View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Queen Mary Research Online

Scalable Person Re-Identification by Harmonious Attention

Wei Li, Xiatian Zhu & Shaogang Gong

International Journal of Computer Vision

ISSN 0920-5691

Int J Comput Vis
DOI 10.1007/s11263-019-01274-1

Springer

Your article is published under the Creative Commons Attribution license which allows users to read, copy, distribute and make derivative works, as long as the author of the original work is cited. You may self-archive this article on your own website, an institutional repository or funder's repository and make it publicly available immediately.

International Journal of Computer Vision
https://doi.org/10.1007/s11263-019-01274-1

Scalable Person Re-Identification by Harmonious Attention

Wei Li[1] · **Xiatian Zhu**[2] · **Shaogang Gong**[1]

Received: 5 October 2018 / Accepted: 3 December 2019
© The Author(s) 2019

Abstract
Existing person re-identification (re-id) deep learning methods rely heavily on the utilisation of large and computationally expensive convolutional neural networks. They are therefore *not scalable* to large scale re-id deployment scenarios with the need of processing a large amount of surveillance video data, due to the lengthy inference process with high computing costs. In this work, we address this limitation via jointly learning re-id attention selection. Specifically, we formulate a novel *harmonious attention network* (HAN) framework to jointly learn soft pixel attention and hard region attention alongside simultaneous deep feature representation learning, particularly enabling more discriminative re-id matching by *efficient* networks with more scalable model inference and feature matching. Extensive evaluations validate the cost-effectiveness superiority of the proposed HAN approach for person re-id against a wide variety of state-of-the-art methods on four large benchmark datasets: CUHK03, Market-1501, DukeMTMC, and MSMT17.

Keywords Person re-identification · Scalable search · Compact model · Attention learning · Local and global representation learning

1 Introduction

Person re-identification (re-id) aims to search people across non-overlapping surveillance camera views deployed at different locations by matching auto-detected person bounding box images. With the 24/7 operating nature of surveillance cameras, person re-id is *intrinsically* a large scale search problem with a fundamental requirement for developing systems with both *fast data throughput* (i.e. low inference cost) and *high matching accuracy*. This is because, model accuracy and inference efficiency *both* are key enabling factors for affordable real-world person re-id applications. In this paper, we define this *cost-effectiveness* measure as the *scalability* of a person re-id system, taking into account model accuracy and computational cost *jointly*, *rather than optimising either alone*.

Earlier person re-id methods in the literature rely on slow-to-compute high-dimensional hand crafted features with inferior model performance, yielding unsatisfactory solutions (Zheng et al. 2013; Liao et al. 2015; Matsukawa et al. 2016; Zhang et al. 2016; Wang et al. 2018b). The recent introduction of large scale person re-id datasets (Wei et al. 2018; Zheng et al. 2015a; Li et al. 2014; Ristani et al. 2016) allows for a natural utilisation of increasingly powerful deep neural networks (He et al. 2016; Szegedy et al. 2017; Huang et al. 2017), substantially improving person re-id accuracy in a single system pipeline.

However, typical existing deep learning re-id methods remain *large sized* and *computationally expensive* therefore unfavourable for real deployments in scalability. This is due to the adoption of deep and wide neural network architectures with a huge number of parameters and exhaustive multiply-add operations. For example, the often-selected CNN architecture ResNet50 (He et al. 2016) consists of 25.1 million parameters consuming 3.80×10^9 FLoating-point OPerations (FLOPs) in forwarding a single person image through the network. While the offline training of large neural networks can be reasonably afforded using industrial-sized or cloud computing clusters with rich high-performance graph-

Communicated by T.E. Boult.

✉ Xiatian Zhu
eddy.zhuxt@gmail.com

Wei Li
w.li@qmul.ac.uk

Shaogang Gong
s.gong@qmul.ac.uk

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

[2] Vision Semantics Limited, London E1 4NS, UK

Published online: 23 December 2019

ics processing units (GPUs), deploying them to process *big video data* suffers from low inference efficiency and expensive energy consumption. There is an intrinsic need for designing cost-effective deep learning re-id methods, which is currently less investigated with insufficient research efforts and attempts.

One intuitive approach to large scale person re-id is to train efficient small neural network models. This is made more attractive by the development of lightweight architectures, e.g. MobileNet (Howard et al. 2017), ShuffleNet (Zhang et al. 2018a), and CondenseNet (Huang et al. 2018b). These methods are based on the observation that there exist highly redundant weights in large neural networks (Denil et al. 2013). However, such networks, originally designed for generic object classification and detection, are less effective for visually fine-grained and subtle person re-id matching (Table 7). It is non-trivial to *simultaneously* achieve both generalisation performance and inference efficiency by a single deep learning person re-id model.

In this work, we investigate *the under-studied scalability and cost-effectiveness problem in deep learning person re-identification*. To this end, we explore the potential of *person attention selection learning* in a single neural network architecture. The rationale is that detecting the fine-grained salient parts of person images not only allows to preserve the model matching performance, but also favourably simplifies the re-id matching due to noise suppression, therefore *rendering small networks sufficient* to induce this simplified target matching function. It is this re-id attention selection learning strategy that distinguishes our method from existing purpose-generic network compression techniques (Howard et al. 2017; Zhang et al. 2018a; Huang et al. 2018b), enabling uniquely a simultaneous realisation of model efficiency and generalisation performance. Owing to the conceptual orthogonality, existing network compression techniques can be naturally integrated as complementary designs into our approach to achieve further model efficiency and scalability.

There have been a number of existing attempts at learning re-id attention selection. Nevertheless, their primary purpose is to address the person misalignment issue for higher model generalisation capability. This is because in practical re-id scenarios, person images are usually automatically detected with arbitrary cropping errors for scaling up to large video data (Zheng et al. 2015a; Li et al. 2014; Ristani et al. 2016). Additionally, uncooperative people are often captured in various poses across open space and time. There is consequently an inevitable need for attention selection within arbitrarily-aligned bounding boxes as an integral part of model learning for re-id.

A common earlier strategy for re-id attention selection is local patch calibration and saliency weighting in pairwise image matching (Zhao et al. 2013; Shen et al. 2015; Zheng et al. 2015b; Wang et al. 2014). This approach relies on pre-fixed hand-crafted features without deep learning jointly more expressive representations and a matching metric in an end-to-end manner. A number of more advanced attention deep learning models for person re-id have been recently developed (Li et al. 2017a; Zhao et al. 2017; Su et al. 2017; Lan et al. 2017; Xu et al. 2018; Wang et al. 2018a; Qian et al. 2018; Suh et al. 2018). Most of these methods consider only coarse region-level attention whilst ignoring the fine-grained pixel-level saliency. Moreover, such methods depend on heavy network architectures therefore suffering the drawbacks of high computational complexity and low model inference efficiency. Our work addresses the weaknesses and limitations of these existing methods for scalable person re-id with both superior matching accuracy and inference efficiency.

We make three *contributions* in this work.

(I) We investigate the under-studied model cost-effectiveness and scalability issue in deep learning person re-id, including model accuracy, inference cost, and matching efficiency. This differs substantially from the existing methods usually ignoring the model efficiency problem whilst only focusing on improving re-id accuracy rates. Through studying this problem, we aim for addressing large scale person re-id deployments typical in practical applications.

(II) We formulate a novel idea of jointly learning multi-granularity attention selection and feature representation for optimising person re-id cost-effectiveness in deep learning. To our knowledge, this is the first attempt at jointly deep learning multiple complementary attention for solving the person re-id scalability problem. The proposed approach is technically orthogonal to existing designs of efficient neural networks therefore allowing for implementing complementary strengths by concurrent integration in a hybrid architecture.

(III) We propose a *harmonious attention network* (HAN) framework to simultaneously learn hard region-level and soft pixel-level along with re-id feature representations for maximising the correlated complementary information between attention selection and feature discrimination in a compact architecture. This is achieved by devising an efficient Harmonious Attention module capable of efficiently and effectively learning different types of attention from the re-id feature representation hierarchy in a multi-task and end-to-end learning fashion. In the harmonious attention, we introduce a cross-attention interaction learning scheme to further enhance the compatibility between attention selection and feature representation, subject to the same re-id discriminative training constraints.

Extensive comparative evaluations demonstrate the cost-effectiveness superiority of the proposed HAN approach over a wide variety of state-of-the-art person re-id models and efficient neural networks on four large benchmark datasets: CUHK03 (Li et al. 2014), Market-1501 (Zheng et al. 2015a),

DukeMTMC (Ristani et al. 2016), and MSMT17 (Wei et al. 2018). Preliminary versions of this study have been presented on concurrently learning global and local feature representations (Li et al. 2017b) and jointly leaning harmonious re-id attention (Li et al. 2018b). Built on the two previous studies, this work presents a further more cost-effective person re-id deep learning framework that enables more efficient model deployments therefore more scalable to processing large surveillance video data.

2 Related Work

Person Re-ID The state-of-the-art person re-id deep methods are typically concerned with supervised learning of identity-discriminative representations (Qian et al. 2017; Chen et al. 2017a; Xiao et al. 2016; Wang et al. 2016a; Li et al. 2017b; Chen et al. 2017b; Kalayeh et al. 2018; Song et al. 2018; Chang et al. 2018a; Wei et al. 2018). Although unsupervised learning and transfer learning based techniques are progressively advancing (Wang et al. 2018c; Zhong et al. 2018a; Li et al. 2018a; Chen et al. 2018b; Kodirov et al. 2015; Peng et al. 2016; Wang et al. 2016b), their re-id performances are significantly inferior therefore less satisfactory and reliable in practical use.

With the emergence of large benchmark datasets (Li et al. 2014; Zheng et al. 2015a, 2017; Wei et al. 2018), more powerful and computationally expensive neural networks like ResNet50 (He et al. 2016), originally designed for object image classification, have been increasingly adopted in building person re-id model architectures. The use of stronger and heavier networks yields significant gains in performance, but simultaneously sacrifices largely the deployment efficiency due to the need for high memory and computing consumption apart from lengthy model inference. Such inefficient systems suffer from low data throughput, therefore limiting the possible application scenarios (*undesired* in processing a large pool of surveillance videos).

