Harmonious Attention Network for Person Re-Identification

Wei Li[1] Xiatian Zhu[2] Shaogang Gong[1]
Queen Mary University of London[1] Vision Semantics Ltd.[2]
{wei.li, s.gong}@qmul.ac.uk eddy@visionsemantics.com

Abstract

Existing person re-identification (re-id) methods either assume the availability of well-aligned person bounding box images as model input or rely on constrained attention selection mechanisms to calibrate misaligned images. They are therefore sub-optimal for re-id matching in arbitrarily aligned person images potentially with large human pose variations and unconstrained auto-detection errors. In this work, we show the advantages of jointly learning attention selection and feature representation in a Convolutional Neural Network (CNN) by maximising the complementary information of different levels of visual attention subject to re-id discriminative learning constraints. Specifically, we formulate a novel Harmonious Attention CNN (HA-CNN) model for joint learning of soft pixel attention and hard regional attention along with simultaneous optimisation of feature representations, dedicated to optimise person re-id in uncontrolled (misaligned) images. Extensive comparative evaluations validate the superiority of this new HA-CNN model for person re-id over a wide variety of state-of-the-art methods on three large-scale benchmarks including CUHK03, Market-1501, and DukeMTMC-ReID.

1. Introduction

Person re-identification (re-id) aims to search people across non-overlapping surveillance camera views deployed at different locations by matching person images. In practical re-id scenarios, person images are typically automatically detected for scaling up to large visual data [44, 18]. Auto-detected person bounding boxes are typically not optimised for re-id due to misalignment with background clutter, occlusion, missing body parts (Fig. 1). Additionally, people (uncooperative) are often captured in various poses across open space and time. These give rise to the notorious image matching *misalignment* challenge in cross-view re-id [7]. There is consequently an inevitable need for *attention selection* within arbitrarily-aligned bounding boxes as an integral part of model learning for re-id.

There are a few attempts in the literature for solving the problem of re-id attention selection within person bounding boxes. One common strategy is local patch calibration and saliency weighting in pairwise image matching [43, 25, 46, 36]. However, these methods rely on handcrafted features without deep learning jointly more expressive feature representations and matching metric holistically (end-to-end). A small number of attention deep learning models for re-id have been recently developed for reducing the negative effect from poor detection and human pose change [17, 42, 27]. Nevertheless, these deep methods implicitly assume the availability of large labelled training data by simply adopting existing deep architectures with high complexity in model design. Additionally, they often consider only coarse region-level attention whilst ignoring the fine-grained pixel-level saliency. Hence, these techniques are ineffective when only a small set of labelled data is available for model training whilst also facing noisy person images of arbitrary misalignment and background clutter.

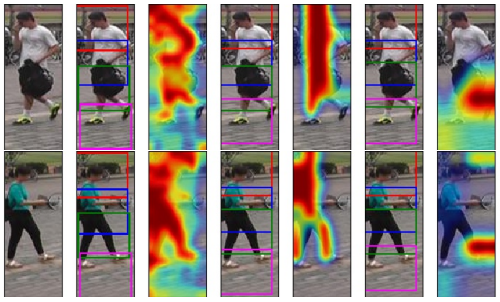


Figure 1. Examples of attention selection in auto-detected person bounding boxes used for person re-id matching.

In this work, we consider the problem of jointly deep learning attention selection and feature representation for optimising person re-id in a more lightweight (with less parameters) network architecture. The **contributions** of this work are: **(I)** We formulate a novel idea of jointly learning multi-granularity attention selection and feature representation for optimising person re-id in deep learning. To our knowledge, this is the first attempt of jointly deep learning multiple complementary attention for solving the person re-id problem. **(II)** We propose a *Harmonious Attention Con-*

volutional Neural Network (HA-CNN) to simultaneously learn hard region-level and soft pixel-level attention within arbitrary person bounding boxes along with re-id feature representations for maximising the correlated complementary information between attention selection and feature discrimination. This is achieved by devising a lightweight Harmonious Attention module capable of efficiently and effectively learning different types of attention from the shared re-id feature representation in a multi-task and end-to-end learning fashion. **(III)** We introduce a cross-attention interaction learning scheme for further enhancing the compatibility between attention selection and feature representation given re-id discriminative constraints. Extensive comparative evaluations demonstrate the superiority of the proposed HA-CNN model over a wide range of state-of-the-art re-id models on three large benchmarks CUHK03 [18], Market-1501 [44], and DukeMTMC-ReID [47].

2. Related Work

Most existing person re-id methods focus on supervised learning of identity-discriminative information, including ranking by pairwise constraints [23, 38], discriminative distance metric learning [13, 45, 40, 20, 41, 4], and deep learning [24, 18, 3, 39, 35, 19]. These methods assume that person images are well aligned, which is largely invalid given imperfect detection bounding boxes of changing human poses. To overcome this limitation, attention selection techniques have been developed for improving re-id by localised patch matching [25, 46] and saliency weighting [36, 43]. These are inherently unsuitable by design to cope with poorly aligned person images, due to their stringent requirement of tight bounding boxes around the whole person and high sensitivity of the hand-crafted features.

