Image and Vision Computing 100 (2020) 103931

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Cross-Correlated Attention Networks for Person Re-Identification

Jieming Zhou [a,*,1], Soumava Kumar Roy [a,1], Pengfei Fang [a], Mehrtash Harandi [b], Lars Petersson [c]

[a] Australian National University, Canberra, Acton 2601, Australia
[b] Monash University, Wellington Rd, Clayton, VIC 3800, Australia
[c] Data61, CSIRO, Canberra, Acton 2601, Australia

ARTICLE INFO

Article history:
Received 3 May 2020
Accepted 10 May 2020
Available online 25 May 2020

Keywords:
Attention
Feature extraction
Cross correlation
Person Re-Identification
Surveillance

ABSTRACT

Deep neural networks need to make robust inference in the presence of occlusion, background clutter, pose and viewpoint variations -to name a few- when the task of person re-identification is considered. Attention mechanisms have recently proven to be successful in handling the aforementioned challenges to some degree. However previous designs fail to capture inherent inter-dependencies between the attended features; leading to restricted interactions between the attention blocks. In this paper, we propose a new attention module called Cross-Correlated Attention (*CCA*); which aims to overcome such limitations by maximizing the information gain between different attended regions. Moreover, we also propose a novel deep network that makes use of different attention mechanisms to learn robust and discriminative representations of person images. The resulting model is called the Cross-Correlated Attention Network (*CCAN*). Extensive experiments demonstrate that the *CCAN* comfortably outperforms current state-of-the-art algorithms by a tangible margin.
Modeling the inherentspatial relations between different attended regions within the deep architecture. Joint end-to-end cross correlated attention and representational learning. State-of-the-art results in terms of mAP and Rank-1 accuracies across several challenging datasets.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

In this paper, we propose a Cross-Correlated Attention Network (CCAN) to jointly learn a holistic attention selection mechanism along with discriminative feature representations for person Re-IDentification (Re-ID). To this end, we make use of complementary attentional information along a global and a local branch (or feature extractor), in order to localize and focus on the discriminative regions of the input image.

Person Re-ID refers to the task of judging whether two images, depicting people, belong to the same individual or not. In general, the two images are obtained from two distinct cameras without any overlapping views. More specifically, given a query image containing the person of interest (or probe), Re-ID aims to find all the images that contain the same identity (id), as that of the query image, from a large gallery set [1].

Any robust Re-ID algorithm is required to address the following challenges: (1) viewpoint variations in visual appearance and environmental conditions due to different non-overlapping camera views, (2) significant pose changes for the same probe across time, space and camera views, (3) background clutter and occlusions, (4) different individuals may have similar appearance across different cameras or vice versa, (5) low resolution of the images limiting the use of face based biometric systems [2]. All these factors lead to significant visual deformations across the multiple camera views for the same person of interest.

In order to overcome these challenges, most of the early works focused on (1) designing discriminative *hand-engineered* feature representations which are invariant to lighting, pose and viewpoint changes, and occlusion or clutter [1,3]; (2) learning a robust *distance metric* for similarity measurement such that the embedded feature vectors belonging to the same class are closer to each other compared to the ones from different classes [4,5].

With the success of Deep Learning (DL) algorithms [6] across a large number of tasks in computer vision, recent deep Re-ID algorithms combine both the aforementioned aspects together into a unified end-to-end framework. While some deep algorithms address Re-ID by developing distinct global feature extraction units [7,8], others use a hybrid model which holistically combines the global and local features for an improved performance [9,10]. Body-part detectors have been predominantly used to extract local features that are distinct, discriminative and compatible with global features [11,12]. Similarly, pose estimation, correction and normalization networks [13,14,15] have also shown

* Corresponding author.
E-mail addresses: Jieming.Zhou@anu.edu.au (J. Zhou), Soumava.KumarRoy@anu.edu.au (S.K. Roy), Pengfei.Fang@anu.edu.au (P. Fang), mehrtash.harandi@monash.edu (M. Harandi), lars.petersson@data61.csiro.au (L. Petersson).
[1] Equal contribution.

https://doi.org/10.1016/j.imavis.2020.103931
0262-8856/© 2020 Elsevier B.V. All rights reserved.

great potential with, or without, part detectors in handling misalignment and viewpoint variations prevalent in the Re-ID datasets. The use of such special purpose auxiliary information tend to improve upon the methods it is applied to.

Attention based person Re-ID models have also been showing promising results as of late. Attention, as the name suggests, is comprised of two basic conceptual functionalities: "*where to look*" and "*how carefully to look*". *Hard-attention* often uses a window produced by, *e.g*, a Spatial Transformer Network (STN) [16] that models the former with a binary mask over the input features, whereas *soft-attention* simulates the latter by importance weighting of the input features [17].

Both these attention based learning approaches have been successfully integrated when addressing the person Re-ID task [11,12]. However, these models do not capture spatial inter-dependencies (*i.e*, *self-attention*) within the input features, thereby failing to recognize and perceive spatially distant, yet visually similar regions. They also do not capture (or improve) any inter- (or cross-correlated) dependencies between the separately attended regions, thus failing to boost the overall Signal-to-Noise Ratio (SNR) in the learnt feature maps. Moreover, convolutional based soft-attention blocks are not able to capture the inherent contextual information that exist in the input features.

To address the aforementioned drawbacks, we design the CCAN, a novel yet intuitive *Cross-Correlated* Attention based deep network. CCAN consists of a novel attention module which aims to *exploit* and *explore* the correlation between different regions at various levels of a deep model. It also benefits from a top-down interaction scheme between the global and local feature extractors through the different attention modules to automatically focus and extract distinct regions in the input image for enhanced feature representation learning.

The major contributions of our work are as follows:

- A novel *Cross-Correlated* Attention (CCA) module to model the inherent spatial relations between different attended regions within the deep architecture.
- A novel deep architecture for joint *end-to-end* cross correlated attention and representational learning.
- State-of-the-art results in terms of mAP and Rank-1 accuracies across several challenging datasets such as Market-1501 [18] and DukeMTMC-reID [19], CUHK03 [8] and MSMT17 [20].

2. Related Work

Much of the earlier work in person Re-ID was focused on hand-engineered feature representations [21,22,23,24,1] or learning a robust metric [25,5,26] to overcome the associated challenges. Recent studies employ Deep Neural Networks (DNNs) for joint learning of the discriminative features and similarity measures in end-to-end frameworks [7, 27]. Since we are chiefly interested in attention methods for person Re-ID in this paper, we will not cover part/pose-based solutions here and refer interested readers to [13,14,28].