Model Efficiency In the literature, model efficiency is an under-studied and critical problem in person re-id. Zheng et al. (2015a) employed a KD-tree based approximate nearest neighbour (ANN) method to expedite the re-id matching process. As another ANN strategy, the learning-to-hash idea has been explored with hand-crafted (Zhu et al. 2018) and deep learning (Zhang et al. 2015; Zhu et al. 2017) models. These methods quantise the feature representations so that the hamming distance metric can be applied to rapidly compute matching scores at the cost of significant performance degradation due to limited expressive capacity. Recently, Wang et al. (2018e) proposed to conduct re-id matching subject to given computation budgets. The hypothesis is that feature representations of easy samples can be computed at lower costs which makes room for computation reduction. However, it is intrinsically difficult and ambiguous to measure the sample easiness degree given the poor-quality surveillance data and the nature of pairwise matching (*not* per-sample inference).

Unlike all these existing strategies, we explore differently the potential of person attention learning for model efficiency and cost-effectiveness. Conceptually, our method is complementary to the prior techniques with extra possible performance benefits.

Attention Learning There exist *learning-to-attend* algorithms developed for improving re-id particularly in misaligned person bounding boxes, e.g. those generated by automatic detection. Earlier approaches are based on localised patch matching (Shen et al. 2015; Zheng et al. 2015b) and saliency weighting (Wang et al. 2014; Zhao et al. 2013). These solutions are mostly ineffective to cope with poorly aligned person images, due to the stringent requirement of tight bounding boxes around the whole person and high dependence on weak hand-crafted features. Besides, such algorithms are usually more computationally expensive with a need for explicit and complex patch processing.

To overcome the aforementioned limitation, more advanced re-id attention deep learning methods have been recently proposed (Li et al. 2017a; Zhao et al. 2017; Su et al. 2017; Lan et al. 2017; Xu et al. 2018; Wang et al. 2018a; Qian et al. 2018; Suh et al. 2018). A common strategy taken by these methods is to incorporate a regional attention (i.e. *hard attention*) selection sub-network into a deep re-id model. For example, Su et al. (2017) integrated a pose detection model separately learned from auxiliary pose ground-truth into a part-based re-id model. Li et al. (2017a) designed an end-to-end trainable part-aligning CNN for extracting latent discriminative regions and exploiting these regional features to perform re-id. Zhao et al. (2017) exploited a Spatial Transformer Network (Jaderberg et al. 2015) as a hard attention module to search re-id discriminative parts given a pre-defined spatial constraint. Lan et al. (2017) formulated a reinforcement attention selection model for salient region refinement under identity discriminative constraints. Qian et al. (2018) rotated persons to canonical poses through pose-specific image synthesising.

A common weakness of these above models is the lack of handling noisy pixel information within selected regions, i.e. no *soft attention* modelling. This issue was considered in (Liu et al. 2017b). However, this model assumes tight person bounding boxes therefore not suitable for processing poor detections. In parallel to our work, Xu et al. (2018) considered a joint end-to-end learning of both body parts and regional saliency. Along with continuously improved matching performance, all the attention learning re-id methods come with significantly increased model complexity and inference costs for realising strong model generalisation capability. This dra-

matically limits their scalability and usability in large scale re-id deployments.

Beyond the conventional advantage from the attentional weighing for more discriminative person matching, the proposed HAN approach is specially designed to simultaneously address the efficiency weaknesses of existing re-id attention methods. This is achieved by formulating a novel *harmonious attention* module that enables a efficient joint learning of both soft and hard attention in compact CNN architectures whilst preserving the model generalisation capability. The results of this design are a class of cost-effective harmonious attention networks dedicated for scalable re-id matching with state-of-the-art accuracy performance. 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 further enhancing the complementary effect between different levels of attention subject to re-id discriminative constraints. This is impossible to do for most existing methods due to their inherent single level attention modelling design. We show the benefits of joint modelling multi-level attention in person re-id in our experiments.

Efficient Neural Networks Conceptually, our work is related to the approaches for designing generic compact networks, e.g. weight pruning (LeCun et al. 1990; Hassibi et al. 1993; Li et al. 2016; He et al. 2017; Luo et al. 2017; Liu et al. 2017c; Huang and Wang 2018), model quantisation (Courbariaux et al. 2015; Rastegari et al. 2016; Hubara et al. 2016; Faraone et al. 2018), and efficient network architectures (Szegedy et al. 2015; He et al. 2016; Iandola et al. 2016; Chollet 2017; Zhang et al. 2017; Zoph et al. 2018; Howard et al. 2017; Zhang et al. 2018a; Mehta et al. 2018). Instead of aiming to formulate some tight building blocks or identifying redundant parameters as these existing methods, the proposed method differently exploits the attention learning mechanism to densely concentrate and more efficiently mine the *intrinsic* learning capacity of a network model. Therefore, our method enables to complement prior efficient model designs. For instance, to control model computational complexity we also use the depth-wise separable convolution, a type of building block commonly selected for making existing lightweight networks (Sifre and Mallat 2014; Szegedy et al. 2015; Chollet 2017; Mehta et al. 2018). Importantly, our method often surpasses the existing purpose-generic efficient networks in re-id performance at the similar computational budgets, thanks to the unique ability of detecting identity relevant information in person images against distracting backgrounds.

More broadly, other related techniques include dynamic networks (Figurnov et al. 2017; Bolukbasi et al. 2017; Huang et al. 2018a) that automatically adjust the model inference for each test sample, and model distillation (Ba and Caruana 2014; Hinton et al. 2015; Zhang et al. 2018b; Lan et al. 2018a, b) that enhances the training of small networks by transferring the knowledge of a larger-capacity teacher model. Conceptually, these techniques are also orthogonal to the proposed attention based model compression, therefore enabling them to complete each other in a single neural network architecture.

3 Scalable Person Re-Identification

Problem Definition Suppose there are 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 class labels $\mathcal{Y} = \{y_i\}_{i=1}^{n}$ where $y_i \in [1, \ldots, n_{\mathrm{id}}]$. We aim to learn a deep feature representation model optimal for person re-id matching under significant viewing condition variations with high computational efficiency.

To that end, we formulate a *lightweight* (less parameters and multiplication-addition operations) *harmonious attention network* (HAN). This takes a principle of attention learning particularly for achieving cost-effective person re-id. The objective of HAN is to concurrently learn a set of harmonious attention along with both global and local feature representations for maximising their complementary benefit and compatibility between the model components in terms of both the discriminative capability and inference efficiency.

Formulation Rationale The HAN model design is based on two motivating considerations: (1) The human visual system that leverages both global contextual and local saliency information concurrently in conjunction with the evolution attention search capability (Navon 1977; Torralba et al. 2006); (2) The divide-and-conquer algorithm design principle (Cormen et al. 2009) that decomposes the highly non-linear re-id feature learning task at different levels of granularity (global & local) and significance (salient or not), simplifying the target problem formulation and enabling efficient small networks suffice to model the desired representations. Intuitively, joint learning of global-local feature representations with attention extracts correlated complementary information in different context, hence efficiently achieving more reliable recognition due to such selective discrimination learning.

For bounding box image based re-id, we consider the entire person in the image as a *global scene context* and body parts of the person as *local information sources*, *both* subject to the surrounding background clutter, potentially also misalignment and partial occlusion due to poor detections. Typically, the location information of body parts is not provided in person re-id image annotation, i.e. only weakly labelled without fine-grained part labels. Therefore, the person attention learning is a *weakly supervised* learning task in the context of optimising the re-id performance.

Under the global-local concurrent design, we consider a multi-branch network architecture. The overall objective of this multi-branch scheme and the architecture composition is to minimise the target function modelling complexity in a divide-and-conquer model decomposition. This enables reducing the network parameter size whilst still maintaining the model representation learning capacity.

HAN Overview The overall design of our HAN architecture is depicted in Fig. 1. In particular, the attention model contains two branches with hierarchically distributed attention modules: (1) One *local branch* consisting of T streams with an identical structure: Each stream aims to learn the most discriminative visual features for one of T local regions of a person bounding box image. To reduce the model parameter size, we share the layer parameters among all local streams. (2) One *global branch*: This aims to learn the optimal global level features from the entire person images. (3) Hierarchical *harmonious attention learning*: This aims to discover and exploit re-id discriminant saliency regions (hard attention) and pixels (soft attention) concurrently in a synergistic interaction with global and local feature representations in an end-to-end learning manner. Next, we describe the main designs of the proposed HAN model.

3.1 Multi-Task Global-Local Feature Learning

The HAN model is designed to perform *multi-task global-local representation learning subject to the same identity label constraints* by allocating each branch with a separate objective loss function derived from the per-batch training person classes. As a consequence, the learning behaviour of each branch is conditioned respectively on their own feature representations.

Loss Function For model training, we use the softmax cross-entropy loss function. Formally, we start by predicting the class posterior probability $\tilde{y}_i$ of a person image $\boldsymbol{I}_i$ over the ground-truth identity class label y_i:

$$p(\tilde{y}_i = y_i|\boldsymbol{I}_i) = \frac{\exp(\boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i)}{\sum_{k=1}^{|n_{\text{id}}|} \exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)} \tag{1}$$

where $\boldsymbol{x}_i$ refers to the feature vector of $\boldsymbol{I}_i$ from the corresponding branch, and $\boldsymbol{w}_k$ the prediction function parameter of training identity class k. The cross-entropy loss for a mini-batch of n_{bs} training images is then defined as:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \log\Big(p(\tilde{y}_i = y_i|\boldsymbol{I}_i)\Big) \tag{2}$$

Sharing Low-Level Feature Learning In a HAN model, we construct the global and local branches on a common low-level conv layer, in particular the first conv layer. This is for facilitating the purpose of inter-branch common representation learning (Fig. 1). The intuition is that, the bottom conv layer captures elementary features such as edges and corners shared by both global and local patterns of person images. This design is in spirit to multi-task learning (Argyriou et al. 2007), where the local and global feature learning branches are viewed as two correlated learning tasks.

Besides, sharing the low-level layer reduces the model parameter size, not only mitigating the model overfitting risk but also improving the model inference efficiency. This is critical in learning person re-id models especially when the labelled training data is limited.