Recently, a few attention deep learning methods have been proposed to handle the matching misalignment challenge in re-id [17, 42, 27, 16]. The common strategy of these methods is to incorporate a regional attention selection sub-network into a deep re-id model. For example, Su et al. [27] integrate a separately trained pose detection model (from additional labelled pose ground-truth) into a part-based re-id model. Li et al. [17] design an end-to-end trainable part-aligning CNN network for locating latent discriminative regions (i.e. hard attention) and subsequently extract and exploit these regional features for performing re-id. Zhao et al. [42] exploit the Spatial Transformer Network [11] as the hard attention model for searching re-id discriminative parts given a pre-defined spatial constraint. However, these models fail to consider the noisy information within selected regions at the pixel level, i.e. no soft attention modelling, which can be important. While soft attention modelling for re-id is considered in [22], this model assumes tight person boxes thus less suitable for poor detections.

The proposed HA-CNN model is designed particularly to address the weaknesses of existing deep methods as above by formulating a joint learning scheme for modelling both soft and hard attention in a single re-id deep model. This is the first attempt of modelling multi-level correlated attention in deep learning for person re-id to our knowledge. In addition, we introduce cross-attention interaction learning for enhancing the complementary effect between different levels of attention subject to re-id discriminative constraints. This is impossible to do for existing methods due to their inherent single level attention modelling. We show the benefits of joint modelling multi-level attention in person re-id in our experiments. Moreover, we also design an efficient attention CNN architecture for improving the model deployment scalability, an under-studied but practically important issue for re-id.

3. Harmonious Attention Network

Given n training bounding box images $\mathcal{I} = \{\boldsymbol{I}_i\}_{i=1}^n$ from n_{id} distinct people captured by non-overlapping camera views together with the corresponding identity labels as $\mathcal{Y} = \{y_i\}_{i=1}^n$ (where $y_i \in [1, \cdots, n_{\text{id}}]$), we aim to learn a deep feature representation model optimal for person re-id matching under significant viewing condition variations. To this end, we formulate a *Harmonious Attention Convolutional Neural Network* (HA-CNN) that aims to concurrently learn a set of harmonious attention, global and local feature representations for maximising their complementary benefit and compatibility in terms of both discrimination power and architecture simplicity. Typically, person parts location information is not provided in person re-id image annotation (i.e. only weakly labelled without fine-grained). Therefore, the attention model learning is *weakly supervised* in the context of optimising re-id performance. Unlike most existing works that simply adopting a standard deep CNN network typically with a large number of model parameters (likely overfit given small size labelled data) and high computational cost in model deployment [15, 26, 30, 8], we design a *lightweight* (less parameters) yet deep (maintaining strong discriminative power) CNN architecture by devising a *holistic* attention mechanism for locating the most discriminative pixels and regions in order to identify optimal visual patterns for re-id. We avoid simply stacking many CNN layers to gain model depth.

This is particularly critical for re-id where the label data is often sparse (large models are more likely to overfit in training) and the deployment efficiency is very important (slow feature extraction is not scalable to large surveillance video data).

HA-CNN Overview We consider a multi-branch network architecture for our purpose. The overall objective of this multi-branch scheme and the overall architecture composi-

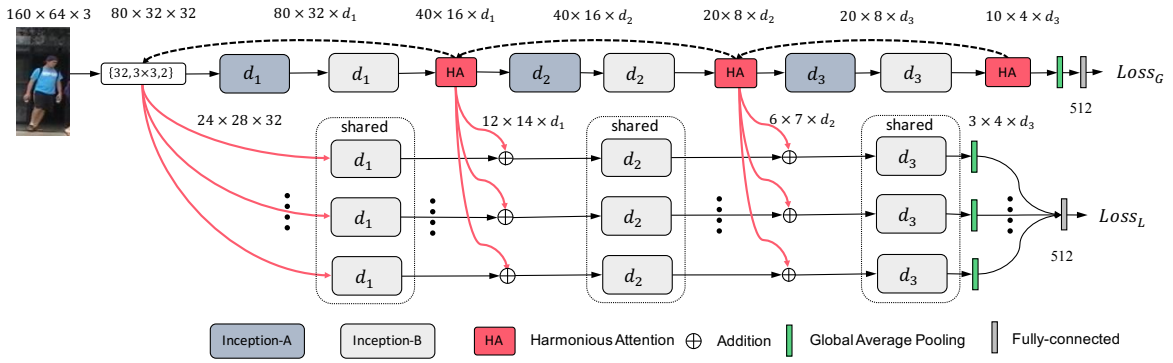


Figure 2. The Harmonious Attention Convolutional Neural Network. The symbol d_l ($l \in \{1, 2, 3\}$) denotes the number of convolutional filter in the corresponding Inception unit at the l-th block.

tion is to minimise the model complexity therefore reduce the network parameter size whilst maintaining the optimal network depth. The overall design of our HA-CNN architecture is shown in Fig. 2. This HA-CNN model contains two branches: **(1)** One *local branch* (consisting of T streams of an identical structure): Each stream aims to learn the most discriminative visual features for one of T local image regions of a person bounding box image. **(2)** One *global branch*: This aims to learn the optimal global level features from the entire person image. For both branches, we select the Inception-A/B units [39, 29] as the basic building blocks[1].

In particular, we used 3 Inception-A and 3 Inception-B blocks for building the global branch, and 3 Inception-B blocks for each local stream. The width (channel number) of each Inception is denoted by d_1, d_2 and d_3. The global network ends with a *global average pooling* layer and a *fully-connected* (FC) feature layer with 512 outputs. For the local branch, we also use a 512-D FC feature layer which fuses the global average pooling outputs of all streams. To reduce the model parameter size, we share the first conv layer between global and local branches and the same-layer Inceptions among all local streams. For our HA-CNN model training, we utilise the *cross-entropy classification loss* function for both global and local branches, which optimise person identity classification.