To address the viewpoint/pose variations and misalignment issues commonly present in a Re-ID system, a profound idea is to benefit from the use of attention techniques in DNNs [11,12,29–33,68]. Li *et al.* [11] used a Spatial Transformer Network (STN) [16] as a basis for creating a form of *hard-attention* to search and focus on the discriminative regions in the image, subject to a pre-defined spatial constraint. Zhao *et al.* [29] designed a novel hard-attention module (with components similar to STN) and integrated it into a CNN. This helped to focus on more discriminative regions. Subsequently, by extracting and processing features from the attention regions, improvements to the overall performance were observed. AANet [33] proposed a Part Feature Network by cropping body parts according to the location of the peak activation in the feature maps. Arguably, hard-attention modules fail to capture the coherence between image pixels within the attention windows due to their inflexible modelling nature. The Comparative Attention Network (CAN) [31] employs LSTMs to perform soft-attention at a holistic scale and identify discriminative regions in Re-ID images. Liu *et al.* [30] proposed *HydraPlus-Net* (HPN) which utilizes soft-attention across multiple scales and levels to learn discriminative representations. Dual ATtention Matching networks (DuATMs) [34] use spatial bi-directional attentions along sequence matching to learn context-aware feature representations. Wang *et al.* proposed *Mancs* [32] and designed a soft-attentional block and a novel curriculum sampling method to learn focused attention masks. In contrast to the aforementioned algorithms, HA-CNN [12] uses both hard and soft attention modules to efficiently learn *where to look* and *how carefully to look* simultaneously.

Recently, Zhou *et al.* [35] propose a novel attention regularizer along with a novel triplet loss which consistently learns correlated attention masks from low, mid and higher level feature maps within an interactive loop. DGNet [36] proposed coupling person re-id learning and image generation in a unified joint learning framework such that the re-id learning stage can benefit from the generated data with an inherent feedback loop to learning a superior embedding space. CAMA [37] enhances learning of traditional global representations for person Re-ID by learning class activation maps to discover discriminative and distinct visual features. CASN [38] designed a new siamese framework in order to learn discriminative attention masks and enforce attention consistency among images of the same person. Likewise, OSNet [39] designed a new aggregation gate that dynamically fuses features at multiple different scales with channel-wise attentional weights. MHAN [40] proposed the High-Order Attention (HOA) to integrate complex and higher order statistical information in learning an attention mask so as to capture and distinguish subtle differences between the pedestrian and the background.

In contrast to the aforementioned techniques, CCAN makes use of a novel, yet intuitive, *cross-correlated* attention module which discovers and exploits inter-correlated spatial dependencies in the learnt feature maps. It then propagates these learnt dependencies along the feature extraction units to inherently learn robust and discriminative features and attention maps; thereby improving the overall information gain in a data-driven fashion.

3. Cross-Correlated Attention Networks

Let $\boldsymbol{x}_i \in \mathcal{X}$ be an image, with $\mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ denoting the image-space, where H, W and C indicate its rows, columns and channels, respectively. In person Re-ID, we are provided with N pairs of the form $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$ with $y_i \in \{1, \cdots, K\}$ representing the identity of the person depicted in $\boldsymbol{x}_i$. The aim, here, is to learn a generic non-linear mapping $\Psi : \mathcal{X} \rightarrow \mathscr{H}$ from the image space $\mathcal{X}$ onto a latent feature space $\mathscr{H}$ such that, in $\mathscr{H}$, embeddings coming from the same identity are closer to each other than those of different identities. We achieve this by exploiting the complementary nature of global and local information in Re-ID images using a combination of two different, and complementary, learnable attention modules. We first provide a detailed overview of the attention modules (§3.1); followed by the overall structure of CCAN (§3.2).

3.1. Attention Layers

In CCAN, we introduce a variation of self-attention named *Cross-Correlated* Attention. The Cross-Correlated Attention mechanism aims to capture, exploit and boost spatial inter-dependencies (or cross-correlation) between different selected regions.

The Cross-Correlated Attention (CC-Attention or CCA) module which aims to model the cross-correlation (or inter-dependencies) between different feature maps as a means to construct the attention mask. Each CCA module accepts two inputs and calculates the attention as a weighted combination of the input feature maps (see Fig. 1 for a conceptual diagram). This, as will be shown empirically, captures the

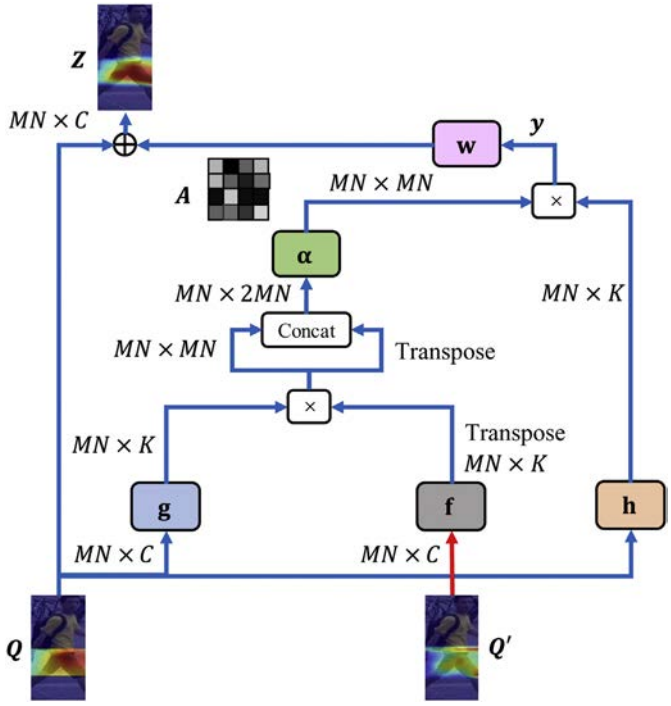

Fig. 1. The architecture of Cross-Correlated Attention (CC-Attention) used in our model (blue blocks in Fig. 3). CC-Attention is able to find correlated spatial locations in its two different input feature maps, which are further processed by the subsequent processing layers for discriminative feature learning.

inter-dependencies between the spatial regions in various feature maps with only a small computational overhead.