Attention in Hierarchy We take a *block-wise* attention module design, that is, each attention module is specifically optimised to attend the input feature representations at its own

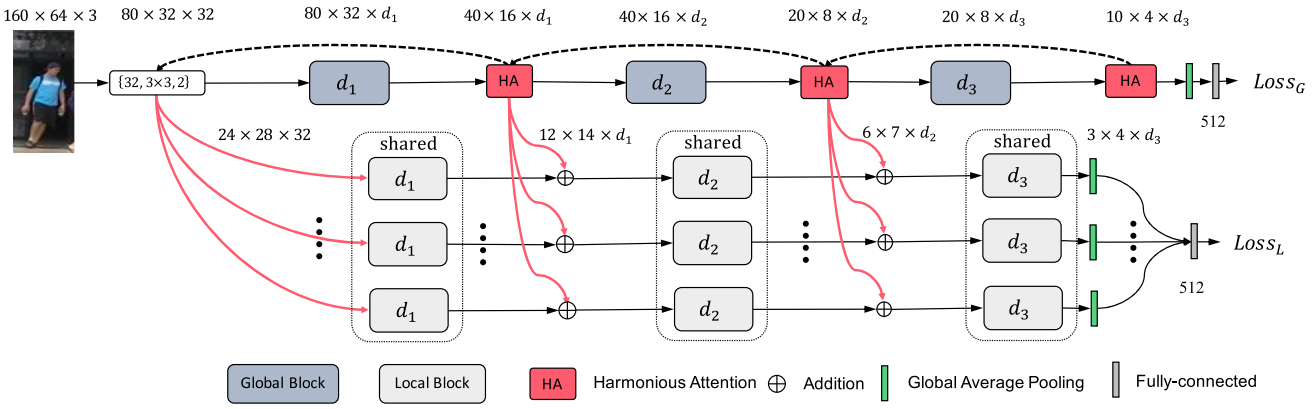


Fig. 1 Schematic architecture of the proposed harmonious attention network (HAN). The design of HAN is characterised by high cost-effectiveness for maximising the model scalability. This is enabled by the introduction of a lightweight but effective harmonious attention (HA) module (see Figs. 2, 3) and a computationally efficient depthwise separable convolution based building block (see Fig. 4 and Table 1). The symbol d_l ($l \in \{1, 2, 3\}$) denotes the number of convolutional filter in the corresponding l-th block

level *alone*. In the CNN hierarchical framework, this naturally allows for *hierarchical* multi-level attention learning to progressively refine the attention maps, again in a spirit of the divide-and-conquer design (Cormen et al. 2009). As a result, we can significantly reduce the attention search space (i.e. the model optimisation complexity) whilst allowing 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 varying, and (2) a re-id model typically needs robust (generalisable) model learning given very limited training data (significantly less than common image classification tasks).

Unlike most existing attention selection based person re-id works that simply adopt a standard CNN network with a large number of model parameters and high computational cost in model deployment (Krizhevsky et al. 2012; Szegedy et al. 2015; He et al. 2016; Xu et al. 2018), our HAN design is more efficient (faster inference in deployment) whilst still having deep CNN architectures to maintain strong discriminative power. 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 important for practical applications at scales (slow feature extraction is not scalable to large surveillance video data).

Remarks The HAN model aims to learn concurrently multiple re-id discriminative feature representations for different local image regions and the entire image. All these representations are optimised by maximising the *same* identity classification tasks individually and collectively at the same time. Concurrent multi-task learning in a multi-loss design enables to preserve both local saliency in feature selection and global coverage in image representation.

In terms of loss function design, while many existing person re-id methods (Shen et al. 2018a; Song et al. 2018; Wang et al. 2018e; Chen et al. 2018a; Wang et al. 2018d; Varior et al. 2016; Subramaniam et al. 2016; Ahmed et al. 2015; Li et al. 2014) suggest the importance of using pairwise comparison based loss objectives, e.g. the triplet and contrastive functions, we empirically found that the simpler cross-entropy loss suffices to achieve satisfied discriminative learning without any extra complexity introduced from hard sample mining. We partly attribute this to the strong capability of the HAN model in automatically attending re-id discriminative information, simplifying the loss design complexity.

Importantly, using only the classification loss formulation brings about a couple of practical benefits:

(i) This significantly *simplifies* the training mini-batch data construction, e.g. only random sampling without any notorious tricks required. This makes our HAN model more scalable to situations when very large training population is available and/or the training data from different camera views are highly imbalanced.

(ii) This also eliminates the *undesirable* need for carefully forming pairs and/or triplets in preparing re-id training samples, as in these existing methods, due to the inherent imbalanced negative and positive pair size distributions.

We consider that the key to person re-id is about model generalisation to unseen test identity classes given the training data from *independent* seen classes. This loss choice is supported by previous visual psychophysical findings that representations optimised for classification tasks generalise well to novel categories (Edelman 1998). We exploit this general classification learning principle beyond the stringent pairwise relative verification loss designs.

3.2 Harmonious Attention Learning

To perform attention selection within person bounding box images with unknown misalignment, we formulate a *harmonious attention learning* scheme. This is the core module component of the proposed model. Specifically, this scheme jointly learns a collection 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. Besides, we 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. Next, we describe the design details of the Harmonious Attention module.

(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:

$$\boldsymbol{A}^l = \boldsymbol{S}^l \times \boldsymbol{C}^l \tag{3}$$

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. 2a): 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 $\boldsymbol{C}^l$

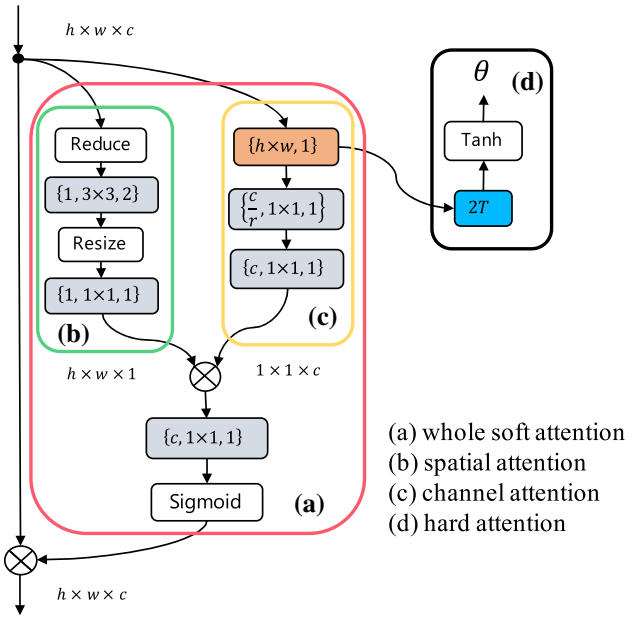


Fig. 2 Structure of a harmonious attention module consists of **a** Soft Attention which includes **b** Spatial Attention (pixel-wise) and **c** Channel attention (scale-wise), and **d** Hard regional attention (part-wise). Layer type is indicated by background colour: grey for *convolutional* (conv), 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. ReLU (Krizhevsky et al. 2012) and Batch Normalisation (Ioffe and Szegedy 2015) (applied to each conv layer) are not shown for brevity

and $\boldsymbol{S}^l$ with a tensor multiplication. Our design is more efficient than common tensor factorisation algorithms (Kolda and Bader 2009) since heavy matrix operations are eliminated.

(i) *Spatial Attention* We model the spatial attention by a tiny (10 parameters) 4-layers sub-network (Fig. 2b). 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 average pooling, formulated as

$$\boldsymbol{S}_{\text{input}}{}^{l} = \frac{1}{c}\sum_{i=1}^{c} \boldsymbol{X}^{l}_{1:h,1:w,i} \tag{4}$$

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 a scaling layer for automatically learning an adaptive fusion scale in order to optimally combining the channel attention described next.

(ii) *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 component (Fig. 2c). We first perform a *squeeze* operation via an averaging pooling layer (0 parameters) to aggregate the feature information distributed across the spatial space into a channel signature as

$$\boldsymbol{C}_{\text{input}}{}^{l} = \frac{1}{h \cdot w}\sum_{i=1}^{h}\sum_{j=1}^{w} \boldsymbol{X}^{l}_{i,j,1:c} \tag{5}$$

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

$$\boldsymbol{C}_{\text{excitation}}{}^{l} = \texttt{ReLU}\big(\boldsymbol{W}_2^{\text{ca}} \times \texttt{ReLU}(\boldsymbol{W}_1^{\text{ca}}\boldsymbol{C}_{\text{input}}{}^{l})\big) \tag{6}$$

where $\boldsymbol{W}_1^{\text{ca}} \in \mathcal{R}^{\frac{c}{r}\times c}$ ($\frac{c^2}{r}$ parameters) and $\boldsymbol{W}_2^{\text{ca}} \in \mathcal{R}^{c\times\frac{c}{r}}$ ($\frac{c^2}{r}$ 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 $(\frac{c^2}{r}+\frac{c^2}{r})$, 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 $1\times1\times c$ conv (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 a 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) (Wang et al. 2017) and Squeeze-and-Excitation (SE) (Hu et al. 2018) 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 (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:

$$\boldsymbol{A}^{l} = \begin{bmatrix} s_h & 0 & t_x \\ 0 & s_w & t_y \end{bmatrix} \tag{7}$$

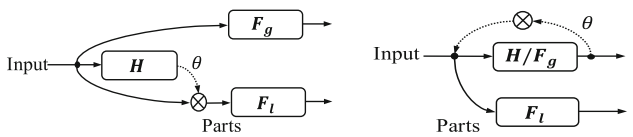
which enables image cropping, translation, and isotropic scaling operations by varying two scale factors (s_h, s_w) and

the 2-D spatial position (t_x, t_y). We pre-define the region size by fixing s_h and s_w for limiting the model complexity. 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. 2d). We exploit the first layer output (a c-D vector) of the channel attention (Eq. (5)) 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 multi-task learning (Evgeniou and Pontil 2004). 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 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. 1).

Remarks The proposed hard attention modelling is conceptually similar to the Spatial Transformer Network (STN) (Jaderberg et al. 2015) because both are designed to learn a transformation matrix for discriminative region identification and alignment. 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), but also provides additional attention refinement in a sequential manner. (2) The STN utilises a separate large sub-network for attention modelling whilst the HAN exploits a much smaller sub-network by sharing the majority model learning with the target-task network using a multi-task learning design (Fig. 3), therefore superior in both higher efficiency and lower overfitting risk. (3) The STN considers only hard attention whilst HAN models both soft and hard attention in an end-to-end fashion so that additional complementary benefits are exploited.



