International Journal of Computer Vision
https://doi.org/10.1007/s11263-021-01518-z

Joint Bilateral-Resolution Identity Modeling for Cross-Resolution Person Re-Identification

Wei-Shi Zheng[1,5] · Jincheng Hong[2] · Jiening Jiao[1,6] · Ancong Wu[1] · Xiatian Zhu[3] · Shaogang Gong[4] · Jiayin Qin[2] · Jianhuang Lai[1]

Received: 10 August 2020 / Accepted: 16 July 2021

Abstract

Person images captured by public surveillance cameras often have low resolutions (LRs), along with uncontrolled pose variations, background clutter and occlusion. These issues cause the *resolution mismatch* problem when matched with high-resolution (HR) gallery images (typically available during collection), harming the person re-identification (re-id) performance. While a number of methods have been introduced based on the joint learning of super-resolution and person re-id, they ignore specific discriminant identity information encoded in LR person images, leading to ineffective model performance. In this work, we propose a novel joint bilateral-resolution identity modeling method that concurrently performs HR-specific identity feature learning with super-resolution, LR-specific identity feature learning, and person re-id optimization. We also introduce an adaptive ensemble algorithm for handling different low resolutions. Extensive evaluations validate the advantages of our method over related state-of-the-art re-id and super-resolution methods on cross-resolution re-id benchmarks. An important discovery is that leveraging LR-specific identity information enables a simple cascade of super-resolution and person re-id learning to achieve state-of-the-art performance, without elaborate model design nor bells and whistles, which has not been investigated before.

Keywords Person re-identification · Low-resolution · Visual surveillance

1 Introduction

Person re-identification (re-id) matches identity classes in person bounding box images extracted from nonoverlapping camera views in open surveillance spaces (Gong et al. 2014). Existing re-id methods typically focus on addressing variations in illumination, occlusion, and background clutter by designing feature representations (Liao et al. 2015; Matsukawa et al. 2016; Wu et al. 2016; Qian et al. 2017; Kalayeh

Communicated by Diane Larlus.

✉ Jianhuang Lai
stsljh@mail.sysu.edu.cn

Wei-Shi Zheng
wszheng@ieee.org ; zhwshi@mail.sysu.edu.cn

Jincheng Hong
hongjch@mail2.sysu.edu.cn

Jiening Jiao
jiaojn@mail2.sysu.edu.cn

Ancong Wu
wuancong@mail2.sysu.edu.cn

Xiatian Zhu
eddy.zhuxt@gmail.com

Shaogang Gong
s.gong@qmul.ac.uk

Jiayin Qin
issqjy@mail.sysu.edu.cn

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China

[2] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510275, China

[3] Centre for Vision, Speech and Signal Processing (CVSSP), Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, England

[4] Shaogang Gong is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, England

[5] Peng Cheng Laboratory, Shenzhen 518005, China

[6] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

Published online: 22 November 2021

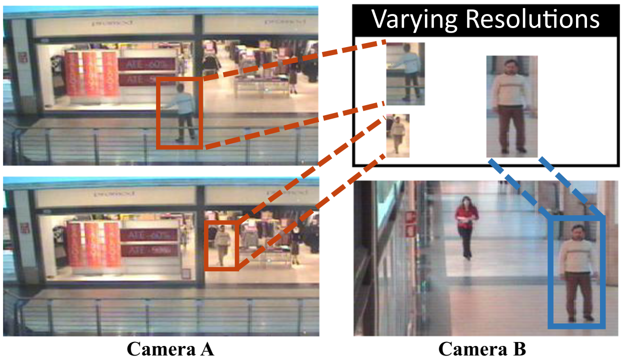

Fig. 1 Illustration of person images with varying resolutions for the open-space person re-identification (re-id) task. Three images of a person were captured by two camera views at different locations in a shopping center. The image captured by camera B has a higher resolution than the two images from camera A. This cross-resolution property makes person re-id more challenging

et al. 2018; Sun et al. 2018; Guo et al. 2019; Zhou et al. 2019) or learning matching distance metrics (Zheng et al. 2013; Wang et al. 2014; He et al. 2016b; Zhang et al. 2016; Fan et al. 2018; Zheng et al. 2018; Yu et al. 2018) or their combination (Li et al. 2014; Ahmed et al. 2015; Xiao et al. 2016; Li et al. 2017; Zheng et al. 2019b; Dai et al. 2019; Zheng et al. 2019a). The designed re-id models of these works often do not take the impact of low resolution images into account. However, surveillance person images often have varying resolutions due to variations in the distance from the camera to the person and the camera deployment settings (Fig. 1), which gives rise to the *resolution mismatch* problem (Jiao et al. 2018; Cheng et al. 2020) (Fig. 2).

It is challenging to reliably match low-resolution (LR) probe images against high resolution (HR) gallery images across both camera views and resolutions [1]. This requires addressing the *discrepancy in the amount of information* in cross-resolution matching since LR images contain much less information than HR images with discriminative appearance details largely lost in the image acquisition process. It is called **cross-resolution person re-identification**.

To address the nontrivial resolution mismatch problem in cross-resolution person re-id, a number of early methods have been proposed (Jing et al. 2015; Wang et al. 2016; Li et al. 2015). These methods, however, share a few common weaknesses: (1) Instead of recovering the missing discriminative appearance information, they perform cross-resolution representation transformation in a predefined feature space. However, these works do not solve the *discrepancy in the amount of information* challenge, *i.e.*, as potential discriminant information from HR images may not be effectively used when matching between low-resolution and high-resolution images directly. (2) Some of them (Jing et al. 2015; Wang et al. 2016; Li et al. 2015) rely on handcrafted visual features without using deep learning for mining the complementary advantages of feature learning and matching metric joint optimization.

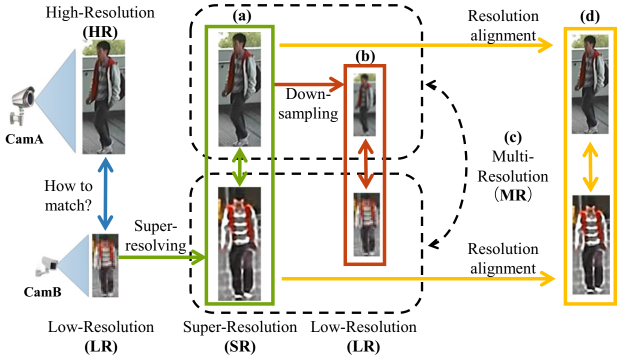

Fig. 2 Four different strategies for cross-resolution person re-identification (re-id): **(a)** Super-resolving LR images to extract the HR-specific identity features, **(b)** downsampling HR images to extract the LR-specific identity features, **(c)** using multi-resolution images to extract joint bilateral resolution features for re-id matching, and **(d)** aligning image features across resolution to learn resolution-invariant representations. Our approach is a multi-resolution solution exploiting both high-resolution specific and low-resolution specific features, considering the strategies of (a), (b) and (c) jointly. It differs from existing works that belong to the category of (a) (Wang et al. 2018; Cheng et al. 2020; Jiao et al. 2018) or (d) (Huang et al. 2020; Li et al. 2019a; Mao et al. 2019)

Another intuitive to solve the resolution mismatch problem in cross-resolution person re-id is to super-resolve the LR images directly so that super-resolved (SR) images can serve as a bridge to realistic HR images for identification. Image super-resolution should offer an effective solution to mitigate the discrepancy in the amount of information challenge due to its capability of synthesizing high-frequency details. However, a direct combination of super-resolution and re-id may be suboptimal in compatibility: Generic-purpose super-resolution methods are designed to improve the image visual fidelity rather than the re-id matching performance, with visual artifacts generated in the super-resolution reconstruction process typically irrelevant and problematic to re-id matching.

Recently, several works, including our preliminary work (Jiao et al. 2018), have shown that joint learning of the image super-resolution and person re-id is a simple yet effective method for cross-resolution person re-id (Cheng et al. 2020; Chen et al. 2017c; Huang et al. 2020; Li et al. 2019a).

[1] Note that in terms of visual surveillance definitions, the quality of so-called high resolution (HR) images is poorer than that of social media photos taken by professional photographers. In this context, we define LR and HR in a relative sense for surveillance-quality image data. By default, we define the "resolution" as *the underlying resolution* (Wang et al. 2010) rather than *the image spatial size (scale)*. A given image can be arbitrarily resized with little change to its underlying resolution (Fig. 3). Hence, the image spatial size is not an accurate indicator of the underlying resolution.

However, these methods ignore the exploitation of LR discriminant information for person re-id and do not attempt to formulate a joint learning framework for exploring the discriminant re-id features in both HR and LR images.

We argue that either HR-specific or LR-specific identity features alone are not sufficient for cross-resolution person re-id. This assertion is inspired by human visual systems, which take advantage of multiscale visual information, including feature representations at both small (global contextual) and large (local saliency) scales (Navon 1977). Therefore, we develop a joint bilateral identity modeling (JBIM) framework. Specifically, JBIM combines HR-specific identity modeling (HIM) and LR-specific identity modeling (LIM): HIM aims to improve the integration compatibility between image super-resolution and person re-id by learning identity-sensitive high-frequency appearance information, and LIM is for learning the complementary LR-specific identity information. The framework therefore learns discriminant bilateral-resolution joint features from SR and LR images, along with person re-id on the joint features. In the presence of different LRs, we further present a multi-resolution adaptive ensemble mechanism by aggregating a set of anchor JBIM network models (each optimized for a reference resolution) in a probe-specific manner. As shown in Fig. 2, our approach is a multi-resolution solution exploiting both high-resolution specific and low-resolution specific features, jointly considering the strategies of (a), (b) and (c) in Fig. 2. It differs from existing works that belong to the category of (a) (Wang et al. 2018; Cheng et al. 2020; Jiao et al. 2018) or (d) (Huang et al. 2020; Li et al. 2019a; Mao et al. 2019).

We have conducted extensive evaluations to verify the superiority of our JBIM approach over related state-of-the-art re-id and image super-resolution methods on five person re-id benchmarks: the Context Aware Vision Using Image-based Active Recognition (CAVIAR) dataset (Cheng et al. 2011), the third Chinese University of Hong Kong (CUHK03) dataset (Li et al. 2014), the Sun Yat-sen University (SYSU) dataset (Chen et al. 2017a), the Viewpoint Invariant Pedestrian Recognition (VIPeR) dataset (Gray and Tao 2008) and the Market-1501 dataset (Zheng et al. 2015). An interesting and significant finding is that with the assistance of LIM, cascading super-resolution and person re-id directly suffices to achieve satisfactory performance without elaborate model design and parameter tuning, which reduces the burden of image super-resolution, enabling our method to effectively and flexibly integrate with different existing super-resolution models.

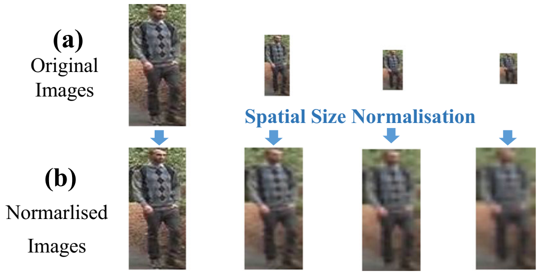

Fig. 3 (**a**) Images with different *underlying resolutions* and (**b**) these normalized to the same *spatial size* without changing the underlying resolution

2 Related Work

Person re-id has attracted extensive research over the past 10 years (Gray and Tao 2008; Zheng et al. 2013; Liao et al. 2015; Ahmed et al. 2015; Zheng et al. 2015; Xiao et al. 2016; Zheng et al. 2016; Zhang et al. 2016; Ristani et al. 2016; Li et al. 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. cross-resolution re-id has also become a research hotspot (Li et al. 2015; Jing et al. 2015; Wang et al. 2016, 2018; Mao et al. 2019; Chen et al. 2017c; Li et al. 2019a; Cheng et al. 2020; Huang et al. 2020). In the literature, existing cross-resolution re-id methods can be categorized into two groups: (1) methods for learning resolution-invariant features and (2) methods for learning joint models for both image super-resolution and person re-identification.

When learning resolution-invariant features, it is assumed in (Li et al. 2015) that images of the same person should be distributed similarly under different resolutions, and a method of simultaneously optimizing cross-resolution image alignment and distance metric modeling was designed. 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 are explored by varying the scale of LR images when matching with HR images. A common limitation of these methods is the inability to synthesize high-frequency and the loss of discriminative appearance information during image acquisition.

Recently, several convolution neural network (CNN) -based methods have been proposed for learning joint models. In (Wang et al. 2018), a cascaded multiple generative adversarial network (GAN) for image super-resolution (SRGAN) was proposed to recover the details of LR images progressively. In (Mao et al. 2019), the research focused on person foreground and learning different feature extractors for HR and LR images. In (Chen et al. 2017c), a model with GAN and autoencoder was used to learn cross-resolution deep image

Table 1 Comparing state-of-the-art cross-resolution re-id methods. "Using GAN" means that the method applies GAN in the joint model. "Learning HR-specific Discriminant Information" implies that the method uses HR-specific representations for person identity classification. In contrast, "Learning LR-specific Discriminant Information" signifies that the method uses LR-specific representations for person identity classification

Model	Using GAN	Learning HR-specific discriminant information	Learning LR-specific discriminant information
CSR-GAN (Wang et al. 2018)	✓	✓	×
RIPR (Mao et al. 2019)	×	✓	×
CAD-Net (Li et al. 2019a)	✓	✓	×
INTACT (Cheng et al. 2020)	✓	✓	×
DI-REID (Huang et al. 2020)	✓	✓	×
JBIM (Ours)	×	✓	✓

representations for re-id. In (Li et al. 2019a), the authors aim to learn resolution-invariant representations and meanwhile ensure recovering re-id-oriented HR details. In (Cheng et al. 2020), an association between the features for re-id and the features for resolution discrimination was introduced as joint learning regularization for cross-resolution person image matching. In (Huang et al. 2020), a degradation invariance learning framework was developed for extracting identity-related robust features.

