The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Deep Low-Resolution Person Re-Identification

Jiening Jiao,[1,2] **Wei-Shi Zheng,**[3,4*] **Ancong Wu,**[1] **Xiatian Zhu,**[5] **Shaogang Gong**[5]
[1] School of Electronics and Information Technology, Sun Yat-sen University, China
[2] Collaborative Innovation Center of High Performance Computing, NUDT, China
[3] School of Data and Computer Science, Sun Yat-sen University, China
[4] Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China
[5] School of Engineering and Computer Science, Queen Mary University of London, UK
jiaojn@mail2.sysu.edu.cn, wszheng@ieee.org, wuancong@mail2.sysu.edu.cn,
xiatian.zhu@qmul.ac.uk, s.gong@qmul.ac.uk

Abstract

Person images captured by public surveillance cameras often have low resolutions (LR) in addition to uncontrolled pose variations, background clutters and occlusions. This gives rise to the *resolution mismatch* problem when matched against the high resolution (HR) gallery images (typically available in enrolment), which adversely affects the performance of person re-identification (re-id) that aims to associate images of the same person captured at different locations and different time. Most existing re-id methods either ignore this problem or simply upscale LR images. In this work, we address this problem by developing a novel approach called ***S**uper-resolution and **I**dentity joi**N**t learnin**G*** (**SING**) to simultaneously optimise image super-resolution and person re-id matching. This approach is instantiated by designing a hybrid deep Convolutional Neural Network for improving cross-resolution re-id performance. We further introduce an adaptive fusion algorithm for accommodating multi-resolution LR images. Extensive evaluations show the advantages of our method over related state-of-the-art re-id and super-resolution methods on cross-resolution re-id benchmarks.

Introduction

Person re-identification (re-id) is a task of matching identity classes in person bounding box images extracted from non-overlapping camera views in open surveillance spaces (Gong et al. 2014). Existing re-id methods typically focus on addressing the variations in illumination, occlusion, and background clutter by designing feature representation (Liao et al. 2015; Matsukawa et al. 2016) or learning matching distance metrics (Zheng, Gong, and Xiang 2013; Wang et al. 2014; He, Chen, and Lai 2016; Zhang, Xiang, and Gong 2016) or their combinations (Li et al. 2014; Ahmed, Jones, and Marks 2015; Xiao et al. 2016; Li, Zhu, and Gong 2017) under the assumption that all person images have similar and sufficiently high resolutions. However, surveillance person images often have varying resolutions due to variations in the person-camera distance and camera deployment settings (Fig. 1). This gives rise to the *resolution mismatch* problem. Specifically, human operators often enrol high resolution (HR) images[1] of target people into the gallery set. As such, it is challenging to reliably match low-resolution (LR) probe images against the HR gallery images across both camera views and resolutions. This requires to address the *information amount discrepancy* challenge in cross-resolution matching since LR images contain much less information with discriminative appearance details largely lost in the image acquisition process. We call this setting as *Low-Resolution person re-identification*.

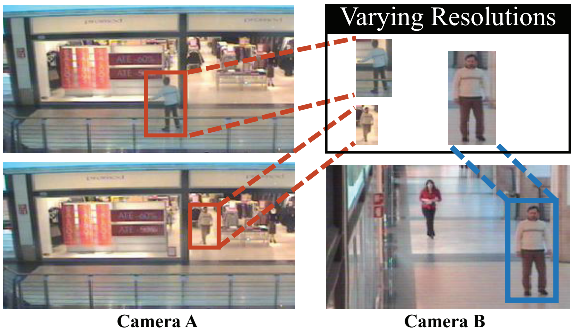

Figure 1: Illustration of person images with varying resolutions in the open-space person re-identification task. Three images of a person were captured by two camera views at different locations of a shopping centre. The image captured by camera B has a higher resolution than the other two from camera A. Person re-id across different resolutions and disjoint camera views is challenging.

While most existing methods ignore the resolution mismatch problem and normalise all images to a single size, a couple of works have recently been proposed to address this LR re-id problem (Jing et al. 2015; Wang et al. 2016; Li et al. 2015). However, these methods share a few com-

[*] Corresponding author

[1] Note that in visual surveillance, the definition quality of so-called high resolution images is poorer than social media photos taken by professional photographers. In this context, we define low and high resolutions in a relative sense for surveillance quality image data. By default, we refer "resolution" to *the underlying resolution* (Wong et al. 2010) rather than *the image spatial size (scale)*. A given image can be arbitrarily resized, but with little change in its underlying resolution (Fig. 2). Hence, the image spatial size is not an accurate indicator of the underlying resolution.

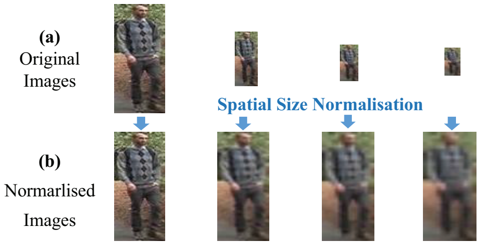

Figure 2: (a) Example images of varying *underlying resolutions*, (b) normalised to the same *spatial size* without changing the underlying resolution.

mon weaknesses: (1) Instead of recovering the missing discriminative appearance information, they perform cross-resolution representation transformation in a pre-defined feature space. This does not solve the *information amount discrepancy* challenge. (2) They rely on hand-crafted visual features without exploring deep learning for mining the complementary advantages of feature learning and matching metric joint optimisation. Intuitively, image super-resolution (SR) should offer an effective solution to mitigate the resolution mismatch problem due to its capability of synthesising high frequency details (Dong et al. 2016). However, a direct combination of SR and re-id may suffer from suboptimal compatibility. Generic-purpose SR methods are designed to improve image visual fidelity rather than the re-id matching performance, with visual artefacts generated in the SR reconstruction process negative to re-id matching.