(a) STN (Jaderberg et al., 2015) **(b)** HAN Hard Attention

Fig. 3 Schematic comparison between **a** spatial transformer network (STN) (Jaderberg et al. 2015) and **b** HAN 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; "$\boldsymbol{F_l}$": Local feature module

(III) *Cross-Attention Interaction Learning* Given the joint learning of soft and hard attention as 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

$$\tilde{\boldsymbol{X}}_L^{(l,k)} = \boldsymbol{X}_L^{(l,k)} + \boldsymbol{X}_G^{(l,k)} \tag{8}$$

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. 1). 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

$$\Delta \boldsymbol{W}_G^{(l)} = \frac{\partial \mathcal{L}_G}{\partial \boldsymbol{X}_G^{(l)}} \frac{\partial \boldsymbol{X}_G^{(l)}}{\partial \boldsymbol{W}_G^{(l)}} + \sum_{k=1}^{T} \frac{\partial \mathcal{L}_L}{\partial \tilde{\boldsymbol{X}}_L^{(l,k)}} \frac{\partial \tilde{\boldsymbol{X}}_L^{(l,k)}}{\partial \boldsymbol{W}_G^{(l)}} \tag{9}$$

So, the global $\mathcal{L}_G$ and local $\mathcal{L}_L$ loss quantities concurrently function 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.

Conceptually, our Harmonious Attention (HA) is a principled union of hard *regional* attention (Jaderberg et al. 2015), soft *spatial* (Wang et al. 2017) and *channel* attention (Hu et al. 2018). This simulates functionally the dorsal and ventral attention mechanism of human brain (Vossel et al. 2014) in the sense of modelling soft and hard attention simultaneously. Soft attention learning aims at selecting *fine-grained* important pixels, whilst hard attention learning at searching *coarse* latent (weakly supervised) discriminative regions. They are thus largely complementary with high compatibility to each other in functionality. Intuitively, their combination and interaction can relieve the modelling burden of challenging soft attention learning, resulting in more discriminative and efficient models.

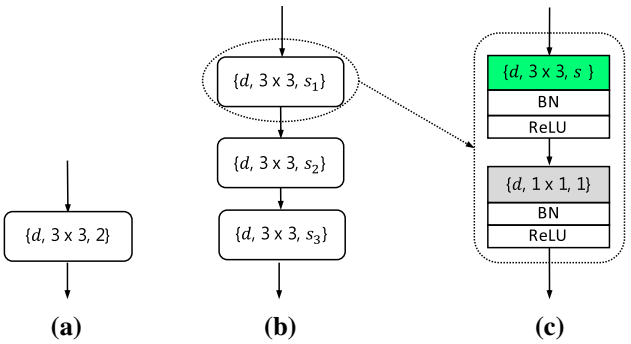


Fig. 4 Structure of **a** local block and **b** global block. Each block **c** consists of two conv layers. Layer type is indicated by background colour: grey for *normal conv*, and green for *depthwise separable conv* layers. The three items in the bracket of a conv layer are: filter number, filter shape, and stride

3.3 HAN Model Instantiation

To instantiate HAN models, we build up on the state-fo-the-art computationally efficient depthwise separable conv units (Sifre and Mallat 2014) in the main implementation[1]. In particular, we use 9 depthwise separable conv units to build the global branch, and 3 for each local stream. We set $T = 4$ regions for hard attention, e.g. a total of 4 local streams. We consider a 3-level attention hierarchy design (Fig. 1). The global branch 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 local streams.

To provide diverse options in model efficiency, we explore three HAN models with different inference computational costs. We realise this through varying the stride parameter s of the building block units in the relatively heavier global branch (Fig. 4b). More specifically, the computational cost (FLOPs) of a HAN model is largely determined by the size of input feature maps per conv layer in each block unit. The stride parameter controls the shifting step size the conv filters travel across the input feature maps, therefore the size of output feature maps and the computational cost of subsequent conv layers. Given the context of hierarchical CNN structure, larger stride values at earlier layer lead to smaller feature maps and lower FLOPs. In our designs, we adopt two strides $\{1, 2\}$ and manage the overall computational complexity of HAN models by positioning the larger stride "2" to different layers. That is, placing the stride "2" to earlier layers yields HANs with higher computational costs, and vice verse. The configurations of the three stride parameters in a global black are summarised in Table 1. We will evaluate all these HAN models in our experiments.

[1] Besides, we also consider other building block designs to evaluate the generalisation of the proposed method in our experiments (Table 12).

Table 1 The stride configuration for the building block units in the global branch of three varying-efficient HAN models

Model	s_1	s_2	s_3	FLOPs
HAN (Small)	2	1	1	3.68×10^8
HAN (Medium)	1	2	1	5.33×10^8
HAN (Large)	1	1	2	7.01×10^8

3.4 Scalable Person Re-ID by HAN

Given a trained HAN model, we obtain a 1,024-D joint feature vector (deep feature representation) by concatenating the local (512-D) and the global (512-D) branch 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 the corresponding 1,024-D feature representation vectors by forward-feeding the images to a trained HAN 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 apply 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 rank all gallery images in ascendant order by the L2 distances against the probe image. The percentage of true matches of the probe person image in top ranks indicate the goodness of the learned HAN deep features for person re-id matching.

4 Experiments

Datasets and Evaluation Protocol. For evaluation, we selected four large-scale re-id benchmark datasets: Market-1501 (Zheng et al. 2015a), DukeMTMC (Ristani et al. 2016), CUHK03 (Li et al. 2014), and MSMT17 (Wei et al. 2018). Figure 5 shows example person bounding box images. To make a fair comparison against existing methods, we adopted the standard person re-id evaluation setting including the training and testing ID split as summarised in Table 2. For DukeMTMC, we followed the person re-id evaluation setup as (Zheng et al. 2017). These datasets present diverse re-id test scenarios with varying training and testing scales under realistic viewing conditions exposed to large variations in human pose and strong similarities among different people, therefore enabling extensive model evaluations in both model learning and generalisation capabilities. We also considered

(a) CUHK03 **(b)** Market-1501 **(c)** DukeMTMC **(d)** MSMT17

Fig. 5 Example cross-view matched person image pairs randomly selected from four person re-id benchmark datasets

Table 2 Data statistics of person re-id datasets

Dataset	# ID	# Train	# Test	# Image	# Cam
CUHK03	1467	767	700	28,192	2
Market-1501	1501	751	750	32,668	6
DukeMTMC	1404	702	702	36,411	8
MSMT17	4101	1041	3060	126,441	15

the model performance with re-ranking (Zhong et al. 2017a) as a post-processing. To measure re-id accuracy performance, we used the cumulative matching characteristic (CMC) and mean average precision (mAP) metrics. For model efficiency measure, we used the FLoating-point OPerations (FLOPs), i.e. the number of multiply-adds, consumed in forwarding a person image through the testing network.

Implementation Details We implemented our HAN model in the tensorflow framework (Abadi et al. 2017). All person images were resized to 160×64, unless stated otherwise. For all HAN models, we set the width of building block units at the 1st/2nd/3rd levels as: $d_1 = 128$, $d_2 = 256$ and $d_3 = 384$ (Fig. 1). In each stream, we fixed the size of three levels of hard attention as 24×28, 12×14 and 6×7. For model training, we used SGD algorithm at the initial learning rate 3×10^{-2} with momentum of 0.9, learning rate decay of 0.1, and weight decay of 0.0005. We set the batch size to 32 and the epoch number to 300 with learning rate decayed at epochs of 100, 200, and 250. To enable efficient and scalable model training, we did *not* adopt any data argumentation methods (e.g. scaling, rotation, flipping, and colour distortion), *neither* ImageNet model pre-training. Existing deep re-id methods typically benefit significantly from these operations, suffering much higher computational cost, notoriously difficult and time-consuming model tuning, and the implicit undesired dependence on the source data.

4.1 Comparing State-of-the-Art Re-ID Methods

Evaluation on Market-1501 We evaluated the HAN models in comparison to recent state-of-the-art methods on the Market-1501 dataset. Table 3 shows clear superiority and advantages of the proposed HAN in model cost-effectiveness. Specifically, in the standard model training setting, G-SCNN (Varior et al. 2016) is featured with the best FLOPs, but far inferior to many alternative methods including all HAN models in terms of re-id performance. HAN (Small) is on par with the recent art MLFN (Chang et al. 2018b) in re-id matching accuracy whilst simultaneously achieving an efficiency advantage of $7\times$ cheaper inference.

With a re-ranking based post-processing, re-id models generally can further improve the accuracy performance. Note, this benefit comes with a higher computational cost, e.g. multiple times standard search expense. Interestingly, the fastest model HAN (Small) benefits the most, achieving superior model efficiency and discrimination simultaneously against other existing alternative methods.

As a training data augmentation strategy, random erasing is shown to be effective for improving re-id model generalisation. For example, the strong models GCS (Chen et al. 2018a) and SGGNN (Shen et al. 2018b) benefit significantly, achieving the best re-id matching rates. However, this model is also largely expensive in the computational cost, e.g. more than $10\times$ cost of HAN (Small).

When applying both random easing and re-ranking, a complementary benefit is likely to be obtained. In this setting, our HAN (Small) suffices to outperform the competitor RW (Shen et al. 2018a) in both accuracy performance and inference cost. If more computational budge is allowed, we can further improve the model performance by deploying HAN (Large).

Evaluation on DukeMTMC We compared the HAN models with recent state-of-the-art re-id methods on the DukeMTMC dataset. Compared to Market-1501, this benchmark provides a similar training and testing data scale, but the person images have more variations in resolution and background clutter. This is due to wider camera views and more complex scene layout, therefore presenting a more challenging re-id task.