For attention selection within each bounding box of some unknown misalignment, we consider a *harmonious attention learning* scheme that aims to jointly learn a set of complementary attention maps including hard (regional) attention for the local branch and soft (spatial/pixel-level and channel/scale-level) attention for the global branch.

We further introduce a *cross-attention interaction learning* scheme between the local and global branches for further enhancing the harmony and compatibility degree whilst simultaneously optimising per-branch discriminative feature representations.

We shall now describe more details of each component of the network design as follows.

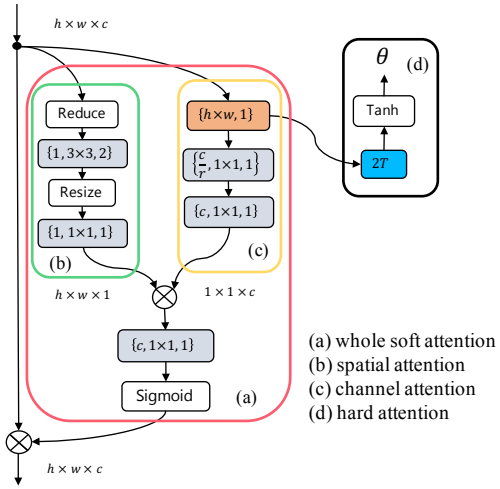


Figure 3. The structure of each Harmonious Attention module consists of **(a)** Soft Attention which includes **(b)** Spatial Attention (pixel-wise) and **(c)** Channel Attention (scale-wise), and **(d)** Hard Regional Attention (part-wise). Layer type is indicated by background colour: **grey** for *convolutional* (conv), **brown** for *global average pooling*, and **blue** for *fully-connected* layers. The three items in the bracket of a conv layer are: filter number, filter shape, and stride. The ReLU [15] and Batch Normalisation (BN) [10] (applied to each conv layer) are not shown for brevity.

3.1. Harmonious Attention Learning

Conceptually, our Harmonious Attention (HA) is a principled union of hard *regional* attention [11], soft *spatial* [34] and *channel* attention [9]. This simulates functionally the dorsal and ventral attention mechanism of human brain [33] in the sense of modelling soft and hard attention simultaneously. The soft attention learning aims at selecting the *fine-grained* important pixels, while the hard attention learning is dedicated to searching the *coarse* latent (weakly supervised) discriminative regions. They are

[1]This choice is independent of our model design and others can be readily considered such as AlexNet [15], ResNet [8] and VggNet [26].

thus largely complementary with high compatibility to each other in functionality. Intuitively, their combination can relieve the modelling burden of soft attention and resulting in more discriminative and robust model learning from the same (particularly small) training data. We propose a novel Harmonious Attention joint learning strategy to unite the three distinct types of attention with only a small number of additional parameters. We take a *block-wise* (module-wise) attention design, that is, each HA module is specifically optimised to attend the input feature representations at its own level alone. In the CNN hierarchical framework, this naturally allows for *hierarchical* multi-level attention learning to progressively refine the attention maps, in the spirit of the divide and conquer design [5]. As a result, we can significantly reduce the attention search space (i.e. the model optimisation complexity) whilst allow multi-scale selectiveness of hierarchical features to enrich the final feature representations. Such progressive and holistic attention modelling is both intuitive and essential for person re-id due to that (1) the surveillance person images often have cluttered background and uncontrolled appearance variations therefore the optimal attention patterns of different images can be highly variable, and (2) a re-id model typically needs robust (generalisable) model learning given very limited training data (significantly less than common image classification tasks). Next, we describe the design of our Harmonious Attention module in details.

(I) Soft Spatial-Channel Attention The input to a Harmonious Attention module is a 3-D tensor $\boldsymbol{X}^l \in \mathcal{R}^{h\times w\times c}$ where h, w, and c denote the number of pixel in the height, width, and channel dimensions respectively; and l indicates the level of this module in the entire network (multiple such modules). Soft spatial-channel attention learning aims to produce a saliency weight map $\boldsymbol{A}^l \in \mathcal{R}^{h\times w\times c}$ of the same size as $\boldsymbol{X}$. Given the largely independent nature between spatial (inter-pixel) and channel (inter-scale) attention, we propose to learn them in a *joint* but *factorised* way as:

$$\boldsymbol{A}^l = \boldsymbol{S}^l \times \boldsymbol{C}^l \tag{1}$$

where $\boldsymbol{S}^l \in \mathcal{R}^{h\times w\times 1}$ and $\boldsymbol{C}^l \in \mathcal{R}^{1\times 1\times c}$ represent the spatial and channel attention maps, respectively. We perform the attention tensor factorisation by designing a two-branches unit (Fig. 3(a)): One branch to model the spatial attention $\boldsymbol{S}^l$ (shared across the channel dimension), and another branch to model the channel attention $\boldsymbol{C}^l$ (shared across both height and width dimensions).

By this design, we can compute *efficiently* the full soft attention $\boldsymbol{A}^l$ from $\boldsymbol{C}^l$ and $\boldsymbol{S}^l$ with a tensor multiplication. Our design is more efficient than common tensor factorisation algorithms [14] since heavy matrix operations are eliminated.