Contributions. In this work, we solve the LR re-id problem by exploring image SR and person re-id techniques in a novel unified formulation. We call the new formulation ***S**uper-resolution and **I**dentity joi**N**t learnin**G** (**SING**).* Our SING is designed to improve the integration compatibility between SR and re-id by achieving identity sensitive high frequency appearance enhancement, therefore addressing the information amount discrepancy problem in cross-resolution re-id matching. To realise SING, we propose a joint loss function on optimsing a hybrid Convolutional Neural Network (CNN) architecture capable of bridging SR and re-id model learning. In addition, given LR person images with different resolutions in practice, we further present a multi-resolution adaptive fusion mechanism by aggregating a set of anchor SING CNN models (each optimised for a reference resolution) in a probe specific manner. We extensively performed comparative evaluations to show the superiority of our SING approach over related state-of-the-art re-id and image SR methods on four person re-id benchmarks CAVIAR (Cheng et al. 2011), VIPeR (Gray and Tao 2008), CUHK03 (Li et al. 2014), and SYSU (Chen et al. 2017a).

Related Work

Person re-identification has attracted extensive research in the past 10 years (Gray and Tao 2008; Zheng, Gong, and Xiang 2013; Liao et al. 2015; Ahmed, Jones, and Marks 2015; Zheng et al. 2015; Xiao et al. 2016; Zheng, Gong, and Xiang 2016; Zhang, Xiang, and Gong 2016; Li, Zhu, and Gong 2017; Chen et al. 2017b) The dominant focus is on handling the re-id challenges arising from uncontrolled variations in illumination, background clutter and human pose. The resolution mismatch problem in LR re-id, however, is under-studied, with only a few works (Li et al. 2015; Jing et al. 2015; Wang et al. 2016) proposed. In (Li et al. 2015), it is assumed that images of the same person should be distributed similarly under different resolutions and propose simultaneously optimising cross-resolution image alignment and distance metric modelling in a joint learning framework. In (Jing et al. 2015), a semi-coupled low-rank dictionary learning approach was proposed to uncover the feature relationship between LR and HR images. In (Wang et al. 2016), the characteristics of the scale-distance function space is explored by varying the scale of LR images when matching with HR ones. These methods are limited due to the incapability of synthesising discriminative appearance information lost in image acquisition.

On the other hand, the related LR face recognition methods have been developed in the literature (Wang and Tang 2005; Hennings-Yeomans, Baker, and Kumar 2008; Huang and He 2011). Their basic idea is to synthesise HR faces by image super-resolution (SR) techniques with the need of dense feature point alignment. While being feasible for structure-constrained face images, it is difficult to align person images due to the greater degree of unknown variations in body parts, e.g., aligning a back view LR person image with a side view HR person image against other clutters. These SR-based LR face matching methods are therefore not suitable for the LR re-id problem (Li et al. 2015). In the mean time, generic-purpose SR methods have achieved remarkable success in synthesising missing appearance fidelity from LR input images thanks to the powerful modelling capacity of deep learning algorithms (Kim, Kwon Lee, and Mu Lee 2016; Dong et al. 2016; Lai et al. 2017; Tai, Yang, and Liu 2017). They may generate HR person images with higher visual quality, but remain ineffectiveness for LR re-id as shown in our evaluations. This is because they are designed for improving low-level pixel values but not high-level identity discrimination when learning to reconstruct the HR images.

In contrast to all the existing methods above, the proposed SING method is particularly designed to address the LR re-id problem uniquely characterised by the capability of synthesising HR images highly discriminative for cross-resolution identity matching without the need for exhaustive dense alignment across images. Our approach is built upon the idea of dedicating image SR for discriminative re-id so that the two processes are seamlessly integrated to maximise their compatibility and complementary advantages. Unlike existing LR re-id methods which depend on hand-crafted features, the proposed SING realises a joint deep learning formulation capable of simultaneously achieving re-id purposed image super-solving, discriminative re-id feature learning, and optimal re-id matching model induction.

Jointly Learning Super-Resolution and Re-ID

We want to reliably match an Low Resolution (LR) probe person image against a set of High Resolution (HR) gallery

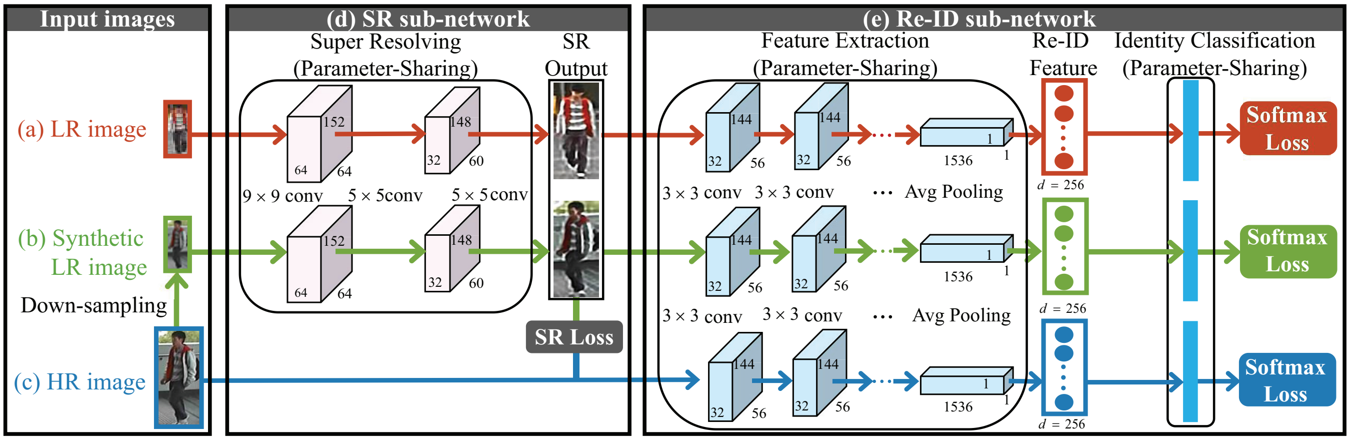

Figure 3: An overview of the proposed SING deep model for joint learning of image super-resolution and person identity classification. The SING CNN consists of two parts: SR sub-network (d) and Re-ID sub-network (e). In model training, we deploy three streams taking as input the LR image (a), synthetic LR image (b), and HR image (c), respectively. The middle stream (b) acts as a bridge for joining image SR (d) and person re-id (e) learning tasks.

images. To that end, we propose a joint learning approach of image Super-Resolution (SR) and person identity classification in order to correlate the two learning tasks and maximise their compatibility and complementary advantages.