Table 4 shows that all HAN methods again present superior model cost-effectiveness as compared to alternative state-of-the-art methods. Overall, we have similar compara-

Table 3 Performance evaluation on Market-1501

Query type	SQ		MQ		FLOPs
Metric (%)	R1	mAP	R1	mAP	
CRAFT (Chen et al. 2017c)	68.7	42.3	77.0	50.3	N/A
CAN (Liu et al. 2017a)	60.3	35.9	72.1	47.9	$>1.55\times10^{10}$
G-SCNN (Varior et al. 2016)	65.8	39.5	76.0	48.4	$\approx$ **1.11×10^{8}**
HPN (Liu et al. 2017b)	76.9	–	–	–	$\approx1.82\times10^{10}$
SVDNet (Sun et al. 2017)	82.3	62.1	–	–	$> 3.80\times10^{9}$
MSCAN (Li et al. 2017a)	80.3	57.5	86.8	66.7	1.36×10^{9}
DLPA (Zhao et al. 2017)	81.0	63.4	–	–	$> 7.29\times10^{8}$
PDC (Su et al. 2017)	84.1	63.4	–	–	$\gg 9.82\times10^{9}$
GLAD (Wei et al. 2017)	89.9	73.9	–	–	$\gg 7.99\times10^{9}$
DPFL (Chen et al. 2017b)	88.9	73.1	92.3	80.7	$\approx 1.2\times10^{10}$
AACN (Xu et al. 2018)	85.9	66.9	89.8	75.1	$> 1.57\times10^{9}$
DML (Zhang et al. 2018b)	89.3	70.5	92.8	79.0	5.69×10^{8}
DaRe-D201(Wang et al. 2018e)	86.0	69.9	–	–	$>4.00\times10^{9}$
PT-D169 (Liu et al. 2018)	87.7	68.9	–	–	$>3.00\times10^{9}$
AOS (Huang et al. 2018c)	86.5	70.4	91.3	78.3	$\approx 3.80\times10^{9}$
BraidNet (Wang et al. 2018d)	83.7	69.5	–	–	$>2.26\times10^{9}$
MLFN (Chang et al. 2018b)	90.0	74.3	92.3	82.4	$\approx 2.60\times10^{9}$
CamStyle (Zhong et al. 2018b)	88.1	68.7	–	–	$\approx 3.80\times0^{9}$
PSE (Saquib et al. 2018)	87.7	69.0	–	–	$>3.80\times10^{9}$
KPM (Shen et al. 2018a)	90.1	75.3	–	–	$>3.80\times10^{9}$
PoseNorm (Qian et al. 2018)	89.4	72.6	92.9	80.2	$> 3.80\times10^{9}$
HAP2S (Yu et al. 2018)	84.6	69.4	90.2	76.8	$\approx3.80\times10^{9}$
HAN (Small) (Ours)	89.4	73.2	93.2	80.8	3.68×10^{8}
HAN (Medium) (Ours)	90.7	74.5	93.9	81.9	5.33×10^{8}
HAN (Large) (Ours)	**91.6**	**76.7**	**94.2**	**83.4**	7.01×10^{8}
AACN[a](Xu et al. 2018)	88.7	83.0	92.2	87.3	$>1.57\times10^{9}$
DaRe-D201[a](Wang et al. 2018e)	88.6	82.2	–	–	$> 4.00\times10^{9}$
MGCAM[a](Song et al. 2018)	83.8	74.3	–	–	3.76×10^{8}
AOS[a](Huang et al. 2018c)	88.7	83.3	92.5	88.6	$\approx3.80\times10^{9}$
PSE[a,b] (Saquib et al. 2018)	90.3	84.0	–	–	$>3.80\times10^{9}$
HAN (Small)[a]	91.2	85.5	93.7	90.1	**3.68×10^{8}**
HAN (Medium)[a]	92.0	86.9	94.1	90.8	5.33×10^{8}
HAN (Large)[a]	**92.0**	**87.5**	**94.3**	**90.9**	7.01×10^{8}
SGGNN[c] (Shen et al. 2018b)	92.3	**82.8**	–	–	$>3.80\times10^{9}$
GCS[c] (Chen et al. 2018a)	**93.5**	81.6	–	–	$>3.80\times10^{9}$
HAN (Small)[c]	90.0	75.3	93.2	82.3	**3.68×10^{8}**
HAN (Medium)[c]	90.9	77.9	93.7	84.0	5.33×10^{8}
HAN (Large)[c]	91.1	78.1	94.1	84.7	7.01×10^{8}
RW[a,b,c] (Shen et al. 2018a)	92.7	82.5	–	–	$>3.80\times10^{9}$
HAN (Small)[a,c]	92.3	87.5	93.8	91.0	**3.68×10^{8}**
HAN (Medium)[a,c]	92.9	88.8	94.5	92.0	5.33×10^{8}
HAN (Large)[a,c]	**93.1**	**89.6**	**94.8**	**92.5**	7.01×10^{8}

Bold values indicate the best result for the respective groups
D201, DenseNet201; D169, DenseNet169; R50, ResNet50; SQ, Single-Query; MQ, Multi-Query
[a]Using the re-ranking method in (Zhong et al. 2017a)
[b]Using a newly proposed re-ranking method
[c]Using random erasing data augmentation (Zhong et al. 2017b)

Table 4 Performance evaluation on DukeMTMC

Metric (%)	R1	mAP	FLOPs
LSRO-R50 (Zheng et al. 2017)	67.7	47.1	$>3.80\times10^9$
SVDNet-R50 (Sun et al. 2017)	76.7	56.8	$>3.80\times10^9$
DPFL (Chen et al. 2017b)	79.2	60.6	$\approx1.2\times10^{10}$
AACN (Xu et al. 2018)	76.8	59.3	$>1.57\times10^9$
PT-R50 (Liu et al. 2018)	78.5	56.9	$>3.80\times10^9$
DaRe-R50 (Wang et al. 2018e)	75.2	57.4	$>3.80\times10^9$
BraidNet (Wang et al. 2018d)	76.4	59.5	$>2.26\times10^9$
MLFN (Chang et al. 2018b)	**81.0**	62.8	$\approx2.60\times10^9$
CamStyle (Zhong et al. 2018b)	75.3	53.5	$\approx3.80\times10^9$
AOS (Huang et al. 2018c)	79.2	62.1	$\approx3.80\times10^9$
PSE (Saquib et al. 2018)	79.8	62.0	$>3.80\times10^9$
KPM (Shen et al. 2018a)	80.3	63.2	$>3.80\times10^9$
PoseNorm (Qian et al. 2018)	73.6	53.2	$>3.80\times10^9$
HAP2S (Yu et al. 2018)	75.9	60.6	$\approx3.80\times10^9$
HAN (Small)	78.9	61.9	$\mathbf{3.68\times10^8}$
HAN (Medium)	80.0	63.4	5.33×10^8
HAN (Large)	80.6	**64.1**	7.01×10^8
DaRe-R50[a] (Wang et al. 2018e)	80.4	74.5	$>3.80\times10^9$
AOS[a] (Huang et al. 2018c)	84.1	78.2	$\approx3.80\times10^9$
PSE[a,b] (Saquib et al. 2018)	**85.2**	**79.8**	$>3.80\times10^9$
HAN (Small)[a]	83.2	77.9	$\mathbf{3.68\times10^8}$
HAN (Medium)[a]	84.0	78.8	5.33×10^8
HAN (Large)[a]	84.0	79.5	7.01×10^8
SGGNN[b] (Shen et al. 2018b)	81.1	68.2	$>3.80\times10^9$
GCS[b] (Chen et al. 2018a)	**84.9**	**69.5**	$>3.80\times10^9$
HAN (Small)[c]	79.4	64.0	$\mathbf{3.68\times10^8}$
HAN (Medium)[c]	80.5	64.7	5.33×10^8
HAN (Large)[c]	80.7	65.9	7.01×10^8
RW[a,b,c] (Shen et al. 2018a)	80.7	**82.5**	$>3.80\times10^9$
HAN (Small)[a,c]	83.9	79.6	$\mathbf{3.68\times10^8}$
HAN (Medium)[a,c]	84.2	80.2	5.33×10^8
HAN (Large)[a,c]	**84.6**	81.3	7.01×10^8

Bold values indicate the best result for the respective groups
R50, ResNet50; D201, DenseNet201
[a]Using the re-ranking method in Zhong et al. (2017a)
[b]Using a newly proposed re-ranking method
[c]Using random erasing data augmentation (Zhong et al. 2017b)

tive observations as on Market-1501. In the standard training setting, our HAN models are the most efficient solutions whilst achieving top performances. With re-ranking, PSE (Saquib et al. 2018) slightly outperforms HAN models with up to 10× more expensive in the computational cost. Similar contrasts between HAN and the competitors are observed when using random erasing based data augmentation alone or along with re-ranking.

Evaluation on CUHK03 We evaluated the HAN models on both manually labelled and auto-detected (more severe misalignment) person bounding boxes on the CUHK03 benchmark. We adopted the 767/700 identity split rather than 1367/100 since the former defines a more realistic and challenging re-id task. In the standard setting, the training set is rather small, with only 7,365 training images vs 12,936 and 16,522 on Market-1501 and DukeMTMC. This generally imposes an extreme challenge to the training of deep models, particularly in case of using *no* large auxiliary data (e.g. ImageNet) for model pre-training like our HAN models.

Table 5 shows that the HAN models still achieve competitive re-id matching accuracy, although outperformed by two recent computationally expensive approaches MLFN (Chang et al. 2018b) and DaRe-R50 (Wang et al. 2018e) which benefit substantially from ImageNet in model pre-training. Among all competitors, our models are most efficient therefore more scalable to large scale re-id deployments in practical use. If more computational resource is available, re-ranking can be applied for all methods to further improve the re-id performance.

Evaluation on MSMT17 We evaluated the HAN models on the new large scale MSMT17 benchmark tested by only a few methods. Having more training data typically benefits larger neural networks due to a reduced model fitting risk, and lightweight networks may be therefore less competitive in accuracy due to relatively inferior representative capabilities. This facilitates a more extensive test on both model learning capacity and generalisation of our lightweight HAN against existing more elaborative and "heavier" deep re-id models, given the larger training and testing sets in terms of the number of images, identities, and cameras. This test is not only complementary to the other re-id benchmark tests, but also a good evaluation on small networks like HAN models in order to evaluate the models' learning capacity when larger training and test data are given whilst having less parameters.

Table 6 shows that the heavy model GLAD (Wei et al. 2017) achieves the best results in the standard setting, but only *slightly* outperforming the HAN models whilst at over 10× more computational costs. Besides, HAN (Large) matches the accuracy performance of ResNet50 with merely 18% inference cost. These suggest that the cost-effectiveness advantages of our HAN models remain on larger scale re-id learning and deployments, and importantly the absolute performances of HAN models are still competitive. The advantages of HAN are similar in case of using re-ranking. This test broadly examines the ability of neural network models in compromising between model complexity (learning capacity) and computational efficiency (processing speed) often required in large scale re-id deployments.