The CCA block works with the so-called positional matrices $\boldsymbol{q}$, $\boldsymbol{q}' \in \mathbb{R}^{MN\times C}$. In our application, the positional matrices are constructed from two feature maps $\boldsymbol{Q}, \boldsymbol{Q}' \in \mathbb{R}^{M\times N\times C}$ via reshaping through spacial dimension, *i.e* $\mathbb{R}^C \ni \boldsymbol{q}_i = \boldsymbol{Q}(m,n,:)\ \ \forall i \in \{1,\ldots,MN\}, \forall m \in \{1,\ldots,M\}, \forall n \in \{1,\ldots,N\}$. The matrices $\boldsymbol{q}$ and $\boldsymbol{q}'$ are then transformed into two feature spaces using independent non-linear mappings $\boldsymbol{g}$ and $\boldsymbol{f}$, respectively. The non-linear mappings are realized through $\mathbf{f}(\boldsymbol{q}_{i'}) = \phi(\boldsymbol{q}_{i'}\boldsymbol{W}_{\mathbf{f}})$ and $\mathbf{g}(\boldsymbol{q}_i) = \phi(\boldsymbol{q}_i \boldsymbol{W}_{\mathbf{g}})$, where $\boldsymbol{W}_{\mathbf{f}}, \boldsymbol{W}_{\mathbf{g}} \in \mathbb{R}^{K\times C}$, where the non-linearity $\phi : \mathbb{R} \to \mathbb{R}$ acts element-wise on $\mathbf{f}$ and $\mathbf{g}$. In our experiments, we choose $\phi(x) = \text{ReLU}(x) = \max(0,x)$. These two spaces are then used to calculate a primary attention map between the inputs at the different spatial locations as follows:

$$\boldsymbol{A} = \phi\left(\left[\boldsymbol{A}', \boldsymbol{A}'^{T}\right]\boldsymbol{W}_{\boldsymbol{\alpha}}\right)\ , \tag{1}$$

where $\boldsymbol{A}' = \mathbf{g}(\boldsymbol{q})\mathbf{f}(\boldsymbol{q}')^T$, $[.,.]$ denotes the concatenation operation along the width. Furthermore, $\boldsymbol{\alpha}$ is a linear layer with weight $\boldsymbol{W}_{\boldsymbol{\alpha}} \in \mathbb{R}^{2MN\times MN}$. $[\boldsymbol{A}]_{ij}$ is a measure of spatial dependencies between the i^{th} and the j^{th} spatial locations of the positional matrices $\boldsymbol{q}$ and $\boldsymbol{q}'$ respectively; thereby realizing a measure of cross-correlation between them. The symmetric operation described above guides the CCA module to focus on the correlated positions in both the $\boldsymbol{q}$ and $\boldsymbol{q}'$, which is processed by the subsequent layers of the network. The resultant map $\boldsymbol{A}$ is then used to generate $\boldsymbol{y} \in \mathbb{R}^{MN\times K}$ for input $\boldsymbol{q}$ as follows:

$$\boldsymbol{y}_i = \frac{1}{MN}\sum_{j=1}^{MN}\left([\boldsymbol{A}]_{ij} \odot \phi\left(\mathbf{h}\left(\boldsymbol{q}_j\right)\right)\right), \forall i = 1....MN, \tag{2}$$

where $\odot$ is Hadamard (element-wise) product, $\boldsymbol{y}_i$ is a weighted combination of the responses at all positions denoted by j, and $\mathbf{h}$ is also a non-linear layer with its weight $\boldsymbol{W}_{\mathbf{h}} \in \mathbb{R}^{K\times C}$ such that $\mathbf{h}(\boldsymbol{q}_i) = \phi(\boldsymbol{q}_i \boldsymbol{W}_{\mathbf{h}})$. We further pass $\boldsymbol{y}$ through a linear layer $\mathbf{w}$ to obtain the final output of the CC-Attention module as follows

$$z_i = \mathbf{w}(\boldsymbol{y}_i) + \boldsymbol{q}_i, \quad \forall i = 1, \ldots, mn \tag{3}$$

with $z = [\boldsymbol{z}_1; \cdots; z_{mn}] \in \mathbb{R}^{MN\times C}$ and $z_i \in \mathbb{R}^C$, and $\mathbf{w}(\boldsymbol{y}_i) = \phi(\boldsymbol{y}_i \boldsymbol{W}_{\mathbf{w}})$ such that $\boldsymbol{W}_{\mathbf{w}} \in \mathbb{R}^{C\times K}$. The output $\mathbf{z}$ is reshaped to $\mathbb{R}^{M\times N\times C}$ to match that of input $\boldsymbol{Q}$. In all our experiments, we have fixed the value of K to be $C/8$.

An intuitive way of thinking about the CCA module is to see $\mathbf{g}$ and $\mathbf{f}$ as non-linear signatures of elements $\boldsymbol{q}$ and $\boldsymbol{q}'$. The cross-correlation between the non-linear signatures acts as a gate and controls the information flow based on inter-correlation for generating the mask. The information, here, is encoded through $\mathbf{h}$. The result is further pruned by $\mathbf{w}$ and generates the attention map in an additive form. The additive form resembles the residual computing which is proven to be beneficial in training deep architectures.

Remark 1. *In the CCA module, we have introduced a symmetric cross-correlation operation between its input feature maps* $\boldsymbol{q}$ *and* $\boldsymbol{q}'$ *to generate the attention map* $\boldsymbol{A}$ *(see Eq. 1). It thereby encapsulates symmetrical inter-dependencies between its inputs. The standard cross-correlation operation does not take into account such symmetric relationships between the inputs. We believe that this subtle change makes CCA attend to highly correlated regions in both of its input feature maps.*

Remark 2. *When* $\boldsymbol{Q} = \boldsymbol{Q}'$*, the overall structure represents a form of Symmetric Self-Attention (SS-Attention or SSA) that aims to model highly correlated regions within itself. This form of symmetric self-attention is applied in the global branch, (i.e,* SS_1^G*) which models the intra-dependencies within the input. Further simplification of the SS-Attention module by removing the Concat and "*$\boldsymbol{\alpha}$*" block leads to the Non-Local Self-Attention module which is shown in Fig. 2. Thus we equip the traditional self-attention module with these two important changes to model symmetric cross-correlation attention between its two different inputs.*

3.2. Structure of the CCAN

A CCAN consists of two main branches (*i.e*, streams or feature extractors), namely the *global*, *G*, and the *local*, *L*, branch (see Fig. 3 for an overview of the architecture of CCAN). The purpose of the global branch is to capture and encode the overall appearance of a person, while the local branch encodes part information. The local branch, itself, has k_p sub-branches (or part-streams).

