Journal of Systems Engineering and Electronics
Vol. 34, No. 3, June 2023, pp.682 – 695

Dual-stream coupling network with wavelet transform for cross-resolution person re-identification

SUN Rui [1,2], YANG Zi [1,2,*], ZHAO Zhenghui [1,2], and ZHANG Xudong [1,2]

1. Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education), Hefei University of Technology, Hefei 230601, China; 2. School of Computer and Information, Hefei University of Technology, Hefei 230601, China

Abstract: Person re-identification is a prevalent technology deployed on intelligent surveillance. There have been remarkable achievements in person re-identification methods based on the assumption that all person images have a sufficiently high resolution, yet such models are not applicable to the open world. In real world, the changing distance between pedestrians and the camera renders the resolution of pedestrians captured by the camera inconsistent. When low-resolution (LR) images in the query set are matched with high-resolution (HR) images in the gallery set, it degrades the performance of the pedestrian matching task due to the absent pedestrian critical information in LR images. To address the above issues, we present a dual-stream coupling network with wavelet transform (DSCWT) for the cross-resolution person re-identification task. Firstly, we use the multi-resolution analysis principle of wavelet transform to separately process the low-frequency and high-frequency regions of LR images, which is applied to restore the lost detail information of LR images. Then, we devise a residual knowledge constrained loss function that transfers knowledge between the two streams of LR images and HR images for accessing pedestrian invariant features at various resolutions. Extensive qualitative and quantitative experiments across four benchmark datasets verify the superiority of the proposed approach.

Keywords: cross-resolution, feature invariant learning, person re-identification, residual knowledge transfer, wavelet transform.

DOI: 10.23919/JSEE.2023.000028

1. Introduction

Person re-identification [1,2] refers to the search of particular pedestrians across surveillance cameras, which is broadly deployed in video surveillance, driverless scenarios, etc. In recent years, homogeneous person re-identification [3−9] based on deep learning architecture has seen more noticeable achievements in various complex scenarios. However, there are still urgent dilemmas in the heterogeneous person problem in cross-resolution situations. The image resolution captured by the camera is not exactly consistent owing to the influence of force majeure factors such as camera parameters, location, pedestrian proximity, and weather conditions. It is worthwhile to explore this image matching task with a large disparity in resolution, which is a challenging issue to overcome.

As far as we know, there is a lack of articles focusing on tackling cross-resolution person re-identification. The majority of available solutions fall into two main categories: utilizing image super-resolution (SR) [10−12] and learning resolution invariant representation [13−15]. The first category of solutions is the one that exploits the SR reconstruction task to assist the low-resolution (LR) image to perform SR transformation and carry out person recognition following the resolution upgrading. However, reconstructed high-resolution (HR) image lacks the objectivity of the image, which is a pathological problem with SR reconstruction. It may add certain "constructed details" when reconstructing the image, rather than the details lacked in LR image itself, aiming at raising the subjective perception of the image. This has a detrimental influence on the subsequent recognition tasks. The second category of solutions is to perform learning the invariant feature subspace of HR images and LR images. While learning invariant features of both HR images and LR images, the model fails to reinforce LR images accordingly. The absence of pedestrian details in LR images means that the

Manuscript received July 20, 2021.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61471154;61876057), and the Key Research and Development Program of Anhui Province-Special Project of Strengthening Science and Technology Police (202004D07020012).

model can barely extract valid information for identity screening. Both types of schemes have certain limitations.

In this paper, we adopt the former solution to the above problem, which is the combination of SR reconstruction and network for feature extraction. In view of the drawbacks of existing methods, we consider how to highlight the missing information in LR images itself and obtain strong pedestrian features. The wavelet multi-resolution analysis principal offers a breakthrough. As the wavelet transform (WT) enables to highlight certain aspects in the problem adequately, we decide to adopt WT to emphasize the missing detailed information from LR images. For the aspect of extracting robust features, we believe that there are still inter-domain differences between LR images and HR images. It is significant to narrow the inter-domain differences to obtain effective pedestrian information.

Based on the above mentioned thinking, we introduce a deep architecture called dual-stream coupling network with WT (DSCWT) for tackling the cross-resolution person re-identification problem. DSCWT consists of two parts, the progressive enhancement network (PEN) and the heterogeneous residual network (HRN). By the WT, the PEN decomposes LR images into subgraphs that contain high-frequency and low-frequency information of the image. High-frequency and low-frequency areas are simultaneously reinforced accordingly. The high-frequency areas are restored in detail to emphasize the person screening information. And the enhancement of the low-frequency regions is performed to boost the saliency of the pedestrian appearance to maintain the identity information. The method performs dual augmentation based upon the rein-forcement findings of high-frequency and low-frequency components for narrowing the feature gap between LR and SR images. HRN is employed as identification that learns LR person invariance features. To further compensate for the absent details within LR images, we devise a constraint function to acquire the residual information. It calculates the information discrepancy between the two streams of SR images and HR images so that the pedestrian validity information in HR can be transferred to SR image as much as possible. Ultimately, stronger discriminative pedestrian feature representation is obtained to secure the performance accuracy of cross-resolution person re-identification.

Overall, the proposed method has the following advantages over previous methods: the proposed method uses the multi-resolution analysis principle of WT to recover the original details lost in LR images, rather than the SR algorithm which constructs non-existent details for the sake of image fidelity. WT provides sufficient benefits for the subsequent feature extraction and recognition tasks. Contributions are summed up as follows:

(i) The proposed DSCWT can address the cross-resolution person re-identification task effectively. Inspired by the principle of multi-resolution analysis, we use WT to highlight the absence of detailed information in LR images at high frequencies while preserving identity information. The performance of the subsequent recognition task is thus guaranteed.

(ii) We design a residual knowledge con-strained loss function to transfer knowledge between images at different resolutions. It shrinks the discrepancy in data distribution between LR images and HR images, hence ensuring that the network earns feature invariant person representations.

(iii) We experiment on CAVIAR, MLR-Market1501, MLR-CUHK03, and MLR-DukeMTMC-REID datasets, and the results show that our method obtains a relatively excellent performance, which is competitive among many advanced methods.

2. Related work

2.1 General person re-identification

Over recent years, the task of person re-identification has gained more and more attention due to the increasing public safety needs. A large amount of work has been devoted to solving the challenges of efficiency, performance, and practicality of person re-identification in real-world scenarios. In terms of efficiency, in order to obtain lightweight networks, Wang et al. [16] used a "soft" pruning model for channel decay with a localized correlation of pre-trained network channels, which maintains the channel distribution of the original network based on redundant pruning of local clusters. Quan et al. [17] used the neural architecture search (NAS) method for a special re-identification search space, and the proposed body-part-aware components can capture the correlation between parts, thus reducing the number of network parameters. In terms of performance, in order to capture more discriminatory person information, Zhang et al. [18] proposed a dense semantic comparison approach to solving the person image misalignment problem using fine-grained semantics. Sun et al. [19] designed a novel loss function named circle loss to screen person identities from high-dimensional redundant features to capture fea-

ture representations with higher discriminative power. In terms of practicality, in order to make the model with efficient generalization ability, Fan et al. [20] used self-paced learning method to assign pseudo-labels to new data by clustering, and the number of reliable samples increases with the quality of the model. Zhong et al. [21] combined supervised and unsupervised learning, while introducing three basic invariants of the re-identification model, using prior knowledge and soft-label to study the intra-domain variation in the target domain.