(1) Spatial Attention We model the spatial attention by a tiny (10 parameters) 4-layers sub-network (Fig. 3(b)). It consists of a global cross-channel averaging pooling layer (0 parameter), a conv layer of 3×3 filter with stride 2 (9 parameters), a resizing bilinear layer (0 parameter), and a scaling conv layer (1 parameter). In particular, the global averaging pooling, defined as,

$$\boldsymbol{S}^l_{\text{input}} = \frac{1}{c}\sum_{i=1}^{c} \boldsymbol{X}^l_{1:h,1:w,i} \tag{2}$$

is designed especially to compress the input size of the subsequent conv layer with merely $\frac{1}{c}$ times of parameters needed. This cross-channel pooling is reasonable because in our design all channels share the identical spatial attention map. We finally add the scaling layer for automatically learning an adaptive fusion scale in order to optimally combining the channel attention described next.

(2) Channel Attention We model the channel attention by a small ($2\frac{c^2}{r}$ parameters, see more details below) 4-layers squeeze-and-excitation sub-network (Fig. 3(c)). Specifically, we first perform a *squeeze* operation via an averaging pooling layer (0 parameters) for aggregating feature information distributed across the spatial space into a channel signature as

$$\boldsymbol{C}^l_{\text{input}} = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w} \boldsymbol{X}^l_{i,j,1:c} \tag{3}$$

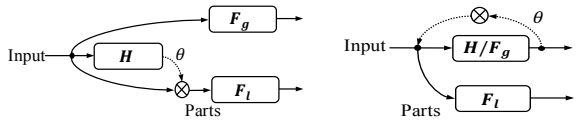
This signature conveys the per-channel filter response from the whole image, therefore providing the complete information for the inter-channel dependency modelling in the subsequent *excitation* operation, formulated as

$$\boldsymbol{C}^l_{\text{excitation}} = \texttt{ReLU}(\boldsymbol{W}^{\text{ca}}_2 \times \texttt{ReLU}(\boldsymbol{W}^{\text{ca}}_1 \boldsymbol{C}^l_{\text{input}})) \tag{4}$$

where $\boldsymbol{W}^{\text{ca}}_1 \in \mathcal{R}^{\frac{c}{r}\times c}$ ($\frac{c^2}{r}$ parameters) and $\boldsymbol{W}^{\text{ca}}_2 \in \mathcal{R}^{c\times\frac{c}{r}}$ ($\frac{c^2}{r}$ parameters) denote the parameter matrix of 2 conv layers in order respectively, and r (16 in our implementation) represents the bottleneck reduction rate. Again, this bottleneck design is for reducing the model parameter number from c^2 (using one conv layer) to $(\frac{c^2}{r}+\frac{c^2}{r})$, e.g. only need $\frac{1}{8}$ times of parameters when $r = 16$.

For facilitating the combination of the spatial attention and channel attention, we further deploy a $1\times1\times c$ convolution (c^2 parameters) layer to compute blended full soft attention after tensor multiplication. This is because the spatial and channel attention are not mutually exclusive but with a co-occurring complementary relationship. Finally, we use the sigmoid operation (0 parameter) to normalise the full soft attention into the range between 0.5 and 1.

Remarks Our model is similar to the Residual Attention (RA) [34] and Squeeze-and-Excitation (SE) [9] concepts but with a number of essential differences: **(1)** The RA requires to learn a much more complex soft attention subnetwork which is not only computationally expensive but



(a) STN [11] (b) HA-CNN Hard Attention

Figure 4. Schematic comparison between (a) STN [11] and (b) HA-CNN Hard Attention. Global feature and hard attention are jointly learned in a multi-task design. "$\boldsymbol{H}$": Hard attention module; "$\boldsymbol{F_g}$": Global feature module; "$\boldsymbol{F_g}$": Local feature module.

also less discriminative when the training data size is small typical in person re-id. **(2)** The SE considers only the channel attention and implicitly assumes non-cluttered background, therefore significantly restricting its suitability to re-id tasks under cluttered surveillance viewing conditions. **(3)** Both RA and SE consider no hard regional attention modelling, hence lacking the ability to discover the correlated complementary benefit between soft and hard attention learning.

(II) Hard Regional Attention The hard attention learning aims to locate latent (*weakly supervised*) discriminative T regions/parts (e.g. human body parts) in each input image at the l-th level. We model this regional attention by learning a transformation matrix as:

$$\boldsymbol{A}^l = \begin{bmatrix} s_h & 0 & t_x \\ 0 & s_w & t_y \end{bmatrix} \tag{5}$$

which allows for image cropping, translation, and isotropic scaling operations by varying two scale factors (s_h, s_w) and the 2-D spatial position (t_x, t_y). We use pre-defined region size by fixing s_h and s_w for limiting the model complex. Therefore, the effective modelling part of $\boldsymbol{A}^l$ is only t_x and t_y, with the output dimension as $2 \times T$ (T the region number). To perform this learning, we introduce a simple 2-layers ($2 \times T \times c$ parameters) sub-network (Fig. 3(d)). We exploit the first layer output (a c-D vector) of the channel attention (Eq. (3)) as the first FC layer ($2 \times T \times c$ parameters) input for further reducing the parameter size while sharing the available knowledge in spirit of the multi-task learning principle [6]. The second layer (0 parameter) performs a *tanh* scaling (the range of $[-1, 1]$) to convert the region position parameters into the percentage so as to allow for positioning individual regions outside of the input image boundary. This specially takes into account the cases that only partial person is detected sometimes. Note that, unlike the soft attention maps that are applied to the input feature representation $\boldsymbol{X}^l$, the hard regional attention is enforced on that of the corresponding network block to generate T different parts which are subsequently fed into the corresponding streams of the *local* branch (see the dashed arrow on the top of Fig 2).