Approach Overview. Assume that $X^l = \{(\boldsymbol{x}_i^l, y_i^l)\}_{i=1}^N$ is an LR person image set from one camera view and $X^h = \{(\boldsymbol{x}_i^h, y_i^h)\}_{i=1}^N$ an HR image set from another view, where $\boldsymbol{x}_i^l$ and $\boldsymbol{x}_i^h$ denote LR and HR images of identity class y_i^l and y_i^h, respectively. We wish to learn (1) an image super-resolution function $F_{sr}(\cdot)$ that can compensate effectively re-id information for the LR image $\boldsymbol{x}_i^l$, and (2) an identity discriminant feature extraction (FE) function $F_{fe}(\cdot)$ that can be performed on both super-resolved $F_{sr}(\boldsymbol{x}_i^l)$ and realistic $\boldsymbol{x}_i^h$ HR images, with the objective that $F_{fe}(F_{sr}(\boldsymbol{x}_i^l))$ is close to $F_{fe}(\boldsymbol{x}_j^h)$ in the feature space when they share the identity label (i.e., $y_j^h = y_i^l$) and vice verse. Formally, by learning $F_{sr}(\cdot)$ and $F_{fe}(\cdot)$ through joint formulation, we aim to obtain a re-id similarity matching metric:

$$S\Big(F_{fe}(F_{sr}(\boldsymbol{x}_i^l)), F_{fe}(\boldsymbol{x}_j^h)\Big) \tag{1}$$

respecting that after a proper image SR enhancement, an LR image captured in one camera view can be associated correctly with an HR image of the same person captured in another camera view.

Super-Resolution Formulation. We compensate the desired discriminative information missing in the LR images through super-resolution. To facilitate SR model training, we generate a *synthetic* LR version $X^{h2l} = \{(\boldsymbol{x}_i^{h2l}, y_i^h)\}_{i=1}^N$ of X^h by down-sampling, where $\boldsymbol{x}_i^{h2l}$ is the *synthetic* LR image corresponding to HR image $\boldsymbol{x}_i^h$. The X^{h2l} allows to optimise the following Mean Square Error (MSE) which measures the quality of image super-resolution:

$$L_{sr}\Big(\{\boldsymbol{x}_i^h\}_{i=1}^N\Big) = \frac{1}{N}\sum_{i=1}^N \|F_{sr}(\boldsymbol{x}_i^{h2l}) - \boldsymbol{x}_i^h\|_F^2. \tag{2}$$

Minimising the L_{sr} enforces the super-resolved image $F_{sr}(\boldsymbol{x}_i^{h2l})$ of $\boldsymbol{x}_i^{h2l}$ to close to the ground truth HR image $\boldsymbol{x}_i^h$. High-resolution appearance information is critical for obtaining reliable re-id features (Li et al. 2015). This optimisation (Eq. (2)) establishes the underlying relationship between LR and HR images in the image pixel space, but without a guarantee that the synthetic HR images are suitable for computing features discriminant for re-id matching. Reasons are: (1) It is very challenging if possible to train a perfect image SR model given that it is a non-convex and difficult-to-optimise problem with complex correlations involved among local and global pixels (Dong et al. 2016). (2) Artefacts are probably generated, which may negatively affect the subsequent re-id matching.

To address this limitation, we propose enforcing an identity constraint to guide the SR optimisation towards an image enhancement solution optimal for identity discrimination. This design differs from the typical SR objective that intrinsically seeks for a pixel-level mapping from LR input images to HR groundtruth without a semantic top-down learning guidance. Interestingly, by merging the re-id learning on $\boldsymbol{x}_i^h$ and $\boldsymbol{x}_i^l$ with this semantic constraint on $\boldsymbol{x}_i^{h2l}$, we simultaneously accomplish the re-id learning task.

Re-ID Formulation. More specifically, we concurrently optimise the classification of discriminative features w.r.t the same person label on HR and synthetic LR images, along with the cross-view LR images together. Formally, we formulate re-id classification constraint in the context of different images as:

$$\begin{aligned} L_{reid}\Big(\{(\boldsymbol{x}_i^l, \boldsymbol{x}_i^h, y_i^l, y_i^h)\}_{i=1}^N\Big) &= \frac{1}{N}\sum_{i=1}^N \Big(L_s\big(F_c(\boldsymbol{f}_i^h), y_i^h\big) \\ &+ L_s\big(F_c(\boldsymbol{f}_i^{h2l}), y_i^h\big) + L_s\big(F_c(\boldsymbol{f}_i^l), y_i^l\big)\Big), \end{aligned} \tag{3}$$

where $(\boldsymbol{x}_i^l, \boldsymbol{x}_i^h, y_i^l, y_i^h)$ consists of an LR image from X^l and an HR image from X^h as well as their corresponding identity labels. $F_c(\cdot)$ and $L_s(\cdot)$ represent a classification and loss function, respectively. All $\boldsymbol{f}$ notations denote the re-id feature vectors obtained from the FE function as:

$$\begin{aligned}\boldsymbol{f}_i^h = F_{fe}(\boldsymbol{x}_i^h), \boldsymbol{f}_i^{h2l} = F_{fe}\big(F_{sr}(\boldsymbol{x}_i^{h2l})\big),\\ \boldsymbol{f}_i^l = F_{fe}\big(F_{sr}(\boldsymbol{x}_i^l)\big).\end{aligned} \tag{4}$$

As such, the SR $F_{sr}(\cdot)$ and FE $F_{fe}(\cdot)$ functions are jointly constrained in the re-id optimisation.