2.2 Cross-resolution person re-identification

Cross-resolution person re-identification mainly consists of two aspects: First, the features of LR images and HR images are learned simultaneously to learn the resolution invariant feature representation. For example, Jing et al. [22] adopted a semi-coupled low-rank discriminant dictionary learning methodology to convert LR features in the query set into discriminative HR features. Chen et al. [13] used an adversarial learning strategy to align and capture feature represent-ations at different resolutions, while it is not subject to the input LR constraints. Huang et al. [15] proposed a self-supervised untangled representation learning strategy to eliminate degeneracy in the real world in an unsupervised. Second, by cascading two tasks, SR and person re-identification, a higher resolution of LR images is achieved. For example, Jiao et al. [11] improved the compatibility between SR and re-identification by enhancing the pedestrian appearance information. Mao et al. [12] used foreground attention to amplify people in the output image of SR network and suppress irrelevant background before the person re-identification task. Cheng et al. [10] introduced a regularization method for the inter-task association mechanism to smooth the SR reconstruction task and the person re-identification task in joint learning, thus intensifying the compatibility between the two tasks.

2.3 SR networks

With the continuous rise of deep learning, SR models based on convolutional neural networks have been proposed. Dong et al. [23] presented an end-to-end SR algorithm employing convolutional neural network (CNN) architecture for the first time. Tong et al. [24] applied dense blocks to the SR problem. Ledig et al. [25] used a generative adversarial network strategy, which utilizes perceptual loss and adversarial loss to recover the resolution as the image. Chu et al. [26] proposed a hybrid controller that combines evolutionary computation and rein-forcement learning to support microscopic search and macroscopic search with an elastic search strategy for SR problems. Zhou et al. [27] proposed a novel image degradation framework to solve the SR problem by estimating various fuzzy kernels and the real noise distribution.

2.4 WT in image decomposition

In recent years, WT has been applied in several fields, including remote sensing and underwater scenes, and is developing more rapidly. He et al. [28] proposed a low bit rate underwater video image compression coding method based on wavelet decomposition for alleviating the disadvantage of limited bandwidth for underwater video image transmission. Ramamonjisoa et al. [29] proposed to use wavelet decomposition and integrate a fully predictable encoder-decoder architecture to reconstruct high-fidelity depth maps. Shahrezaei et al. [30] used the discrete WT (DWT) to decompose the KOMPSAT-5 SAR image of noncoherent sea ice texture, and then performed further fractal analysis.

3. Proposed method

3.1 Overview

As shown in Fig. 1, the DSCWT is programmed to handle the cross-resolution person re-identification issue. When training the model, we hypothesize that the input is a set of images $\{I_{\mathrm{lr}}, I_{\mathrm{hr}}, p^i\}$. I_{lr} and I_{hr} are LR images and HR images respectively, and p^i is the ground-truth person identity label. Our intention is that LR images in the query set can be closely matched with HR images in the gallery set during in the testing phase. With the above purpose, we design the PEN, which performs multi-resolution analysis of pedestrian images by WT. WT of LR image highlights the missing detail information of the image and thus upgrades the resolution, alleviating the dilemma of mismatched human resolution. After that, we raise the HRN which decides to narrow the feature gap between the two streams of SR images and HR images to acquire resolution invariant person features. Subsequently, we are going to elaborate on the DSCWT. Concretely, the HRN is composed of efficient feature extraction blocks (EFEB), which is divided into two training networks using a similar architecture, thereby catching the feature maps of I_{sr} as well as I_{hr}.

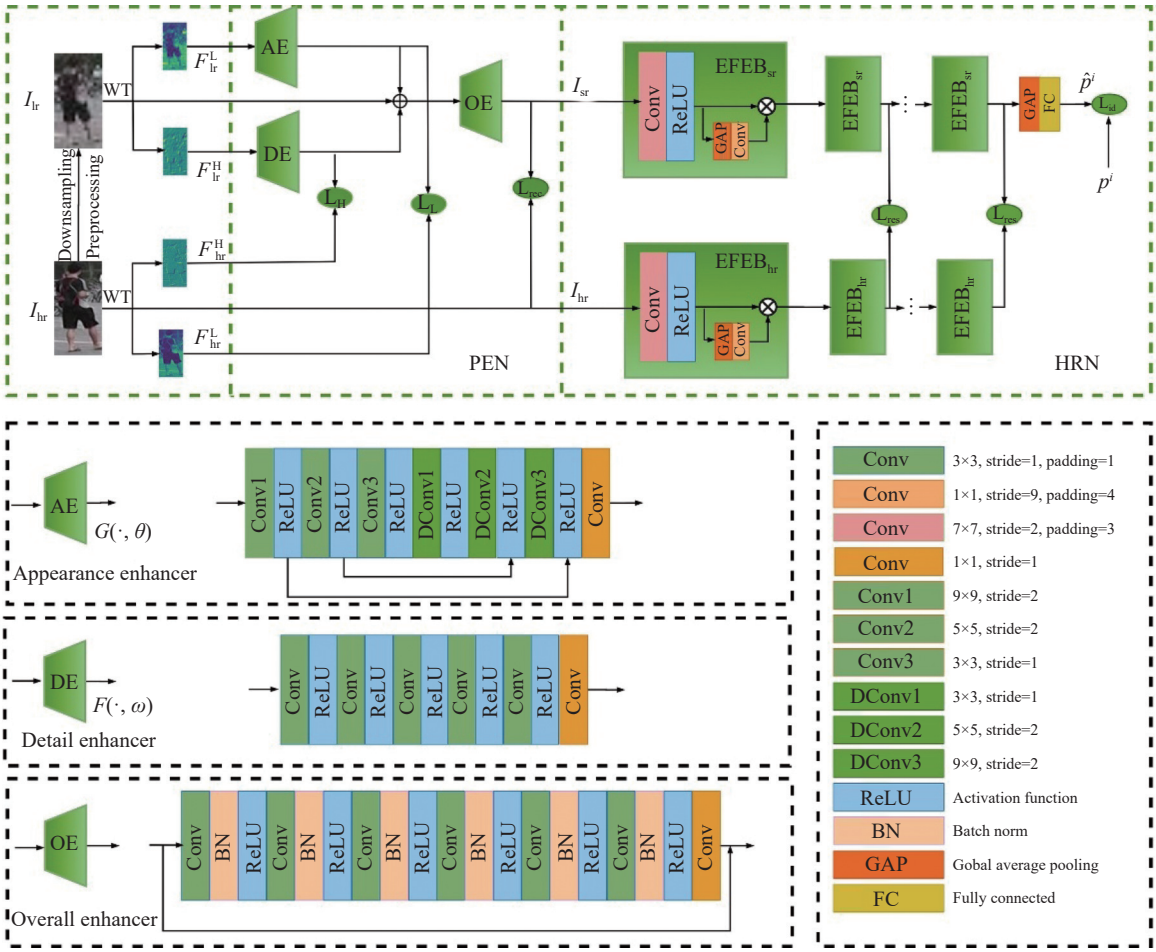


Fig. 1 Overview of the proposed DSCWT for addressing cross-resolution person re-identification

3.2 Pedestrian image multi-resolution decoupling using WT

WT [31] is the basis of multi-resolution theory analysis. Its main feature is that the transform can sufficiently highlight specific aspects to the problem. In other words, the superiority of WT is the fact that features that cannot be detected at a certain resolution can be easily detected at a certain resolution.