In summary, most existing cross-resolution re-id methods adopt the strategy of learning invariant cross-resolution features, regardless of whether image super-resolution is integrated. This approach nonetheless ignores the usefulness of LR image specific information, *i.e.*, the low-resolution specific discriminant features are not explored.

As we found in experiments (see Table 3), this strategy is rather limited in learning discriminating identity features. The proposed learning method solves this limitation by introducing an LR-specific feature learning component. Consequently, the HR-specific and LR-specific identity features are jointly learned so that they can be made highly complementary to improve the cross-resolution re-id results more effectively than with other methods.

In our preliminary work (Jiao et al. 2018), for the first time, we introduced the idea of jointly learning image super-resolution and person re-id. Since then, our method has been frequently adopted in a number of followup works (Cheng et al. 2020; Chen et al. 2017c; Huang et al. 2020; Li et al. 2019a; Mao et al. 2019) that introduce various variations to continuously verify this idea and improve re-id performance. In this work, we revisit the same joint learning idea for HR-specific identity feature learning on top of our early model and make a couple of novel contributions: (1) By exploiting LR-specific identity feature, which is largely ignored by all previous studies, we find that taking use of low-resolution specific identity feature could be a more effective and significant way as compared to pursuing elaborate model design (*e.g.*, using GAN-based design); and (2) with the assistance of LIM, we empirically find that our joint learning method can be further improved significantly in terms of not only the generalizability but also the design flexibility and robustness. For instance, the choice of super-resolution models is less performance-sensitive, hence superior yet hard-to-train super-resolution methods can be bypassed, *e.g.*, GAN models (see Table 1).

In addition to works involving image super-resolution in person re-id, LR face recognition methods are also relevant and have been advanced in the literature (Wang and Tang 2005; Hennings-Yeomans et al. 2008; Huang and He 2011; Cheng et al. 2018). Their underlying idea is to synthesize HR faces by image super-resolution techniques without the need for dense feature point alignment. While feasible for structure-constrained face images, it is difficult to align person images due to greater degrees of unknown variations in body parts, *e.g.*, aligning a back-view LR person image with a side-view HR person image against other clutter. These super-resolution-based LR face-matching methods are therefore not suitable for cross-resolution re-id (Li et al. 2015). In the meantime, generic-purpose super-resolution methods have achieved remarkable success in synthesizing missing appearance fidelity from LR input images, mainly due to the powerful modeling capacity of deep learning algorithms (Dong et al. 2014, 2016b; Kim et al. 2016a, b; Lai et al. 2017; Tai et al. 2017; Ledig et al. 2017; Lim et al. 2017; Haris et al. 2018; Zhang et al. 2018; Li et al. 2019b). They may generate HR person images with higher visual quality but remain ineffective for cross-resolution re-id, as validated by our evaluations, because they are designed for improving low-level pixel values but not high-level identity discrimination when learning to reconstruct HR images.

3 Approach

We need to match an LR probe person image with a set of HR gallery images. To that end, we propose a bilateral-resolution learning approach (Fig. 4). We aim to not only acquire super-resolution person images discriminative for re-identification,

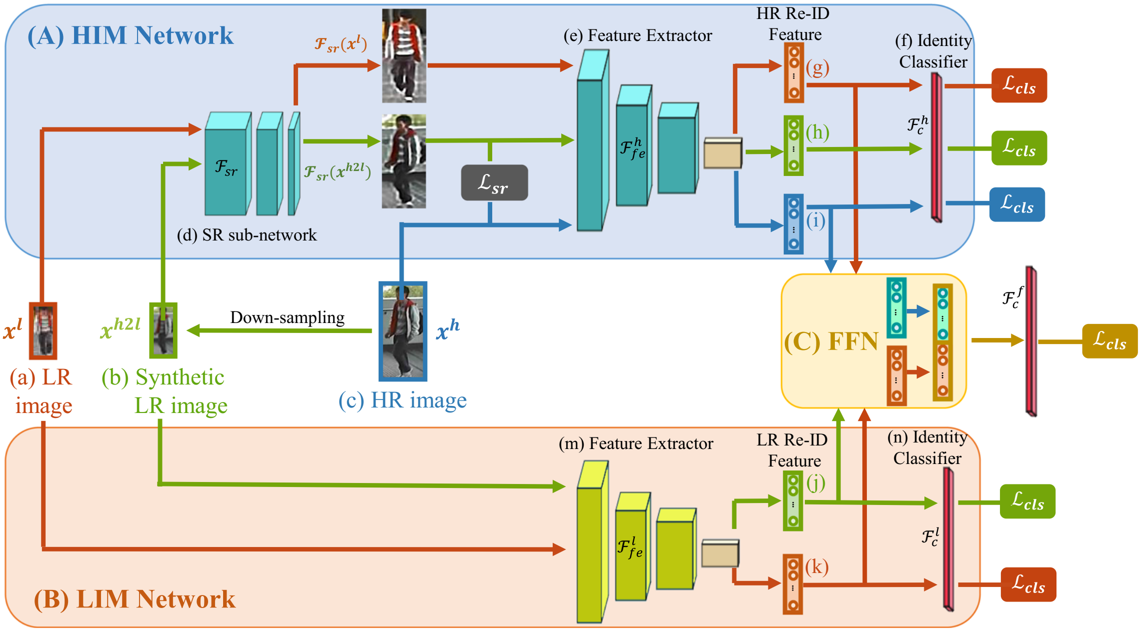

Fig. 4 An overview of the proposed joint bilateral resolution identity modeling (JBIM) framework. The JBIM framework consists of three components: (**A**) a high-resolution (HR) specific identity modeling (HIM) network, (**B**) a low-resolution (LR) specific identity modeling (LIM) network, and (**C**) a feature fusion network (FFN). Specifically, the HIM network contains two modules: (**d**) a super-resolution module and (**e and f**) a person re-identification (re-id) module. In training the HIM network, we deploy three streams taking (**a**) LR images, (**b**) synthetic LR images, and (**c**) HR images as input. The middle stream (**b**) acts as a bridge for joining (**d**) the image super-resolution and (**e and f**) person re-id learning tasks. Besides, the LIM network takes as input (**a**) the LR images and (**b**) synthetic LR images, with (**m and n**) person re-id as the learning objective. The FFN takes as input **(g) and (i)** HR re-id features and **(j) and (k)** LR re-id features, and outputs two feature vectors: a fusion of **(g) and (k)**, and a fusion of **(i) and (j)**. At the test time, we use both HIM and LIM. With HIM, we apply the super-resolution module to resolve the LR probe images and use the re-id module to extract features of both the SR probe images and HR gallery images. With LIM, we employ the re-id module to extract features from both the LR probe images and the *downsampled* gallery images. To obtain the final fused features, we separately concatenate the features from the LR probe images and those of the corresponding SR images and the features of the HR gallery images and those of the corresponding *downsampled* images. Finally, we utilize the L_2 distance of the fused features to perform cross-resolution re-id matching

but also to explore identity-sensitive LR specific features for a more effective solution which is not explored before. To maximize the model performance, we propose to jointly learn HR-specific and LR-specific identity features with optimal compatibility.

Suppose $X^l = \{(\boldsymbol{x}^l_i, y^l_i)\}^{N_l}_{i=1}$ is an LR person image set with N_l images from one camera view, $X^h = \{(\boldsymbol{x}^h_i, y^h_i)\}^{N_h}_{i=1}$ is an HR image set with N_h images from another view, where $\boldsymbol{x}^l_i$ and $\boldsymbol{x}^h_i$ denote LR and HR images captured with different camera views of identity classes y^l_i and y^h_i, respectively. To extract LR-specific identity features from HR images, we generate a *synthetic* LR set $X^{h2l} = \{(\boldsymbol{x}^{h2l}_i, y^h_i)\}^{N_h}_{i=1}$ of X^h by downsampling, where $\boldsymbol{x}^{h2l}_i$ is a *synthetic* LR image w.r.t. an HR image $\boldsymbol{x}^h_i$.

To learn the discriminative joint bilateral-resolution features, we want to obtain the following key components: (1) an image super-resolution function $\mathcal{F}_{sr}(\cdot)$ that can effectively compensate for re-id information in the LR images $\boldsymbol{x}^l_i$; (2) an HR-specific identity discriminant feature extraction function $\mathcal{F}^h_{fe}(\cdot)$ for both super-resolved LR images $\boldsymbol{x}^{sr}_i$ and realistic HR images $\boldsymbol{x}^h_j$, where $\boldsymbol{x}^{sr}_i = \mathcal{F}_{sr}(\boldsymbol{x}^l_i)$, with the objective that $\mathcal{F}^h_{fe}(\boldsymbol{x}^{sr}_i)$ is close to $\mathcal{F}^h_{fe}(\boldsymbol{x}^h_j)$ in the feature space when they share the identity label (*i.e.*, $y^l_i = y^h_j$), and vice versa; (3) an LR-specific identity discriminant feature extraction function $\mathcal{F}^l_{fe}(\cdot)$ that can extract the LR-specific identity features $\mathcal{F}^l_{fe}(\boldsymbol{x}^l_i)$ and $\mathcal{F}^l_{fe}(\boldsymbol{x}^{h2l}_j)$ from the realistic LR images $\boldsymbol{x}^l_i$ and the synthetic LR images $\boldsymbol{x}^{h2l}_j$, respectively, and when $y^l_i = y^h_j$, the corresponding LR-specific identity features should be similar, and vice versa; and (4) a feature fusion function $\mathcal{F}_{fus}(\cdot)$ that can fuse HR-specific and LR-specific identity features to obtain more discriminative joint bilateral resolution features. Similarly, we require the fusion features $\mathcal{F}_{fus}(\mathcal{F}^h_{fe}(\boldsymbol{x}^{sr}_i), \mathcal{F}^l_{fe}(\boldsymbol{x}^l_i))$ and $\mathcal{F}_{fus}(\mathcal{F}^h_{fe}(\boldsymbol{x}^h_j), \mathcal{F}^l_{fe}(\boldsymbol{x}^{h2l}_j))$ to be re-id discriminant.

Formally, by learning $\mathcal{F}_{sr}(\cdot)$, $\mathcal{F}^h_{fe}(\cdot)$, $\mathcal{F}^l_{fe}(\cdot)$ and $\mathcal{F}_{fus}(\cdot)$ through some joint formulation, we aim to obtain a re-id similarity matching metric:

$$\mathcal{S}\Big(\mathcal{F}_{fus}(\mathcal{F}^h_{fe}(\boldsymbol{x}^{sr}_i), \mathcal{F}^l_{fe}(\boldsymbol{x}^l_i)), \quad \mathcal{F}_{fus}(\mathcal{F}^h_{fe}(\boldsymbol{x}^h_j), \mathcal{F}^l_{fe}(\boldsymbol{x}^{h2l}_j))\Big), \tag{1}$$

subject to that after joint learning of image super-resolution, HR-specific and LR-specific identity feature extraction, and feature fusion, 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.

In the following, we first describe the HIM component for deriving discriminant HR-specific identity features by joint learning of super-resolution and identity classification. Then, we expound on the LIM component for learning discriminative LR-specific identity features and combine the HIM and LIM results in the Feature Fusion Networks (FFN) to acquire more discriminative re-id representations for joint bilateral resolution identity learning. An overview of the JBIM framework is shown in Fig. 4.

3.1 High-Resolution Specific Identity Modeling: Super-Resolution and Identity Joint Learning

Since generic-purpose super-resolution methods are designed to improve image visual fidelity rather than the re-id matching performance, a direct combination of independently trained image super-resolution and person re-id for HR-specific identity features might be suboptimal for re-id. Our HIM framework is hence formulated as joint learning of image super-resolution and person identity classification to correlate the two learning tasks as well as maximize their compatibility and complementary advantages.

- Super-Resolving Image. We first compensate for the desired discriminative information missing in the LR images through super-resolution. To facilitate super-resolution model training, we construct the image super-resolution loss with the help of the *synthetic* LR version X^{h2l} downsampled by X^h. Specifically, X^{h2l} allows optimizing the mean squared error (MSE), which measures the image super-resolution quality:

$$\mathcal{L}_{sr}\left(\{\boldsymbol{x}_i^h\}_{i=1}^{N_h}\right) = \frac{1}{N_h}\sum_{i=1}^{N_h} \|\mathcal{F}_{sr}(\boldsymbol{x}_i^{h2l}) - \boldsymbol{x}_i^h\|_F^2. \tag{2}$$

Minimizing loss $\mathcal{L}_{sr}$ enforces the super-resolved image $\mathcal{F}_{sr}(\boldsymbol{x}_i^{h2l})$ of $\boldsymbol{x}_i^{h2l}$ to be similar to the ground-truth HR image $\boldsymbol{x}_i^h$. HR appearance information is critical for obtaining reliable re-id features (Li et al. 2015). This optimization scheme (Eq. (2)) establishes the underlying relationship between the LR and HR images in the image pixel space, but without guaranteeing that the synthetic HR images are suitable for computing features discriminant for re-id matching. The reasons are as follows.