Overall Formulation. After combining the SR and re-id formulation designs as above, we formulate the overall SING loss function as:

$$\begin{aligned}&L\Big(\{(\boldsymbol{x}_i^l, \boldsymbol{x}_i^h, y_i^l, y_i^h)\}_{i=1}^N\Big) =\\ &L_{reid}\Big(\{(\boldsymbol{x}_i^l, \boldsymbol{x}_i^h, y_i^l, y_i^h)\}_{i=1}^N\Big) + \alpha L_{sr}\Big(\{\boldsymbol{x}_i^h\}_{i=1}^N\Big),\end{aligned} \tag{5}$$

where the parameter α controls the balance between image SR loss and re-id loss. Optimising the joint loss L allows guiding the $F_{sr}(\cdot)$ to compensate semantically appearance details of the LR images towards identity salient fidelity synthesis and concurrently driving the $F_{fe}(\cdot)$ to extract accordingly identity discriminative features in a harmonious manner. Such a multi-task joint learning formulation is supposed to suit the LR person re-id problem.

Remark. A key characteristic of the proposed SING formulation (Eq. (5)) is the *seamless joining* of a restoration quantisation SR loss (Eq. (2)) and a person re-id loss (Eq. (3)), *both* subject to the same synthetic LR training image $\boldsymbol{x}_i^{h2l}$ (Fig. 3(b)) in the context of concurrent identity discriminant supervision on all three types of training images. That is, the synthetic LR image $\boldsymbol{x}_i^{h2l}$ and its re-id feature $\boldsymbol{f}_i^{h2l}$ together bridge and correlate the image SR (Fig. 3(d)) and person re-id (Fig. 3(e)) learning tasks. Without this connection, the two loss functions L_{sr} and L_{reid} will be optimised independently, rather than jointly and concurrently.

SING Instantiation

We choose to realise our SING formulation by deep CNN models. This is because deep CNN model has the following merits: (1) Good at learning discriminative representations from training data with successful demonstrations on both image SR (Dong et al. 2016; Wang et al. 2015) and person re-id (Li et al. 2014; Xiao et al. 2016); (2) Strong capability of learning non-convex tasks therefore suitable for handling complex appearance variations from lighting, occlusions and background clutters; (3) High flexibility of reformulating the network architecture with the possibility of avoiding the optimisation algorithm modification. The proposed SING CNN architecture is depicted in Fig. 3.

Network Architecture. Specifically, the SING CNN consists of two sub-networks: **(I)** *SR sub-network* which aims to compensate and recover the information loss in LR images, i.e., realising $F_{sr}(\cdot)$. It has two parameter-sharing streams taking as input $\boldsymbol{x}_i^l$ (LR image) and $\boldsymbol{x}_i^{h2l}$ (synthetic LR image), respectively. Following the SRCNN in (Dong et al. 2016), our SR sub-network is constructed by two convolutional (conv) layers followed by a ReLU non-linear layer and a reconstruction conv layer. The MSE loss function Eq. (2) is used for quantifying pixel level alignment degree between the groundtruth HR $\boldsymbol{x}_i^h$ and the SR output of $\boldsymbol{x}_i^{h2l}$ in training. **(II)** *Re-ID sub-network* which aims to learn identity discriminant features, i.e., realising $F_{fe}(\cdot)$, and also impose re-id constraints, i.e., realising $F_c(\cdot)$. It has three parameter-sharing streams taking as input the SR outputs of $\boldsymbol{x}_i^{h2l}$, LR image $\boldsymbol{x}_i^l$, and HR image $\boldsymbol{x}_i^h$, respectively. In our implementation, we adopt the DGD network (Xiao et al. 2016). In each stream, the penultimate fully connected (FC) layer outputs the re-id feature, which is then fed into the last FC layer for identity classification. The summation of all three stream's softmax losses (Eq. (3)) is used as the supervision signal for jointly qualifying the identification of all inputs during model training. In implementation, we upscale the LR images to an appropriate size (160×72 in our experiments) by bicubic interpolation as (Dong et al. 2016).

SR and Re-ID Joint Deep Learning. We achieve an end-to-end joint learning of image SR and person re-id in the proposed CNN by the multi-purposed synthetic LR image $\boldsymbol{x}_i^{h2l}$ (Fig. 3(b)). Formally, $\boldsymbol{x}_i^{h2l}$ and its re-id feature $\boldsymbol{f}_i^{h2l}$ function to join four losses: one SR loss on ($\boldsymbol{x}_i^{h2l}$, $\boldsymbol{x}_i^h$) correlated with three re-id losses on $\boldsymbol{f}_i^{h2l}$, $\boldsymbol{f}_i^l$ and $\boldsymbol{f}_i^h$. It is this loss connection design that brings more re-id discrimination awareness into the jointly optimised image SR model. We will evaluate the effect of this new modelling in our experiments.

LR Re-ID Deployment. In LR re-id deployment, we extract the re-id features for both LR probe and HR gallery images and then use the generic L_2 distance metric (Eq. (1)) for re-id matching. For HR images, we directly apply the jointly learned Re-ID sub-network to compute the re-id features. For LR images, we apply SR sub-network to super-resolve them before performing feature extraction as HR ones. We resize both LR and HR images to the input scale before feature computation as required by the SING CNN model.