The human visual system is adaptive to the existence of the scale with the object. If the size of the object is small or the contrast is low, in which we only see the overall outline of the object, afterwards we usually study them at a higher resolution. If the size of the object is greater or the contrast is taller so that we are able to observe more detailed information about the object, then our cursory observation is sufficient. If smaller or larger objects are present simultaneously, then it will be more advantageous to study them at multiple resolutions. This is the basic motivation for wavelet multi-resolution processing.

In the open world, the constant change between pedestrian and camera positions makes the size and resolution of the captured pedestrians never similar. If we intend to determine whether the pedestrians captured by cross-camera are the same person, the trend is to zoom in on pedestrian images with lower resolution or smaller size. The idea fits perfectly with the motivation of wavelet multi-resolution analysis. Moreover, LR images are seized by the camera lack discriminative information about the pedestrians, while the missing detail information can be highlighted at high frequencies using WT. The two advantages are what make us choose to apply WT for cross-resolution person re-identification tasks.

A one-dimensional (1D) DWT (1D-DWT) of discrete signal $x[n]$ is defined as

$$\mathrm{WT}_x(k,l)=\sum_{n=-\infty}^{+\infty}x[n]\psi_{k,l}(n) \tag{1}$$

where $\psi_{k,l}(n)$ is the dilated and translated version of the mother wavelet ψ which can be calculated as

$$\psi_{k,l}(n)=2^{-\frac{k}{2}}\psi[2^{-k}n-l]. \tag{2}$$

The DWT feeds the signal to the low-pass filter $l(e)$ and high-pass filter $h(e)$ for downsampling with a scale factor of 2. It ultimately yields the high-frequency and low-frequency components of the signal. For the Harr wavelet, $l(e)$ and $h(e)$ are defined as

$$l(e)=\begin{cases}1, & e=0,1\\ 0, & \text{otherwise}\end{cases}, \tag{3}$$

$$h(e)=\begin{cases}1, & e=0\\ -1, & e=1\\ 0, & \text{otherwise}\end{cases}. \tag{4}$$

For a two-dimensional (2D) image $I(x,y)$, the process of wavelet decomposition is as follows. First, perform 1D-DWT row by row in $I(x,y)$, and then perform 1D-DWT column by column in the row transformed image, and finally four components of $I(x,y)$ are obtained. The four components of $I(x,y)$ are the approximate component cA, the horizontal component cH, the vertical component cV, and the diagonal component cD. Among them, cA is the low-frequency component $L(x,y)$ of $I(x,y)$. cH, cV, and cD are the high-frequency components $H(x,y)$ of $I(x,y)$. Fig. 2 plots the results of the image subjected to 2D-DWT.

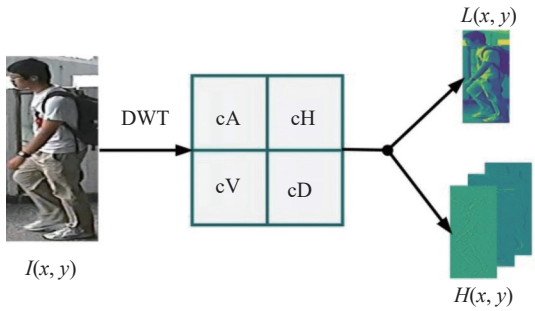


Fig. 2 Image $I(x,y)$ for 2D-DWT results graph

3.3 Progressive augmentation network

Using exclusively LR images as input can have a considerable impact on recognition accuracy given the critical absence of detailed high-frequency information about the person in LR image. This is typically overcome by performing a resolution tune-up of LR probes prior to the recognition task. Unlike the general case where a single-stage CNN is used for SR reconstruction, it is a two-stage CNN that we employ for this work. The reason is that it is hard to coordinate the enhancement of color and texture components of the image by merely one CNN. In its resolution promotion, WT is utilized to reinforce the image at two levels. We put forward a two paths, two-stage network called PEN. The two paths are intended to reinforce the appearance and high-frequency crucial features of the person respectively and the two stages further enrich the personal information. Thus, it alleviates the issue of resolution mismatch between the query set and the gallery set. We then introduce PEN in details.

As shown in Fig. 1, the appearance enhancer (AE), the detail enhancer (DE), and the overall enhancer (OE) are composed of the PEN. The ultimate intention of DSCWT is to highlight the absence of detailed information in LR images in the query set, thus boosting the image resolution for matching with HR images in the gallery set. We use a cascading approach to connect PEN and HRN. This is done by the following steps. Firstly, for input image pair $\{I_{\text{lr}}, I_{\text{hr}}, p^i\}$, WT is used to decompose the I_{lr} to the high frequency F_{lr}^H and the low frequency F_{lr}^L, and the I_{hr} is also decomposed to F_{hr}^H and F_{hr}^L. Secondly, F_{lr}^L and F_{lr}^H are input to AE and DE networks respectively. Next, the AE and OE outputs are amplified to the desired size and composed to I_{lr}. Then the coupling results are fed into OE to obtain SR image I_{sr}. Finally, the validity information in I_{hr} is gradually learned and integrated into I_{sr} to obtain $\hat{p}^i$. The complete operation is depicted as follows: inspired by time-frequency analysis [31], we resort to WT, which decomposes the LR image and the HR image into cA, cH, cV, and cD, correspondingly.

$$[\text{cA},\text{cD},\text{cH},\text{cV}]=\text{DWT}(I(x,y),'\text{Harr}'), \tag{5}$$

$$\begin{cases}L(x,y)=\text{cA}\\ H(x,y)=[\text{cH},\text{cV},\text{cD}]\end{cases}, \tag{6}$$

where Harr means Harr wavelet.

The high-frequency portion reflects the partial detail information of the image which can be empowered with key features of pedestrians to capture discriminative pedestrians features. Realizing the reinforcement of high-frequency minutiae, we designate the DE following a residual knowledge learning strategy. It not only guarantees the consistency of contextual information and semantics but also is intended to perform the mapping relationship between the high-frequency detail component $F_{\text{lr}}^{H(x,y)}$ of LR images and the high-frequency detail component $F_{\text{hr}}^{H(x,y)}$ of HR images. We adopt L1 parametric as the loss function of DE.

$$L_H=\frac{1}{n}\sum_{i=1}^{n}\left|F_{\text{hr}(i)}^{H(x,y)}-F(F_{\text{lr}(i)}^{H(x,y)},\omega)\right| \tag{7}$$

where $F(\cdot,\omega)$ is the DE mapping function with parameters.

The low-frequency portion indicates the worldwide spontaneity of the image is chosen to emphasize the appearance of the person while preserving the identity invariance information. Motivated by the objective toward intensification of LR image appearance features,

we conceive an appearance enhancer. It is adopted to find the mapping relationship between the low-frequency appearance component $F_{\mathrm{lr}}^{L(x,y)}$ of LR image and the low-frequency appearance component $F_{\mathrm{hr}}^{L(x,y)}$ of HR image. Moreover, the mean square error (MSE) is served as the loss function of AE, which is formulated as follows:

$$L_L = \frac{1}{n}\sum_{i=1}^{n}\left(F_{\mathrm{hr}(i)}^{L(x,y)} - G(F_{\mathrm{lr}(i)}^{L(x,y)}, \theta)\right)^2 \tag{8}$$

where $G(\cdot,\theta)$ is the AE mapping function with parameters.