(1) It is very challenging to train a perfect image super-resolution model given that it is a highly nonconvex and difficult-to-optimize problem with extremely complex correlations among the local and global pixels (Dong et al. 2016a).

(2) Artifacts are likely generated particularly for low-quality surveillance person images, which may negatively affect the subsequent re-id matching results.

- Quantifying Identification. To address the above limitation, we propose enforcing an identity constraint to guide the super-resolution optimization behavior toward an image enhancement solution optimal for identity discrimination. This design differs from a typical super-resolution objective that seeks pixel-level mapping from LR input images to HR ground-truth images without a semantic top-down learning constraint.

Specifically, we concurrently optimize the classification of discriminative features w.r.t. the same person label in the HR and synthetic LR images, along with the cross-view LR images. Formally, we formulate the re-id classification constraint in the context of different images as:

$$\begin{aligned}\mathcal{L}_{reid}\left(\{(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{x}_i^h, y_i^h)\}_{i=1}^N\right) = \frac{1}{N}\sum_{i=1}^{N}\Big(&\mathcal{L}_{cls}\left(\mathcal{F}_c^h(\boldsymbol{f}_{\boldsymbol{h}i}^h), y_i^h\right)\\ &+ \mathcal{L}_{cls}\left(\mathcal{F}_c^h(\boldsymbol{f}_{\boldsymbol{h}i}^{h2l}), y_i^h\right)\\ &+ \mathcal{L}_{cls}\left(\mathcal{F}_c^h(\boldsymbol{f}_{\boldsymbol{h}i}^l), y_i^l\right)\Big),\end{aligned} \tag{3}$$

where $(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{x}_i^h, y_i^h)$ consists of an LR image from X^l and an HR image from X^h as well as their corresponding identity labels. We can construct N groups from X^l and X^h. All $\boldsymbol{f_h}$ notations denote the HR-specific re-id feature vectors obtained from the following feature extraction function:

$$\boldsymbol{f}_{\boldsymbol{h}i}^h = \mathcal{F}_{fe}^h(\boldsymbol{x}_i^h),\ \boldsymbol{f}_{\boldsymbol{h}i}^{h2l} = \mathcal{F}_{fe}^h(\mathcal{F}_{sr}(\boldsymbol{x}_i^{h2l})),\ \boldsymbol{f}_{\boldsymbol{h}i}^l = \mathcal{F}_{fe}^h(\mathcal{F}_{sr}(\boldsymbol{x}_i^l)). \tag{4}$$

And $\mathcal{F}_c^h(\cdot)$ represents an HR-specific classification function. $\mathcal{L}_{cls}(\cdot)$ is the integration of the identity loss $\mathcal{L}_{id}(\cdot)$ and the triplet loss $\mathcal{L}_{tri}(\cdot)$, which is defined as:

$$\mathcal{L}_{cls} = \mathcal{L}_{id} + \mathcal{L}_{tri}. \tag{5}$$

- Simultaneous Super-Resolution Learning and Re-id. After combining the super-resolution and re-id formulation designs as above, we formulate the overall HR-specific re-id loss function as:

$$\begin{aligned}&\mathcal{L}_{HIM}\left(\{(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{x}_i^h, y_i^h)\}_{i=1}^N\right)\\ &\quad= \mathcal{L}_{reid}\left(\{(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{x}_i^h, y_i^h)\}_{i=1}^N\right) + \alpha\mathcal{L}_{sr}\left(\{\boldsymbol{x}_i^h\}_{i=1}^N\right),\end{aligned} \tag{6}$$

where the parameter α controls the balance between the image super-resolution loss and the re-id loss. Optimizing the joint loss $\mathcal{L}_{HIM}$ allows guiding the $\mathcal{F}_{sr}(\cdot)$ to compensate the semantic appearance details of the LR images toward identity-salient fidelity synthesis and concurrently drives $\mathcal{F}_{fe}^h(\cdot)$ to accordingly extract identity discriminative features in a harmonious manner. Such a multitask joint learning formulation is supposed to mitigate the resolution mismatch problem in cross-resolution person re-id.

A key characteristic of the HIM formulation (Eq. (6)) is the *seamless joining* of a restoration quantization super-resolution loss (Eq. (2)) and a person re-id loss (Eq. (3)), *both* subject to the same synthetic LR training images $\boldsymbol{x}_i^{h2l}$ (Fig. 4b) in the context of concurrent identity discriminant supervision on all three types of training images. That is, the synthetic LR images $\boldsymbol{x}_i^{h2l}$ and its re-id features $\boldsymbol{f}_i^{h2l}$ together bridge and correlate the image super-resolution (Fig. 4d) and person re-id (Fig. 4 e and f) learning tasks. Without this connection, the two loss functions $\mathcal{L}_{sr}$ and $\mathcal{L}_{reid}$ will be optimized independently, rather than jointly and concurrently.

3.2 Learning Bilateral-Resolution Identity Features with Low-Resolution Identity Modeling

Although the information of LR images may be incomplete, a fraction of it may be absent in the HR-specific representations but may be useful for re-id, *i.e.*, LR-specific re-id information. Under this consideration, we propose further learning discriminative features in the LR image space, in a similar manner as HR-specific re-id formulation above. We ground this learning component on every pair of a synthetic LR image (*i.e.*, $\boldsymbol{x}_i^{h2l}$ downsampled from HR image $\boldsymbol{x}_i^h$) and a cross-view realistic LR image (*i.e.*, $\boldsymbol{x}_i^l$).

Formally, we introduce an LR re-id loss function as:

$$\mathcal{L}_{LIM}\Big(\{(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{x}_i^h, y_i^h)\}_{i=1}^N\Big) = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathcal{L}_{cls}\big(\mathcal{F}_c^l(\boldsymbol{f}_{li}^l), y_i^l\big) + \mathcal{L}_{cls}\big(\mathcal{F}_c^l(\boldsymbol{f}_{li}^{h2l}), y_i^h\big)\Big), \quad (7)$$

where $\mathcal{F}_c^l(\cdot)$ represents the LR-specific identity classification. Both $\boldsymbol{f_l}$ notations denote LR-specific re-id feature vectors by the LR-specific feature extraction function as:

$$\boldsymbol{f}_{li}^l = \mathcal{F}_{fe}^l\big(\boldsymbol{x}_i^l\big), \quad \boldsymbol{f}_{li}^{h2l} = \mathcal{F}_{fe}^l\big(\boldsymbol{x}_i^{h2l}\big). \quad (8)$$

By minimizing the classification loss $\mathcal{L}_{LIM}$, we encourage $\mathcal{F}_{fe}^l(\cdot)$ to learn discriminative LR-specific identity features from realistic and synthetic LR images.

- JBIM: Joint Bilateral-Resolution Identity Modeling. Collaboratively fusing HR-specific identity learning and LR-specific identity learning leads to the proposed JBIM framework. This framework aims to derive more discriminate feature representations from the LRs and HRs together.

Specifically, with the image super-resolution function $\mathcal{F}_{sr}(\cdot)$ and the HR-specific feature extraction function $\mathcal{F}_{fe}^h(\cdot)$, we can obtain the HR-specific identity feature $\mathcal{F}_{fe}^h(\mathcal{F}_{sr}(\boldsymbol{x}_i^l))$ of an LR image $\boldsymbol{x}_i^l$ and the HR-specific identity feature $\mathcal{F}_{fe}^h(\boldsymbol{x}_i^h)$ of an HR image $\boldsymbol{x}_i^h$. Moreover, the LR-specific identity feature $\mathcal{F}_{fe}^l(\boldsymbol{x}_i^l)$ and $\mathcal{F}_{fe}^l(\boldsymbol{x}_i^{h2l})$ can be extracted by the discriminative LR-specific feature extraction function $\mathcal{F}_{fe}^l(\cdot)$. We conjecture that HR representations and LR representations capture different characteristics for re-id, *i.e.*, they are complementary. To maximize this complementary effect, we optimize the fusing of the features by minimizing the following re-id classification loss as:

$$\mathcal{L}_{fus}\Big(\{(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{x}_i^h, y_i^h)\}_{i=1}^N\Big) = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathcal{L}_{cls}\big(\mathcal{F}_c^f(\boldsymbol{f}_{fi}^l), y_i^l\big) + \mathcal{L}_{cls}\big(\mathcal{F}_c^f(\boldsymbol{f}_{fi}^h), y_i^h\big)\Big), \quad (9)$$

where $\mathcal{F}_c^f(\cdot)$ represents the classification function of the fused feature vectors $\boldsymbol{f_f}$, obtained by the feature fusion function as:

$$\begin{aligned}\boldsymbol{f}_{fi}^l &= \mathcal{F}_{fus}(\mathcal{F}_{fe}^h(\mathcal{F}_{sr}(\boldsymbol{x}_i^l)),\ \mathcal{F}_{fe}^l(\boldsymbol{x}_i^l)),\\ \boldsymbol{f}_{fi}^h &= \mathcal{F}_{fus}(\mathcal{F}_{fe}^h(\boldsymbol{x}_i^h),\ \mathcal{F}_{fe}^l(\boldsymbol{x}_i^{h2l})).\end{aligned} \quad (10)$$

Finally, we derive the overall objective function as:

$$\mathcal{L}_{overall} = \mathcal{L}_{HIM} + \mathcal{L}_{LIM} + \beta\mathcal{L}_{fus}, \quad (11)$$

where the parameter β is a balance parameter. We set an identical weight to the HR re-id loss ($\mathcal{L}_{HIM}$) and the LR re-id loss ($\mathcal{L}_{LIM}$).

Optimizing the overall loss $\mathcal{L}_{overall}$ allows: (1) guiding the super-resolution $\mathcal{F}_{sr}(\cdot)$ to recover the appearance information lost in low-resolution identity features, which is beneficial for identity classification; (2) forcing the high-resolution feature extraction function $\mathcal{F}_{fe}^h(\cdot)$ to extract discriminative high-resolution identity features; (3) compelling the low-resolution feature extraction function $\mathcal{F}_{fe}^l(\cdot)$ to extract the discriminative low-resolution identity features. Besides, reducing the loss $\mathcal{L}_{fus}$ makes the high-resolution identity mapping and low-resolution identity mapping jointly learned for obtaining resolution-specific discriminant features, which could be complementary. For example, LIM and HIM attend to different regions of a person image, making their features complementary, as shown in Fig. 5.

4 Model Instantiation

We consider a deep CNN for model instantiation due to its strong merits such as (1) its ability learning discriminative representations from training data with great success on both image super-resolution (Dong et al. 2016a; Wang et al. 2015) and person re-id (Li et al. 2014; Xiao et al. 2016); (2) its strong capability of learning highly non-convex tasks and

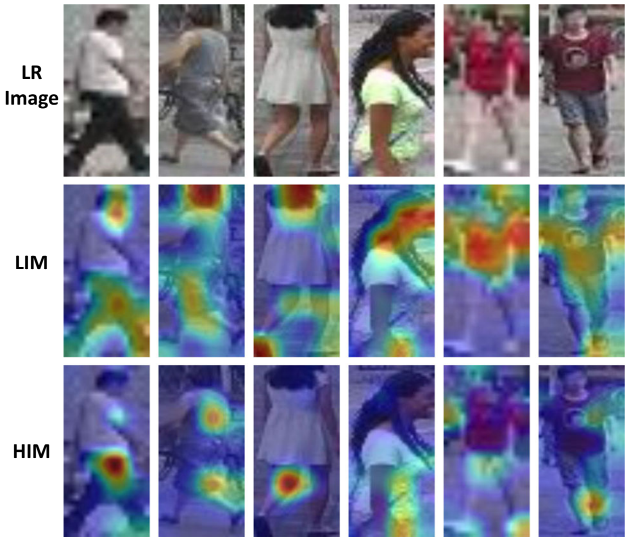

Fig. 5 A large difference in the class activation maps derived by low-resolution specific identity mapping (LIM, second row) and high-resolution specific identity mapping (HIM, third row) suggests that the learning module could extract different complementary appearance feature information from the LR and HR person images

its suitability for learning complex appearance variations in lighting, occlusion and background clutter; and (3) its high flexibility of reformulating the network architecture without the need for redesigning the optimization algorithm.

In particular, we design a hybrid deep neural network to realize the nonlinear functions involved, including the super-resolution function $\mathcal{F}_{sr}(\cdot)$, the HR-specific feature extraction function $\mathcal{F}^{h}_{fe}(\cdot)$, the LR-specific feature extraction function $\mathcal{F}^{l}_{fe}(\cdot)$ and the feature fusion function $\mathcal{F}_{fus}(\cdot)$. The entire framework is made up of three parts: the HIM network (Fig. 4A) to compute the discriminative HR-specific identity features, the LIM network (Fig. 4B) to obtain the discriminative LR-specific identity features, and the FFN module (Fig. 4C) to acquire the final representations fused by the HR-specific and LR-specific identity features.