Multi-Resolution Adaptive Fusion

The SING CNN model formulated as above assumes that all LR images have similar underlying resolutions. This is due to that the SR sub-network is optimised to super-resolve images by ratio m or around – the resolution ratio between synthetic LR and HR images. Consequently, the learned SING model may be suboptimal when LR-HR image resolution ratio is far from ratio m as possible in practice since there exist multiple different resolutions in real-world LR person images[2]. To address this problem, we propose to create φ anchor SING models $\{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_\varphi\}$ with each responsible for optimising a reference SR ratio in $\{m_1, m_2, \cdots, m_\varphi\}$ accordingly, and use them jointly to accommodate various resolutions involved in LR re-id matching. Each model $\boldsymbol{M}_i$ can be similarly learned as described above by the corresponding synthetic LR images X^{h2l} gen-

[2] While HR images also have different resolutions, we focus on handling the LR images in this work. This is because LR images suffer more significant information loss and therefore the major cause of degraded re-id matching performance. We assume HR images share a similar resolution for simplicity. However, the strategy proposed here can be similarly applied to deal with HR images of different underlying resolutions.

erated by ratio m_i down-sampling, along with LR and HR training images. In our evaluations, we used three models corresponding to down-sampling ratio $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$. In deployment, given an LR probe image, we firstly compute φ distance vectors $\{D_i\}_{i=1}^{\varphi}$ between the probe image and all the gallery images with each anchor SING model, where D_i denotes the distance by $\boldsymbol{M}_i$, $i \in \{1, 2, \cdots, \varphi\}$. Then, we form a multi-resolution fused distance vector as:

$$D_{mra} = \sum_{i=1}^{\varphi} w_i D_i, \tag{6}$$

where $\{w_i\}_{i=1}^{\varphi}$ represent the weights of the corresponding distances. To make D_{mra} resolution adaptive, we consider the similarity in underlying resolution among the LR probe image, all HR gallery images, and each SING model. We quantify the resolution similarity between the LR probe and HR gallery images as:

$$r = \sqrt{\frac{A_p}{\tilde{A}_g}}, \tag{7}$$

where A_p denotes the spatial area (i.e., pixel number) of the LR probe and $\tilde{A}_g$ the mean spatial area of all HR gallery images. They are computed on the genuine resolution scales without resizing. We then combine the model super-resolving ratio m_i as:

$$w_i = \exp\{-\sigma^{-2} \cdot (r - m_i)^2\}, \tag{8}$$

where σ is a scaling parameter estimated by cross validation.

Experiments

Datasets. We performed evaluations on three *simulated* and one *genuine* LR person re-id datasets (Fig. 4). Instead of assuming a single underlying resolution for all LR images, we consider Multiple Low Resolutions (*MLR*) as in real-world situations. Therefore, we used different down-sampling rates when simulating LR images by low-resolving HR ones.

(1) MLR-VIPeR was constructed from the VIPeR (Gray and Tao 2008) dataset. VIPeR contains 632 person image pairs captured by two cameras. Each image is of high resolution 128×48 in pixel. To make this dataset suitable for LR person re-id evaluation, we down-sampled all images from one camera view by a ratio randomly picked from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$, whilst remaining images of another view the same. This results in a simulated Multiple Low Resolutions VIPeR (MLR-VIPeR) dataset.

(2) MLR-SYSU is based on the SYSU dataset (Chen et al. 2017a). SYSU has totally 24,446 images of 502 people captured by two cameras. We randomly selected three images per person per camera in our evaluations and created an LR re-id dataset MLR-SYSU as for VIPeR.

(3) MLR-CUHK03 was built from the CUHK03 (Li et al. 2014) dataset. CUHK03 consists of five different pairs of camera views, and has more than 14,000 images of 1,467 pedestrians. Following the settings in (Xiao et al. 2016), both the manually cropped and automatically detected images were used in our evaluations. For each camera pair, we randomly selected one as LR probe image source by performing similar multi-resolution down-sampling. This results in a simulated LR re-id dataset MLR-CUHK03.

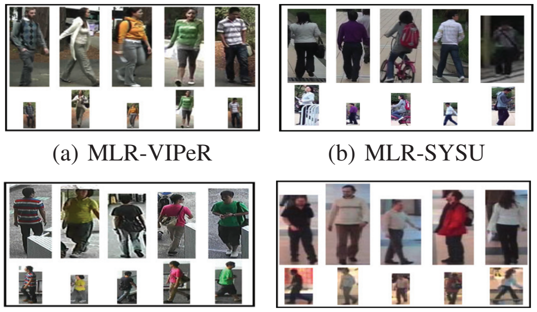
(a) MLR-VIPeR (b) MLR-SYSU (c) MLR-CUHK03 (d) CAVIAR

Figure 4: Examples of HR (1st row) and LR (2nd row) person images from four datasets.

(4) CAVIAR is a genuine LR person re-id dataset (Cheng et al. 2011). It contains 1,220 images of 72 persons captured from two camera views. Albeit in small scale, this dataset is suitable for evaluating LR re-id because the resolution of images from one camera (distant) is much lower than that from the other (close). We discard 22 people who appeared only in the close camera with HR images. For each of the remaining 50 used in our experiments, there are 10 HR and 10 LR images, i.e., a total of 1,000 images. Unlike other simulated datasets, LR images in CAVIAR involves multiple *realistic* resolutions.

Table 1: Comparing state-of-the-art LR re-id methods (%). The 1st/2nd best results are indicated in red/blue.

CAVIAR	r=1	r=5	r=10	r=20
JUDEA	**22.0**	**60.1**	**80.8**	**98.1**
SLD^{2}L	18.4	44.8	61.2	83.6
SDF	14.3	37.5	62.5	95.2
SING	**33.5**	**72.7**	**89.0**	**98.6**
MLR-CUHK03	r=1	r=5	r=10	r=20
JUDEA	**26.2**	**58.0**	**73.4**	**87.0**
SLD^{2}L	-	-	-	-
SDF	22.2	48.0	64.0	80.0
SING	**67.7**	**90.7**	**94.7**	**97.4**
MLR-SYSU	r=1	r=5	r=10	r=20
JUDEA	18.3	**41.9**	**54.5**	**68.0**
SLD^{2}L	**20.3**	34.8	43.4	55.4
SDF	13.3	26.7	42.9	66.7
SING	**50.7**	**75.4**	**83.1**	**88.1**
MLR-VIPeR	r=1	r=5	r=10	r=20
JUDEA	**26.0**	**55.1**	**69.2**	**82.3**
SLD^{2}L	20.3	44.0	62.0	**78.2**
SDF	9.52	38.1	52.4	68.0
SING	**33.5**	**57.0**	**66.5**	76.6

Evaluation Protocol. We adopted the standard single-shot re-id setting in our experiments. All datasets except MLR-CUHK03 were randomly divided into two halves, one for training and one for testing. That is, there are $p = 25, 316$ and 251 persons in the testing set of CAVIAR, MLR-VIPeR and MLR-SYSU, respectively. Following (Xiao et al. 2016), we utilised the benchmarking 1,367/100 training/test identity split. On the testing data, we constructed the probe set

Table 2: Comparing combinations of image super-resolution and person re-id schemes (%).