Remarks The proposed hard attention modelling is conceptually similar to the Spatial Transformer Network (STN) [11] because both are designed to learn a transformation matrix for discriminative region identification. However, they differ significantly in design: **(1)** The STN attention is *network-wise* (one level of attention learning) whilst our HA is *module-wise* (multiple levels of attention learning). The latter not only eases the attention modelling complexity (divide-and-conquer design) and but also provides additional attention refinement in a sequential manner. **(2)** The STN utilises a separate large sub-network for attention modelling whilst the HA-CNN exploits a much smaller sub-network by sharing the majority model learning with the target-task network using a multi-task learning design (Fig. 4), therefore superior in both higher efficiency and lower overfitting risk. **(3)** The STN considers only hard attention whilst HA-CNN models both soft and hard attention in an end-to-end fashion so that additional complementary benefits are exploited.

(III) Cross-Attention Interaction Learning Given the joint learning of soft and hard attention above, we further consider a cross-attention interaction mechanism for enriching their joint learning harmony by interacting the *attended* local and global features across branches. Specifically, at the l-th level, we utilise the global-branch feature $\boldsymbol{X}_G^{(l,k)}$ of the k-th region to enrich the corresponding local-branch feature $\boldsymbol{X}_L^{(l,k)}$ by tensor addition as

$$\tilde{\boldsymbol{X}}_L^{(l,k)} = \boldsymbol{X}_L^{(l,k)} + \boldsymbol{X}_G^{(l,k)} \tag{6}$$

where $\boldsymbol{X}_G^{(l,k)}$ is computed by applying the hard regional attention of the $(l+1)$-th level's HA attention module (see the dashed arrow in Fig. 2). By doing so, we can simultaneously reduce the complexity of the local branch (fewer layers) since the learning capability of the global branch can be partially shared. During model training by back-propagation, the global branch takes gradients from both the global and local branches as

$$\Delta \boldsymbol{W}_G^{(l)} = \frac{\partial \mathcal{L}_G}{\partial \boldsymbol{X}_G^{(l)}} \frac{\partial \boldsymbol{X}_G^{(l)}}{\partial \boldsymbol{W}_G^{(l)}} + \sum_{k=1}^{T} \frac{\partial \mathcal{L}_L}{\partial \tilde{\boldsymbol{X}}_L^{(l,k)}} \frac{\partial \tilde{\boldsymbol{X}}_L^{(l,k)}}{\partial \boldsymbol{W}_G^{(l)}} \tag{7}$$

Therefore, the global $\mathcal{L}_G$ and local $\mathcal{L}_L$ loss quantities are concurrently utilised in optimising the parameters $\boldsymbol{W}_G^{(l)}$ of the global branch. As such, the learning of the global branch is interacted with that of the local branch at multiple levels, whilst both are subject to the same re-id optimisation constraint.

Remarks By design, cross-attention interaction learning is subsequent to and complementary with the harmonious attention joint reasoning above. Specifically, the latter learns soft and hard attention from the same input feature representations to maximise their compatibility (*joint attention generation*), whilst the former optimises the correlated complementary information between attention refined global and local features under the person re-id matching constraint

(*joint attention application*). Hence, the composition of both forms a complete process of joint optimisation of attention selection for person re-id.

3.2. Person Re-ID by HA-CNN

Given a trained HA-CNN model, we obtain a 1,024-D joint feature vector (deep feature representation) by concatenating the local (512-D) and the global (512-D) feature vectors. For person re-id, we deploy this 1,024-D deep feature representation using *only* a generic distance metric *without* any camera-pair specific distance metric learning, e.g. the L2 distance. Specifically, given a test probe image $\boldsymbol{I}^p$ from one camera view and a set of test gallery images $\{\boldsymbol{I}_i^g\}$ from other non-overlapping camera views: (1) We first compute their corresponding 1,024-D feature vectors by forward-feeding the images to a trained HA-CNN model, denoted as $\boldsymbol{x}^p = [\boldsymbol{x}_g^p; \boldsymbol{x}_l^p]$ and $\{\boldsymbol{x}_i^g = [\boldsymbol{x}_g^g; \boldsymbol{x}_l^g]\}$. (2) We then compute L2 normalisation on the global and local features, respectively. (3) Lastly, we compute the cross-camera matching distances between $\boldsymbol{x}^p$ and $\boldsymbol{x}_i^g$ by the L2 distance. We then rank all gallery images in ascendant order by their L2 distances to the probe image. The probabilities of true matches of probe person images in Rank-1 and among the higher ranks indicate the goodness of the learned HA-CNN deep features for person re-id tasks.

4. Experiments

(a) CUHK03 (b) Market-1501 (c) DukeMTMC

Figure 5. Example cross-view matched pairs from three datasets.