4.1 Architecture for High-Resolution Identity Modeling

The network architecture for HIM is depicted in Fig. 4A. Specifically, it consists of the two following modules.

-The Super-Resolution Module aims to compensate for and recover the information lost in the LR image acquisition, *i.e.*, realizing $\mathcal{F}_{sr}(\cdot)$. It has two parameter-sharing streams, taking $\boldsymbol{x}^{l}_{i}$ (an LR image) and $\boldsymbol{x}^{h2l}_{i}$ (a cross-view synthetic LR image) as input . Following the super-resolution CNN (SRCNN) in (Dong et al. 2014), our super-resolution module is constructed by two convolutional layers followed by a non-linear rectified linear unit (ReLU) layer and a reconstruction convolutional layer. The MSE loss function (Eq. 2) is used for quantifying the pixel-level alignment degree between the ground-truth HR $\boldsymbol{x}^{h}_{i}$ and the super-resolution output of $\boldsymbol{x}^{h2l}_{i}$ during training. The super-resolution outputs of $\boldsymbol{x}^{l}_{i}$ are not involved in calculating the reconstruction loss.

- The HR Re-ID Module aims to learn HR-specific identity discriminant features, *i.e.*, realizing $\mathcal{F}^{h}_{fe}(\cdot)$, and imposing HR-specific re-id constraints, *i.e.*, realizing $\mathcal{F}^{h}_{c}(\cdot)$. It has HR parameter-sharing streams taking as input the super-resolution output of the realistic LR image $\boldsymbol{x}^{l}_{i}$ and the super-resolution output of the synthetic LR image $\boldsymbol{x}^{h2l}_{i}$ as well as the HR image $\boldsymbol{x}^{h}_{i}$. In our implementation, we adopt a 50-layer residual neural network (ResNet-50) (He et al. 2016a), which has achieved effective results in classification and detection tasks. 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 streams' identity losses (Eq. 3) is used as the supervision signal for jointly qualifying the identification of all inputs during model training. In the implementation, we upscale the LR images to an appropriate size (256×128 in our experiments) by bicubic interpolation as (Dong et al. 2016a).

We achieve joint learning of image super-resolution and person re-id in the proposed CNN by using multipurposed synthetic LR image $\boldsymbol{x}^{h2l}_{i}$ (Fig. 4b), *i.e.*, $\boldsymbol{x}^{h2l}_{i}$ is used for both training super-resolution module and person re-id module. Formally, $\boldsymbol{x}^{h2l}_{i}$ and its super-resolution re-id feature vector $\boldsymbol{f}^{h2l}_{\boldsymbol{hi}}$ ground four loss quantities: one super-resolution loss on $(\boldsymbol{x}^{h2l}_{i}, \boldsymbol{x}^{h}_{i})$ correlated with three re-id losses on three computed features $\boldsymbol{f}^{h2l}_{\boldsymbol{hi}}$, $\boldsymbol{f}^{l}_{\boldsymbol{hi}}$ and $\boldsymbol{f}^{h}_{\boldsymbol{hi}}$. This loss connection design injects more re-id discrimination awareness into a jointly optimized image super-resolution model. We will evaluate the effect of our model design in our experiments.

4.2 Architecture for Low-Resolution Identity Modeling

With LIM, we realize the learning of the LR-specific feature function, *i.e.*, realizing $\mathcal{F}^{l}_{fe}(\cdot)$ and imposing LR-specific re-id constraints $\mathcal{F}^{l}_{c}(\cdot)$, by a deep CNN as well (Fig. 4m and n). Two parameter-sharing streams are involved, taking a realistic LR image $\boldsymbol{x}^{l}_{i}$ and a synthetic LR image $\boldsymbol{x}^{h2l}_{i}$ as input. In the same manner as the HR-specific re-id network, we adopt ResNet-50 (He et al. 2016a). In each stream, the penultimate FC layer outputs the re-id feature, which is then fed into the last FC layer for LR-specific identity classification. The two streams' identity losses (Eq. 7) are summed up and used as supervision signals for jointly qualifying the identification of both the realistic and synthetic LR inputs during training, *i.e.*, two LR re-id losses on LR-specific identity features $\boldsymbol{f}^{l}_{\boldsymbol{li}}$ and $\boldsymbol{f}^{h2l}_{\boldsymbol{li}}$. The size of the LR images is fixed as 256×128 to obtain the discriminative LR-specific identity features. Note that the images still have the same LRs, although the spatial size is enlarged.

4.3 High- and Low-Resolution Collaborative Learning

We design a FFN to jointly learn the HR-specific and LR-specific identity features simultaneously.It collaboratively fuses the discriminative HR-specific identity features (Sect. 4.1) and the LR-specific identity features (Sect. 4.2). The FFN module consists of a concatenation layer, a fully-connected layer, a batch normalization layer, a ReLU layer and a dropout layer.

Specifically, taking the HR-specific feature $\boldsymbol{f}^{l}_{\boldsymbol{h}i}$ and the LR-specific feature $\boldsymbol{f}^{l}_{\boldsymbol{l}i}$ as input, the FFN outputs a fused feature $\boldsymbol{f}^{l}_{\boldsymbol{f}i}$ for an LR image $\boldsymbol{x}^{l}_{i}$; Moreover, taking the HR-specific feature $\boldsymbol{f}^{h}_{\boldsymbol{h}i}$ and the LR-specific feature $\boldsymbol{f}^{h2l}_{\boldsymbol{l}i}$ as input, the FFN concurrently outputs another fused feature $\boldsymbol{f}^{h}_{\boldsymbol{f}i}$ for a cross-view HR image $\boldsymbol{x}^{h}_{i}$. Subsequently, the two fused features (*i.e.*, $\boldsymbol{f}^{l}_{\boldsymbol{f}i}$ and $\boldsymbol{f}^{h}_{\boldsymbol{f}i}$) are further fed into an FC layer for separate identity classification. Finally, the two identity losses (Eq. 9) are summed up to jointly supervise the identification of both the LR and HR inputs during training.

For feature fusion, we consider the following operation:

$$\boldsymbol{f_f} = \mathcal{F}_{fus}(\boldsymbol{f_h}, \boldsymbol{f_l}) \in \mathbb{R}^{n_h+n_l} \tag{12}$$

$$\boldsymbol{f_h} = [f_{h,1}, f_{h,2}, ..., f_{h,j}, ..., f_{h,n_h}] \in \mathbb{R}^{n_h} \tag{13}$$

$$\boldsymbol{f}_l = [f_{l,1}, f_{l,2}, ..., f_{l,k}, ..., f_{h,n_l}] \in \mathbb{R}^{n_l} \tag{14}$$

where $\boldsymbol{f_h}$ is an HR feature vector (*e.g.*, $\boldsymbol{f_h}^{h}$, $\boldsymbol{f_h}^{l}$) and $\boldsymbol{f}_l$ is an LR feature vector (*e.g.*, $\boldsymbol{f_l}^{h}$, $\boldsymbol{f_l}^{l}$). The scalars n_h and n_l are the dimensions of $\boldsymbol{f_h}$ and $\boldsymbol{f_l}$, respectively. $\boldsymbol{f_f}$ denotes the final joint bilateral-resolution features generated by $\boldsymbol{f_h}$ and $\boldsymbol{f_l}$. The back propagation operation of the concatenation layer is as follows:

$$\frac{\partial f_{concat,i}}{\partial f_{h,j}} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{15}$$

$$\frac{\partial f_{concat,i}}{\partial f_{l,k}} = \begin{cases} 1 & \text{if } i = k + n_h \\ 0 & \text{if } i \neq k + n_h \end{cases} \tag{16}$$

where $\boldsymbol{f_{concat}} = [\boldsymbol{f_h}, \boldsymbol{f_l}] \in \mathbb{R}^{n_h+n_l}$. The symbol [,] means the concatenation of the two vectors. $f_{concat,i} \in \boldsymbol{f_{concat}} (i = 1, 2, ..., n_h + n_l)$, $f_{h,j} \in \boldsymbol{f_h} (j = 1, 2, ..., n_h)$, $f_{l,k} \in \boldsymbol{f_l} (k = 1, 2, ..., n_l)$. Since the concatenation operation will only keep the gradient of each neuron, the partial derivative is equal to 1 at corresponding position, otherwise it is equal to 0. These two equations indicate that the gradients of the HR feature vector $\boldsymbol{f_h}$ and the LR feature vector $\boldsymbol{f_l}$ do not influence each other before they are input into the FFN module.

4.4 Model Testing

In the model test phase, we extract joint bilateral-resolution features for both LR probe and HR gallery images. The generic L_2 distance metric is then used for re-id matching.

Specifically, for the HR gallery images, we directly use the jointly learned HR re-id subnet in the HIM network to obtain the HR-specific identity features and compute the LR-specific identity features of the corresponding downsampled LR images by LIM. For the LR probe images, we directly apply the LIM network to compute the LR-specific identity features. To compute the HR-specific identity features, we apply the image super-resolution network in HIM to super-resolve them before performing feature extraction by the re-id subnet in the HIM network. The final joint bilateral resolution features are generated by the FFN module in which the HR-specific identity features and the corresponding LR-specific identity features are concatenated.

5 Multiresolution Adaptive Ensemble

The JBIM framework formulated as above assumes that all the LR images have similar underlying resolutions because the super-resolution network in the HIM network is optimized for super-resolving the LR images by a ratio m, which renders a single JBIM model suboptimal when the resolution ratio is far away from m as typically encountered in practice, where multiple different LRs are present [2].

To address this problem, we propose creating φ anchor JBIM models $\{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_\varphi\}$, where each is responsible for optimizing a reference super-resolution ratio in the set $\{m_1, m_2, \cdots, m_\varphi\}$. These JBIM models are then used jointly to accommodate various resolutions involved in cross-resolution re-id matching. Each model $\boldsymbol{M}_i$ can be similarly learned as described above by the corresponding synthetic LR images X^{h2l} generated by downsampling HR images with a ratio m_i, along with realitic LR and HR training images. In our experiments, we used three models corresponding to the downsampling ratio $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$.

In the test, given an LR probe image, we first compute φ distance vectors $\{D_i\}_{i=1}^{\varphi}$ between the probe image and all the gallery images with each anchor JBIM model, where D_i denotes the distance computed by model $\boldsymbol{M}_i$, $i \in \{1, 2, \cdots, \varphi\}$. Then, we compute a multiresolution fused

[2] While HR images also have different resolutions, we focus on handling the LR images in this work because LR images suffer more significant information loss than HR images during data acquisition and are therefore the major cause of degraded re-id matching performance. We assume that 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.

distance vector as:

$$D_{mra} = \sum_{i=1}^{\varphi} w_i D_i, \tag{17}$$

where $\{w_i\}_{i=1}^{\varphi}$ represents the distance weights.

To make D_{mra} resolution adaptive, we consider the similarity in the underlying resolution among the LR probe image, all the HR gallery images, and each JBIM model. We quantify the resolution similarity between the LR probe and HR gallery images as:

$$r = \sqrt{\frac{A_p}{\tilde{A_g}}}, \tag{18}$$

where A_p denotes the spatial area (*i.e.*, the number of pixels) of the LR probe and $\tilde{A_g}$ is the mean spatial area of all the HR gallery images. They are computed on genuine resolution scales without resizing. We then take into account the super-resolving ratio m_i as:

$$w_i = \exp\{-\sigma^{-2} \cdot (r - m_i)^2\}, \tag{19}$$

where σ is a scaling parameter estimated by cross-validation.

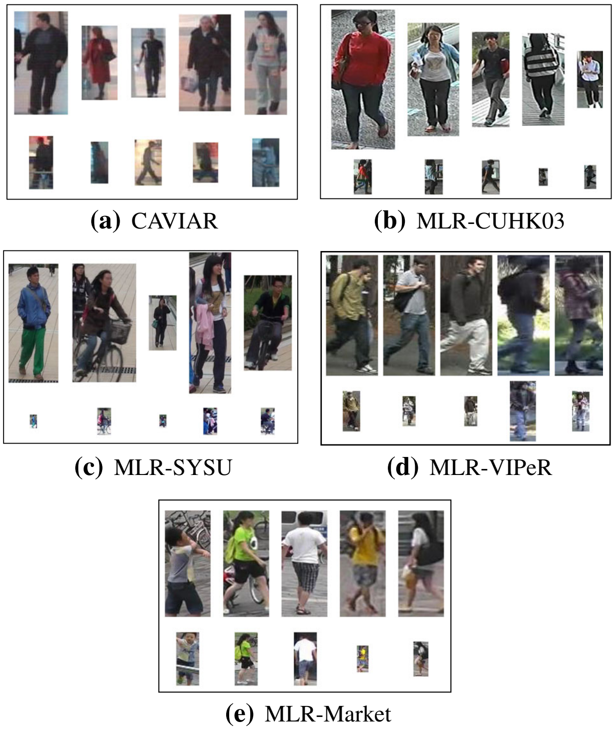
(a) CAVIAR **(b)** MLR-CUHK03 **(c)** MLR-SYSU **(d)** MLR-VIPeR **(e)** MLR-Market

Fig. 6 Examples of HR (1st row) and LR (2nd row) person images from five datasets

6 Experiments

6.1 Datasets

We performed evaluations on one *genuine* and four *simulated* cross-resolution person re-id datasets (Fig. 6). Instead of assuming a single underlying resolution for all the LR images, we consider multiple LRs (*MLR*) as in real-world situations. Therefore, we used different downsampling rates when simulating the LR images from the HR images.