The basic building block of all branches is the *Inception block* of GoogLeNet [42]. The global branch makes use of three Inception blocks, $\{\text{I}_k^G\}_{k=1}^3$ along with a self-attention module SS_1^G to encode the global appearance (I_k^* marks the beginning of the k-th level of processing in CCAN). The Inception blocks in the global stream enable us to analyze the input at various resolutions, thereby realizing a coarse to fine global representation. The local branch, as the name implies, attends to the local and discriminative parts of the input image. The local branch comprises of k_p sub-branches, each intended to extract features belonging to a distinct part in the input image. For the s-th sub-branch, we denote its Inception blocks by $\text{I}_{k,\,s}^L$ with $k \in \{2,3\}$ and $s \in \{1,\cdots,k_p\}$ (see Fig. 3 for

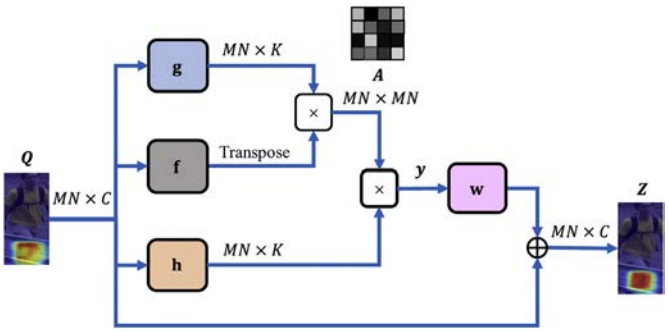

Fig. 2. Schematic of the Non-Local Self-Attention module as defined in [41].

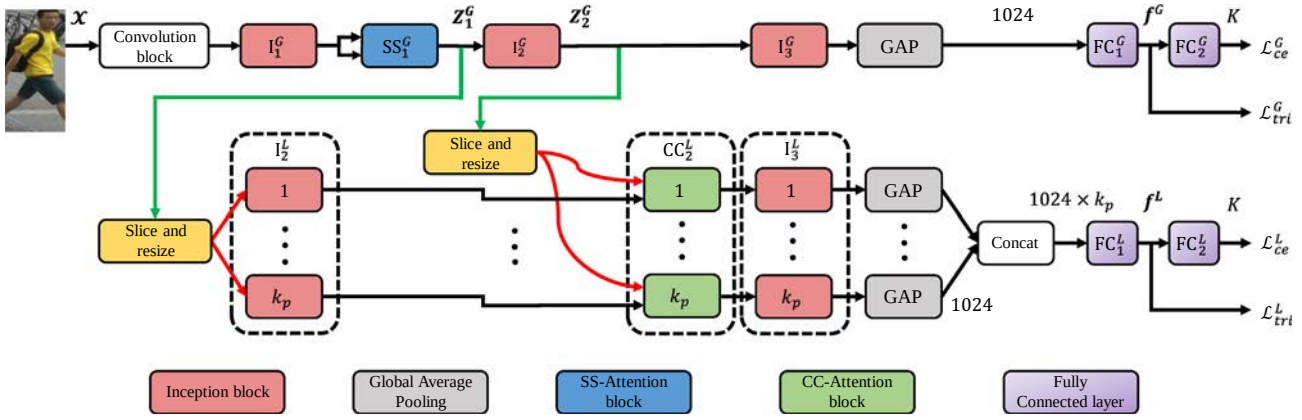

Fig. 3. Architecture of CCAN. *G* and *L* denote global and local branches. The local branch has k_p sub-branches. The local branches receive part patches from the global branch (*i.e* $\boldsymbol{Z}_1^G$ and $\boldsymbol{Z}_2^G$). Building blocks of the sub-branches are shown by dashed boxes (refer to §3.2 for more detail). $\mathscr{L}_{ce}$ and $\mathscr{L}_{tri}$ denote cross-entropy and triplet loss respectively. Green arrows indicate inputs for creating part patches.

details). We emphasize that each $\mathrm{I}_{k,\cdot}^L$ is an independent module, meaning that weights are not shared across the k_p part-streams.

In order to feed part information into local branches, we slice the feature maps at $\boldsymbol{Z}_1^G$ and $\boldsymbol{Z}_2^G$ (*i.e* the input and output of I_2^G) into k_p horizontal equal patches independently. Thereafter, all the sliced patches are resized to the size of their corresponding feature maps using bilinear interpolation. Moreover, each of the sub-branches consist of a cross-correlated attention module (*i.e* $CC_{2,s}^L$) $\forall s \in \{1,\cdots,k_p\}$. Every $CC_{2,s}^L$ calculates the cross-correlation between the sliced part patches of $\boldsymbol{Z}_1^G$ (after having been passed through $\mathrm{I}_{k,\cdot}^L$) and $\boldsymbol{Z}_2^G$ in each of the sub-branches independently. This sharing of feature mapsbetween the attention modules across the global and local branch within CCAN leads to the discovery of highly correlated regions; thereby realizing a simple but effective CCA scheme within CCAN.

The global branch is appended with a *global average pooling* (GAP) layer and two fully connected (FC_1^G and FC_2^G) layers, with the output of the FC_1^G realizing a *d*-dimensional embedding space. Similarly, the outputs of local sub-branches are passed through GAP layers and concatenated to produce a $1024 \times k_p$ feature vector. This is then passed through FC_1^L to produce the *d*-dimensional embedding vector in the local branch, which is further passed through FC_2^L. It should be noted that the FC_2^G and FC_2^L realize representations suitable for classification (*i.e*, $\mathscr{L}_{ce}^G$ and $\mathscr{L}_{ce}^L$). As such, their output dimensionality is *K*, the number of identities in the training set. We will discuss this in more detail later.

3.3. Loss function

Following the common practice in learning embeddings [43,44,45,46], we make use of a combination of classification and ranking losses (cross entropy loss with Label-Smoothing Regularization (LSR) [47] and the semi-hard triplet loss [48,49], respectively), to jointly optimize the global and the local branch. The overall loss is defined as follows:

$$\mathscr{L}_{tot} = \mathscr{L}_{ce}^G + \mathscr{L}_{tri}^G + \mathscr{L}_{ce}^L + \mathscr{L}_{tri}^L \ , \tag{4}$$

where the subscripts *ce* and *tri* denote the cross-entropy and triplet loss respectively. Moreover, the superscripts *G* and *L* indicate the global and local branch. We briefly describe the semi-hard triplet mining strategy used in our algorithm for calculating the triplet loss.

Semi-hard Triplet Mining.