Datasets and Evaluation Protocol For evaluation, we selected three large-scale person re-id benchmarks, Market-1501 [37], DukeMTMC-ReID [47] and CUHK03 [18]. Figure 5 shows several example person bounding box images. We adopted the standard person re-id setting including the training/test ID split and test protocol (Table 1). For performance measure, we use the cumulative matching characteristic (CMC) and mean Average Precision (mAP) metrics.

Table 1. Re-id evaluation protocol. TS: Test Setting; SS: Single-Shot; MS: Multi-Shot. SQ: Single-Query; MQ: Multi-Query.

Dataset	# ID	# Train	# Test	# Image	Test Setting
CUHK03	1,467	767	700	14,097	SS
Market-1501	1,501	751	750	32,668	SQ/MQ
DukeMCMT-ReID	1,402	702	702	36,411	SQ

Implementation Details We implemented our HA-CNN model in the Tensorflow [1] framework. All person images are resized to 160×64. For HA-CNN architecture, we set the width of Inception units at the 1st/2nd/3rd levels as: $d_1 = 128$, $d_2 = 256$ and $d_3 = 384$. Following [19], we use $T = 4$ regions for hard attention, e.g. a total of 4 local streams. In each stream, we fix the size of three levels of hard attention as 24×28, 12×14 and 6×7. For model optimisation, we use the ADAM [12] algorithm at the initial learning rate 5×10^{-4} with the two moment terms $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the batch size to 32, epoch to 150, momentum to 0.9. Note, we do ***not*** adopt any data argumentation methods (e.g. scaling, rotation, flipping, and colour distortion), ***neither*** model pre-training. Existing deep re-id methods typically benefit significantly from these operations at the price of not only much higher computational cost but also notoriously difficult and time-consuming model tuning.

4.1. Comparisons to State-of-the-Art Methods

Evaluation on Market-1501 We evaluated HA-CNN against 13 existing methods on Market-1501. Table 2 shows the clear performance superiority of HA-CNN over all state-of-the-arts with significant Rank-1 and mAP advantages. Specifically, HA-CNN outperforms the 2nd best model JLML (pre-defined hard attention based) by 6.1% (91.2-85.1) (SQ) and 4.1% (93.8-89.7) (MQ) in Rank-1; 10.2% (75.7-65.5) (SQ) and 8.3% (82.8-74.5) (MQ) in mAP. Compared to the only soft attention alternative HPN, our model improves the Rank-1 by 14.3% (91.2-76.9) (SQ). This indicates the superiority of our factorised spatial and channel soft attention modelling over HPN's multi-directional attention mechanism. HA-CNN also surpasses recent hard attention re-id methods (MSCAN, DLPA and PDC), boosting the Rank-1 by 10.9%, 10.2% and 7.1%, mAP by 18.2%, 12.3% and 12.3% (SQ), respectively. These validate the significant advantage of our harmonious soft/hand attention joint and interaction learning over existing methods replying on either hard or soft attention at a single level.

Evaluation on DukeMTMC-ReID We evaluated HA-CNN on the recently released DukeMTMC-ReID dataset[2]. Compared to Market-1501, person images from this benchmark have more variations in resolution and background due to wider camera views and more complex scene layout, therefore presenting a more challenging re-id task. Table 3 shows that HA-CNN again outperforms all compared state-of-the-arts with clear accuracy advantages, surpassing the 2nd best SVDNet-ResNet50 (without attention modelling) by 3.8% (80.5-76.7) in Rank-1 and 7.0% (63.8-56.8) in mAP. This suggests the importance of attention modelling in re-id and the efficacy of our attention joint learning approach in a more challenging re-id scenario. Importantly, the performance advantage by our method is achieved at a

[2]Only a small number of methods (see Table 3) have been evaluated and reported on DukeMTMC-ReID.

Table 2. Market-1501 evaluation. $1^{st}/2^{nd}$ best in red/blue.

Query Type	Single-Query		Multi-Query	
Measure (%)	R1	mAP	R1	mAP
XQDA[20]	43.8	22.2	54.1	28.4
SCS[2]	51.9	26.3	-	-
DNS[41]	61.0	35.6	71.5	46.0
CRAFT[4]	68.7	42.3	77.0	50.3
CAN[21]	60.3	35.9	72.1	47.9
S-LSTM[32]	-	-	61.6	35.3
G-SCNN[31]	65.8	39.5	76.0	48.4
HPN [22]	76.9	-	-	-
SVDNet [28]	82.3	62.1	-	-
MSCAN [17]	80.3	57.5	86.8	66.7
DLPA [42]	81.0	63.4	-	-
PDC [27]	84.1	63.4	-	-
JLML [19]	85.1	65.5	89.7	74.5
HA-CNN	91.2	75.7	93.8	82.8

Table 3. DukeMTMC-ReID evaluation. $1^{st}/2^{nd}$ best in red/blue.

Measure (%)	R1	mAP
BoW+KISSME [37]	25.1	12.2
LOMO+XQDA [20]	30.8	17.0
ResNet50 [8]	65.2	45.0
ResNet50+LSRO [47]	67.7	47.1
JLML [19]	73.3	56.4
SVDNet-CaffeNet [28]	67.6	45.8
SVDNet-ResNet50 [28]	76.7	56.8
HA-CNN	80.5	63.8

lower model training and test cost through an much easier training process. For example, the performance by SVDNet relies on the heavy ResNet50 CNN model (23.5 million parameters) with the need for model pre-training on the ImageNet data (1.2 million images), whilst HA-CNN has only 2.7 million parameters with no data augmentation.