AE and DE are paralleled to reconstruct the SR images. However, the single-stage CNN training is not adequate so that LR images could be elevated to high-quality HR images, yielding unsatisfactory reinforcement performances. Therefore, we insert a second-stage CNN to be further subtle LR image. The outputs of AE and DE are merged as input, and the multi-scale structural similarity (MS-SSIM) [32] loss function is selected as the loss function of OE. It is easier to reserve pedestrian details and edge information as well as to acquire the reconstruction loss of the SR image to HR image, rendering the SR image equivalent to real HR image.

$$L_{\mathrm{rec}} = \mathrm{MS}-\mathrm{SSIM}(I_{\mathrm{sr}}, I_{\mathrm{hr}}) \tag{9}$$

A detailed description and calculation procedure of the MS-SSIM loss function can be found in [32].

3.4 HRN

Consider that state-of-the-art person re-identification models will only be capable of delivering favorable recognition effects when the model is employed on ideal HR images. Even if LR images are trained on the model, the retrieval function is diminished due to the resolution mismatch between the query set and gallery set images. What is the essential point to settle the above matter is how it is possible to seize the meaningful pedestrian information from LR images, which is analogous to that in HR images. According to the aforesaid challenges, we draw up a novel network called HRN. It implements two primary functions: (i) mitigating the inconsistency of data distri-bution between the query set LR images and the gallery set HR images; (ii) retrieving the distinguishable person information while preserving the identity information.

As for $\mathrm{EFEB}_{\mathrm{sr}}^{i}$ and $\mathrm{EFEB}_{\mathrm{hr}}^{i}$, the corresponding generated feature maps for I_{sr} and I_{hr} are f_{sr}^{i} and f_{hr}^{i}. Owing to the shared background of pedestrians filmed within the identical camera, when in situations that are detrimental to the recognition task, such as indigent weather conditions like cloudy days with poor lighting, it seems that the model may prefer to concentrate on the common background areas in the images. It is a fact that various pedestrians are identified as the same person, which deteriorates the recognition accuracy. An attention module is integrated to address this problem. It enables the feature extraction module to concentrate further on the pedestrian part from the image while excluding the interference from other factors such as background. The final feature map is

$$\begin{cases} f_{\mathrm{sr}}^{i'} = \omega_{\mathrm{sr}}^{i} \times f_{\mathrm{sr}}^{i} \\ f_{\mathrm{hr}}^{i'} = \omega_{\mathrm{hr}}^{i} \times f_{\mathrm{hr}}^{i} \end{cases} \tag{10}$$

where ω_{sr}^{i} and ω_{hr}^{i} are obtained from the global average pooling (GAP) and 1×1 kernel size convolutional (Conv) layers. The formula for GAP, ω_{sr}^{i}, and ω_{hr}^{i} are as follows:

$$f_{\mathrm{GAP}}^{d} = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H} f_{d,i,j}, \tag{11}$$

$$\omega_{\mathrm{sr}}^{i} = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(f_{\mathrm{sr}}^{i}))), \tag{12}$$

$$\omega_{\mathrm{hr}}^{i} = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(f_{\mathrm{hr}}^{i}))), \tag{13}$$

where $f_{d,i,j}$ is the pixel value of row i and column j of the feature map in channel d. σ is the sigmoid activation function. The general calculation formula of various convolution layers is as follows:

$$f = \mathrm{relu}\left(\sum_{d=1}^{D}\sum_{i=1}^{W}\sum_{j=1}^{H}\omega_{d,i,j} \times f_{d,i,j} + \omega_{\mathrm{bias}}\right), \tag{14}$$

$$\mathrm{relu}(f) = \max(0, f), \tag{15}$$

where $\omega_{d,i,j}$ and ω_{bias} are respectively convolution kernel weight and bias, and relu(·) is the activation layer function.

The HRN obtained from EFEB stacking has brilliant feature learning power. Nevertheless, accounting for the diversity in the learning capability of the network for dissimilar inputs, there is an inter-domain shift in I_{sr} compared to I_{hr}. It also means that I_{sr} fails to guarantee that the network will necessarily pick up sufficient discriminative information about pedestrians. We thus engineer a residual knowledge loss function, which attunes the output to shrink the inter-domain discrepancy between SR images and HR images by evaluating the residual information between the two streams of SR images and HR images. It learns invariant feature representations with different resolutions in the presence of domain shifts, which allows further refinement of person features. It is specified that the loss function of residual information is as follows:

$$L_{\mathrm{res}} = \frac{1}{K}\sum_{i=1}^{K}\left\|f_{\mathrm{sr}}^{i'} - f_{\mathrm{hr}}^{i'}\right\|_2^2. \tag{16}$$

Ultimately, the label information obtained from the

training data is exploited for the cross-resolution person re-identification task. We perform the standard cross-entropy loss function, which is designed to accelerate the performance of person identity classification. The cross-entropy loss function is formulated below:

$$L_{\text{id}} = \sum_{i=1}^{C} p^i \log_2(\hat{p}^i) + (1-p^i)\log_2(1-\hat{p}^i) \tag{17}$$

where p^i is the ground truth label of the selected sample, and $\hat{p}^i$ is the predicted probability of the sample.

3.5 Loss function

Ultimately, the overall loss function of the DSCWT can be depicted as

$$L_{\text{total}} = L_{\text{id}} + \lambda_H L_H + \lambda_L L_L + \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{res}} L_{\text{res}} \tag{18}$$

where, $\lambda_H, \lambda_L, \lambda_{\text{rec}}$, and λ_{res} are hyper-parameters whose relevance is measured to control the significance of the corresponding loss functions. Alternatively, we remarke that L_H, L_L, and L_{rec} jointly renew the PEN which is spared for LR image resolution enrichment task. While L_{res} and L_{id} are conceived to refresh the HRN that implements the person identification task. Algorithm 1 gives a summary of the entire network training process.

Algorithm 1 Training procedure of the proposed method

Input: Training data $\{I^i_{\text{lr}}, I^i_{\text{hr}}, p^i\}$

Output: A cross-resolution person re-identification model composed of PEN and HRN

Initialization: Batchsize B; Training epoch E; Learning rate lr;

Phase 1 (Preparation)

Take $\{I^i_{hr}, p^i\}$ as input

for i=1 to B **do**

Update the single HRN with the loss L_{id} with (16)

end for

Phase 2 (Joint PEN and HRN learning)

Take $\{I^i_{\text{lr}}, I^i_{\text{hr}}, p^i\}$ as input

Import the first phase trained HRN module

for i=1 to B **do**

Update the joint PEN and HRN learning loss with the loss L_{total} with (17)

Optimization:

1. Update the model parameter ω of the detail enhancer $F(,\omega)$ by Adam with the loss L_H with (6)

$m \leftarrow \beta_1 m + (1-\beta_1)\nabla_\omega L_H$ (Update biased first moment estimate)

$v \leftarrow \beta_2 v + (1-\beta_2)\nabla^2_\omega L_H$ (Update biased second raw moment estimate)

$\hat{m} \leftarrow m/(1-\beta_1)$, $\hat{v} \leftarrow v/(1-\beta_2)$ (Correction term)

$\omega \leftarrow \text{Adam}(\nabla_\omega L_H, \text{lr}) \leftarrow \omega - \text{lr}/(\sqrt{\hat{v}}+\epsilon)\cdot \hat{m}$ (Update ω)

2. Update the model parameter θ of the appearance enhancer $G(,\theta)$ by Adam with the loss L_L with (7)

$m \leftarrow \beta_1 m + (1-\beta_1)\nabla_\theta L_L$ (Update biased first moment estimate)

$v \leftarrow \beta_2 v + (1-\beta_2)\nabla^2_\theta L_L$ (Update biased second raw moment estimate)

$\hat{m} \leftarrow m/(1-\beta_1)$, $\hat{v} \leftarrow v/(1-\beta_2)$ (Correction term)

$\theta \leftarrow \text{Adam}(\nabla_\theta L_L, lr) \leftarrow \theta - lr/(\sqrt{\hat{v}}+\epsilon)\hat{m}$ (Update θ)

end for

4. Experiments

4.1 Datasets and evaluations

4.1.1 MLR-Market1501

MLR-Market1501 is a simulated multi-resolution dataset of 1 501 people captured by six cameras containing 32 668 images. This dataset is simulated from the Market1501 [33] dataset by randomly selecting a scale from $\{1/2, 1/3, 1/4\}$ to downsample all images under one camera, while the images of the same person under another view are not changed. Moreover, the original training set and test set division criterion is retained, i.e., 751 person identities are used for training, and 750 person identities are used for testing.