In each mini-batch of *N* training samples, we *mine* $|P|$ triplets of the form $\{(\boldsymbol{x}_i^a, \boldsymbol{x}_i^p, \boldsymbol{x}_i^n)\}_{i=1}^{|P|}$, with the constraint that $(\boldsymbol{x}_i^a, \boldsymbol{x}_i^p)$ are in the same category, while $(\boldsymbol{x}_i^a, \boldsymbol{x}_i^n)$ are not. We also use the semi-hard mining strategy [48] to generate robust triplets for training the network. More specifically, given the anchor $\boldsymbol{x}^a$ and its positive example $\mathbf{x}^p$, we obtain the top *r* semi-hard negative triplets as follows

$$\boldsymbol{x}^n = \left\{ \boldsymbol{x}^j : \underset{D_a^p < D_a^j, \forall j=1,\cdots,r}{\arg\min} \ D_a^j ; s.t. D_a^j < D_a^{j+1} \right\} \ ,$$

where $D_a^j = \| \boldsymbol{x}^a - \boldsymbol{x}^j \|^2$. *r* is set to 10 for all the datasets. Moreover, to avoid any degeneracy, we randomly pick *v* different identities and sample *N*/*v* random images from each of the selected identities to create the mini-batch. These triplets are then used to compute the triplet embedding loss:

$$\mathscr{L}_{tri} = \frac{1}{|P|} \sum_{i=1}^{|P|} \left[\|\boldsymbol{x}_i^a - \boldsymbol{x}_i^p\|^2 - \|\boldsymbol{x}_i^a - \boldsymbol{x}_i^n\|^2 + \tau \right]_+ \ , \tag{5}$$

where $[y]_+ = \max(0, y)$ is the hinge loss, and $\tau > 0$ is a user-specified margin.

3.4. Person Re-ID by CCAN

Given a trained CCAN model and an input image $\boldsymbol{x}_i$; we first obtain its *d* dimensional global feature $\boldsymbol{f}_i^G$ and *d* dimensional local feature $\boldsymbol{f}_i^L$. We perform L2 normalization on each of them separately, and then proceed to concatenate them to obtain the joint 2*d* feature vector $\boldsymbol{f}_i^A = [\boldsymbol{f}_i^G, \boldsymbol{f}_i^L]$. Thus, given a probe image $\boldsymbol{x}_p$ from one camera view and all the gallery images $\{\bar{\boldsymbol{x}}_j\}$ from the other camera views, we obtain $\boldsymbol{f}_p^A$ and $\{\bar{\boldsymbol{f}}_j^A\}$ and compute the between-camera matching distances using the Euclidean distance. We then rank all $\{\bar{\boldsymbol{x}}_j\}$ in ascending order based on their distances given $\mathbf{x}_i$ and use that to evaluate the identity of $\boldsymbol{x}_p$.

4. Experiments

Datasets and Evaluation Protocol In this section, we show the effectiveness of our proposed algorithm through an extensive set of experiments across three well known person Re-ID datasets; (a) Market-1501 [18], (b) DukeMTMC-reID (or DukeMTMC) [19], (c) CUHK03 [8] and (d) MSMT [20]. Market-1501 has 751/750 train/test identity split, and 32,668 images in total. DukeMTMC-reID has 702/702 train/test identity split, and 36,411 images in total. CUHK03 has 14,097 images in total. In order to make the re-identification task more challenging on CUHK03, we use the 767/700 train/test identity split [50] instead of the 1367/100 standard split. The train/test id split and the test protocol are shown in Table 1. The **MSMT17** [20] dataset consists of 126,441 person images from 4,101 identities, thus constituting the largest person Re-ID dataset at present. All person images are detected using a Faster R-CNN [51]. This dataset is collected using 15 different cameras; and

the images were captured over 4 different days experiencing different weather conditions during a month. The training set consists of 32,621 images belonging to 1,041 identities, whereas the test set contains 93,820 images belonging to the remaining 3,060 identities. The test set is further randomly divided into 11,659 and 82,161 images for query and gallery sets respectively. Both mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) metrics are used for measuring performance on these datasets.

Table 1: The details of evaluated datasets. *Dis* refers to the distractor images of the DukeMTMC-reID dataset. TS, SQ, MQ and SS stand for Test Setting, Single Query, Multiple Query and Single Shot, respectively.

Dataset	Images	IDs	Train	Test	TS
Market1501	32,668	1501	751	750	SQ/MQ
DukeMTMC-reID	36,411	1404 + 408 *dis*	702	702	SQ
CUHK03	14,097	1467	767	700	SS
MSMT17	126,441	4101	1041	3060	SQ

Table 2: Comparison results on Market-1501 [18] dataset.

Method	SVDNet [52]	MHAN [40]	Dare [53]	AOS [54]	MLFN [55]	SGGNN [56]
mAP	62.1	85.0	69.9	70.4	74.3	82.8
R1	82.3	95.1	86.0	86.5	90.0	92.3
Method	IANet [57]	PCB [28]	MSCAN [11]	JLML [10]	PBR [58]	MGCAM [59]
mAP	83.1	81.6	57.5	65.5	76.0	74.3
R1	94.4	93.1	80.3	85.1	90.2	83.8
Method	AANet [33]	HPN [30]	DKPM [60]	DuATM [34]	Mancs [32]	HA-CNN [12]
mAP	83.4	–	75.3	76.6	82.3	75.7
R1	93.9	76.9	90.1	91.4	93.1	91.2
Method	CASN [38]	CAR [35]	OSNet [39]	DGNet [36]	CAMA [37]	CCAN (Ours)
mAP	82.8	84.7	84.9	**86.0**	84.5	**87.0**
R1	94.4	**96.1**	94.8	94.8	94.7	**94.6**

Table 3: Comparison results on DukeMTMC [19] dataset.

Method	SVDNet [52]	IDE [1]	Dare [53]	AOS [54]	MLFN [55]	SGGNN [56]
mAP	56.8	64.2	56.3	62.1	62.8	68.2
R1	76.7	80.1	74.5	79.2	81.0	81.1
Method	IANet [57]	PCB [28]	MSCAN [11]	JLML [10]	PBR [58]	MGCAM [59]
mAP	73.4	69.7	–	56.4	64.2	–
R1	87.1	83.9	–	73.3	82.1	–
Method	AANet [33]	HPN [30]	DKPM [60]	DuATM [34]	Mancs [32]	HA-CNN [12]
mAP	74.3	–	63.2	64.6	71.8	63.8
R1	87.7	–	80.3	81.8	84.9	80.5
Method	CASN [38]	CAR [35]	OSNet [39]	DGNet [36]	CAMA [37]	CCAN (Ours)
mAP	73.7	73.1	73.5	**74.8**	72.9	**76.8**
R1	87.7	86.3	**88.6**	86.6	85.8	**87.2**