Evaluation on CUHK03 We evaluated HA-CNN on both manually labelled and auto-detected (more misalignment) person bounding boxes of the CUHK03 benchmark. We utilise the 767/700 identity split rather than 1367/100 since the former defines a more realistic and challenging re-id task. In this setting, the training set is small with only about 7,300 images (*versus* 12,936/16,522 in Market-1501/DukeMCMT-ReID). This generally imposes a harder challenge to deep models, particularly when our HA-CNN does not benefit from any auxiliary data pre-training (e.g. ImageNet) nor data augmentation. Table 4 shows that HA-CNN still achieves the best re-id accuracy, outperforming hand-crafted feature based methods significantly and deep competitors less so. Our model achieves a small margin (+0.2% in Rank-1 and +1.3%) over the best alternative SVDNet-ResNet50 on the detected set. However, it is worth pointing out that SVDNet-ResNet50 benefits additionally from not only large ImageNet pre-training but also a much

Table 4. CUHK03 evaluation. The setting is 767/700 training/test split. $1^{st}/2^{nd}$ best in red/blue.

Measure (%)	Labelled		Detected	
	R1	mAP	R1	mAP
BoW+XQDA [37]	7.9	7.3	6.4	6.4
LOMO+XQDA [20]	14.8	13.6	12.8	11.5
IDE-C [48]	15.6	14.9	15.1	14.2
IDE-C+XQDA [48]	21.9	20.0	21.1	19.0
IDE-R [48]	22.2	21.0	21.3	19.7
IDE-R+XQDA [48]	32.0	29.6	31.1	28.2
SVDNet-CaffeNet [28]	-	-	27.7	24.9
SVDNet-ResNet50 [28]	-	-	41.5	37.3
HA-CNN	44.4	41.0	41.7	38.6

Table 5. Evaluating individual types of attention in our HA model. *Setting*: SQ. SSA: Soft Spatial Attention; SCA: Soft Channel Attention; HRA: Hard Regional Attention.

Dataset	Market-1501		DukeMTMC-ReID	
Metric (%)	R1	mAP	R1	mAP
No Attention	84.7	65.3	72.4	53.4
SSA	85.5	65.8	73.9	54.8
SCA	86.8	67.9	73.7	53.5
SSA+SCA	88.5	70.2	76.1	57.2
HRA	88.2	71.0	75.3	58.4
All	**91.2**	**75.7**	**80.5**	**63.8**

larger network and more complex training process. In contrast, HA-CNN is much more lightweight on parameter size with the advantage of easy training and fast deployment. This shows that our attention joint learning can be a better replacement of existing complex networks with time-consuming model training.

4.2. Further Analysis and Discussions

Effect of Different Types of Attention We further evaluated the effect of each individual attention component in our HA model: Soft Spatial Attention (SSA), Soft Channel Attention (SCA), and Hard Regional Attention (HRA). Table 5 shows that: **(1)** Any of the three attention *in isolation* brings person re-id performance gain; **(2)** The combination of SSA and SCA gives further accuracy boost, which suggests the complementary information between the two soft attention discovered by our model; **(3)** When combining the hard and soft attention, another significant performance gain is obtained. This shows that our method is effective in identifying and exploiting the complementary information between coarse hard attention and fine-grained soft attention.

Effect of Cross-Attention Interaction Learning We also evaluated the benefit of cross-attention interaction learning (CAIL) between global and local branches. Table 6 shows that CAIL has significant benefit to re-id matching, improving the Rank-1 by 4.6%(91.2-86.6) / 6.5%(80.5-74.0), mAP by 9.5%(75.7-66.2) / 8.4%(63.8-55.4) on Market-1501 /

Table 6. Evaluating cross-attention interaction learning (CAIL). *Setting*: SQ.

Dataset	Market-1501		DukeMTMC-ReID	
Metric (%)	R1	mAP	R1	mAP
w/o CAIL	86.6	66.2	74.0	55.4
w/ CAIL	**91.2**	**75.7**	**80.5**	**63.8**

DukeMTMC-ReID, respectively. This validates our design is rational that it is necessary to jointly learn the *attended* feature representations across soft and hard attention subject to the same re-id label constraint.

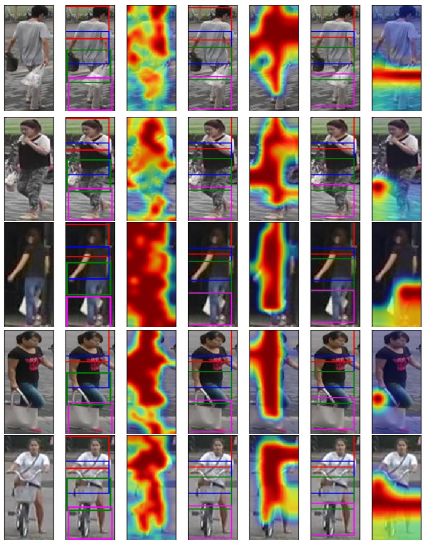
Effect of Joint Local and Global Features We evaluated the effect of joint local and global features by comparing their individual re-id performances against that of the joint feature. Table 7 shows: **(1)** Either feature representation *alone* is already very discriminative for person re-id. For instance, the global HA-CNN feature outperforms the best alternative JLML [19] (Table 2) by 4.8%(89.9-85.1) in Rank-1 and by 7.0%(72.5-65.5) in mAP (SQ) on Market-1501. **(2)** A further performance gain is obtained by joining the two representations, yielding 6.1%(91.2-85.1) in Rank-1 boost and 10.2%(75.7-65.5) in mAP increase. Similar trends are observed on the DukeMCMT-ReID (Table 3). These validate the complementary effect of jointly learning local and global features in harmonious attention context by our HA-CNN model.