Super-Resolution Method	Re-ID Method	CAVIAR				MLR-CUHK03				MLR-SYSU				MLR-VIPeR			
		r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
Bilinear	XQDA	22.7	61.4	81.9	98.8	45.5	78.0	87.8	93.7	39.3	67.4	77.2	85.6	37.6	65.7	78.5	89.7
Bicubic	XQDA	24.3	63.4	83.1	**99.2**	45.1	78.1	87.7	93.3	40.0	66.9	77.2	85.5	37.8	66.0	78.9	89.3
SRCNN	XQDA	24.8	64.3	84.0	**99.1**	44.7	77.8	87.5	93.1	40.3	67.3	77.4	85.6	36.5	65.1	78.9	89.8
Bilinear	NFST	23.3	60.5	82.2	99.0	48.0	47.9	46.2	49.0	41.6	69.0	79.5	87.7	**39.7**	**68.4**	**81.0**	**90.4**
Bicubic	NFST	24.5	61.1	82.1	99.2	47.9	74.8	83.6	92.8	42.4	69.3	**80.5**	**88.0**	**39.2**	**67.9**	**80.3**	**90.7**
SRCNN	NFST	25.0	61.2	82.9	99.2	49.0	74.8	85.0	92.1	**43.2**	**69.6**	80.3	88.0	38.6	67.1	79.5	90.1
Bilinear	DGD	25.3	61.0	82.6	98.4	58.5	86.0	92.2	96.0	39.6	66.4	74.8	82.5	23.1	45.9	56.6	67.7
Bicubic	DGD	27.4	63.4	83.0	98.5	62.5	88.7	93.7	96.5	41.5	67.4	76.9	84.7	25.0	51.3	59.2	69.3
SRCNN	DGD	**28.4**	**66.3**	**85.9**	98.5	**63.8**	**89.3**	**93.9**	**96.8**	42.6	68.2	77.1	85.5	25.3	48.4	57.3	66.5
SING		**33.5**	**72.7**	**89.0**	98.6	**67.7**	**90.7**	**94.7**	**97.4**	**50.7**	**75.4**	**83.1**	**88.1**	33.5	57.0	66.5	76.6

with all LR images per person, and the gallery set with one randomly selected HR image per person. We repeated 10 times of the above random data split. For performance evaluation, we used the average Cumulative Match Characteristic (CMC) to measure the LR re-id matching performance.

Implementation of SING. We initialised the SR and Re-ID sub-networks by SRCNN (Dong et al. 2016) pre-trained on ImageNet-1K and DGD (Xiao et al. 2016) pre-trained on the training data of Market-1501 (Zheng et al. 2015), respectively. The scaling parameter σ in Eq. (8) was set by cross validation on the validation set. We set the balance coefficient $\alpha = 1$ (Eq. (5)) which assumes equal importance between image SR and re-id feature learning.

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

We compared the proposed SING method with three existing state-of-the-art LR re-id methods: (1) JUDEA (Li et al. 2015) – a cross-scale discriminative distance metric learning model, (2) SLD^{2}L (Jing et al. 2015) – a feature transformation or alignment model, (3) SDF (Wang et al. 2016) – a scale-distance function learning model. For both baselines, we used the codes provided by the authors. It is evident from Table 1 that the SING method outperforms both competitors in most cases, for example, surpassing the best alternative JUDEA by 11.5%, 41.5%, 32.4%, 7.5% at rank-1 on CAVIAR, MLR-CUHK03, MLR-SYSU, and MLR-VIPeR respectively. The performance margins of SING over the SLD^{2}L and SDF models are larger still[3]. This indicates the advantages of the proposed SING model in handling both simulated and genuine LR re-id. The performance superiority is mainly due to: **(1)** The capability of jointly super-resolving person images and learning re-id discriminant features, which allows to maximise their mutual correlation. Compared to cross-resolution alignment based competitor, our model is able to synthesise high-frequency missing in LR images by re-id discriminative super-resolution and therefore extract richer representation. This not only directly mitigates the information amount discrepancy problem but also fills the hard-to-bridge matching gap between different resolutions with appearance pattern divergence involved. **(2)** The deep learning advantages in modelling non-convex SR and re-id optimisation by learning from multi-sourced image data in a unified model.

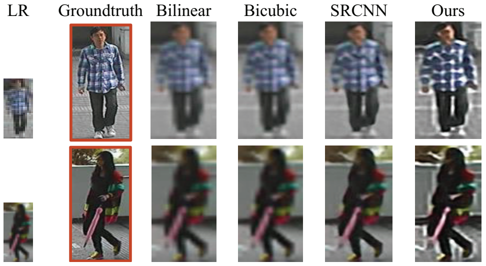

Figure 5: Qualitative examples of super-resolved person images. The groundtruth HR images are indicated by red bounding boxes.

Comparing Super-Resolution + Re-ID Scheme

We further evaluated the LR person re-id performance by deploying a straightforward combination of the super-resolution and person re-id scheme. While conventional re-id methods assume HR images, we utilise state-of-the-art SR models when LR images are given to meet their requirement. We used the same training images as the proposed SING to fine-tune the SR models. The proposed multi-resolution adaptive fusion algorithm was applied to all the compared methods in a fair comparison principle.