4.1.2 MLR-DukeMTMC-REID

MLR-DukeMTMC-REID is a simulated multi-resolution dataset of 1 404 people captured by eight cameras, containing 36 411 images. This dataset is simulated from the DukeMTMC-REID [34] dataset in the same way as MLR-Market1501. The partitioning criterion of the original dataset is retained, and 720 pedestrian identities are included in the training and test sets respectively.

4.1.3 MLR-CUHK03

MLR-CUHK03 is a simulated multi-resolution dataset of more than 13 000 images taken by five pairs of cameras, containing 1 367 people. This dataset is simulated from CUHK03 [35], which is the same as MLR-Market1501 while keeping the original test protocol. That is, 1 367 pedestrians are randomly selected as the training set and 100 pedestrians as the test set.

4.1.4 CAVIAR

CAVIAR [36] is a real-world multi-resolution dataset containing 72 pedestrians, with a total of 1 220 images, captured by one HR and one LR camera. We use only 1 000 images from 50 of these people, and then perform a non-overlapping partition of the dataset to obtain a query set of LR images and a gallery set of HR images.

We employ the cumulative match characteristic (CMC) [37] at Rank1 and Rank5 as well as the mean average precision (mAP) [33] to evaluate all methods.

CMC (also known as rank-n matching accuracy), indicates the accuracy of the top-n images of the search results.

$$\mathrm{CMC}(n)=\frac{1}{N}\sum_{i=1}^{N}\begin{cases}1, & n_i\leqslant n\\ 0, & n_i>n\end{cases} \tag{19}$$

where n_i denotes the nth matching result for the ith pedestrian. When n=1 or n=5, the accuracy of the first or the first five images in the test set and the query set being the same label is calculated.

mAP indicates that multiple classes are used to measure the average retrieval function, which is given by

$$\mathrm{mAP}=\frac{\sum_{i=1}^{n}\mathrm{AP}_i}{C}$$

where AP_i denotes the accuracy of each class and C denotes the number of classes. In addition, the statistics of the dataset applied are plotted in Fig. 3, and the representative images are plotted in Fig. 4.

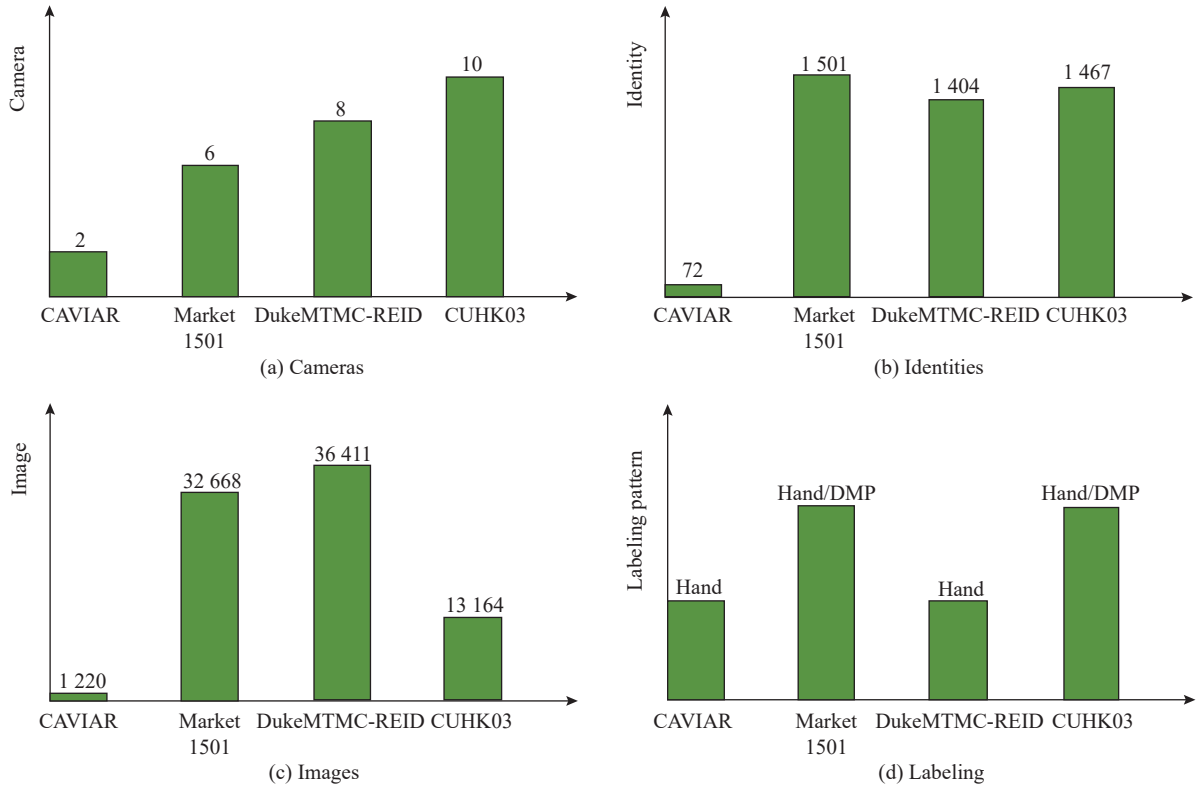


Fig. 3 Fundamental facts in four datasets about the number of cameras, number of identities, number of images, and annotation method

Fig. 4 Examples of HR (the first row) images and LR (the second row) images of people from four datasets

4.2 Implementation details

The PEN is comprised of a U-net [38], a residual concatenation module, and a simplified CNN module. Three step convolutions and three transpose convolutions are employed in U-net for feature map sampling of LR images, which expands the perceptual field while preventing information loss. The CNN module again adopts the residual connection structure with batch normalization (BN) [39] layers, which consists of six convolutional layers to ensure that the input information can be propagated through all parameter layers. The HRN is built by overlaying the efficient feature extraction block EEFB with a backbone of ResNet50 [40] pre-trained on ImageNet.

It is trained through the Pytorch framework. Along with random horizontal flipping and shifting, all input images are adjusted to 257×129×3 with the combined loss function of the PEN and HRN. We utilize the Adam optimizer [41]. The epoch is set to 60 and the batchsize is set to 16. Initial learning rate lr is set to 2e-04, which decreases to 2e-05 and 2e-06 at the 20th epoch and 40th epoch. The other hyperparameters are set to: $\lambda_H = 0.02$, $\lambda_L = 0.03$, $\lambda_{\text{rec}} = 0.1$, and $\lambda_{\text{res}} = 0.05$. All experiments involved in our method are done on a machine with a GeForce RTX 2 070 and an Intel® Core™ i7-9700K CPU @3.60 GHz × 8. There is a time cost for training the model as follows: In the case of MLR-Market1501, the training time is approximately five hours. As for CAVIAR, the training time is in the range of half an hour. For MLR-CUHK03, the training time is around eight hours. To MLR-DukeMTMC-REID, the training time is within about seven hours.