(1) CAVIAR is an cross-resolution person re-id dataset (Cheng et al. 2011). It contains 1220 images of 72 persons captured from two camera views in a shopping mall. Albeit of small scale, this dataset is particularly suitable for evaluating cross-resolution re-id because the resolution of images from one camera (the distant camera) is much lower than that from the other camera (the close camera). We discarded 22 persons who appeared only in the close camera (HR images). For each of the remaining 50 persons used in our experiments, there are 10 HR and 10 LR images, *i.e.*, a total of 1,000 images. Unlike other simulated datasets, the LR images in CAVIAR involve multiple *realistic* resolutions.

(2) MLR-CUHK03 was built from the CUHK03 (Li et al. 2014) dataset. The CUHK03 dataset consists of five different pairs of camera views and contains more than 14,000 images of 1,467 pedestrians. By following the settings outlined 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 the LR probe image source by performing downsampling by a ratio randomly selected from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$. This procedure results in a simulated multiple LRs (MLRs) re-id dataset MLR-CUHK03.

(3) MLR-SYSU is based on the SYSU dataset (Chen et al. 2017a), which has 24,446 images of a total of 502 persons captured by two cameras. We randomly selected three images per person per camera in our evaluations and created an MLR re-id dataset MLR-SYSU as for CUHK03.

(4) MLR-VIPeR was constructed from the VIPeR (Gray and Tao 2008) dataset, which contains 632 person image pairs captured by two cameras. Each image is of HR (128×48 pixels). To make this dataset suitable for cross-resolution person re-id evaluation, we performed similar multiresolution downsampling on all the images from one camera view, while the remaining images from the other view remained the same. This procedure resulted in the MLR-VIPeR dataset.

(5) MLR-Market was constructed from the Market-1501 (Zheng et al. 2015) dataset, which contains 1501 people captured by 6 cameras with varying viewpoints and lighting conditions. Each person is captured by at least two cameras, and each camera may obtain more than 10 pictures. To make this dataset suitable for cross-resolution person re-id evalu-

ation, we downsampled half of the images of a person by a ratio randomly selected from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$, while the remaining images of the other half remained the same.

6.2 Settings and Implementation

- Evaluation Protocol. We adopted the standard single-shot re-id setting in our experiments. The CAVIAR, MLR-VIPeR and MLR-SYSU datasets were randomly divided into two halves: one-half for training and the other half for testing. That is, there are $p = 25$, $p = 316$ and $p = 251$ persons in the testing sets of CAVIAR, MLR-VIPeR and MLR-SYSU, respectively. Following (Xiao et al. 2016), we utilized the benchmarking 1,367/100 training/test identity split for the MLR-CUHK03 dataset. In addition, we applied the 751/750 training/test identity split setting on the MLR-Market dataset following (Zheng et al. 2015). For the testing data, we constructed the probe set with all LR images per person and the gallery set with one randomly selected HR image per person. For performance evaluation, we used the average cumulative match characteristic (CMC) and mean average precision (mAP) to measure the cross-resolution re-id matching performance.

- Implementation Details. We first trained the HIM and LIM networks on each target cross-resolution re-id dataset separately and then trained the whole JBIM model with the HIM and LIM networks initialized with the independently trained parameters and the randomly initialized FFN. The backbone of our network was trained with random horizontal flipping, random cropping and batch normalization neck (BNNeck) (Luo et al. 2019). For the HIM network, we initialized the super-resolution network by the SRCNN (Dong et al. 2014) with a padding operation for all the convolution layers and the ResNet-50-based re-id network (He et al. 2016a) with ImageNet weights. We find that using other super-resolution networks yields little performance improvement. We initialized the ResNet-50-based LIM network with ImageNet weights. The σ (Eq. (19)) was set to 0.5. We set the balancing coefficient $\alpha = 1$ (Eq. (6)) and $\beta = 3$ (Eq. (11)). The parameters α and β will be discussed in the experiments.

6.3 Comparing Existing Low-Resolution Re-Identification Models

We compared the proposed JBIM method with eight existing state-of-the-art cross-resolution re-id methods, including three traditional methods and five deep CNN-based methods: (1) joint multiscale discriminant component analysis (JUDEA) (Li et al. 2015): a cross-scale discriminative distance metric learning model; (2) semi-coupled low-rank discriminant dictionary learning (SLD^{2}L) (Jing et al. 2015): a feature transformation or alignment model; (3) scale-distance function (SDF) (Wang et al. 2016): a scale-distance function learning model; (4) cascaded SR-GAN (CSR-GAN) (Wang et al. 2018): a joint learning of the person re-id and multiple cascaded SR-GANs; (5) resolution-invariant person re-identification (RIPR) (Mao et al. 2019): a network jointly training a foreground focus super-resolution module and a reso-lution-invariant feature extractor by end-to-end CNN learning; (6) cross-resolution adversarial dual network (CAD-Net) (Li et al. 2019a): a generative dual model for cross-resolution person re-id. (7) inter-task association critic (INTACT) (Cheng et al. 2020): a model leveraging the association between image SR and person re-id tasks; (8) degradation-invariant re-id (DI-REID) (Huang et al. 2020): a degradation-invariant learning framework.

Table 2 Comparison of the complexity of state-of-the-art cross-resolution re-id methods that are based on ResNet-50 as the backbone. We calculate the complexity of these models according to the description in their papers. “Params” is the number of parameters of model, and “MACs” is the number of fixed-point multiplication and accumulation operations per second

Model	Params (M)	MACs (G)
CSR-GAN (Wang et al. 2018)	187.50	48.68
CAD-Net (Li et al. 2019a)	115.83	48.52
DI-REID (Huang et al. 2020)	472.86	734.63
JBIM	52.96	13.96

It is evident from Table 3 that our JBIM method outperforms all the competitors in most cases. For example, the JBIM method surpassed the best alternative traditional method JUDEA by 30.0%, 62.1%, 40.8%, and 23.7% in terms of the rank-1 matching rate on the CAVIAR, MLR-CUHK03, MLR-SYSU, and MLR-VIPeR datasets, respectively. The performance margins of the JBIM method over the SLD^{2}L and SDF models are still larger.

Moreover, our JBIM method surpasses the four deep CNN-based cross-resolution re-id methods on each dataset except the DI-REID model on the VIPeR dataset. Specifically, the strongest competitors are namely DI-REID (Huang et al. 2020) and INTACT (Cheng et al. 2020), particularly on the MLR-VIPeR and MLR-Market dataset. Whilst our model’s result is on par with DI-REID, the complexity of our method is largely smaller (9 times smaller), as reported in Table 2, meaning that our model is more cost-effective and more computationally scalable. In comparison with INTACT, we found that it uses OSNet (Zhou et al. 2019) as the backbone which is stronger than ResNet-50 (He et al. 2016a) as our model used for re-id. We thus tested our model with OSNet. Table 3 shows that our method can outperform INTACT on all datasets with a clear margin using the same backbone model.

Table 3 Comparison of the performance of the state-of-the-art re-id methods(%). The 1*st*/2*nd* best results are indicated in bold/italic

Methods	CAVIAR				MLR-CUHK03				MLR-SYSU				MLR-VIPeR				MLR-Market			
	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
JUDEA (Li et al. 2015)	22.0	60.1	80.8	–	26.2	58.0	73.4	–	18.3	41.9	54.5	–	26.0	55.1	69.2	–	–	–	–	–
SLD^{2}L (Jing et al. 2015)	18.4	44.8	61.2	–	–	–	–	–	20.3	34.8	43.4	–	20.3	44.0	62.0	–	–	–	–	–
SDF (Wang et al. 2016)	14.3	37.5	62.5	–	22.2	48.0	64.0	–	13.3	26.7	42.9	–	9.25	38.1	52.4	–	–	–	–	–
CSR-GAN (Wang et al. 2018)	34.7	72.5	87.4	36.4*	71.3	92.1	97.4	72.1*	–	–	–	-	37.2	62.3	71.6	50.1*	76.4	88.5	91.9	58.3*
RIPR (Mao et al. 2019)	36.4	72.0	–	38.5*	73.3	92.1	–	73.7*	–	–	–	–	41.6	64.9	–	50.9*	75.4*	88.7*	91.0*	59.5*
CAD-Net (Li et al. 2019a)	42.8	76.2	91.5	–	82.1	97.4	98.8	–	–	–	–	–	43.1	68.2	77.5	–	83.7	92.7	95.8	–
INTACT (Cheng et al. 2020)	44.0	81.8	93.9	–	86.4	*97.4*	98.5	–	–	–	–	–	46.2	73.1	81.6	–	88.1	95.0	96.9	–
DI-REID (Huang et al. 2020)	51.2	*83.6*	94.4	–	85.7	97.1	98.6	–	–	–	–	–	*50.3*	*77.9*	**87.3**	–	–	–	–	–
OSNet (Zhou et al. 2019)	26.6	62.5	81.4	29.2	83.8	94.9	96.8	85.4	50.2	75.4	82.5	55.6	33.2	59.2	72.5	45.8	84.3	93.6	95.6	67.9
ABD-Net (Chen et al. 2019a)	31.4	69.9	87.4	35.2	84.3	95.6	97.9	85.9	57.3	80.0	88.2	62.0	44.3	70.3	83.5	56.4	74.0	87.4	91.2	60.6
AGW (Ye et al. 2021)	25.8	64.9	83.3	29.7	87.3	96.4	98.2	88.4	46.0	70.1	79.0	51.8	30.1	48.4	61.1	39.7	85.0	92.9	95.6	68.2
JBIM	*52.0*	83.1	*94.4*	*50.1*	*88.3*	*97.2*	*98.7*	*89.9*	*59.1*	*82.3*	88.9	*65.7*	49.7	72.5	81.3	*58.0*	*88.1*	*95.1*	*96.9*	*73.5*
JBIM(OSNet)	**53.1**	**84.0**	**95.2**	**51.2**	**88.7**	**97.5**	**99.0**	**90.3**	**61.7**	**85.2**	**90.7**	**67.3**	**52.7**	**78.6**	*87.1*	**63.1**	**89.6**	**96.2**	**97.7**	**74.7**

* These results are produced by ourselves using codes provided by the authors

The above results indicate the advantages of the proposed JBIM model in handling both simulated and genuine cross-resolution re-id. The performance superiority is mainly due to **(1)** the capability of jointly super-resolving images and learning discriminative person re-id features, which allows us to maximize their mutual correlation. Compared to the cross-resolution alignment-based competitor, our model is able to synthesize high-frequency missing LR images by re-id discriminative super-resolution and thus extracts richer representations, which not only directly mitigates the information amount discrepancy problem but also fills the hard-to-bridge matching gap between different resolutions with the appearance pattern divergence involved. **(2)** By jointly learning the bilateral-resolution identity feature, the presented multiresolution (HR and LR) features better characterize the salience of a person in cross-resolution re-id than all the traditional and deep CNN-based methods using only single-resolution features.

6.4 Comparing Existing State-of-the-Art Re-Identification Methods

To validate the necessity of a specially designed framework for the cross-resolution re-id problem, we have implemented three SOTA re-id methods (i.e. OSNet (Zhou et al. 2019), ABD-Net (Chen et al. 2019a), and AGW (Ye et al. 2021)) in our problem setting, which are not specially designed for cross-resolution re-id. The same training and testing settings are used as our method. As shown in Table 3, our method surpasses all these alternative re-id methods for cross-resolution person re-id. It demonstrates that the resolution mismatch problem need targeted solutions.

6.5 Comparing the Super-resolution + Re-Identification Scheme

We further evaluated the cross-resolution person re-id performance by deploying a straightforward combination of the super-resolution and person re-id scheme. While conventional re-id methods assume using HR images, we utilize state-of-the-art super-resolution models when LR images are given to meet their requirement. We used the same training images as the proposed JBIM method to fine-tune the super-resolution models. The proposed multiresolution adaptive ensemble algorithm was applied to all the compared methods for a fair comparison.

- Compared Methods. *The conventional Re-ID methods* considered in our evaluations are as follows: (1) cross-view quadratic discriminant analysis (XQDA) (Liao et al. 2015): a supervised Mahalanobis metric learning method; (2) domain-guided dropout (DGD) : a widely used deep CNN re-id model; (3) ResNet-50 (He et al. 2016a): a state-of-the-art deep CNN classification model. We utilized the contemporary local maximal occurrence (LOMO) handcrafted features (Liao et al. 2015) for XQDA.

The image super-resolution methods we selected for evaluation include two standard algorithms and five state-of-the-art algorithms: (1) Bilinear: a popular linear interpolation-based super-resolution model effective in handling generic image scaling; (2) Bicubic: another widely used image super-resolution method, which is an extension of cubic interpolation; (3) SRCNN (Dong et al. 2014): an existing state-of-the-art deep CNN-based super-resolution model; (4) fast SRCNN (FSRCNN) (Dong et al. 2016b): an accelerated deep CNN-based super-resolution model; (5) very deep super-resolution (VDSR) (Kim et al. 2016a): a super-resolution method based on a very deep CNN; (6) deeply-recursive convolutional network (DRCN) (Kim et al. 2016b): a deeply-recursive convolutional network-based super-resolution method; (7) SRGAN (Ledig et al. 2016): a super-resolution method using a GAN to compensate for the image details.

