Deep Low-Resolution Person Re-Identification

Jiening Jiao,1,2 Wei-Shi Zheng,3,4∗ Ancong Wu,1 Xiatian Zhu,5 Shaogang Gong5
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 Super-resolution and Identity joint learning 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 of-

∗Corresponding author

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
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 images by learning the integration 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 purposes image super-solving, discriminative re-id feature learning, and optimal re-id matching model induction.
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 $X^i = \{(x^i_l, y^i_l)\}_{i=1}^N$ is an LR person image set from one camera view and $X^h = \{(x^h_l, y^h_l)\}_{i=1}^N$ an HR image set from another view, where $x^i_l$ and $x^h_l$ denote LR and HR images of identity class $y^i_l$ and $y^h_l$, 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 $x^i_l$, and (2) an identity discriminant feature extraction (FE) function $F_{fe}(\cdot)$ that can be performed on both super-resolved $F_{sr}(x^i_l)$ and realistic $x^h_l$ HR images, with the objective that $F_{fe}(F_{sr}(x^i_l))$ is close to $F_{fe}(x^h_l)$ in the feature space when they share the identity label (i.e., $y^h_l = y^i_l$) and vice versa. Formally, by learning $F_{sr}(\cdot)$ and $F_{fe}(\cdot)$ through joint formulation, we aim to obtain a re-id similarity matching metric:

$$S \left( F_{fe}(F_{sr}(x^i_l)), F_{fe}(x^h_l) \right)$$

(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 $X^{h2l} = \{(x^{h2l}_l, y^i_l)\}_{i=1}^N$ of $X^h$ by down-sampling, where $x^{h2l}_l$ is the synthetic LR image corresponding to HR image $x^h_l$. The $X^{h2l}$ allows to optimise the following Mean Square Error (MSE) which measures the quality of image super-resolution:

$$L_{sr} \left( \{x^h_l\}_{i=1}^N \right) = \frac{1}{N} \sum_{i=1}^{N} \| F_{sr}(x^{h2l}_l) - x^h_l \|_F^2.$$  

(2)

Minimising the $L_{sr}$ enforces the super-resolved image $F_{sr}(x^{h2l}_l)$ of $x^{h2l}_l$ to close to the ground truth HR image $x^h$. 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 $x^h_l$ and $x^i_l$ with this semantic constraint on $x^{h2l}_l$, 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:

$$L_{reid} \left( \{(x^i_l, x^h_l, y^i_l)\}_{i=1}^N \right) = \frac{1}{N} \sum_{i=1}^{N} \left( L_s \left( F_{fe}(x^h_l), y^i_l \right) \right) + L_s \left( F_{fe}(x^{h2l}_l), y^h_l \right) + L_s \left( F_{fe}(x^i_l), y^i_l \right),$$

(3)

where $(x^i_l, x^h_l, y^i_l)$ consists of an LR image from $X^i$ and an HR image from $X^h$ as well as their corresponding iden-

---

**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.
tity labels. \( F_\epsilon(\cdot) \) and \( L_s(\cdot) \) represent a classification and loss function, respectively. All \( f \) notations denote the re-id feature vectors obtained from the FE function as:

\[
f^h_i = F_\epsilon(x^h_i), \quad f^{2h}_i = F_\epsilon(F_{sr}(x^{2h}_i)),
\]

\[
f^l_i = F_\epsilon(F_{sr}(x^l_i)).
\]  

(4)

As such, the SR \( F_{sr}(\cdot) \) and FE \( F_\epsilon(\cdot) \) 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:

\[
L \left( \{(x^l_i, x^h_i, y^l_i, y^h_i)\}_{i=1}^N \right) =
L_{re-id} \left( \{(x^l_i, x^h_i, y^l_i, y^h_i)\}_{i=1}^N \right)
+ \alpha L_{sr} \left( \{x^h_i\}_{i=1}^N \right),
\]  

(5)

where the parameter \( \alpha \) 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 \( F_\epsilon(\cdot) \) 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 