4.3 Ablation study

4.3.1 Effectiveness of PEN

We evaluate the validity of PEN on the dataset MLR-CUHK03, where the experimental data are shown in Table 1. Initially, we hold the HRN and transpose the PEN into super resolution convolutional neural networks (SRCNN) and very deep super resolution (VDSR) respectively. Meanwhile, to validate the necessity of the two-stage enhancement of PEN, we add a set of validation experiments on the PEN variant, which contains only the first stage of person appearance and detail enhancement. We reveal that Rank1 degrades from 86.0% to 69.0% when merely a single CNN enhancement network is employed to enhance the network. This indicates that the two-stage enhancement can effectively boost the resolution of LR images so that it can be better adapted to the person re-identification task. Alternatively, we obtain a dramatic advantage of the proposed method when comparing with other SR models. It demonstrates that our model plays a promising role in reinforcing the detailed information in LR images. We in turn briefly verify the performance of PEN using peak signal-to-noise ratio (PSNR), an image quality evaluation metric. If we consider Table 2, it may be that the PSNR of PEN is not the highest. We speculate that the potential reason is that SR itself is a pathological problem in which high PSNR in no way implies the effectiveness of the SR algorithm. Moreover, instead of the SR reconstruction task, our aim is to render the boosted resolution LR images which are tailored better to the recognition task.

Table 1 Performance of each SR module on MLR-CUHK03

Method	Rank1/%	Rank5/%	mAP/%	PSNR/dB
SRCNN [23]	83.8	91.2	78.6	23.447 0
VDSR [42]	84.9	91.3	80.6	24.724 4
PEN w/o OE	69.0	82.7	63.8	20.330 0
PEN	86.0	93.3	82.3	20.342 8

Table 2 Performance of various feature extraction blocks on MLR-CUHK03 %

Method	Rank1	Rank5	mAP
Single ResNet50	82.4	89.3	78.5
Two ResNet50	83.2	90.2	78.7
HRN	86.0	93.3	82.3

4.3.2 Effectiveness of HRN

Equally, we authenticate the efficiency of HRN on the dataset MLR-CUHK03. The experimental results are listed in Table 2. The first step is to anchor the PEN and permute the EFEB in HRN into the following feature extraction blocks: (i) a single ResNet50; (ii) a dual ResNet50 with HR images for supervision. This is followed by comparing the results obtained from the above feature extraction block on MLR-CUHK03 against the ones obtained from EFEB. It is observed that for a single ResNet50 baseline, Rank1 and mAP are 82.4% and 78.5%, separately. Meanwhile, for the supervised dual ResNet50, Rank1 and mAP respectively are 83.2% and 78.7%. We perform notably well compared to the better-resulting dual ResNet50 method, with 2.8% and 3.6% higher Rank1 and mAP, correspondingly. The reasons for the proposed approach to work well are perhaps: firstly, joining the attention module to highlight the pedestrian region; secondly, L_{res} is gained to narrow the feature gap between LR images and HR images, resulting in weakening the likelihood of inconsistent data distribution.

4.3.3 Effectiveness of the loss functions

We execute ablation experiments on the dataset MLR-CUHK03, which is a tool designed to prove the effectiveness of the proposed loss function. The accuracy scores of Rank1, Rank5, and mAP in the cross-resolution recognition task are presented in Table 3. In Table 3, w/o represents “without”.

Table 3 Ablation experiments to confirm the effectiveness of each loss function on MLR-CUHK03 %

Method	Rank1	Rank5	mAP
Proposed method w/o L_L	84.5	89.9	80.4
Proposed method w/o L_H	84.2	91.4	80.6
Proposed method w/o L_{rec}	83.3	90.9	79.5
Proposed method w/o L_{res}	85.0	90.6	79.9
Proposed method w/o L_{id}	1.0	3.1	3.1
Proposed method	86.0	93.3	82.3

(i) Image recovery constraint function L_{rec}

When L_{rec} is removed from the model, the PEN runs out of valid loss functions from which to constrain the graph reconstruction task. This not only has a deleterious effect on the image quality, but also the potential loss of discriminative information for pedestrian features captured by the EFEB module on subsequent recognition tasks.

(ii) The high-frequency constraint function L_H and the low-frequency constraint function L_L

When L_H and L_L are discarded, we observe that the high-frequency and low-frequency components of the HR image will no longer be available to supervise the high-frequency and low-frequency components of LR image. It makes that the PEN not allowed to recover the detailed sections of the pedestrians from the real-world HR image, and thus has an insufficient power to restore the detailed portion of LR image.

(iii) Residual information constrains the function L_{res}

It is assumed that throwing away L_{res} yields the following consequences. There is a pronounced slowdown in performance throughout the task, which implies that our model does not earn resolution-invariant person features, rendering the mismatch of resolutions in LR image in the query set and HR image in the gallery set intact.

(iv) Identity loss function L_{id}

Should L_{id} be thrown away, then our model can only accomplish the image SR reconstruction task. With respect to personal identification, there is no screening information capability from our model resulting from the shortage of identity labels in the training process. It turns out that the model suffers a significant performance degradation in performing the person identification process.

4.3.4 Effectiveness of WT

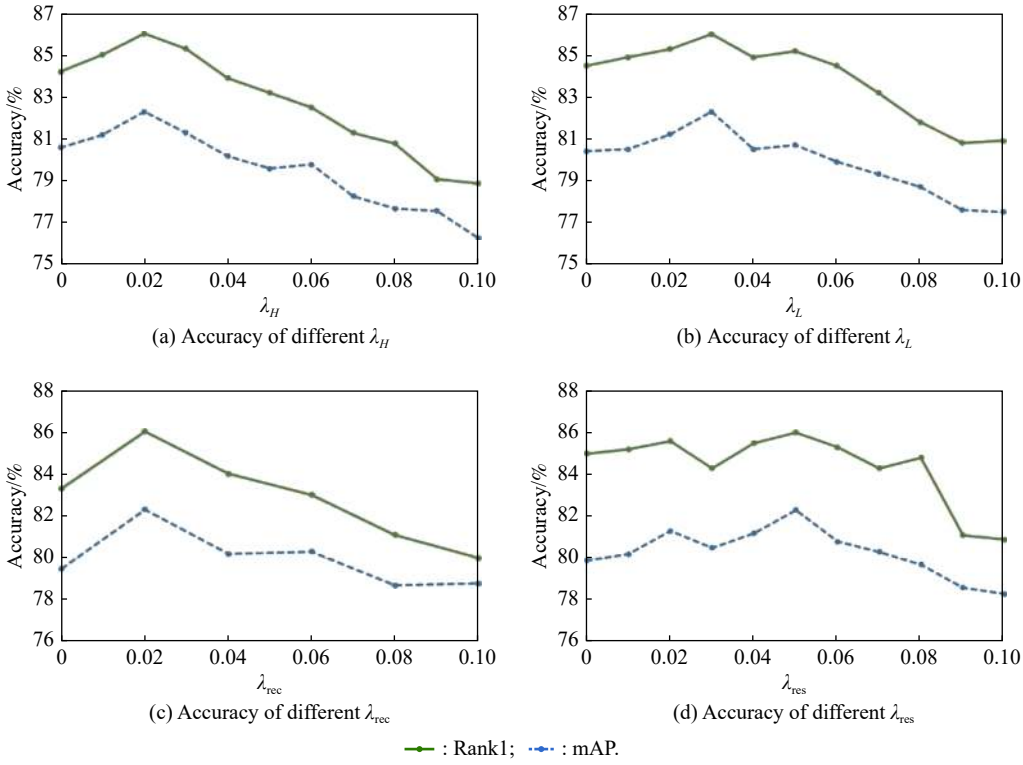
In this subsection, we verify the impact of WT on LR person re-identification task from the perspective of different network inputs. The experimental results on the MLR-CUHK03 dataset are shown in Table 4. We divide the network input into $I_{\text{lr}}, I_{\text{lr}}+F_{\text{lr}}^L, I_{\text{lr}}+F_{\text{lr}}^H$, and $I_{\text{lr}}+F_{\text{lr}}^L+F_{\text{lr}}^H$. When the input is I_{lr}, Rank1 is 79.8%. We can conclude that low resolution images are unfavorable for pedestrian identification. When the input is $I_{\text{lr}}+F_{\text{lr}}^L$ and $I_{\text{lr}}+F_{\text{lr}}^H$, Rank1 is 84.9% and 83.2%, respectively. When the input is $I_{\text{lr}}+F_{\text{lr}}^L+F_{\text{lr}}^H$, we see that Rank1 is the highest. The above data indicate the following: (i) The rank accuracy of the model with input $I_{\text{lr}}+F_{\text{lr}}^L+F_{\text{lr}}^H, I_{\text{lr}}+F_{\text{lr}}^L$, and $I_{\text{lr}}+F_{\text{lr}}^H$ verifies that the network enhances F_{lr}^L and F_{lr}^H at the same time better than the network enhances F_{lr}^L and F_{lr}^H alone. (ii) The rank accuracy of the I_{lr} input model is compared with that of the other three inputs, indicating that WT is beneficial to the person re-identification task in the case of low resolution.