- Results & Analysis. For the other methods, we show different straightforward combinations of the super-resolution methods and re-id methods. Table 4 shows that the proposed JBIM method significantly outperformed all of the combinations of super-resolution+re-id methods. Specifically, the rank-1 matching rate over all the competitors by the JBIM method can reach 12.9% (52.0-39.1), 14.5%(88.3-73.8), 6.3%(59.1-52.8), 17.4%(49.7-32.3) and 22.9%(88.1-65.2) on the CAVIAR, MLR-CUHK03, MLR-SYSU, MLR-VIPeR and MLR-Market datasets, respectively. These results show that the joint bilateral resolution features outperformed the single HR-specific identity features of the images super-resolved by the generic super-resolution methods.

- Qualitative Evaluation of Different Super-Resolution Methods. We qualitatively compared the super-resolved person images produced by Bilinear, Bicubic, SRCNN, FSRCNN, VDSR, DRCN, SRGAN and our JBIM. Two examples are presented in Fig. 7. We make the following observations: (1) Super-resolved images by bilinear and bicubic interpolation are more blurry than those produced by the CNN-based super-resolution methods and our proposed JBIM method. (2) More edge/contour elements and better structured texture patterns are recovered by the proposed JBIM method. In addition, the color distributions of the images produced by the JBIM method are more similar to the ground-truth color distributions than those produced by the other methods. This difference visually indicates the advantages of JBIM over super-resolution + re-id methods due to the capability of recovering missing/enhancing appearance details while ensuring better re-id discrimination.

7 Further Analysis of the Proposed Method

7.1 Evaluation of the Individual Components

We provide detailed model component analysis in terms of performance contribution. The comparisons on the five cross-resolution re-id datasets are summarized in Table 5. In particular, “JBIM w/o HIM” means that we train the LIM branch (Fig. 4B) independently by using only the LR-specific identity features for re-id. Similarly, “JBIM w/o LIM” means that we train the HIM branch (Fig. 4A) independently by using only the HR-specific identity features for re-id.

- Super-Resolution *Versus* Low-resolution Specific Identity Features. We evaluate the cross-resolution re-id performance of the HR-specific identity features learned from the JBIM w/o LIM model (*i.e.*, using the HIM network only) in comparison to the LR-specific identity features from the JBIM w/o HIM model (*i.e.*, using the LIM network only). The results are shown on five datasets in Table 5. Although it is intuitive that using the HR features outperformed the LR features by 1.7%(48.7-47.0), 0.3%(87.1-86.8), 2.3%(58.0-55.7), 4.8%(45.6-40.8), 1.2%(87.7-86.5) in terms of the rank-1 matching rate on the CAVIAR, MLR-CUHK03, MLR-SYSU, MLR-VIPeR and MLR-Market datasets, respectively, it is also clear that the JBIM method without HIM is not weak; *i.e.*, discriminant identity features exist in the LR person image.

- Single-Resolution *Versus* Multiresolution Features. We further evaluate the cross-resolution re-id performance advantages of our multiresolution features in comparison to independently learned single-resolution features. Table 5 reports the matching results of the JBIM method, the JBIM method without LIM, and the JBIM method without HIM. From the experimental results, it can be observed that the joint learned multiresolution features achieve a better performance than using only one of them. Without the LIM network, the rank-1 matching rate of the JBIM method drops by 3.3%(52.0-48.7), 1.2%(88.3-87.1), 1.1%(59.1-58.0), 4.1% (49.7-45.6), 0.4% (88.1-87.7) on the CAVIAR, MLR-CUHK03, MLR-SYSU, MLR-VIPeR and MLR-Market datasets, respectively, which suggests that the LR-specific identity features are still important for constructing the joint bilateral-resolution features, although the LR-specific identity features contain less information than the HR-specific identity features. The performance margins of the JBIM method over the JBIM method without HIM (*i.e.*, LIM) are even larger. This finding validates the effectiveness of our proposed multiresolution feature learning method in improving cross-resolution re-id matching.

Table 4 Comparing combinations of image super-resolution and person re-id schemes (%). The best and 2nd-best results are indicated in bold and italic, respectively.

SR	Re-ID	CAVIAR			MLR-CUHK03			MLR-SYSU			MLR-VIPeR			MLR-Market		
Method	Method	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10
Bilinear	XQDA (Liao et al. 2015)	24.2	60.7	80.7	35.6	69.7	83.3	31.1	57.3	69.1	33.2	61.1	77.5	30.6	52.7	63.1
Bicubic		23.0	60.4	81.2	35.4	69.9	83.2	31.4	58.4	69.2	35.4	62.0	76.0	31.2	54.6	64.7
SRCNN (Dong et al. 2014)		24.2	59.4	81.3	35.3	69.5	82.8	29.6	56.0	67.9	31.7	60.1	75.6	29.8	52.1	63.0
FSRCNN (Dong et al. 2016b)		25.6	60.6	81.1	36.1	70.8	83.6	30.8	56.5	67.2	33.9	61.7	75.3	30.5	51.9	62.7
VDSR (Kim et al. 2016a)		23.4	59.3	80.9	37.3	70.9	83.9	30.6	56.8	67.5	32.3	64.9	76.3	30.5	54.0	64.2
DRCN (Kim et al. 2016b)		23.8	59.5	80.3	38.1	71.6	83.4	31.1	57.0	67.6	33.2	62.3	74.1	29.9	53.0	63.4
SRGAN (Ledig et al. 2016)		23.2	59.8	80.4	35.7	69.5	82.7	29.1	55.6	66.8	28.8	60.4	71.8	25.1	47.2	57.4
Bilinear	DGD (Xiao et al. 2016)	20.4	61.0	82.2	68.7	88.4	91.6	35.5	61.7	73.4	15.5	31.1	45.3	57.7	77.3	84.4
Bicubic		22.6	61.3	81.7	71.4	89.1	92.0	35.9	61.6	74.0	13.9	34.8	48.1	58.3	77.7	84.6
SRCNN		21.9	58.4	80.0	69.6	88.9	92.1	33.7	58.1	69.7	13.6	29.4	41.1	56.0	77.3	83.9
FSRCNN		21.6	61.2	80.3	69.8	89.2	92.2	35.7	61.3	72.6	13.9	33.5	45.3	57.1	76.9	84.5
VDSR		21.6	58.2	78.5	70.2	89.1	92.5	35.1	60.7	72.1	14.8	32.6	45.9	59.3	80.1	86.1
DRCN		21.4	59.0	79.1	69.9	89.1	92.2	35.1	61.4	71.5	14.6	31.0	48.4	59.1	78.5	84.9
SRGAN		21.5	59.0	80.7	69.8	89.5	92.2	34.0	61.8	73.7	14.9	32.3	47.5	56.6	76.4	83.9
Bilinear	ResNet-50 (He et al. 2016a)	35.5	71.0	87.8	72.8	92.2	95.9	52.3	77.1	84.6	28.5	52.2	68.0	*65.2*	*81.8*	87.0
Bicubic		*39.1*	77.5	*92.7*	72.3	92.2	96.0	*52.8*	*77.5*	*85.1*	28.8	50.0	63.0	63.6	81.0	86.5
SRCNN		38.6	*78.2*	92.5	72.7	92.0	95.7	52.2	76.0	83.7	28.2	56.0	68.4	62.6	80.9	87.2
FSRCNN		38.4	71.2	87.9	73.7	92.3	95.7	50.4	75.7	82.9	28.5	56.3	66.8	64.1	81.3	*87.3*
VDSR		37.6	76.9	92.8	*73.8*	*92.4*	95.7	52.4	77.2	84.7	*32.3*	*60.4*	*69.9*	63.5	80.3	87.0
DRCN		37.3	77.6	92.3	73.2	92.0	*95.9*	51.3	77.0	83.9	26.6	55.7	66.5	65.0	81.3	87.1
SRGAN		38.4	74.2	90.1	71.8	91.8	95.6	50.6	76.1	83.1	24.1	48.4	63.0	64.8	81.3	87.1
JBIM		**52.0**	**83.1**	**94.4**	**88.3**	**97.2**	**98.7**	**59.1**	**82.3**	**88.9**	**49.7**	**72.5**	**81.3**	**88.1**	**95.1**	**96.9**

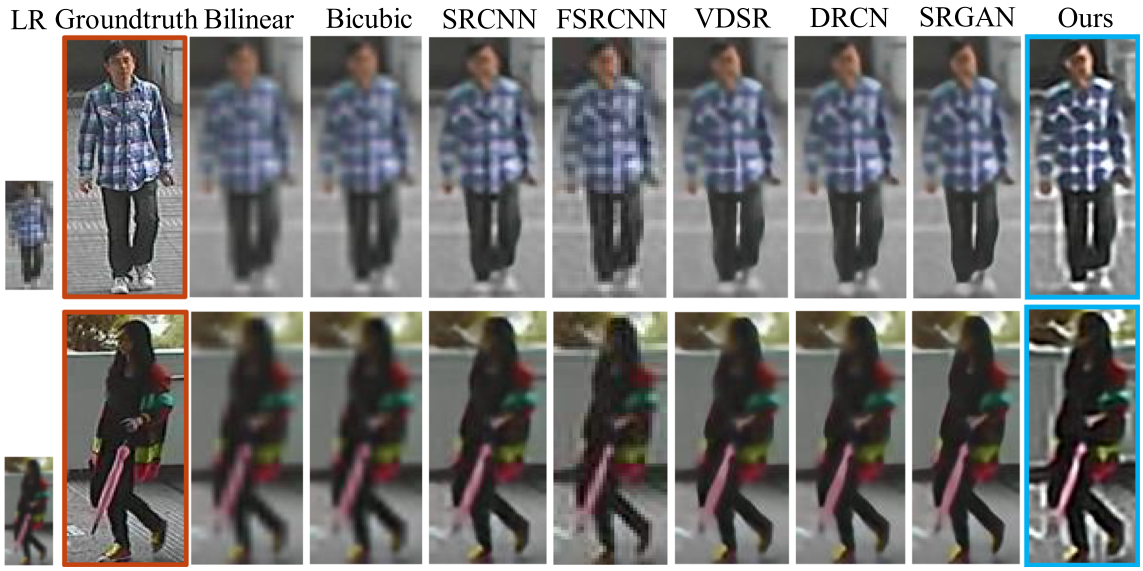

Fig. 7 Qualitative evaluations of the super-resolved person images by different methods. The ground-truth normal-resolution images (2nd column) are indicated by red bounding boxes (Color figure online)

Table 5 Matching rate (%): Evaluation of the individual components

CAVIAR	Rank1	Rank5	Rank10	mAP
JBIM w/o HIM	47.0	81.6	92.8	48.6
JBIM w/o LIM	48.7	81.0	92.6	49.0
JBIM	**52.0**	**83.1**	**94.4**	**50.1**
MLR-CUHK03	Rank1	Rank5	Rank10	mAP
JBIM w/o HIM	86.8	96.3	98.1	88.2
JBIM w/o LIM	87.1	96.3	98.2	88.5
JBIM	**88.3**	**97.2**	**98.7**	**89.9**
MLR-SYSU	Rank1	Rank5	Rank10	mAP
JBIM w/o HIM	55.7	79.8	86.7	62.2
JBIM w/o LIM	58.0	81.2	88.0	63.9
JBIM	**59.1**	**82.3**	**88.9**	**65.7**
MLR-VIPeR	Rank1	Rank5	Rank10	mAP
JBIM w/o HIM	40.8	68.0	79.1	49.8
JBIM w/o LIM	45.6	68.0	79.7	52.6
JBIM	**49.7**	**72.5**	**81.3**	**58.0**
MLR-Market	Rank1	Rank5	Rank10	mAP
JBIM w/o HIM	86.5	94.3	96.4	68.5
JBIM w/o LIM	87.7	94.8	96.8	70.7
JBIM	**88.1**	**95.1**	**96.9**	**73.5**

7.2 Jointly Learning the Multiresolution Features

JBIM can be considered a joint hybrid model of the HIM network, the LIM network and the FFN to learn a multiresolution feature fused by HR-specific and LR-specific identity features. To validate the effectiveness of jointly learned multiresolution features, we conduct a comparison experiment with "HIM+LIM" , which learns the HR-specific and LR-specific identity features independently and then concatenates the HR-specific and LR-specific identity features into the final features. As shown in Table 6, such a feature fusion method is effective: the JBIM method yields rank-1 matching rate improvement over "HIM+LIM" by 2.8%(52.0-49.2), 0.8%(88.3-87.5), 0.4%(59.1-58.7), 2.9% (49.7-46.8) and 0.4%(88.1-87.7) on the CAVIAR, MLR-CUHK03, MLR-SYSU and MLR-Market datasets, respectively. This finding suggests that the joint learning of the HR-specific and LR-specific identity features achieves better results than concatenating the independent LR-specific and HR-specific identity features. The results validate the effectiveness of FFN of the JBIM framework.