The conventional Re-ID methods considered in our evaluations are: (1) XQDA (Liao et al. 2015): A supervised Mahalanobis metric learning method, (2) NFST (Zhang, Xiang, and Gong 2016): A null subspace learning method, (3) DGD (Xiao et al. 2016): A widely used deep CNN re-id model. We utilised the contemporary LOMO hand-crafted feature (Liao et al. 2015) for the XQDA and NFST methods.

Image SR methods we selected for evaluation include two standard algorithms and one state-of-the-art: (1) Bilinear: A popular linear interpolation based SR model which is effective to handle generic image scaling; (2) Bicubic: Another widely used image SR method which is an extension of cubic interpolation; (3) SRCNN (Dong et al. 2016): An existing state-of-the-art deep CNN based SR model.

Evaluating Overall Performance. Table 2 shows that the proposed SING outperformed all SR+Re-ID methods on all datasets except MLR-VIPeR. When compared with SR+DGD, the SING is consistently superior on all datasets even on MLR-VIPeR. Specifically, the rank-1 matching gain over all competitors by the SING can reach 5.1%(33.5-28.4), 3.9%(67.7-63.8), and 7.5%(50.7-43.2) on CAVIAR,

[3] SLD^{2}L fails to run on MLR-CUHK03 due to out of memory on a modern workstation with 256 GB memory.

MLR-CUHK03 and MLR-SYSU, respectively. On MLR-VIPeR, the best performers are hand-crafted feature LOMO based models XQDA and NFST. This is reasonable considering that the MLR-VIPeR training data is sparse (632 images from 316 person classes). Nevertheless, our SING surpassed the deep alternative SRCNN+DGD by 8.2%(33.5-25.3) rank-1, which validates the benefits of joint learning SR and Re-ID in the proposed approach.

Effect of Image SR. We examine the effect of only performing image SR for conventional re-id methods on LR re-id matching performance. It is found in Table 2 that an independent preprocessing of super-resolving LR person images can only bring marginal benefits. For example, when using the DGD re-id model, as compared to Bilinear, SRCNN yields merely 3.1%(28.4-25.3) / 5.3%(63.8-58.5) / 3.0%(42.6-39.6) / 2.2%(25.3-23.1) additional rank-1 rates on CAVIAR / MLR-CUHK03 / MLR-SYSU / MLR-VIPeR, respectively. The positive effect of SR for XQDA and NFST is even more limited. In comparison, the performance advantages of the SING over SRCNN+DGD can be observed on all datasets, achieved by learning the two models jointly with a single multi-task loss optimisation in an end-to-end manner. This suggests that directly applying existing SR models cannot solve the LR re-id problem, although they can produce visually favourable HR images (Fig. 5).

Qualitative Evaluation. We compared the super-resolved person images produced by Bilinear, Bicubic, SRCNN and our SING. Two examples are shown in Fig. 5. We have the following observations: (1) Super-resolved images by Bilinear and Bicubic are more blurry than those by SRCNN and SING. (2) More edge/contour elements and better texture patterns are recovered by SING. (3) The colour distributions of resolved images by SING are most similar to the ground truth. This visually indicates the advantages of SING over SR+Re-ID methods – due to the capability of recovering missing appearance details whilst ensuring high re-id discrimination. Note that the PSNR scores for the top/bottom images in Fig. 5 are 19.04/21.25 by bilinear, 19.24/21.60 by bicubic, 19.86/22.72 by SRCNN, 17.95/18.97 by ours. Among them, our method achieves lower PSNR in contrast to the re-id performance comparison. This confirms that the PSNR is not a high-level perceptual quality measurement, but a low-level pixel-wise metric.

Table 3: Effect of jointly super-resolving and classifying synthetic LR images (%).

Models	CAVIAR	MLR-CUHK03	MLR-SYSU	MLR-VIPeR
	r=1	r=1	r=1	r=1
SING(No Synthetic LR)	25.8	57.1	38.7	23.1
SING	**33.5**	**67.7**	**50.7**	**33.5**

Further Analysis of SING

Effect of Synthetic LR Images in SING. We evaluated the contribution of joint super-resolving the synthetic LR images by the MSE loss (Eq. (2)), in conjunction with classifying the resolved image (Eq. (3)). To this end, we evaluate a stripped-down SING without the stream of the synthetic LR images (see the "green" arrows in Fig. 3). As such, the MSE SR loss is removed due to no LR-HR training image pairs available. Table 3 shows that inferior LR re-id performance will be generated without this joint learning stream. For example, the rank-1 rate drops from 33.5% to 25.8% on CAVIAR, from 67.7% to 57.1% on MLR-CUHK03, from 50.7% to 38.7% on MLR-SYSU, from 33.5% to 23.1% on MLR-VIPeR, respectively. This drop suggests the positive impact of the proposed joint learning approach in guiding the image SR model towards generating HR images with re-id discriminative visual information.

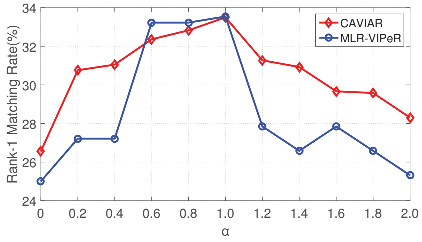

Figure 6: Effect of balancing image SR and person Re-ID loss.

Table 4: Effect of scale adaptive low-resolution fusion (%).