Table 4 Ablation experiments to confirm effectiveness of WT on MLR-CUHK03 %

Input	Rank1	Rank5	mAP
I_{lr}(noWT)	79.8	84.3	80.5
$I_{\text{lr}}+F_{\text{lr}}^L$	84.9	91.7	80.3
$I_{\text{lr}}+F_{\text{lr}}^H$	83.2	89.6	78.9
$I_{\text{lr}}+F_{\text{lr}}^L+F_{\text{lr}}^H$	86.0	93.3	82.3

4.3.5 Hyperparametric analysis

In this subsection, the role of the selection of hyperparameters $\lambda_H, \lambda_L, \lambda_{\text{rec}}$, and λ_{res} on the model performance is analyzed separately as shown in Fig. 5. And the impact of these parameters on the accuracy of Rank1 and mAP on the MLR-CUHK03 dataset is evaluated. Fig. 5 depicts the results of our experiments. We conclude that the model performance achieves promising results when $\lambda_H = 0.02$, $\lambda_L = 0.03$, $\lambda_{\text{rec}} = 0.1$, and $\lambda_{\text{res}} = 0.05$ holding other hyperparameters constant. It is observed that the accuracy of Rank1 and mAP drops when λ_H and λ_L exceed 0.05, which points out that the model fails to adequately learn pedestrian detail information. In contrast, when the value of λ_{res} is around 0.05, the accuracy of the model is relatively stable. To sum up, the hyperparameters should be placed within a specific range to bring out the peak performance of the network. We only roughly analyze the selection of hyper-parameters, and the model performance may be boosted if the hyperparameters are fine-tuned.



(a) Accuracy of different λ_H (b) Accuracy of different λ_L

(c) Accuracy of different λ_{rec} (d) Accuracy of different λ_{res}

Fig. 5 Analysis of hyperparameters

4.4 Comparisons with the state-of-the-arts

We conduct a comparison among our model and two categories that are relatively advanced. (i) Normal person re-identification models: CamStyle, FD-GAN. (ii) Cross-resolution pedestrian re-identification model: i) Legacy manual features driven model: JUDEA, SLD2L, and SDF. ii) Deep learning approach related model: SING, CSR-GAN, FFSR+RIFE, and RAIN. The performance of the dataset on CamStyle, FD-GAN, CSR-GAN, and FFSR+RIFE is implemented with the settings of the parameters and the code offered in the authors' paper. The performance models of the dataset is directly replicated from [11−14,22,43−46]. The comparative findings on the MLR-CUHK03, MLR-Market1501, MLR-DukeMTMC-REID, and CAVIAR datasets are displayed in Table 5.

Table 5 Results of cross-resolution person re-identification on four datasets %

Model	MLR-CUHK03		MLR-Market1501		MLR-DukeMTMC-REID		CAVIAR	
	Rank1	Rank5	Rank1	Rank5	Rank1	Rank5	Rank1	Rank5
CamStyle [43]	69.1	89.6	74.5	88.6	64.0	78.1	32.1	72.3
FD-GAN [44]	73.4	93.8	79.6	91.6	67.5	82.0	33.5	71.4
JUDEA [14]	26.2	58.0	–	–	–	–	22.0	60.1
SLD^2L [22]	–	–	–	–	–	–	18.4	44.8
SDF [45]	22.2	48.0	–	–	–	–	14.3	37.5
SING [11]	67.7	90.7	74.4	87.8	65.2	80.1	33.5	72.7
CSR-GAN [46]	71.3	92.1	76.4	88.5	67.6	81.4	34.7	72.5
RAIN [13]	78.9	97.3	–	–	–	–	42.0	77.3
FFSR+RIFE [12]	73.3	92.6	82.0	92.0	66.0	78.1	36.4	72.0
Our proposed model	86.0	93.3	83.3	92.9	75.9	86.4	40.4	72.8

We have the following conclusions.

(i) All in all, the DSCWT reveals a rather privileged performance. It is found that the Rank1 of DSCWT is 7.1% ahead of the toughest contender on MLR-CUHK03, the Rank1 on MLR-DukeMTMC-REID is 8.3% beyond the strongest contender, the Rank1 on MLR-Market1501 is 1.3% beyond the strongest contender, and the Rank1 on CAVIAR is approximately similar to the strongest contender.

(ii) As compared with the typical person re-identification model (CamStyle, FD-GAN), the Rank1 of FD-GAN is 73.4% on MLR-CUHK03, which is 13.1% weaker than our model. The Rank1 of CamStyle reaches 69.1%, which is 17.4% smaller than our model. It is indicative that the mismatch of resolution in the query set and gallery set is ignored in the typical person re-identification.

(iii) The deep learning-based approach yields far superior performance versus the traditional hand-designed models (JUDEA, SLD2L, SDF). It can be appreciated that the proposed model attains a crushing competitive strength on all four benchmark datasets.

(iv) It achieves decent performance on each dataset in comparison with other deep learning-based methods (SING, CSR-GAN, FFSR+RIFE, and RAIN). The Rank1 is 83.3% on MLR-Market1501, 75.9% on MLR-DukeMTMC-REID, 86.0% on MLR-CUHK03, and 40.4% on CAVIAR.

5. Conclusions

In response to the mismatched resolution issue between query set LR images and gallery set HR images we introduce a fictitious cross-resolution person re-identification model named DSCWT. The uniqueness of the model is that it employs the concept of component divide-and-conquer, using WT in recovering the crucial of the pedestrian in LR images from coarse to fine information. At the same time, the distribution gap between LR image and real HR image is minimized, and the person features of constant resolution are collected. Considerable experimental findings indicate that the performance of the proposed method is competitive compared to the extant state-of-the-art methods on four benchmark datasets. In addition, the experimental results confirm the validity of the proposed method.

