Harmonious Attention Network for Person Re-Identification

Wei Li\textsuperscript{1} \hspace{1cm} Xiatian Zhu\textsuperscript{2} \hspace{1cm} Shaogang Gong\textsuperscript{1}
Queen Mary University of London\textsuperscript{1} \hspace{1cm} Vision Semantics Ltd.\textsuperscript{2}
{w.li, s.gong}@qmul.ac.uk \hspace{1cm} 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 \cite{49, 20, 27}. 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 \cite{9}. 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 \cite{48, 28, 51, 39}. However, these methods rely on hand-crafted 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 \cite{19, 47, 30, 2}. 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 training whilst also facing noisy person images of arbitrary misalignment and background clutter.

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 Convolutional 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 [20], Market-1501 [49], and DukeMTMC-ReID [52].

2. Related Work

Most existing person re-id methods focus on supervised learning of identity-discriminative information, including ranking by pairwise constraints [25, 42, 43], discriminative distance metric learning [15, 50, 45, 22, 46], and deep learning [26, 20, 4, 44, 38, 41, 21, 5]. 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 [28, 51] and saliency weighting [39, 48]. 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 [19, 47, 30, 18]. 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. [30] integrate a separately trained pose detection model (from additional labelled pose ground-truth) into a part-based re-id model. Li et al. [19] 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. [47] exploit the Spatial Transformer Network [13] 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. While soft re-id attention modelling is considered in [24], 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 \( \mathcal{I} = \{ I_i \}_{i=1}^{n} \) from \( n_d \) distinct people captured by non-overlapping camera views together with the corresponding identity labels as \( \mathcal{Y} = \{ y_i \}_{i=1}^{n} \) (where \( y_i \in [1, \ldots, n_d] \)), 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 CNN network typically with a large number of model parameters (likely overfit given small size labelled data) and high computational cost in model deployment [17, 29, 33, 10], 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 architecture composition 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 architec-
Figure 2. The Harmonious Attention Convolutional Neural Network. The symbol \( d_l \) \((l \in \{1, 2, 3\})\) denotes the number of convolutional filters in the corresponding Inception unit at the \( l \)-th block.

ature 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 entire person images\)\: For both branches, we select the Inception-A/B units \([44, 32]\) as the basic building blocks.

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.

3.1. Harmonious Attention Learning

Conceptually, our Harmonious Attention \( (HA) \) is a principled union of hard \( (\)regional\) attention \([13]\), soft spatial \([37]\) and channel attention \([11]\). This simulates functionally the dorsal and ventral attention mechanism of human brain \([36]\) in the sense of modelling soft and hard attention simultaneously. Soft attention learning aims at selecting \( (\)fine-grained\) important pixels, while hard attention learning searching \( (\)coarse\) latent \( (\)weakly supervised\) discriminative regions. They are 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.

In particular, 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 [7]. 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 intuitive and essential for 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.

(1) Soft Spatial-Channel Attention The input to a Harmonious Attention module is a 3-D tensor $X^l \in \mathcal{R}^{h \times w \times c}$ 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 $A^l \in \mathcal{R}^{h \times w \times c}$ of the same size as $X$. 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:

$$A^l = S^l \times C^l \quad (1)$$

where $S^l \in \mathcal{R}^{h \times w \times 1}$ and $C^l \in \mathcal{R}^{1 \times 1 \times c}$ 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 $S^l$ (shared across the channel dimension), and another branch to model the channel attention $C^l$ (shared across both height and width dimensions). By this design, we can compute efficiently the full soft attention $A^l$ from $C^l$ and $S^l$ with a tensor multiplication. Our design is more efficient than common tensor factorisation algorithms [16] 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 $3 \times 3$ 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,

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

is designed especially to compress the input size of the subsequent conv layer with merely $\frac{1}{2}$ 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 ($2 \times \frac{c^2}{r}$ parameters, see more details below) 4-layers squeeze-and-excitation component (Fig. 3(c)). 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

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

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

$$C^l_{\text{excitation}} = \text{ReLU}(W^{ca}_{2} \times \text{ReLU}(W^{l}\cdot C^l_{\text{input}})) \quad (4)$$

where $W^{ca}_{2} \in \mathcal{R}^{\frac{c}{r} \times c}$ ($\frac{c^2}{r}$ parameters) and $W^{l} \in \mathcal{R}^{c \times \frac{c}{r}}$ ($\frac{c^2}{r}$ parameters) denote the parameter matrix of 2 conv layers in order respectively, and $r$ (16 in our implementation) represents the bottleneck reduction rate. This bottleneck design reduces the model parameter number from $c^2$ (using 1 conv layer) to $\left(\frac{c^2}{r} + \frac{c^2}{r}\right)$, e.g. only need $\frac{1}{8}$ parameters when $r=16$.