4.1. Implementation

Our CCAN model is implemented in PyTorch [61]. We use GoogLeNet-V1 [42] with Batch Normalization [62] pretrained on Imagenet [63] as our backbone architecture. The dimensionality of the output feature maps of the global branch (*i.e*, I_1^G, I_2^G and I_3^G) is fixed to 480, 832, and 1024 respectively. Similarly, in the local branch, the dimensionality of the output feature maps of $I_{2,\,s}^L$ and $I_{3,\,s}^L$ is set to 832, and 1024 for every s respectively. The embedding dimension d and the number of local parts (*i.e* k_p) are set to 1024 and 4 across all the four datasets. None of the Inception and FC layers share weights between each other. The ADAM optimizer [64] is used to train the model, with the two moment terms (β_1, β_2), and the weight decay set to (0.9, 0.99) and 1×10^{-4}, respectively. The learning rate is initially set to 5×10^{-4} for Market-1501 and DukeMTMC-reID; and 1×10^{-3} for CUHK03 in both the labeled and detected settings; which is fixed for the first 150 epochs and decayed by a factor of 0.1 after every 50 epochs thereafter. The batch size is set to 64 of 16 identities with 4 images per identity in all the datasets. The smoothing parameter ε of LSR is 0.1. The margin τ for the triplet loss (Refer to Eq. 5) is set to 1 for Market-1501 and DukeMTMC-reID, and 1.5 for CUHK03 in both the dataset settings. The training images are first resized to 288×144 and then randomly cropped to 256×128, followed by a random horizontal flip. Following the protocol of [32], we apply random erasing [65] after the 50th epoch. However, during the test phase, the images are resized to 256×128 without any such data-augmentation techniques. We report the results after 200 epochs of training.

4.2. Comparison to State-of-the-Art Methods[2]

Evaluation on Market-1501.

We have evaluated against a number of recently proposed methods with, or without, the use of attention modules. Table 2 clearly shows the superior performance of CCAN against all the other methods in terms of mAP and Rank-1 accuracies on the Market-1501 dataset. More specifically, CCAN improves over the current state-of-the-art AANet by a prominent margin in the single query setting. We also outperform hard and soft attention based HA-CNN by 11.3/3.4% with respect to mAP and Rank-1 respectively in the single query setting.

Evaluation on DukeMTMC-reID.

We further evaluated our proposed CCAN on the DukeMTMC-reID [19] dataset. More variations in resolution and viewpoints due to wider camera views, and more complex environmental layout make DukeMTMC-reID more challenging compared to the Market-1501 dataset for the task of Re-ID. Table 3 shows that CCAN again outperforms almost all the baseline algorithms, except AANet in terms of Rank-1. However, we achieve higher mAP by a significant margin. We also outperform hard and soft attention based HA-CNN by 13.0/6.7% with respect to mAP and Rank-1 respectively.

Table 4: Comparison results on CUHK03 dataset in both the **Labeled** and the **Detected** settings.

	Labeled		Detected	
Measure (%)	mAP	R1	mAP	R1
MLFN [55]	49.2	54.7	47.8	52.8
IDE [1]	48.5	52.9	46.3	50.4
AOS [54]	–	–	47.1	43.4
Dare (De) [53]	52.2	56.4	50.1	54.3
PCB [28]	56.8	61.9	54.4	60.6
SVDNet [52]	–	–	37.3	41.5
MGCAM[59]	50.2	50.1	46.9	46.7
Mancs[32]	63.9	69.0	60.5	65.5
HA-CNN[12]	41.0	44.4	38.6	41.7
CAMA [37]	–	–	64.2	66.6
OSNet [39]	–	–	**67.8**	**72.3**
CASN [38]	**68.0**	**73.7**	64.4	71.5
CCAN (Ours)	**72.9**	**75.2**	**70.7**	**73.0**

Evaluation on CUHK03.

We have also evaluated CCAN on both the manually *labeled* and *detected* person bounding boxes versions of CUHK03. The 767/700 split results in a small training set with only 7365 images against 12,936/16,522 training images in Market-1501/DukeMTMC-reID datasets respectively. Even with such a constrained training setting, Table 4 clearly shows that notable improvement for CCAN against the baseline methods, including the current state-of-the-art Mancs, in both the labeled and detected settings. Furthermore, we also outperform HA-CNN by 31.9/30.8% and 32.1/31.3% in terms of mAP and Rank-1 in both the settings respectively.

[2] We report our results in **bold**, while we use **red** to report the best previous results obtained so far.

Evaluation on MSMT17.

Table 5 shows the result of our proposed CCAN when trained and evaluated on the new challenging MSMT17 [20] dataset. As can be seen, CCAN achieves a significant performance gain with regard to mAP and Rank-1 over all the baseline algorithms. Specifically, CCAN outperforms the current state-of-the-art algorithm on MSMT, *i.e.* Glad [66], by 19.6/14.9% in terms of mAP and Rank-1 respectively.

Table 5: Comparison results on MSMT17 dataset.

Model	mAP	R-1	R-5	R-10
GLNet [42]	23.0	47.6	65.0	71.8
PDC [67]	29.7	58.0	73.6	79.4
Glad [66]	34.0	61.4	76.8	81.6
PCB [28]	40.4	68.2	–	–
OSNet [39]	**52.9**	**78.7**	–	–
IANet [57]	46.8	75.5	–	–
DGNet [36]	52.3	77.2	–	–
CCAN (Ours)	**53.6**	**76.3**	**86.9**	**90.2**

These results, on all the four challenging datasets mentioned above, clearly demonstrate and validate our proposed approach of cross-correlation based joint attention and discriminative feature learning for person Re-ID. CCAN outperforms all the current methods that rely only on hard, soft, or a combination of these two types of attention.

5. Ablation Study

In this section, we undertake a detailed study of the various aspects of our proposed CCAN framework.

5.1. Dimensionality of the embedding space

We first evaluate CCAN for different values of d on the Market-1501 [18] dataset. As observed in Fig. 4(a), both mAP and R1 continue to increase as d is increased from 128 to 1024, with the highest values obtained when d is set to 1024. Based on this experimental study, we decided to choose 1024 as the embedding dimension d for all the experiments. It is to be noted that even with a smaller d (such as 256), we still outperform all baseline algorithms (Refer to Table??). This clearly shows that CCAN is able to learn discriminative features and achieve state-of-the-art results for a large range of d.

5.2. Number of body parts

We further evaluated the effect of various number of parts, *i.e.*, k_p in CCAN. Fig. 4(b) provides a detailed overview of the following evaluation for five different values of k_p. It can be seen that CCAN performs the best when k_p is set to 4, thereby suggesting that CCAN is able to detect and focus on the 4 distinct regions of the input person image; namely (a) head-shoulder, (b) upper-body, (c) thighs, and (d) crus-foot. It should also be noted that even with 2 different parts, CCAN is able to achieve competitive results against several baseline algorithms. This indeed demonstrates that CCAN is successful in exploiting the complementary nature of the learnt CCA attention modules even when lesser number of parts are specified. Based on this, in all the subsequent experiments, we have fixed the dimensionality of the embedding space (*i.e.* d) to 1024 and the number of parts (*i.e.* k_p) to 4.