Models	CAVIAR	MLR-CUHK03	MLR-SYSU	MLR-VIPeR
	r=1	r=1	r=1	r=1
$M_{\frac{1}{2}}$	29.5	63.1	44.0	28.5
$M_{\frac{1}{3}}$	27.2	62.6	45.1	30.7
$M_{\frac{1}{4}}$	27.4	61.6	44.6	30.1
$M_{\frac{1}{2}}+M_{\frac{1}{3}}$	31.1	**66.8**	48.6	31.3
$M_{\frac{1}{2}}+M_{\frac{1}{4}}$	**31.9**	64.7	48.6	32.0
$M_{\frac{1}{3}}+M_{\frac{1}{4}}$	29.9	63.6	**49.5**	**32.6**
$M_{\frac{1}{2}}+M_{\frac{1}{3}}+M_{\frac{1}{4}}$	**33.5**	**67.7**	**50.7**	**33.5**

Effect of Multi-Resolution Adaptive Fusion. We evaluated the LR re-id performance of 6 combination schemes from 3 different resolution-specific SING models $\{M_{\frac{1}{2}}, M_{\frac{1}{3}}, M_{\frac{1}{4}}\}$. Table 4 shows that fusing more resolutions leads to better results with the best overall performance yielded by fusing all three resolution-specific SING models. More broadly, this finding is consistent in spirit with the classical pyramid matching kernel (Grauman and Darrell 2005; Lazebnik, Schmid, and Ponce 2006) with the difference that our multi-resolution fusion is *uniquely* on modelling multiple resolutions rather than multiple spatial decompositions of a single resolution.

Effect of SR and Re-ID Loss Balancing. We evaluated the balancing effect between image SR and person re-id loss by varying the trade-off parameter α in Eqn. (5) ($\alpha = 1$ in all other experiments). We conducted this analysis on the genuine LR dataset CAVIAR and the simulated MLR-VIPeR. Figure 6 shows that: (1) When setting $\alpha = 0$, the rank-1 performances drop from 33.5% to 26.6%, 33.5% to 25.0% on CAVIAR and MLR-VIPeR, respectively. This is because SR reconstruction is totally ignored and thus there is no interaction between SR and re-id. (2) When setting a large α, e.g., > 1, the image SR reconstruction loss will dominate the joint learning. This adversely affects discriminant feature extraction. This evaluation implies that both SR and re-id modelling can be similarly important for LR re-id.

Conclusion

In this work, we present for the first time an image SR and person re-id joint formulation SING for tackling the under-studied LR re-id matching problem. We realise this approach by designing a hybrid deep CNN architecture for not only achieving highly non-convex SR and re-id functions but also enjoying an end-to-end joint optimisation in order to maximise complementary advantages, i.e., the dedication of image SR for LR re-id matching. Moreover, we introduce an adaptive fusion algorithm for handling the largely ignored multi-resolution problem. By extensive comparative evaluations on both simulated and genuine LR person re-id datasets, we have shown the superiority of our SING approach over a wide variety of state-of-the-art re-id and SR methods. We also provide in-depth component examinations and analysis for giving insights on the SING model design.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC(61522115, 61472456, 61661130157), the Royal Society Newton Advanced Fellowship (NA150459), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157). The principal investigator of this work is Wei-Shi Zheng.

References

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*, 3908–3916.

Chen, Y.-C.; Zheng, W.-S.; Lai, J.-H.; and Yuen, P. C. 2017a. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE TCSVT* 27(8):1661–1675.

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

Cheng, D. S.; Cristani, M.; Stoppa, M.; Bazzani, L.; and Murino, V. 2011. Custom pictorial structures for re-identification. In *BMVC*, 6–10.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2016. Image super-resolution using deep convolutional networks. *IEEE TPAMI* 38(2):295–307.

Gong, S.; Cristani, M.; Yan, S.; and Loy, C. C. 2014. *Person re-identification*. Springer.

Grauman, K., and Darrell, T. 2005. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 1458–1465.

Gray, D., and Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 262–275.

He, W.-X.; Chen, Y.-C.; and Lai, J.-H. 2016. Cross-view transformation based sparse reconstruction for person re-identification. In *ICPR*, 3410–3415.

Hennings-Yeomans, P. H.; Baker, S.; and Kumar, B. V. 2008. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *CVPR*, 1–8.

Huang, H., and He, H. 2011. Super-resolution method for face recognition using nonlinear mappings on coherent features. *IEEE TNN* 22(1):121–130.

Jing, X.-Y.; Zhu, X.; Wu, F.; You, X.; Liu, Q.; Yue, D.; Hu, R.; and Xu, B. 2015. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, 695–704.

Kim, J.; Kwon Lee, J.; and Mu Lee, K. 2016. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 1646–1654.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 5835 – 5843.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2169–2178.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 152–159.

Li, X.; Zheng, W.-S.; Wang, X.; Xiang, T.; and Gong, S. 2015. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 3765–3773.

Li, W.; Zhu, X.; and Gong, S. 2017. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2197–2206.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 1363–1372.

Tai, Y.; Yang, J.; and Liu, X. 2017. Image super-resolution via deep recursive residual network. In *CVPR*, 2790 – 2798.

Wang, X., and Tang, X. 2005. Hallucinating face by eigentransformation. *IEEE TSMC(Part C)* 35(3):425–434.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*, 688–703.

Wang, Z.; Liu, D.; Yang, J.; Han, W.; and Huang, T. 2015. Deep networks for image super-resolution with sparse prior. In *ICCV*, 370–378.

Wang, Z.; Hu, R.; Yu, Y.; Jiang, J.; Liang, C.; and Wang, J. 2016. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, 2669–2675.

Wong, Y.; Sanderson, C.; Mau, S.; and Lovell, B. C. 2010. Dynamic amelioration of resolution mismatches for local feature based identity inference. In *ICPR*, 1200–1203.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 1249–1258.

Zhang, L.; Xiang, T.; and Gong, S. 2016. Learning a discriminative null space for person re-identification. In *CVPR*, 1239–1248.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*, 1116–1124.

Zheng, W.-S.; Gong, S.; and Xiang, T. 2013. Re-identification by relative distance comparison. *IEEE TPAMI* 35(3):653–668.

Zheng, W.-S.; Gong, S.; and Xiang, T. 2016. Towards open-world person re-identification by one-shot group-based verification. *IEEE TPAMI* 38(3):591–606.