7.3 Synthetic Low-Resolution Images

We evaluate 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 JBIM framework in which the HIM network without the streams of the synthetic LR images (see the green arrows in Fig. 4 (A)). As such, the MSE super-resolution loss is removed because no LR-HR train-

Table 6 Matching rate (%): Comparing the independent learning methods. "HIM + LIM" means learning HIM and LIM independently; "SRCNN+ResNet-50+LIM" means learning super-resolution module, HR-specific re-id module and LIM independently

CAVIAR	Rank1	Rank5	Rank10	mAP
SRCNN + ResNet-50 + LIM	46.8	82.6	**95.2**	47.2
HIM + LIM	49.2	82.8	94.3	47.9
JBIM	**52.0**	**83.1**	94.4	**50.1**
MLR-CUHK03	Rank1	Rank5	Rank10	mAP
SRCNN + ResNet-50 + LIM	86.0	96.6	98.0	87.5
HIM + LIM	87.5	96.8	98.3	88.1
JBIM	**88.3**	**97.2**	**98.7**	**89.9**
MLR-SYSU	Rank1	Rank5	Rank10	mAP
SRCNN + ResNet-50 + LIM	57.8	81.5	87.9	63.5
HIM + LIM	58.7	82.4	88.8	64.3
JBIM	**59.1**	**82.3**	**88.9**	**65.7**
VIPeR	Rank1	Rank5	Rank10	mAP
SRCNN + ResNet-50 + LIM	41.6	69.0	80.1	55.2
HIM + LIM	46.8	**72.8**	**82.3**	57.5
JBIM	**49.7**	72.5	81.3	**58.0**
MLR-Market	Rank1	Rank5	Rank10	mAP
SRCNN + ResNet-50 + LIM	83.8	94.0	96.3	70.1
HIM + LIM	87.7	95.4	97.0	71.2
JBIM	**88.1**	**95.1**	**96.9**	**73.5**

ing image pairs are available. Table 7 shows that an inferior cross-resolution re-id performance will be yielded without this joint learning stream. For example, the rank-1 rate drops from 52.0% to 50.1% on the CAVIAR dataset, from 88.3% to 87.9% on the MLR-CUHK03 dataset, from 59.1% to 58.3% on the MLR-SYSU dataset, from 49.7% to 47.2% on the MLR-VIPeR dataset, and from 88.1% to 87.5% on the MLR-Market dataset, respectively.

This performance drop validates the usefulness of the proposed joint learning approach in guiding the image super-resolution model toward generating HR images with re-id discriminative visual information. We further directly evaluate the stripped-down HIM network (JBIM without LIM) without synthetic LR images. From Table 7, we can see that the performance also decreases without joint learning. Specifically, the rank-1 matching rate decreases by 1.8%, 0.2%, 0.8%, 1.5%, 0.8% on CAVIAR, MLR-CUHK03, MLR-SYSU, MLR-VIPeR and MLR-Market datasets, respectively. The results further indicate that super-resolution and re-id joint learning can obtain better discriminative HR-specific identity features than other methods.

7.4 Comparing Models Trained with All Resolution Images

We consider that there are specific features for different resolutions of an image. In order to prove the necessity of LR-specific identity information, we train a ResNet-50 (*i.e.*, our backbone) with all resolutions of training images, and we denote this variant as "All-Resolution". Compared with our JBIM that learns the collaboration of LR-specific and HR-specific information, the "All-Resolution" model achieves suboptimal performance. A plausible reason is that given a large number of different resolutions with the training data, the learned network has to fit all resolutions and could discard the specific resolution information. In contrast, our approach is able to extract different specific resolution information and make them collaborate by optimisation. As shown in Table 8, our method outperforms the model trained with all resolution images. In addition, we conduct a visualization of class activation maps of these two networks, and it can be found in Fig. 8. It is found that our model's class activation maps could capture complementary person appearance.

7.5 Super-Resolution and Re-ID Loss Balancing

We evaluated the balancing effect between image super-resolution and the person re-id loss by varying the trade-off parameter α in Eq. (6) ($\alpha = 1$ in all the other experiments). We conducted this analysis on all the genuine and simulated cross-resolution re-id datasets with $\beta = 3$ fixed. Table 9 shows that the experimental results are insensitive to variations in parameter α. Specifically, the rank-1 matching rate is highest when parameter $\alpha = 1$ in Table 9. Thus, we use $\alpha = 1$ in our comparison experiments. Moreover, when setting $\alpha = 0$, the rank-1 matching rate performance dropped 2.6%, 0.5%, 0.6%, 1.6%, 1.0% on the CAVIAR, MLR-CUHK03, MLR-SYSU, MLR-VIPeR and MLR-Market datasets, respectively, because super-resolution reconstruction is totally ignored, and thus, there is no interaction between super-resolution and re-id. Moreover, the learning constraint on the super-resolution submodel is weak, backpropagated from person identity classification supervision, and therefore results in poor re-id matching.

7.6 Single- and Multi- Resolution Re-ID Loss Balancing

We further evaluated the balancing effect between the single-resolution loss (*i.e.*, the LIM loss and the HIM loss) and the resolution fusion loss by varying the trade-off parameter β in Eq. (11) ($\beta = 3$ in all the other experiments). We conducted this analysis on all the genuine and simulated cross-resolution

Table 7 Effect of jointly super-resolving and classifying synthetic LR images (%)

Models	HIM (no synthetic LR)		HIM		JBIM (no synthetic LR)		JBIM	
Datasets	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
CAVIAR	46.9	47.6	48.7	49.0	50.1	48.6	**52.0**	**50.1**
MLR-CUHK03	86.9	88.0	87.1	88.5	87.9	89.4	**88.3**	**89.9**
MLR-SYSU	57.2	62.6	58.0	63.9	58.3	65.0	**59.1**	**65.7**
MLR-VIPeR	44.1	51.7	45.6	52.6	47.2	56.2	**49.7**	**58.0**
MLR-Market	86.9	69.9	87.7	70.7	87.5	72.8	**88.1**	**73.5**

Table 8 Matching rate (%): "All-Resolution" means training a ResNet-50 with all resolution images (e.g. normal HR training images and LR images obtained by manually down-sampling original training images at the ratios of $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$)

Dataset	Method	Rank1	Rank5	Rank10	mAP
MLR-VIPeR	All-Resolution	36.7	62.3	74.4	48.9
	JBIM	**49.7**	**72.5**	**81.3**	**58.0**
MLR-CUHK03	All-Resolution	52.6	77.3	84.9	56.4
	JBIM	**88.3**	**97.2**	**98.7**	**89.9**
MLR-Market	All-Resolution	53.8	71.1	77.6	38.0
	JBIM	**88.1**	**95.1**	**96.9**	**73.5**

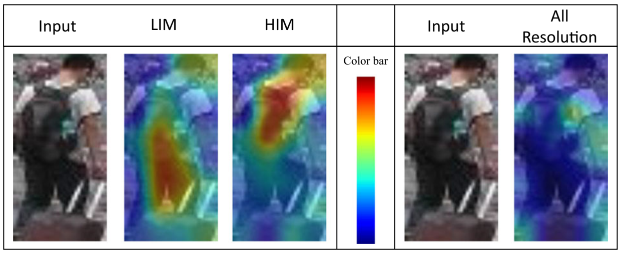

Fig. 8 Visualizations of our JBIM and All-Resolution

re-id datasets with $\alpha = 1$ fixed. Table 10 shows that the experimental results are insensitive to variations in parameter β. Specifically, the rank-1 matching rate of most datasets peaks when $\beta = 3$. Therefore, we use $\beta = 3$ in our comparison experiments.

7.7 Different Basic Super-Resolution or Re-ID Models

We further validate the flexibility of the JBIM framework by choosing different super-resolution or re-id CNN models to construct different variants. In particular, we replaced the default super-resolution model (*i.e.*, SRCNN) with VDSR, DRCN and SRGAN in all the other experiments or replaced the default re-id model (*i.e.*, ResNet-50) with OSNet in all the other experiments for our proposed JBIM framework and JBIM framework without LIM (*i.e.*, using HIM only) models. The results are shown in Tables 11 and 12. In the tables, "JBIM (VDSR)" and "HIM (VDSR)" mean that we use VDSR for super-resolution in the JBIM and HIM frameworks, respectively. Similarly, "JBIM (OSNet)", "HIM (OSNet)" and "LIM (OSNet)" mean that OSNet is used in our model as backbone. The results indicate that our joint learning methods perform stably across different super-resolution and re-id models.

7.8 Scale-Adaptive Low-Resolution Fusion

We evaluated the effect of fusing discriminative feature representations from multiple LR scales for improving person matching using the proposed scale-adaptive ensemble algorithm. To this end, we evaluated the cross-resolution re-id

Table 9 Effect of balancing image super-resolution and the person re-identification loss ($\beta = 3$)

α	0		1		10		100		1000	
Datasets	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
CAVIAR	49.4	47.7	**52.0**	**50.1**	51.5	48.9	51.0	48.7	50.8	48.2
MLR-CUHK03	87.8	89.0	88.3	89.9	88.2	89.5	**88.5**	**90.1**	88.2	89.7
MLR-SYSU	58.5	64.8	**59.1**	**65.7**	59.0	65.2	59.0	65.4	59.0	65.6
MLR-VIPeR	48.1	55.8	**49.7**	**58.0**	48.7	57.1	49.1	56.9	49.1	57.3
MLR-Market	87.1	72.2	**88.1**	73.5	88.0	**73.6**	87.9	73.2	87.9	73.3

Table 10 Effect of balancing the single-resolution and multiresolution re-id losses ($\alpha = 1$)

β	1		2		3		4		5	
Datasets	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
CAVIAR	51.1	49.1	52.0	49.8	**52.0**	**50.1**	50.9	48.7	50.6	48.7
MLR-CUHK03	88.1	89.8	88.1	89.5	**88.3**	**89.9**	88.2	89.9	88.3	89.7
MLR-SYSU	59.1	**65.8**	59.0	65.4	59.1	65.7	**59.2**	65.6	59.1	65.4
MLR-VIPeR	47.5	55.7	48.4	57.2	**49.7**	**58.0**	46.5	55.6	46.8	55.0
MLR-Market	87.9	73.0	88.0	73.3	**88.1**	**73.5**	87.9	72.9	87.8	72.9

Table 11 Effects of super-resolution CNN models (%)

Models	CAVIAR		MLR-CUHK03		MLR-SYSU		MLR-VIPeR		MLR-Market	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
HIM (SRCNN)	48.7	49.0	87.1	88.5	58.0	63.9	45.6	52.6	87.7	70.7
HIM (VDSR)	48.5	48.7	86.4	88.0	57.6	63.0	44.9	51.9	85.6	69.7
HIM (DRCN)	48.9	49.2	87.3	88.6	57.3	63.4	45.3	52.2	85.9	70.0
HIM (SRGAN)	47.6	46.8	86.1	87.7	58.0	63.5	44.4	51.5	85.7	69.6
JBIM (SRCNN)	**52.0**	**50.1**	88.3	89.9	59.1	65.7	**49.7**	**58.0**	**88.1**	**73.5**
JBIM (VDSR)	51.4	49.7	88.0	89.5	59.0	65.5	48.5	56.8	87.9	73.2
JBIM (DRCN)	51.8	49.7	**88.6**	**90.1**	**60.1**	**66.2**	49.2	57.5	88.0	73.5
JBIM (SRGAN)	49.3	48.9	87.4	88.6	60.0	66.0	47.4	56.0	87.9	72.8

Table 12 Effects of Re-ID CNN models (%)

Models	CAVIAR		MLR-CUHK03		MLR-SYSU		MLR-VIPeR		MLR-Market	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
SRCNN+OSNet	43.0	45.6	86.8	88.3	59.6	64.9	46.8	60.4	86.0	70.7
HIM (OSNet)	49.0	49.7	87.5	89.2	60.9	66.7	49.1	61.0	87.4	72.6
HIM (OSNet)+LIM (OSNet)	49.8	50.1	87.6	89.1	61.4	66.6	49.7	62.5	88.1	73.5
JBIM (OSNet)	**53.1**	**51.2**	**88.7**	**90.3**	**61.7**	**67.3**	**52.7**	**63.1**	**89.6**	**74.7**

performance of six combination schemes from three different scale-specific JBIM models ($M_{\frac{1}{2}}$, $M_{\frac{1}{3}}$, $M_{\frac{1}{4}}$). We further evaluated the effect of the proposed scale-adaptive ensemble algorithm on the JBIM framework without LIM (*i.e.*, using only the HIM network). They correspond to 3 downsampling ratios $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$. Table 13 shows that more scales of LR information fused by the proposed method yield better results that we can achieve. The best results over all five cross-resolution re-id datasets are yielded by fusing all three scale-specific models. This finding validates the efficacy of the proposed multiscale fusion algorithm. Moreover, this observation is consistent in spirit with the classical pyramid matching kernel (Grauman and Darrell 2005; Lazebnik et al. 2006), except that our multiscale fusion is *uniquely* on multiple pixel-level resolutions rather than on multiple spatial extents of the same resolution.

7.9 Ablation Study for the Scaling Parameter σ

We have evaluated the effect of the scaling parameter σ in Eq. (19). As shown in Fig. 9, the increase of σ improves the performances on each dataset when σ is smaller than 0.2. When σ is larger than 0.2, the performance does not change a lot. The parameter σ is not an important parameter in our experiments, and we just use it to enlarge the difference between the resolution similarity r and super-resolving ratio m_i. In our experiments, σ is set to 0.5.