5.3. Importance of various attention modules

We perform an ablation study in order to study the importance of various attention modules in CCAN. The results, evaluated on Market1501 dataset [18] single query setting, are shown in Table 6. The following critical insights are observed: **(a)** The performance of the global branch G (Id = 1) and the local branch L (Id = 2) by itself reads as 81.7% and 79.5% mAP respectively. **(b)** Though combination of G and L helps (Id = 3), incorporating only SS_1^G along G (Id = 4) leads to almost similar performance. **(c)** Furthermore, Id = 3 and 5 show the importance of adding a CCA module, *i.e* CC_2^L, along L.**(d)** Finally CCAN improves over Id = 6 with the addition of a SS_1^G along G (Refer to Fig. 3). This indeed verifies the joint interactive learning of the attention modules and feature extractors to obtain a discriminative embedding space for the person images. It is to be noted that in all our experiments, we have kept the final structure of CCAN fixed across all the datasets, suggesting a novel and rich architecture for the task of Re-ID that generalises well.

Table 6: Study of the importance of various attention modules on Market-1501 dataset.

Id	1	2	3	4	5	6
Setting	G	L	G + L	G+SS_1^G	G + L+CC_2^L	**CCAN**
mAP	81.7	79.5	83.6	83.3	85.6	**87.0**
R1	92.7	92.1	93.3	92.9	94.3	**94.6**

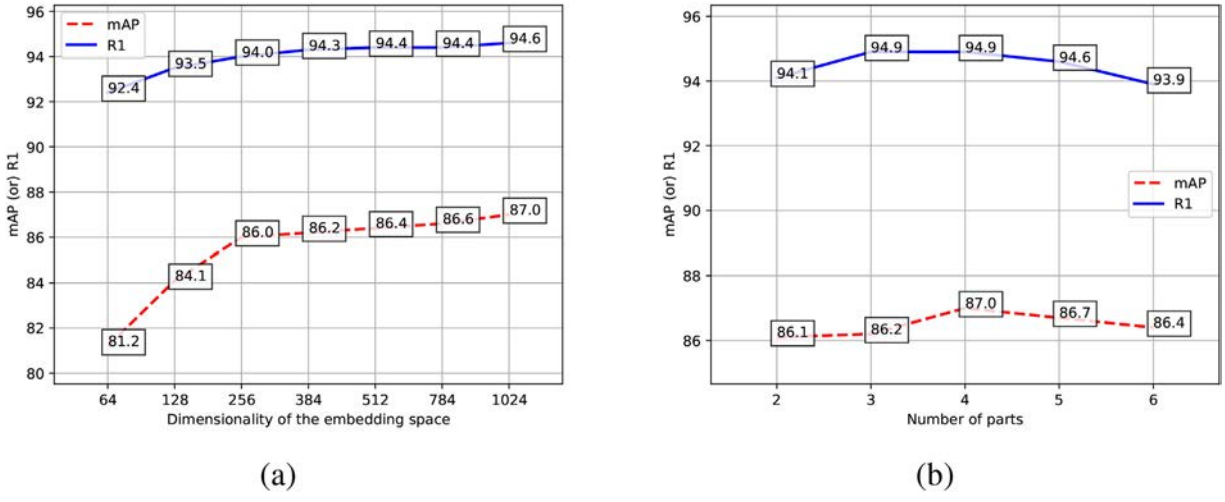

Fig. 4. Ablation study of the (a) dimensionality of the embedding space (*i.e.* d) and (b) number of body parts (*i.e.* k_p). Both the experiments were conducted on Market-1501 in the Single Query setting.

6. Conclusions

In this paper, we propose a new attention module, called *Cross-Correlated Attention* (CCA), which aims to improve the information gain by learning to focus on the correlated regions of the input image. We incorporate CCA into a novel deep attention architecture that we name *Cross-Correlated Attention Network* (CCAN) to achieve state-of-the-art results on three challenging datasets by utilizing the complementary nature of the attention mechanisms. In contrast to most existing attention based Re-ID models that use constrained attention learning algorithms, CCAN is capable of exploring and exploiting correlated interaction among the attention modules to locate and focus on the discriminative regions of the input person image without the need of any part (or pose) based estimator or detector network in a unified end-to-end CNN architecture. In the future, we plan to design and incorporate attention-diversity loss into CCAN to obtain further improvements and better focused attention maps. We also plan to study the effects of augmenting CCAN with additional part/pose estimation or detection networks in the future.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Zheng, Y. Yang, A. G. Hauptmann, Person Re-Identification: Past, Present and Future, arXiv preprint arXiv:1610.02984 (2016). 2, 4, 13, 14

[2] B. DeCann, A. Ross, Modelling Errors in a Biometric Re-Identification System, IET Biometrics 4 (4) (2015) 209–219. 2

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person Re-Identification by Symmetry-Driven Accumulation of Local Features, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 2360–2367. 2

[4] D. Chen, Z. Yuan, B. Chen, N. Zheng, Similarity Learning with Spatial Constraints for Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1268–1277. 2

[5] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large Scale Metric Learning from Equivalence Constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2288–2295. 2, 4

[6] Y. LeCun, Y. Bengio, G. Hinton, Deep Learning, nature 521 (7553) (2015) 436. 2

[7] E. Ahmed, M. Jones, T. K. Marks, An Improved Deep Learning Architecture for Person Re-Identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 2, 4

[8] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: Deep Filter Pairing Neural Network for Person Re-identification, in: CVPR, 2014. 2, 3, 11

[9] M. Tian, S. Yi, L. Hongsheng, L. Shihua, x. Zhang, J. Shi, J. Yan, X. Wang, Eliminating Background-bias for Robust Person Re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2

[10] W. Li, X. Zhu, S. Gong, Person Re-identification by Deep Joint Learning of Multi-Loss Classification, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17, AAAI Press, 2017, pp. 2194–2200. URL http://dl.acm.org/citation.cfm?id=3172077.3172193 2, 12, 13

[11] D. Li, X. Chen, Z. Zhang, K. Huang, Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2, 3, 4, 12, 13

[12] W. Li, X. Zhu, S.Gong, Harmonious Attention Network for Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2285–2294. 2, 3, 4, 12, 13, 14

[13] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-Driven Deep Convolutional Model for Person Re-Identification, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 3980–3989. 2, 4

[14] L. Zheng, Y. Huang, Y. Lu, Huchuan Yang, Pose Invariant Embedding for Deep Person Re-identification, arXiv:1701.07732 [cs.CV] (2017). arXiv:arXiv:1701.07732. URL https://arxiv.org/abs/1701.07732 2, 4