4.2 Comparing State-of-the-Art Efficient Networks

We compared the proposed HAN models with three state-of-the-art compact neural network models: MobileNet (Howard

Table 5 Performance evaluation on CUHK03

Metric (%)	Labelled		Detected		FLOPs
	R1	mAP	R1	mAP	
IDE-C (Zhong et al. 2017a)	15.6	14.9	15.1	14.2	7.25×10^8
XQDA-C (Zhong et al. 2017a)	21.9	20.0	21.1	19.0	7.25×10^8
IDE-R50 (Zhong et al. 2017a)	22.2	21.0	21.3	19.7	3.80×10^9
XQDA-R50(Zhong et al. 2017a)	32.0	29.6	31.1	28.2	3.80×10^9
SVDNet-C (Sun et al. 2017)	–	–	27.7	24.9	$>7.25\times10^8$
SVDNet-R50 (Sun et al. 2017)	–	–	41.5	37.3	$>3.80\times10^9$
DPFL (Chen et al. 2017b)	43.0	40.5	40.7	37.0	$\approx1.2\times10^{10}$
PT-R50 (Liu et al. 2018)	45.1	42.0	41.6	38.7	$>3.80\times10^9$
AOS (Huang et al. 2018c)	–	–	47.1	43.3	$\approx3.80\times10^9$
DaRe-R50 (Wang et al. 2018e)	**58.1**	**53.7**	**55.1**	**51.3**	$>3.80\times10^9$
MLFN (Chang et al. 2018b)	54.7	49.2	52.8	47.8	$>2.26\times10^9$
HAN (Small)	42.7	42.4	40.9	40.0	$\mathbf{3.68\times10^8}$
HAN (Medium)	42.0	42.3	42.8	42.0	5.33×10^8
HAN (Large)	46.5	46.1	47.5	45.5	7.01×10^8
AOS[a] (Huang et al. 2018c)	–	–	54.6	56.1	$\approx3.80\times10^9$
MGCAM[a] (Song et al. 2018)	50.1	50.2	46.7	46.9	3.76×10^8
DaRe-R50[a] (Wang et al. 2018e)	**66.0**	**66.7**	**62.8**	**63.6**	$>3.80\times10^9$
HAN (Small)[a]	49.6	54.3	46.9	51.6	$\mathbf{3.68\times10^8}$
HAN (Medium)[a]	50.1	55.2	49.7	54.0	5.33×10^8
HAN (Large)[a]	53.6	58.7	54.9	57.9	7.01×10^8

Bold values indicate the best result for the respective groups
C, CaffeNet; R50, ResNet50; D201, DenseNet201
[a]Using the re-ranking method in Zhong et al. (2017a)

Table 6 Performance evaluation on MSMT17

Metric (%)	R1	mAP	FLOPs
GoogLeNet (Szegedy et al. 2015)	47.6	23.0	1.57×10^9
PDC (Su et al. 2017)	58.0	29.7	$\gg9.82\times10^9$
GLAD (Wei et al. 2017)	**61.4**	**34.0**	$\gg7.99\times10^9$
ResNet50 (He et al. 2016)	59.7	33.7	3.80×10^9
HAN (Small)	56.3	29.2	$\mathbf{3.68\times10^8}$
HAN (Medium)	57.1	30.3	5.33×10^8
HAN (Large)	60.1	32.6	7.01×10^8
ResNet50[a] (He et al. 2016)	64.6	**47.6**	3.80×10^9
HAN (Small)[a]	61.8	41.8	$\mathbf{3.68\times10^8}$
HAN (Medium)[a]	63.8	42.9	5.33×10^8
HAN (Large)[a]	**66.2**	46.2	7.01×10^8

Bold values indicate the best result for the respective groups
[a]Using the re-ranking method (Zhong et al. 2017a)

et al. 2017), ShuffleNet (Zhang et al. 2018a), and CondenseNet (Huang et al. 2018b). These competitors are general-purpose lightweight neural networks therefore directly applicable for person re-id although not evaluated in the original works.

Table 7 shows that our HAN models achieve the best performances at competitive inference computational costs. In particular, HAN (Small) significantly outperforms MobileNet whilst enjoying more efficient inference. While the CondenseNet and ShuffleNet are more efficient than HAN, their re-id matching performances are the worst. HAN (Large) further improves the model generalisation capability by extra 3.33×10^8 (7.01-3.68) FLOPs per image. These indicate the cost-effectiveness advantages of the proposed HAN models in person re-id over state-of-the-art efficient network designs.

4.3 Further Analysis and Discussions

To provide more detailed examinations and insights, we conducted a sequence of component analysis using HAN (Large) on Market-1501 and DukeMTMC in the Single-Query setting.

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 8 shows that: (1) Either feature representation *alone* is already very discriminative for person re-id. (2) A further performance gain is obtained by joining the two representations, yielding 2.7% (91.6–88.9) in Rank-1 boost and 4.5% (76.7–

Table 7 Comparisons with efficient neural networks

Dataset	Market-1501 (SQ)		Market-1501 (MQ)		DukeMTMC		CUHK03 (L)		CUHK03 (D)		MSMT17		FLOPs
Metric (%)	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	
MobileNet	84.6	64.3	89.3	72.8	72.1	50.9	36.1	35.5	35.0	34.3	53.8	26.4	5.69×10^8
ShuffleNet	80.6	58.4	86.4	67.0	70.0	48.2	35.6	35.1	33.2	33.3	42.8	18.9	$\mathbf{1.40\times10^8}$
CondenseNet	83.5	62.8	88.2	70.8	72.6	51.3	33.4	33.5	32.1	31.8	54.2	26.5	2.74×10^8
HAN (Small)	89.4	73.2	93.2	80.8	78.9	61.9	42.7	42.4	40.9	40.0	56.3	29.2	3.68×10^8
HAN (Medium)	90.7	74.5	93.9	81.9	80.0	63.4	42.0	42.3	42.8	42.0	57.1	30.3	5.33×10^8
HAN (Large)	**91.6**	**76.7**	**94.2**	**83.4**	**80.6**	**64.1**	**46.5**	**46.1**	**47.5**	**45.5**	**60.1**	**32.6**	7.01×10^8

Bold values indicate the best result for the respective groups
SQ single-query; *MQ* multi-query; *L* labelled; *D* detected

Table 8 Evaluating the global and local-level features

Dataset	Market-1501		DukeMTMC	
Metric (%)	R1	mAP	R1	mAP
Global	88.9	72.2	77.9	59.8
Local	89.3	73.0	77.8	59.8
Global + Local	**91.6**	**76.7**	**80.6**	**64.1**

Bold values indicate the best result for the respective groups

Table 9 Comparative evaluation of individual types of attention in our HA model

Dataset	Market-1501		DukeMTMC	
Metric (%)	R1	mAP	R1	mAP
No Attention	85.2	64.2	72.8	50.7
SSA	86.3	65.3	74.6	51.2
SCA	87.2	67.7	74.8	53.0
SSA+SCA	88.9	69.9	77.2	56.1
HRA	87.9	70.3	77.1	60.0
SSA+HRA	89.5	72.1	77.5	60.1
SCA+HRA	90.1	74.9	78.9	62.9
HAN (All)	**91.6**	**76.7**	**80.6**	**64.1**

Bold values indicate the best result for the respective groups
SSA soft spatial attention; *SCA* soft channel attention; *HRA* hard regional attention

72.2) in mAP increase on Market-1501. Similar trends are observed on DukeMTMC (Table 4). These validate the complementary effect of jointly learning local and global features in the harmonious attention context by our HAN model.

Different Types of Attention We tested the effect of individual attention components in the HAN model: Soft Spatial Attention (SSA), Soft Channel Attention (SCA), and Hard Regional Attention (HRA). Table 9 shows that: (1) Each type of attention *in isolation* brings person re-id performance gain; (2) SSA+SCA gives a further accuracy boost, suggesting the complementary information between the two soft attention discovered by our model; (3) When combining the hard and soft attention (SSA, SCA, or both), the model performance can be further improved. This indicates that our method is effective in identifying and exploiting the complementary benefits between coarse-grained hard attention and fine-grained soft attention.

Table 10 Evaluating the cross-attention interaction learning (CAIL) component

Dataset	Market-1501		DukeMTMC	
Metric (%)	R1	mAP	R1	mAP
Global (**w/o** CAIL)	84.1	64.0	73.2	53.3
Global (**w/** CAIL)	**88.9**	**72.2**	**77.9**	**59.8**
Local (**w/o** CAIL)	17.1	7.6	24.3	14.3
Local (**w/** CAIL)	**89.3**	**73.0**	**77.8**	**59.8**
Global + Local (**w/o** CAIL)	72.1	50.9	56.7	40.1
Global + Local (**w/** CAIL)	**91.6**	**76.7**	**80.6**	**64.1**

Bold values indicate the best result for the respective groups

Table 11 Evaluating different types of objective functions

Dataset	Market-1501		DukeMTMC	
Metric (%)	R1	mAP	R1	mAP
Cross-Entropy	**91.6**	**76.7**	**80.6**	**64.1**
Triplet	63.7	41.7	57.8	37.6
Cross-Entropy + Triplet	91.5	76.3	79.9	61.5

Bold values indicate the best result for the respective groups

Table 12 Evaluating different types of building blocks

Dataset	Market-1501		DukeMTMC		FLOPs
Metric (%)	R1	mAP	R1	mAP	
Depthwise	**91.6**	**76.7**	**80.6**	**64.1**	$\mathbf{7.01\times10^8}$
Inception	91.2	75.7	80.5	63.8	1.09×10^9
Residual	91.0	75.0	78.5	62.1	1.34×10^9

Bold values indicate the best result for the respective groups

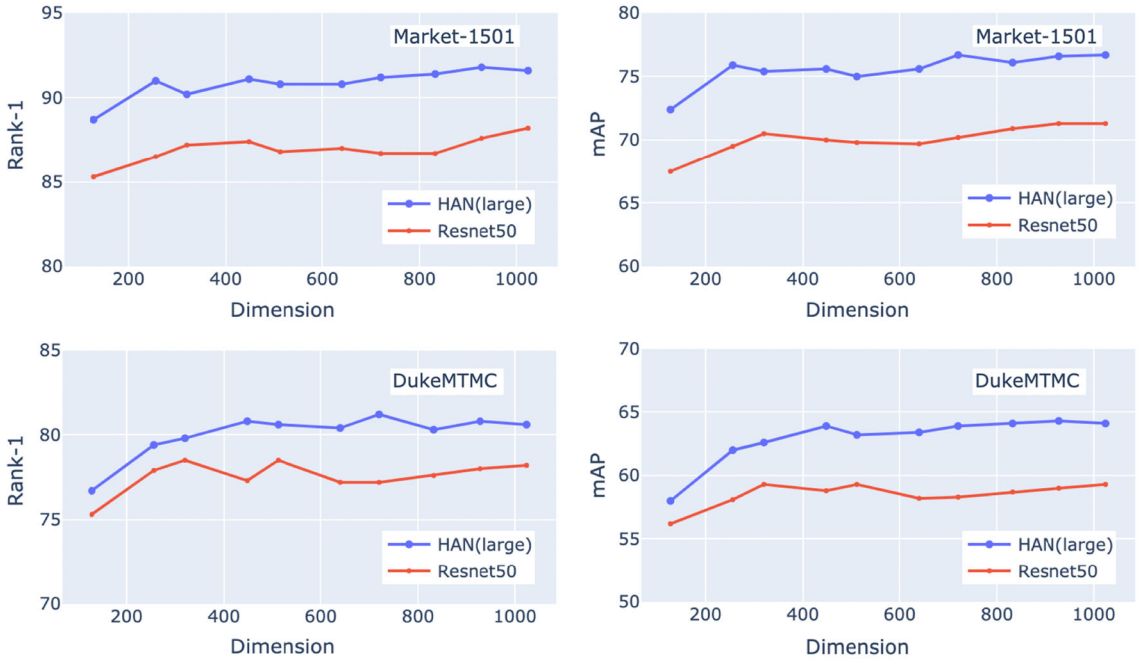


Fig. 6 Evaluating a set of feature dimensions {128, 256, 320, 448, 512, 640, 720, 832, 928, 1024} w.r.t. the re-id performance on Market-1501 and DukeMTMC.