**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 re-formulating 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) \( \text{SR sub-network} \) 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 \( x^l_i \) (LR image) and \( x^{2h}_i \) (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 \( x^h_i \) and the SR output of \( x^{2h}_i \) in training. (II) \( \text{Re-ID sub-network} \) which aims to learn identity discriminant features, i.e., realising \( F_\epsilon(\cdot) \), and also impose re-id constraints, i.e., realising \( F_{sr}(\cdot) \). It has three parameter-sharing streams taking as input the SR outputs of \( x^{2h}_i \), LR image \( x^l_i \), and HR image \( x^h_i \), 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 qualify the identification of all inputs during model training. In implementation, we upscale the LR images to an appropriate size (160 \( \times \) 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 \( x^{2h}_i \) (Fig. 3(b)). Formally, \( x^{2h}_i \) and its re-id feature \( f^{2h}_i \) function to join four losses: one SR loss on \( (x^{2h}_i, x^h_i) \) correlated with three re-id losses on \( f^{h}_i, f^l_i \) and \( f^h_i \). 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 L2 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 \( \varphi \) anchor SING models \{\( M_1, M_2, \ldots, M_\varphi \)\} with each responsible for optimising a reference SR ratio in \{\( m_1, m_2, \ldots, m_\varphi \)\} accordingly, and use them jointly to accommodate various resolutions involved in LR re-id matching. Each model \( M_i \) can be similarly learned as described above by the corresponding synthetic LR images \( X^{2h}_i \) generated by

\(^2\)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 $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{5}\}$. In deployment, given an LR probe image, we firstly compute $\varphi$ 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 $M_i$, $i \in \{1, 2, \cdots, \varphi\}$. Then, we form a multi-resolution fused distance vector as:

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

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:

$$r = \sqrt{\frac{A_p}{A_g}},$$

where $A_p$ denotes the spatial area (i.e., pixel number) of the LR probe and $A_g$ 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:

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

where $\sigma$ 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 $128 \times 48$ 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 $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{5}\}$, 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.

4. **CA VIAR** 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 CA VIAR involves multiple realistic resolutions.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>r=1</th>
<th>r=5</th>
<th>r=10</th>
<th>r=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAVIAR</td>
<td>JUDEA</td>
<td>22.0</td>
<td>60.1</td>
<td>80.8</td>
</tr>
<tr>
<td></td>
<td>SLD$^2$L</td>
<td>18.4</td>
<td>44.8</td>
<td>61.2</td>
</tr>
<tr>
<td></td>
<td>SDF</td>
<td>14.3</td>
<td>37.5</td>
<td>62.5</td>
</tr>
<tr>
<td></td>
<td>SING</td>
<td>33.5</td>
<td>72.7</td>
<td>89.0</td>
</tr>
<tr>
<td>MLR-CUHK03</td>
<td>JUDEA</td>
<td>26.2</td>
<td>58.0</td>
<td>73.4</td>
</tr>
<tr>
<td></td>
<td>SLD$^2$L</td>
<td>22.2</td>
<td>48.0</td>
<td>64.0</td>
</tr>
<tr>
<td></td>
<td>SDF</td>
<td>67.7</td>
<td>90.7</td>
<td>94.7</td>
</tr>
<tr>
<td>MLR-SYSU</td>
<td>JUDEA</td>
<td>18.3</td>
<td>41.9</td>
<td>54.5</td>
</tr>
<tr>
<td></td>
<td>SLD$^2$L</td>
<td>20.3</td>
<td>34.8</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>SDF</td>
<td>13.3</td>
<td>26.7</td>
<td>42.9</td>
</tr>
<tr>
<td></td>
<td>SING</td>
<td>59.7</td>
<td>75.4</td>
<td>83.1</td>
</tr>
<tr>
<td>MLR-VIPeR</td>
<td>JUDEA</td>
<td>26.0</td>
<td>55.1</td>
<td>69.2</td>
</tr>
<tr>
<td></td>
<td>SLD$^2$L</td>
<td>20.3</td>
<td>44.0</td>
<td>62.0</td>
</tr>
<tr>
<td></td>
<td>SDF</td>
<td>9.52</td>
<td>38.1</td>
<td>52.4</td>
</tr>
<tr>
<td></td>
<td>SING</td>
<td>33.5</td>
<td>57.0</td>
<td>66.5</td>
</tr>
</tbody>
</table>

**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
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) SLD²L (Jing et al. 2015) – a feature transformation or alignment model, and (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 SLD²L and SDF models are larger still. 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.

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.

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,
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 SRCCN+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 pre-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.

**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.

**Effect of Multi-Resolution Adaptive Fusion.** We evaluated the LR re-id performance of 6 combination schemes from 3 different resolution-specific SING models \( \{ M_1, M_2, M_3 \} \). 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 \( \alpha \) in Eqn. (5) \( (\alpha = 1 \text{ in all other experiments}) \). 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 \( \alpha \), e.g., > 1, the image SR reconstruction loss will dominate the joint learning. This adversely affects discriminant fea-
ture 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**