Table 7. Evaluating global-level and local-level features. *Setting*: SQ.

Dataset	Market-1501		DukeMTMC-ReID	
Metric (%)	R1	mAP	R1	mAP
Global	89.9	72.5	78.9	60.0
Local	88.9	71.7	77.3	59.5
Global+Local	**91.2**	**75.7**	**80.5**	**63.8**

Visualisation of Harmonious Attention We visualise both learned soft attention and hard attention at three different levels of HA-CNN. Figure 6 shows: **(1)** Hard attention localises four body parts well at all three levels, approximately corresponding to head+shoulder (red), upper-body (blue), upper-leg (green) and lower-leg (violet). **(2)** Soft attention focuses on the discriminative pixel-wise selections progressively in spatial localisation, e.g. attending hierarchically from the global whole body by the 1st-level spatial SA (c) to local salient parts (e.g. object associations) by the 3rd-level spatial SA (g). This shows compellingly the effectiveness of joint soft and hard attention learning.

Model Complexity We compare the proposed HA-CNN model with four popular CNN architectures (Alexnet [15], VGG16 [26], GoogLeNet [30], and ResNet50 [8]) in model size and complexity. Table 8 shows that HA-CNN has the smallest model size (2.7 million parameters) and the 2nd smallest FLOPs (1.09×10^9) and yet, still retains the 2nd deepest structure (39).



(a) (b) (c) (d) (e) (f) (g)

Figure 6. Visualisation of our harmonious attention in person re-id. From left to right, (a) the original image, (b) the 1st-level of HA, (c) the 1st-level of SA, (d) the 2nd-level of HA, (e) the 2nd-level of SA, (f) the 3rdlevel of HA, (g) the 3rd-level of SA.

Table 8. Comparisons of model size and complexity. FLOPs: the number of FLoating-point OPerations; PN: Parameter Number.

Model	FLOPs	PN (million)	Depth
AlexNet [15]	$\mathbf{7.25\times10^8}$	58.3	7
VGG16 [26]	1.55×10^{10}	134.2	16
ResNet50 [8]	3.80×10^9	23.5	**50**
GoogLeNet [30]	1.57×10^9	6.0	22
JLML	1.54×10^9	7.2	39
HA-CNN	1.09×10^9	**2.7**	39

5. Conclusion

In this work, we presented a novel Harmonious Attention Convolutional Neural Network (HA-CNN) for joint learning of person re-identification attention selection and feature representations in an end-to-end fashion. In contrast to most existing re-id methods that either ignore the matching misalignment problem or exploit stringent attention learning algorithms, the proposed model is capable of extracting/exploiting multiple complementary attention and maximising their latent complementary effect for person re-id in a unified *lightweight* CNN architecture. This is made possible by the Harmonious Attention module design in combination with a two-branches CNN architecture. Moreover, we introduce a cross-attention interaction learning mechanism to further optimise joint attention selection and re-id feature learning. Extensive evaluations were conducted on three re-id benchmarks to validate the advantages of the proposed HA-CNN model over a wide range of state-of-the-art methods on both manually labelled and more challenging auto-detected person images. We also provided detailed model component analysis and discussed HA-CNN's model complexity as compared to popular alternatives.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. 6

[2] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 7

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017. 2

[4] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE TPAMI*, 2017. 2, 7

[5] T. H. Cormen. *Introduction to algorithms*. MIT press, 2009. 4

[6] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *SIGKDD*, 2004. 5

[7] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, January 2014. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 7, 8

[9] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv*, 2017. 3, 4

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 3

[11] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. 2, 3, 5

[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[13] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2

[14] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 4

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2, 3, 8

[16] X. Lan, H. Wang, S. Gong, and X. Zhu. Deep reinforcement learning attention selection for person re-identification. In *BMVC*, 2017. 2

[17] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 1, 2, 7

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 6

[19] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017. 2, 6, 7, 8

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2, 7

[21] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE TIP*, 2017. 7

[22] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 2, 7

[23] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 2

[24] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, 2017. 2

[25] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 1, 2

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2, 3, 8

[27] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 1, 2, 7

[28] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 7

[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 3

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2, 8

[31] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 7

[32] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 7

[33] S. Vossel, J. J. Geng, and G. R. Fink. Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, 20(2):150–159, 2014. 3

[34] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *CVPR*, 2017. 3, 4

[35] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 2

[36] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 1, 2

[37] H. Wang, S. Gong, and T. Xiang. Highly efficient regression for scalable person re-identification. In *BMVC*, 2016. 6, 7

[38] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE TPAMI*, 2016. 2

[39] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 2, 3

[40] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014. 2

[41] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 2, 7

[42] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 1, 2, 7

[43] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2

[44] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2

[45] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE TPAMI*, pages 653–668, March 2013. 2

[46] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, pages 4678–4686, 2015. 1, 2

[47] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 2, 6, 7

[48] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 7