Cross-Attention Interaction Learning We evaluated the benefit of cross-attention interaction learning (CAIL) between the global and local branches. Table 10 shows three observations: (1) CAIL benefits clearly the learning of global feature, local feature and their combination. (2) The local branch obtains substantially more performance gain, which is expected since its design is of super-lightweight with insufficient learning capacity on its own; With CAIL, it also simultaneously borrows the representation learning capacity from the global branch. This overall design aims for minimising the model parameter redundancy. (3) Without CAIL, the combined feature is even inferior to the global feature alone, due to the negative impact from the very weak and incompatible local feature. This suggests that CAIL also plays a significant bridging role between the two branches in our model formulation. Overall, this experiment validates our design consideration that it is necessary to jointly learn the *attended* feature representations across soft and hard attention subject to the same label constraint.

Objective Loss Function We evaluated the choice of objective loss function in HAN. In particular, we additionally tested the common triplet ranking loss. To more effectively and efficiently impose useful triplet constraints, we exploited the online triplet selection strategy in a hard sample mining principle (Schroff et al. 2015; Hermans et al. 2017).

Specifically, for each mini-batch training we identify on-the-fly and use only those hard triplets yielding positive loss values while throwing away the remaining ones that fulfil the triplet constraints with zero loss values.

We have some interesting observations in Table 11: (1) Using the triplet loss *alone* in the tiny HAN model gives significantly inferior re-id performance. The plausible reason is that such pairwise comparison based objective has an unnoticed need for large (less efficient though more expressive) neural network models. (2) Combining the triplet and cross-entropy loss functions can only achieve similar results as using the latter alone. This suggests that the triplet ranking loss is hardly able to provide complementary supervision information, due to the strong capability of identifying re-id attention by the HAN. This favourably eliminates the need of detecting subtle discrepancy between visually similar different persons through exhuastive (a quadratic number of identity pairs) pairwise contrasts in the triplet loss. (3) With a strong attention model like HAN, it is likely that the simpler cross-entropy loss suffices to induce a discriminative person re-id model.

Network Building Blocks We examined the generalisation capability of HAN in the incorporation of three state-of-the-art CNN building block designs: (1) Inception-A/B unit (Xiao et al. 2016; Szegedy et al. 2017), (2) Residual bottleneck unit (He et al. 2016), and (3) Depthwise separable conv unit used in our main models (Sifre and Mallat 2014). Table 12 shows that HAN is able to effectively accommodate varying *types* of conv building blocks. Among the three designs, it is interestingly found that the depthwise one is the most cost-effective unit, yielding both the best accuracy and efficiency. This provides an unusual example that the lightweight depthwise conv units are *not necessarily* inferior in recognition per-

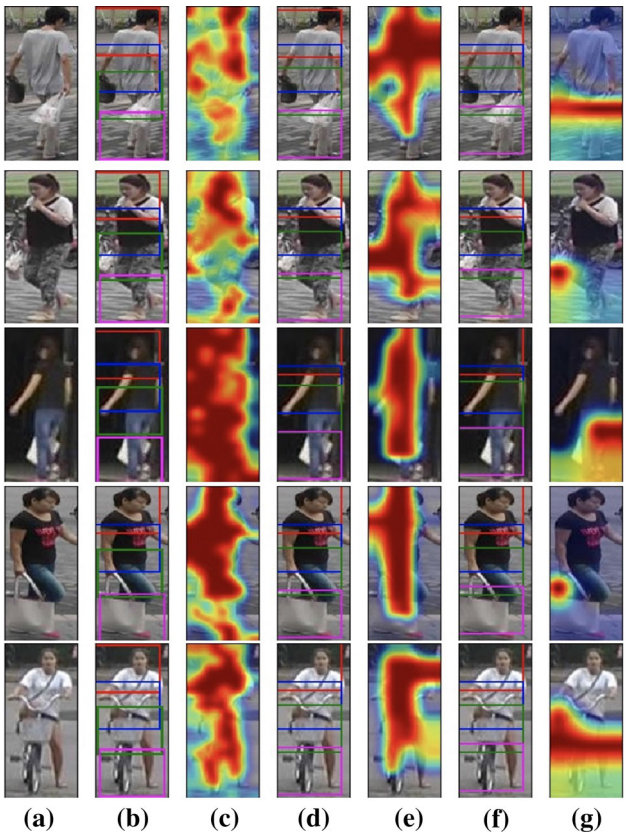


Fig. 7 Visualisation of the harmonious re-id attention. From left to right, **a** the original image, **b** the 1st-level of hard attention, **c** the 1s-level of soft attention, **d** the 2nd-level of hard attention, **e** the 2nd-level of soft attention, **f** the 3rd-level of hard attention, **g** the 3rd-level of soft attention

formance, contrary to the existing finding in coarse-grained object detection and recognition tasks (Howard et al. 2017).

Feature Dimension In addition to feature extraction cost, feature dimension is another scalability factor in the large scale re-id search process, regarding to data transportation, storage size, and matching speed. We evaluated this factor by comparing our method with the golden standard model ResNet50, using a set of feature dimensions ranging from 128 to 1024. As shown in Fig. 6, the HAN (large) model not only delivers consistent performance advantage over all the feature dimensions, but also enables the use of lower-dimensional feature vectors whilst simultaneously yielding a similar or even better re-id performance. This verifies the superior scalability of our method in terms of memory usage and matching efficiency, therefore scalable to small feature vectors for better data transportation and potential deployment in the cloud or at the edge. The margin gets smaller on the more challenging DukeMTMC dataset when using very low feature dimensions (e.g. 128) due to too limited feature representation capacity to fully exploit the learning capability of HAN.

Visualisation of Harmonious Attention We visualised both learned soft attention and hard attention discovered by the HAN re-id model at three different network levels. Figure 7 shows that: (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 soft attention (c) to local salient parts (e.g. object associations) by the 3rd-level spatial soft attention (g). This shows compellingly the effectiveness and collaboration of soft and hard attention joint learning.

5 Conclusion

In this work, we present a cost-effective Harmonious Attention Network (HAN) framework for joint learning of person re-identification attention selection and feature representations. In contrast to existing re-id deep learning methods that typically ignore the model efficiency issue, the HAN model is designed to scale up large deployments whilst simultaneously achieving top re-id matching performances. This is realised by designing a Harmonious Attention mechanism enabling to establish *lightweight* CNN architectures with sufficient discrimination learning capability. Moreover, we introduce a cross-attention interaction learning strategy to enhance the joint optimisation of attention selection and re-id features. Extensive evaluations have been conducted on four large re-id benchmarks with varying training and test scales to validate the cost-effectiveness advantages of our HAN model over state-of-the-art re-id methods and scalable neural network designs. We also provide a series of detailed model component analysis and insightful discussions on the HAN model cost-effectiveness superiority.

Acknowledgements This work was partially supported by the China Scholarship Council, Vision Semantics Ltd, Royal Society Newton Advanced Fellowship Programme (NA150459), and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M., et al. (2017). Tensorflow: A system for large-scale machine learning.

Ahmed, E., Jones, M., & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (Vol. 19, pp. 41–48). MIT Press. http://papers.nips.cc/paper/3143-multi-task-feature-learning.pdf.

Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep? In: Advances in neural information processing systems (pp. 2654–2662).

Bolukbasi, T., Wang, J., Dekel, O., & Saligrama, V. (2017) Adaptive neural networks for fast test-time prediction. In *International conference on machine learning*.

Chang, X., Hospedales, T. M., & Xiang, T. (2018a). Multi-level factorisation net for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Chang, X., Hospedales, T. M., & Xiang, T. (2018b). Multi-level factorisation net for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Chen, W., Chen, X., Zhang, J., & Huang, K. (2017a). A multi-task deep network for person re-identification. In *AAAI conference on artificial intelligence*.

Chen, D., Xu, D., Li, H., Sebe, N., & Wang, X. (2018a). Group consistent similarity learning via deep CRF for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Chen, Y., Zhu, X., & Gong, S. (2017b). Person re-identification by deep learning multi-scale representations. In *Workshop of IEEE international conference on computer vision*.

Chen, Y., Zhu, X., & Gong, S. (2018b). Deep association learning for unsupervised video person re-identification. In *British machine vision conference*.

Chen, Y. C., Zhu, X., Zheng, W. S., & Lai, J. H. (2017c). Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 392–408.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *IEEE conference on computer vision and pattern recognition*.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. Cambridge: MIT Press.

Courbariaux, M., Bengio, Y., & David, J.-P. (2015). BinaryConnect: Training deep neural networks with binary weights during propagations. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28, pp. 3123–3131). Curran Associates, Inc. http://papers.nips.cc/paper/5647-binaryconnect-training-deep-neural-networks-with-binary-weights-during-propagations.pdf.

Denil, M., Shakibi, B., Dinh, L., & De Freitas, N., et al. (2013). Predicting parameters in deep learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26, pp. 2148–2156). Curran Associates, Inc. http://papers.nips.cc/paper/5025-predicting-parameters-in-deep-learning.pdf.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(04), 449–467.

Evgeniou, T., & Pontil, M. (2004). Regularized multi-task learning. In *ACM SIGKDD international conference on knowledge discovery and data mining*.

Faraone, J., Fraser, N., Blott, M., & Leong, P. H. (2018). SYQ: Learning symmetric quantization for efficient deep neural networks. In *IEEE conference on computer vision and pattern recognition*.