[15] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A Pose-Sensitive Embedding for Person Re-Identification With Expanded Cross Neighborhood Re-Ranking, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2018, p. 2. 2

[16] M. Jaderberg, K. Simonyan, A. Zisserman, k. kavukcuoglu, Spatial Transformer Networks, in: Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 2017–2025. 3, 4

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, in: International conference on machine learning, 2015, pp. 2048–2057. 3

[18] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable Person Re-identification: A Benchmark, in: IEEE International Conference on Computer Vision (ICCV), 2015. 3, 11, 12, 15, 16

[19] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking, in: European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking, 2016. 3, 11, 13

[20] L. Wei, S. Zhang, W. Gao, Q. Tian, Person Transfer Gan to Bridge Domain Gap for Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88. 3, 11, 14

[21] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2197–2206. 4

[22] D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A Richly Annotated Dataset for Pedestrian Attribute Recognition, arXiv preprint arXiv:1603.07054. 4

[23] H. Wang, S. Gong, T. Xiang, Highly Efficient Regression for Scalable Person Re-Identification, arXiv preprint arXiv:1612.01341. 4

[24] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-Ranking Person Re-Identification with k-Reciprocal Encoding, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 3652–3661. 4

[25] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by Relative Distance Comparison, IEEE transactions on pattern analysis and machine intelligence 35 (3) (2013) 653–668. 4

[26] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person Re-Identification Using Kernel-Based Metric Learning Methods, in: European conference on computer vision, Springer, 2014, pp. 1–16. 4

[27] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 4

[28] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline), in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480–496. 4, 12, 13, 14, 15

[29] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-Learned Part-Aligned Representations for Person Re-Identification, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 3239–3248. 4

[30] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-Net: Attentive Deep Features for Pedestrian Analysis, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 350–359. 4, 12, 13

[31] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-End Comparative Attention Networks for Person Re-Identification, IEEE Transactions on Image Processing 26 (7) (2017) 3492–3506. 4

[32] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-identification, in: The European Conference on Computer Vision (ECCV), 2018. 4, 12, 13, 14

[33] C.-P. Tay, S. Roy, K.-H. Yap, AAnet: Attribute Attention Network for Person Re-Identifications, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7134–7143. 4, 12, 13

[34] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, G. Wang, Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-Identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 4, 12, 13

[35] S. Zhou, F. Wang, Z. Huang, J. Wang, Discriminative Feature Learning with Consistent Attention Regularization for Person Re-Identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8040–8049. 4, 12, 13

[36] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint Discriminative and Generative Learning for Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2138–2147. 4, 12, 13, 15

[37] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, S. Zhang, Towards Rich Feature Discovery with Class Activation Maps Augmentation for Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1389–1398. 5, 12, 13, 14

[38] M. Zheng, S. Karanam, Z. Wu, R. J. Radke, Re-Identification with Consistent Attentive Siamese Networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 5735–5744. 5, 12, 13, 14

[39] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-Scale Feature Learning for Person Re-Identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3702–3712. 5, 12, 13, 14, 15

[40] B. Chen, W. Deng, J. Hu, Mixed High-Order Attention Network for Person Re-Identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 371–381. 5, 12

[41] X. Wang, R. Girshick, A. Gupta, K. He, Non-local Neural Networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 8

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper with Convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9. 8, 11, 15

[43] K. Q. Weinberger, L. K. Saul, Distance Metric Learning for Large Margin Nearest Neighbor Classification, Journal of Machine Learning Research 10 (Feb) (2009) 207–244. 10

[44] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep Metric Learning via Lifted Structured Feature Embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4004–4012. 10

[45] J. Hu, J. Lu, Y.-P. Tan, Discriminative Deep Metric Learning for Face Verification in the Wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1875–1882. 10

[46] H. O. Song, S. Jegelka, V. Rathod, K. Murphy, Deep Metric Learning via Facility Location, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 2206–2214. 10
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2818–2826. 10
[48] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A Unified Embedding for Face Recognition and Clustering, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 10
[49] R. Manmatha, C.-Y. Wu, A. J. Smola, P. Krähenbühl, Sampling Matters in Deep Embedding Learning, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 2859–2867. 10
[50] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking Person Re-identification with *k*-reciprocal Encoding, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3652–3661. doi:https://doi.org/10.1109/CVPR.2017.389. 11
[51] R. Girshick, Fast R-CNN, in: International Conference on Computer Vision (ICCV), 2015. 11
[52] Y. Sun, L. Zheng, W. Deng, S. Wang, SVDNet for Pedestrian Retrieval, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 3820–3828. 12, 13, 14
[53] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, K. Q. Weinberger, Resource Aware Person Re-Identification Across Multiple Resolutions, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 12, 13, 14
[54] H. Huang, D. Li, Z. Zhang, X. Chen, K. Huang, Adversarially Occluded Samples for Person Re-Identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5098–5107. 12, 13, 14
[55] X. Chang, T. M. Hospedales, T. Xiang, Multi-Level Factorisation Net for Person Re-Identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2018, p. 2. 12, 13, 14
[56] Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person Re-identification with Deep Similarity-Guided Graph Neural Network, in: The European Conference on Computer Vision (ECCV), 2018. 12, 13
[57] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-And-Aggregation Network for Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9317–9326. 12, 13, 15
[58] Y. Suh, J. Wang, S. Tang, T. Mei, K. Mu Lee, Part-Aligned Bilinear Representations for Person Re-Identification, in: The European Conference on Computer Vision (ECCV), 2018. 12, 13
[59] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-Guided Contrastive Attention Model for Person Re-Identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 12, 13, 14
[60] Y. Shen, T. Xiao, H. Li, S. Yi, X. Wang, End-to-End Deep Kronecker-Product Matching for Person Re-Identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 12, 13
[61] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic Differentiation in Pytorch, in: NIPS-W, 2017. 11
[62] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: International Conference on Machine Learning, 2015, pp. 448–456. 11
[63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (3) (2015) 211–252. 11
[64] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980 (2014). 12
[65] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data Augmentation, arXiv preprint arXiv:1708.04896 (2017). 12
[66] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, GLAD: Global–Local-Alignment Descriptor for Pedestrian Retrieval, in: Proceedings of the 25th ACM International Conference on Multimedia, MM '17, 2017. 15
[67] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven Deep Convolutional Model for Person Re-identification, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3960–3969. 15
[68] P. Fang, J. Zhou, S. K. Roy, L. Petersson, M. Harandi, Bilinear Attention Networks for Person Retrieval, in: The IEEE International Conference on Computer Vision (ICCV), 2019. 4