For facilitating the combination of the spatial attention and channel attention, we further deploy a $1 \times 1 \times c$ 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 Residual Attention (RA) [37] and Squeeze-and-Excitation (SE) [11] concepts but with a number of essential differences: (1) The RA requires to learn a much more complex soft attention sub-network which is not only computationally expensive but 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
attention (Eq. (c)) exploit the first layer output (a)
lower overfitting risk.

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) 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 $X_G^{(l,k)}$ of the $k$-th region to enrich the corresponding local-branch feature $X_L^{(l,k)}$ by tensor addition as
\[
\hat{X}_L^{(l,k)} = X_L^{(l,k)} + X_G^{(l,k)}
\] (6)
where $X_G^{(l,k)}$ 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
\[
\Delta W_G^{(l)} = \frac{\partial L_G}{\partial X_G^{(l)}} \frac{\partial X_G^{(l)}}{\partial W_G^{(l)}} + \sum_{k=1}^{T} \frac{\partial L_L}{\partial X_L^{(l,k)}} \frac{\partial X_L^{(l,k)}}{\partial W_G^{(l)}}
\] (7)

So, the global $L_G$ and local $L_L$ loss quantities concurrently function in optimising the parameters $W_G^{(l)}$ 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 $I^p$ from one camera view and a set of test gallery images $\{I^g_i\}$ 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 $x^p = [x_g^p; x_t^p]$ and $x^g = [x_g^g; x_t^g]$. (2) We then compute L2 normalisation on the global and local features, respectively. (3) Lastly, we compute the cross-camera matching distances between $x^p$ and $x^g$ 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

Datasets and Evaluation Protocol For evaluation, we selected three large-scale person re-id benchmarks, Market-1501 [40], DukeMTMC-ReID [52] and CUHK03 [20]. 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.

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

Implementation Details We implemented our HA-CNN model in the Tensorflow [1] framework. All person images are resized to 160x64. For HA-CNN architecture, we set the width of Inception units at the $1^{st}/2^{nd}/3^{rd}$ levels as: $d_1 = 128$, $d_2 = 256$ and $d_3 = 384$. Following [21], 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 $24 \times 28$, $12 \times 14$ and $6 \times 7$. For model optimisation, we use the ADAM [14] algorithm at the initial learning rate $5 \times 10^{-4}$ 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 augmentation 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 the State-of-the-Art

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 JML (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, DLP 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 relying 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. 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 ap-

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

<table>
<thead>
<tr>
<th>Query Type</th>
<th>Single-Query</th>
<th>Multi-Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure (%)</td>
<td>R1 mAP</td>
<td>R1 mAP</td>
</tr>
<tr>
<td>XQDA[22]</td>
<td>43.8</td>
<td>22.2</td>
</tr>
<tr>
<td>SCS[3]</td>
<td>51.9</td>
<td>26.3</td>
</tr>
<tr>
<td>DNS[46]</td>
<td>61.0</td>
<td>35.6</td>
</tr>
<tr>
<td>CRAFT[6]</td>
<td>68.7</td>
<td>42.3</td>
</tr>
<tr>
<td>CAN[23]</td>
<td>60.3</td>
<td>35.9</td>
</tr>
<tr>
<td>S-LSTM[35]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>G-SCNN[34]</td>
<td>65.8</td>
<td>39.5</td>
</tr>
<tr>
<td>HPN[24]</td>
<td>76.9</td>
<td>-</td>
</tr>
<tr>
<td>SVDDNet[31]</td>
<td>82.3</td>
<td>62.1</td>
</tr>
<tr>
<td>MSCAN[19]</td>
<td>80.3</td>
<td>57.5</td>
</tr>
<tr>
<td>DLP[47]</td>
<td>81.0</td>
<td>63.4</td>
</tr>
<tr>
<td>PDC[30]</td>
<td>84.1</td>
<td>63.4</td>
</tr>
<tr>
<td>JML[21]</td>
<td>85.1</td>
<td>65.5</td>
</tr>
<tr>
<td>HA-CNN</td>
<td>91.2</td>
<td>75.7</td>
</tr>
</tbody>
</table>

Figure 5. Example cross-view matched pairs from three datasets.
proach in a more challenging re-id scenario. Importantly, the performance advantage by our method is achieved at a 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 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 / 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 [21] (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 & local-level features. Setting: SQ.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Market-1501</th>
<th>DukeMTMC-ReID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric (%)</td>
<td>R1 mAP</td>
<td>R1 mAP</td>
</tr>
<tr>
<td>Global</td>
<td>89.9</td>
<td>72.5</td>
</tr>
<tr>
<td>Local</td>
<td>88.9</td>
<td>71.7</td>
</tr>
<tr>
<td>Global+Local</td>
<td>91.2</td>
<td>75.7</td>
</tr>
</tbody>
</table>

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 [17], VGG16 [29], GoogLeNet [33], and ResNet50 [10]) 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 \times 10^9) and yet, still retains the 2nd deepest structure (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.

Acknowledgements

This work was partially supported by the China Scholarship Council, Vision Semantics Ltd, Royal Society Newton Advanced Fellowship Programme (NA150459), and InnovateUK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety.
References

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