Figurnov, M., Collins, M. D., Zhu, Y., Zhang, L., Huang, J., Vetrov, D., & Salakhutdinov, R. (2017). Spatially adaptive computation time for residual networks. In *IEEE conference on computer vision and pattern recognition*.

Hassibi, B., Stork, D. G., & Wolff, G. J. (1993). Optimal brain surgeon and general network pruning. In *IEEE international conference on neural networks* (pp. 293–299).

He, Y., Zhang, X., & Sun, J. (2017). Channel pruning for accelerating very deep neural networks. In *IEEE international conference on computer vision* (Vol. 2, pp. 6).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*.

Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. arXiv:1703.07737.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In *Advances in neural information processing systems, deep learning workshop*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, G., Chen, D., Li, T., Wu. F., van der Maaten, L., & Weinberger, K. Q. (2018a). Multi-scale dense convolutional networks for efficient prediction. In *International conference on learning representations*.

Huang, H., Li, D., Zhang, Z., Chen, X., & Huang, K. (2018c). Adversarially occluded samples for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Huang, G., Liu, S., van der Maaten, L., Weinberger, K. Q. (2018b). Condensenet: An efficient densenet using learned group convolutions. In *IEEE conference on computer vision and pattern recognition*.

Huang, G., Liu, Z., Weinberger, K. Q., & van der Maaten, L. (2017). Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition* (Vol. 1, p. 3).

Huang, Z., Wang, N. (2018). Data-driven sparse structure selection for deep neural networks. In *European conference on computer vision*.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 4107–4115). Curran Associates, Inc. http://papers.nips.cc/paper/6573-binarized-neural-networks.pdf.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size. arXiv preprint arXiv:1602.07360.

Ioffe, S., & Szegedy, C. (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Jaderberg, M., Simonyan, K., & Zisserman, A., et al. (2015) Spatial transformer networks. In: Advances in neural information processing systems (pp. 2017–2025).

Kalayeh, M. M., Basaran, E., Gokmen, M., Kamasak, M. E., Shah, M. (2018). Human semantic parsing for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Kodirov, E., Xiang, T., & Gong, S. (2015). Dictionary learning with iterative Laplacian regularisation for unsupervised person re-identification. In *British machine vision conference*.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 1097–1105). Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Lan, X., Wang, H., Gong, S., & Zhu, X. (2017). Deep reinforcement learning attention selection for person re-identification. In *British machine vision conference*.

Lan, X., Zhu, X., & Gong, S. (2018a). Knowledge distillation by on-the-fly native ensemble. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31, pp. 7517–7527). Curran Associates, Inc. http://papers.nips.cc/paper/7980-knowledge-distillation-by-on-the-fly-native-ensemble.pdf.

Lan, X., Zhu, X., Gong, S. (2018b). Self-referenced deep learning. In *Asian conference on computer vision*.

LeCun, Y., Denker, J. S., Solla, S. A. (1990). Optimal brain damage. In *Advances in neural information processing systems* (pp. 598–605).

Li, D., Chen, X., Zhang, Z., & Huang, K. (2017a). Learning deep context-aware features over body and latent parts for person re-identification. In *IEEE conference on computer vision and pattern recognition* (pp. 384–393).

Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. (2016). Pruning filters for efficient convnets. arXiv:1608.08710.

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Li, W., Zhu, X., & Gong, S. (2017b). Person re-identification by deep joint learning of multi-loss classification. In *International joint conference of artificial intelligence*.

Li, M., Zhu, X., & Gong, S. (2018a). Unsupervised person re-identification by deep learning tracklet association. In *European conference on computer vision*.

Li, W., Zhu, X., & Gong, S. (2018b). Harmonious attention network for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015) Person re-identification by local maximal occurrence representation and metric learning. In *IEEE conference on computer vision and pattern recognition*.

Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C. (2017c). Learning efficient convolutional networks through network slimming. In *IEEE international conference on computer vision* (pp. 2755–2763).

Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng. S., & Hu, J. (2018). Pose transferrable person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., & Wang, X. (2017b). Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE international conference on computer vision*.

Liu, H., Feng, J., Qi, M., Jiang, J., & Yan, S. (2017a). End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, *26*, 3492–3506.

Luo, J. H., Wu, J., Lin, W. (2017) Thinet: A filter level pruning method for deep neural network compression. In *The IEEE international conference on computer vision (ICCV)*.

Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y. (2016). Hierarchical Qaussian descriptor for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., & Hajishirzi, H. (2018). Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *European conference on computer vision*.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., & Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Qian, X., Fu, Y., Jiang, Y. G., Xiang, T., & Xue, X. (2017). Multi-scale deep learning architectures for person re-identification. In *IEEE international conference on computer vision*.

Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu. Y., Jiang. Y. G., & Xue, X. (2018). Pose-normalized image generation for person re-identification. In *European conference on computer vision*.

Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision* (pp. 525–542).

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In: *Workshop of European conference on computer vision*.

Saquib, S. M., Schumann, A., Eberle, A., & Stiefelhagen, R. (2018). A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *IEEE conference on computer vision and pattern recognition*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Shen, Y., Li, H., Xiao, T., Yi, S., Chen, D., & Wang, X. (2018a). Deep group-shuffling random walk for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Shen, Y., Li, H., Yi, S., Chen, D., & Wang, X. (2018b). Person re-identification with deep similarity-guided graph neural network. In *European conference on computer vision*.

Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., & Wang, J. (2015). Person re-identification with correspondence structure learning. In *IEEE international conference on computer vision*.

Sifre, L., & Mallat, P. (2014). Rigid-motion scattering for image classification. Ph.D. thesis, Citeseer.

Song, C., Huang, Y., Ouyang, W., & Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. In *ICCV*.

Subramaniam, A., Chatterjee, M., & Mittal, A. (2016). Deep neural networks with inexact matching for person re-identification. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 2667–2675). Curran Associates, Inc. http://papers.nips.cc/paper/6367-deep-neural-networks-with-inexact-matching-for-person-re-identification.pdf.

Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K. (2018). Part-aligned bilinear representations for person re-identification. In *European conference on computer vision*.

Sun, Y., Zheng, L., Deng, W., & Wang, S. (2017). Svdnet for pedestrian retrieval. In *IEEE international conference on computer vision*.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI conference on artificial intelligence*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In: IEEE conference on computer vision and pattern recognition (pp. 1–9).

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766.

Varior, R. R., Haloi, M., Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*.

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

Wang, Y., Chen, Z., Wu, F., & Wang, G. (2018d). Person re-identification with cascaded pairwise convolutions. In *IEEE conference on computer vision and pattern recognition*.

Wang, H., Gong, S., & Xiang, T. (2014). Unsupervised learning of generative topic saliency for person re-identification. In *British machine vision conference*.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017). Residual attention network for image classification. In *IEEE conference on computer vision and pattern recognition*.

Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Huang, G., Hariharan, B., & Weinberger, K. Q. (2018e). Resource aware person re-identification across multiple resolutions. In *IEEE conference on computer vision and pattern recognition* (pp. 8042–8051).

Wang, C., Zhang, Q., Huang, C., Liu, W., & Wang, X. (2018a). Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *European conference on computer vision*.

Wang, J., Zhu, X., Gong, S., & Li, W. (2018c). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Wang, H., Zhu, X., Gong, S., & Xiang, T. (2018b). Person re-identification in identity regression space. *International Journal of Computer Vision*, *126*, 1288–1310.

Wang, H., Zhu, X., Xiang, T., & Gong, S. (2016b). Towards unsupervised open-set person re-identification. In *IEEE international conference on image processing*.

Wang, F., Zuo, W., Lin, L., Zhang, D., & Zhang, L. (2016a). Joint learning of single-image and cross-image representations for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Wei, L., Zhang, S., Yao, H., Gao, W., & Tian, Q. (2017). Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM international conference on multimedia* (pp. 420–428).

Xiao, T., Li, H., Ouyang, W., & Wang, X. (2016). Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Xu, J., Zhao, R., Zhu, F., Wang, H., & Ouyang, W. (2018). Attention-aware compositional network for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y., & Bai, X. (2018). Hard-aware point-to-set deep metric for person re-identification. In *European conference on computer vision*.

Zhang, T., Qi, G. J., Xiao, B., & Wang, J. (2017). Interleaved group convolutions. In *IEEE international conference on computer vision* (pp. 4373–4382).

Zhang, L., Xiang, T., & Gong, S. (2016). Learning a discriminative null space for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Zhang, Y., Xiang, T., Hospedales, T. M., Lu, H. (2018b). Deep mutual learning. In *IEEE conference on computer vision and pattern recognition*.

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018a). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *IEEE conference on computer vision and pattern recognition*.

Zhang, R., Lin, L., Zhang, R., Zuo, W., & Zhang, L. (2015). Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Transactions on Image Processing*, *24*, 4766–4779.

Zhao, L., Li, X., Wang, J., & Zhuang, Y. (2017). Deeply-learned part-aligned representations for person re-identification. In *IEEE international conference on computer vision*.

Zhao, R., Ouyang, W., & Wang, X. (2013). Unsupervised salience learning for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Zheng, W. S., Li, X., Xiang, T., Liao, S., Lai, J., & Gong, S. (2015b). Partial person re-identification. In *IEEE international conference on computer vision* (pp. 4678–4686).

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015a). Scalable person re-identification: A benchmark. In *IEEE international conference on computer vision*.

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *IEEE international conference on computer vision*.

Zheng, W. S., Gong, S., & Xiang, T. (2013). Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 653–668.

Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017a). Re-ranking person re-identification with k-reciprocal encoding. In *IEEE conference on computer vision and pattern recognition*.

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017b). Random erasing data augmentation. arXiv preprint arXiv:1708.04896.

Zhong, Z., Zheng, L., Li, S., & Yang, Y. (2018a). Generalizing a person retrieval model hetero-and homogeneously. In *European conference on computer vision* (pp. 172–188).

Zhong, Z., Zheng, L., Zheng, Z., Li, S., & Yang, Y. (2018b). Camera style adaptation for person re-identification. In *IEEE conference on computer vision and pattern recognition*.

Zhu, F., Kong, X., Zheng, L., Fu, H., & Tian, Q. (2017). Part-based deep hashing for large-scale person re-identification. *IEEE Transactions on Image Processing*, *26*(10), 4806–4817.

Zhu, X., Wu, B., Huang, D., & Zheng, W. S. (2018). Fast open-world person re-identification. *IEEE Transactions on Image Processing*, *27*(5), 2286–2300.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8697–8710).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.