine learning, and control theory.

Jieming Zhou received the MEng degree from the Australian National University, Australia, in 2017. He is currently working toward the PhD degree in AUN in conjunction from the Commonwealth Scientific and Industrial Research Organisation. His research interests include the graph neural networks.

Soumava Kumar Roy received the bachelor's of engineering degree in electronics and communication engineering from the Manipal Institute of Technology, Manipal, India, in 2013, and the master's of technology degree in Information Technology from the Indian Institute of Information Technology, Allahabad, India. He is currently working toward the PhD degree with the College of Engineering and Computer Science, Australian National University. His research interests include deep learning and computer vision.

Pan Ji received the PhD degree from Australian National University, in 2016. He is currently a staff research engineer at InnoPeak Technology (a.k.a., OPPO US Research Center). From 2018 to 2020, he was a researcher with NEC Labs America. Before moving to the U.S., in 2016, he was an ARC senior research associate with the University of Adelaide. His research interests include computer vision and machine learning.

Lars Petersson is currently a principal research scientist at the Imaging and Computer Vision Group, Data61, CSIRO, Australia. He is also leading one of the activities under CSIRO's Machine Learning and Artificial Intelligence Future Science Platform effort where data science problems from the smallest of microscopy scales to the largest of astronomical scales are addressed. Before joining Data61/CSIRO, he was a principal researcher and research leader with NICTA's computer vision research group, where he was leading projects, including the Smart Cars, AutoMap, and Distributed Large Scale Vision.

Mehrtash Harandi is currently a senior lecturer at the Department of Electrical and Computer Systems Engineering, Monash University. He is also a contributing research scientist at the Machine Learning Research Group, Data61/CSIRO and an associated investigator with the Australian Center for Robotic Vision. His current research interests include theoretical and computational methods in machine learning, computer vision, signal processing, and Riemannian geometry.