If we take σ as a variable s and $(r - m_i)^2$ as a constant C, the function in Eq. (19) can be converted to

$$w_i = e^{-C*s^{-2}}, \quad s \in (0, +\infty) \tag{20}$$

Table 13 Matching rate(%): Effect of scale-adaptive LR fusion of HIM and JBIM.

Models	CAVIAR				MLR-CUHK03				MLR-SYSU				MLR-VIPeR				MLR-Market			
	HIM		JBIM		HIM		JBIM		HIM		JBIM		HIM		JBIM		HIM		JBIM	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
$M_{\frac{1}{2}}$	45.1	46.2	46.8	44.9	86.5	87.3	87.6	88.5	48.8	54.3	53.5	60.0	38.7	45.2	40.5	49.8	83.3	66.8	85.7	70.2
$M_{\frac{1}{3}}$	42.9	43.2	48.2	46.5	85.1	86.9	86.6	88.0	54.2	58.1	56.5	61.3	36.7	43.8	42.4	51.3	84.7	67.5	86.5	70.9
$M_{\frac{1}{4}}$	40.2	41.0	43.8	43.7	84.8	86.5	85.3	87.6	52.1	57.2	53.9	59.7	39.6	45.8	42.7	51.2	84.0	67.2	86.0	70.3
$M_{\frac{1}{2}} + M_{\frac{1}{3}}$	48.4	48.7	51.2	49.6	87.1	87.6	88.2	89.5	55.7	62.0	57.7	64.2	44.0	51.2	45.6	54.2	86.8	70.0	87.0	71.7
$M_{\frac{1}{2}} + M_{\frac{1}{4}}$	46.9	47.2	48.0	46.3	86.4	87.3	87.5	88.9	55.0	61.8	56.4	64.1	44.0	51.5	46.5	54.5	86.5	69.5	87.4	72.2
$M_{\frac{1}{3}} + M_{\frac{1}{4}}$	45.6	46.8	48.0	46.0	86.2	87.2	87.0	88.3	57.7	63.5	58.3	65.0	41.5	46.4	45.3	54.0	86.7	69.8	87.4	72.5
$M_{\frac{1}{2}} + M_{\frac{1}{3}} + M_{\frac{1}{4}}$	**48.7**	**49.0**	**52.0**	**50.1**	**87.3**	**88.5**	**88.3**	**89.9**	**58.0**	**63.9**	**59.1**	**65.7**	**45.6**	**52.6**	**49.7**	**58.0**	**87.7**	**70.7**	**88.1**	**73.5**

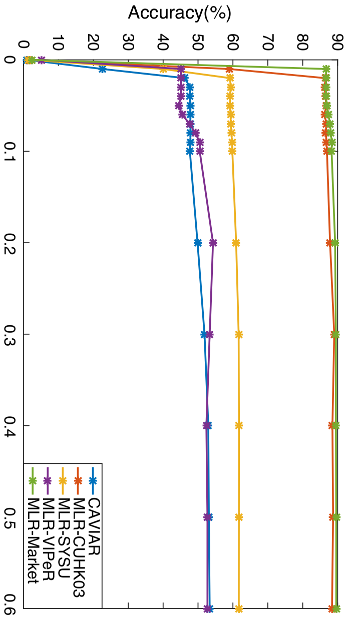

Fig. 9 Ablation study of the scaling parameter σ

which is a monotone increasing bounded function whose upper bound is 1. Therefore, the distance weight w_i does not change a lot if we continue to increase the scaling parameter σ.

7.10 Study of Down-Sampling Ratios

Except the original setting with the down-sampling ratios $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$ for both training and testing sets, we have now evaluated each individual ratio separately. From Table 14, we have several observations: 1) When the down-sampling ratio of test set becomes larger, the performance will drop due to more missing observations; 2) The best performance can be achieved when the same ratio is applied to the training set, suggesting that image resolution is a dimension that matters to model performance.

8 Conclusion

In this work, we present a joint bilateral-resolution identity modeling (JBIM) framework for solving the cross-resolution person re-identification challenge. The JBIM framework collaboratively learns both HR-specific and LR-specific identity features by introducing a synergistic interplay between super-resolution and discriminant re-id feature learning. In particular, we have demonstrated the significance of exploiting LR-specific identity features in joint learning for overcoming cross-view and cross-resolution cross-resolution re-id. Extensive evaluations on five benchmarks show the clear superiority of our JBIM model over existing cross-resolution re-id methods and fusions of state-of-the-art super-resolution and re-id models, without elaborate network architecture design. We have also conducted a full spectrum of model component analysis to validate the effectiveness of individual modules and provide insights into our model formulation.

Table 14 Ablation study of the down-sampling ratio. "Down-sampling ratio for training" refers to the down-sampling ratio of LR images using in the training process, HR images do not change. "Down-sampling ratio for testing" means the down-sampling ratio of LR probe images using in the testing process, HR gallery images stay as normal

Down-sampling ratio for **testing** Down-sampling ratio for **training**	1/2		1/3		1/4		{1/2, 1/3, 1/4}	
MLR-VIPeR	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
1/2	38.6	50.3	36.7	49.8	32.3	44.0	36.4	48.9
1/3	38.3	50.2	37.3	50.7	33.9	46.7	37.0	49.4
1/4	37.0	49.6	35.4	48.9	36.1	48.6	36.4	49.2
{1/2, 1/3, 1/4}	44.6	58.5	44.3	57.8	40.8	55.0	49.7	58.0
MLR-CUHK03	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
1/2	91.3	92.6	89.9	91.4	88.1	89.5	78.3	80.8
1/3	91.2	92.1	91.0	91.7	89.7	90.8	81.5	83.1
1/4	90.0	91.4	89.0	90.6	89.7	90.5	80.6	82.5
{1/2, 1/3, 1/4}	91.2	92.5	90.1	91.5	90.3	91.5	88.3	89.9
MLR-Market	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
1/2	90.1	75.6	87.4	72.7	83.4	67.7	70.7	57.2
1/3	89.0	74.0	88.5	72.8	85.8	69.5	73.3	57.7
1/4	87.5	71.5	86.8	70.8	86.1	70.2	71.1	56.0
{1/2, 1/3, 1/4}	92.3	79.8	91.2	78.4	89.0	75.9	88.1	73.5

Acknowledgements This work was supported partially by the NSFC (U1911401, U1811461), Guangdong NSF Project (No. 2020B1515120085, 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03), and the Key-Area Research and Development Program of Guangzhou (202007030004). The corresponding author for this paper is Jian-Huang Lai.

References

Ahmed, E., Jones, M., & Marks, T. K. (2015). *An improved deep learning architecture for person re-identification* (pp. 3908–3916). CVPR.

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., et al. (2019a). *Abd-net: Attentive but diverse person re-identification*. ICCV.

Chen, Y. C., Zheng, W. S., Lai, J. H., & 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., & Lai, J. H. (2017b). *Person re-identification by camera correlation aware feature augmentation*. IEEE TPAMI.

Chen, Y. C., Li, Y. J., Du, X., & Wang, Y. C. F. (2019c). *Learning resolution-invariant deep representations for person re-identification*. AAAI.

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., & Murino, V. (2011). *Custom pictorial structures for re-identification* (pp. 6–10). BMVC.

Cheng, Z., Zhu, X., & Gong, S. (2018). *Low-resolution face recognition*. CVPR.

Cheng, Z., Dong, Q., Gong, S., & Zhu, X. (2020). *Inter-task association critic for cross-resolution person re-identification*. CVPR.

Dai, Z., Chen, M., Gu, X., Zhu, S., & Tan, P. (2019). *Batch dropblock network for person re-identification and beyond*. ICCV.

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). *Learning a deep convolutional network for image super-resolution* (pp. 184–199). ECCV, Springer.

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

Dong, C., Loy, C. C., & Tang, X. (2016b). *Accelerating the super-resolution convolutional neural network* (pp. 391–407). ECCV, Springer.

Fan, X., Jiang, W., Luo, H., & Fei, M. (2018). *Spherereid: Deep hypersphere manifold embedding for person re-identification*. ECCV.

Gong, S., Cristani, M., Yan, S., & Loy, C. C. (2014). *Person re-identification*. Springer.

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

Gray, D., & Tao, H. (2008). *Viewpoint invariant pedestrian recognition with an ensemble of localized features* (pp. 262–275). ECCV.

Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J., & Han, K. (2019). *Beyond human parts: Dual part-aligned representations for person re-identification* (pp. 3642–3651). ICCV.

Haris, M., Shakhnarovich, G., & Ukita, N. (2018). *Deep back-projection networks for super-resolution* (pp. 1664–1673). CVPR.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). *Deep residual learning for image recognition* (pp. 770–778). CVPR.

He, W. X., Chen, Y. C., & Lai, J. H. (2016b). *Cross-view transformation based sparse reconstruction for person re-identification* (pp. 3410–3415). ICPR.

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

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

Huang, Y., Zha, Z. J., Fu, X., Hong, R., & Li, L. (2020). *Real-world person re-identification via degradation invariance learning*. CVPR.

Jiao, J., Zheng, W. S., Wu, A., Zhu, X., & Gong, S. (2018). *Deep low-resolution person re-identification*. AAAI.

Jing, X. Y., Zhu, X., Wu, F., You, X., Liu, Q., Yue, D., et al. (2015). *Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning* (pp. 695–704). CVPR.

Kalayeh, M. M., Basaran, E., Gokmen, M., Kamasak, M. E., & Shah, M. (2018). *Human semantic parsing for person re-identification*. CVPR.

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016a). *Accurate image super-resolution using very deep convolutional networks* (pp. 1646–1654). CVPR.

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016b). *Deeply-recursive convolutional network for image super-resolution* (pp. 1637–1645). CVPR.

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

Lazebnik, S., Schmid, C., & Ponce, J. (2006). *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories* (pp. 2169–2178). CVPR.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A., Tejani, A., et al. (2016). *Photo-realistic single image super-resolution using a generative adversarial network*. CVPR.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). *Photo-realistic single image super-resolution using a generative adversarial network* (pp. 4681–4690). CVPR.

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

Li, W., Zhu, X., & Gong, S. (2017). *Person re-identification by deep joint learning of multi-loss classification*. IJCAI.

Li, X., Zheng, W. S., Wang, X., Xiang, T., & Gong, S. (2015). *Multi-scale learning for low-resolution person re-identification* (pp. 3765–3773). ICCV.

Li, Y. J., Chen, Y. C., Lin, Y. Y., Du, X., & Wang, Y. C. F. (2019a). *Recover and identify: A generative dual model for cross-resolution person re-identification*. ICCV.

Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., & Wu, W. (2019b). *Feedback network for image super-resolution* (pp. 3867–3876). CVPR.

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

Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). *Enhanced deep residual networks for single image super-resolution* (pp. 136–144). CVPR WS.

Luo, H., Gu, Y., Liao, X., Lai, S., & Jiang, W. (2019). *Bag of tricks and a strong baseline for deep person re-identification*. CVPR WS.

Mao, S., Zhang, S., & Yang, M. (2019). *Resolution-invariant person re-identification*. IJCAI.

Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2016). *Hierarchical gaussian descriptor for person re-identification* (pp. 1363–1372). CVPR.

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

Qian, X., Fu, Y., Jiang, Y. G., Xiang, T., & Xue, X. (2017). *Multi-scale deep learning architectures for person re-identification*. CVPR.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). *Performance measures and a data set for multi-target, multi-camera tracking*. ECCV.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). *Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)* (pp. 501–518). ECCV.

Tai, Y., Yang, J., & Liu, X. (2017). *Image super-resolution via deep recursive residual network* (pp. 2790–2798). CVPR.

Wang, T., Gong, S., Zhu, X., & Wang, S. (2014). *Person re-identification by video ranking* (pp. 688–703). ECCV.

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

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

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

Wang, Z., Ye, M., Yang, F., Bai, X., & Satoh, S. (2018). *Cascaded sr-gan for scale-adaptive low resolution person re-identification* (pp. 3891–3897). IJCAI.

Wong, Y., Sanderson, C., Mau, S., & Lovell, B. C. (2010). *Dynamic amelioration of resolution mismatches for local feature based identity inference* (pp. 1200–1203). ICPR.

Wu, L., Shen, C., & Avd, Hengel. (2016). *Personnet: Person re-identification with deep convolutional neural networks*. CVPR.

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

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. H. (2021). *Deep learning for person re-identification: A survey and outlook*. TPAMI.

Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y., & Bai, X. (2018). *Hard-aware point-to-set deep metric for person re-identification* (pp. 196–212). ECCV.

Zhang, L., Xiang, T., & Gong, S. (2016). *Learning a discriminative null space for person re-identification* (pp. 1239–1248). CVPR.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). *Image super-resolution using very deep residual channel attention networks* (pp. 286–301). ECCV.

Zheng, Z., Zheng, L., & Yang, Y. (2018). *A discriminatively learned cnn embedding for person reidentification*. ACM MM.

Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., et al. (2019a). *Pyramidal person re-identification via multi-loss dynamic training*. CVPR.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). *Scalable person re-identification: A benchmark* (pp. 1116–1124). ICCV.

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

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

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., & Kautz, J. (2019b). *Joint discriminative and generative learning for person re-identification*. CVPR.

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). *Omni-scale feature learning for person re-identification*. ICCV.