References

[1] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: past, present and future. https://arxiv.53yu.com/abs/1610.02984.

[2] YE M, SHEN J B, LIN G L, et al. Deep learning for person re-identification: a survey and outlook. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 44(6): 2872–2893.

[3] BAI S, TANG P, TORR P H S, et al. Re-ranking via metric fusion for object retrieval and person re-identification. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 740–749.

[4] CHANG X, HOSPEDALES T M, XIANG T. Multi-level factorisation net for person re-identification. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2109–2118.

[5] CHEN Y C, ZHU X, ZHENG W S, et al. Person re-identification by camera correlation aware feature augmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 40(2): 392–408.

[6] LI M X, ZHU X T, GONG S G. Unsupervised tracklet person re-identification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 42(7): 1770–1782.

[7] SHEN Y T, LI H S, YI S, et al. Person re-identification with deep similarity-guided graph neural network. Proc. of the European Conference on Computer Vision, 2018: 486–504.

[8] SONG J, YANG Y, SONG Y Z, et al. Generalizable person re-identification by domain-invariant mapping network. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 719–728.

[9] XIAO T, LI H S, OUYANG W L, et al. Learning deep feature representations with domain guided dropout for person re-identification. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1249–1258.

[10] CHENG Z Y, DONG Q, GONG S G, et al. Inter-task association critic for cross-resolution person re-identification. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2605–2615.

[11] JIAO J, ZHENG W S, WU A, et al. Deep low-resolution person re-identification. Proc. of the AAAI Conference on Artificial Intelligence. 2018, 32(1). DOI: https://doi.org/10.1609/aaai.v32i1.12284.

[12] MAO S, ZHANG S, YANG M. Resolution-invariant person re-identification. https://arxiv.53yu.com/abs/1906.09748.

[13] CHEN Y C, LI Y J, DU X, et al. Learning resolution-invariant deep representations for person re-identification. Proc. of the AAAI Conference on Artificial Intelligence, 2019, 33(1). DOI: http://doi.org/10.1609/aaai.v33i01.33018215.

[14] LI X, ZHENG W S, WANG X, et al. Multi-scale learning for low-resolution person re-identification. Proc. of the IEEE International Conference on Computer Vision, 2015: 3765–3773.

[15] HUANG Y, ZHA Z J, FU X, et al. Real-world person re-identification via degradation invariance learning. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 14084–14094.

[16] WANG X D, ZHENG Z D, HE Y, et al. Progressive local filter pruning for image retrieval acceleration. https://arxiv.53yu.com/abs/2001.08878.

[17] QUAN R J, DONG X Y, WU Y, et al. Auto-reid: Searching for a part-aware convnet for person re-identification. Proc. of the IEEE/CVF International Conference on Computer Vision, 2019: 3750–3759.

[18] ZHANG H Z, LAN C L, ZENG W J, et al. Densely semantically aligned person re-identification. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 667–676.

[19] SUN Y F, CHENG C M, ZHANG Y H, et al. Circle loss: a

unified perspective of pair similarity optimization. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 6398–6407.
[20] FAN H H, ZHENG L, YAN C G, et al. Unsupervised person re-identification: clustering and fine-tuning. ACM Trans. on Multimedia Computing, Communications, and Applications, 2018, 14(4): 83.
[21] ZHONG Z, ZHENG L, LUO Z M, et al. Invariance matters: exemplar memory for domain adaptive person re-identification. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 598–607.
[22] JING X Y, ZHU X, WU F, et al. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 695–704.
[23] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015, 38(2): 295–307.
[24] TONG T, LI G, LIU X J, et al. Image super-resolution using dense skip connections. Proc. of the IEEE International Conference on Computer Vision, 2017: 4799–4807.
[25] LEDIG C, THEIS L, HUSZAR F, et al. Photo-realistic single image super-resolution using a generative adversarial network. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4681–4690.
[26] CHU X X, ZHANG B, MA H L, et al. Fast, accurate and lightweight super-resolution with neural architecture search. Proc. of the IEEE 25th International Conference on Pattern Recognition, 2021: 59–64.
[27] ZHOU R F, SUSSTRUNK S. Kernel modeling super-resolution on real low-resolution images. Proc. of the IEEE/CVF International Conference on Computer Vision, 2019: 2433–2443.
[28] HE Y G, BU X Z, JIANG M, et al. Low bit rate underwater video image compression and coding method based on wavelet decomposition. China Communications, 2020, 17(9): 210–219.
[29] RAMAMONJISOA M, FIRMAN M, WATSON J, et al. Single image depth prediction with wavelet decomposition. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11089–11098.
[30] SHAHREZAEI I H, KIM H C. Fractal analysis and texture classification of high-frequency multiplicative noise in SAR sea-ice images based on a transform-domain image decomposition method. IEEE Access, 2020, 8: 40198–40223.
[31] WIRSING K. Time frequency analysis of wavelet and Fourier transform. MOHAMMADY S. Wavelet Theory. London, UK: IntechOpen, 2020.
[32] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment. Proc. of the IEEE 37th Asilomar Conference on Signals, Systems & Computers, 2003, 2: 1398–1402.
[33] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: a benchmark. Proc. of the IEEE International Conference on Computer Vision, 2015: 1116–1124.
[34] ZHENG Z D, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. Proc. of the IEEE International Conference on Computer Vision, 2017: 3754–3762.
[35] LI W, ZHAO R, XIAO T, et al. Deepreid: deep filter pairing neural network for person re-identification. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 152–159.
[36] CHENG D S, CRISTANI M, STOPPA M, et al. Custom pictorial structures for re-identification. Proc. of the British Machine Vision Conference, 2011. DOI: 10.5244/C.25.68.
[37] WANG X, DORETTO G, SEBASTIAN T, et al. Shape and appearance context modeling. Proc. of the IEEE 11th International Conference on Computer Vision, 2007: 1–8.
[38] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation. Proc. of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015: 234–241.
[39] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proc. of the International Conference on Machine Learning, 2015: 448–456.
[40] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770−778.
[41] KINGMA D P, BA J. A method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980.
[42] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1646−1654.
[43] ZHONG Z, ZHENG L, ZHENG Z D, et al. Camera style adaptation for person re-identification. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5157−5166.
[44] GE Y X, LI Z W, ZHAO H Y, et al. Fd-gan: pose-guided feature distilling gan for robust person re-identification. Proc. of the 32nd International Conference on Neural Information Processing Systems, 2018: 1230–1241.
[45] WANG Z, HU R M, YU Y, et al. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. Proc. of the 25th International Joint Conference on Artificial Intelligence, 2016: 2669–2675.
[46] WANG Z, YE M, YANG F, et al. Cascaded SR-GAN for scale-adaptive low resolution person re-identification. Proc. of the 27th International Joint Conference on Artificial Intelligence, 2018: 3891–3897.

Biographies

SUN Rui was born in 1976. He received his B.S. degree from the Central South University of China, in 1998, M.S. degree from Harbin Engineering University of China, in 2000, and Ph.D. degree from Huazhong University of Science and Technology of China, in 2003. He worked as a senior software engineer with TCL Mobile Communication Company, China, from 2003 to 2005. He was a visiting scholar with the Computer Science Department, University of Missouri, Columbia, MO, USA, from 2010 to 2011. He was a postdoctoral researcher with Chery automobile Company, China, from 2012 to 2014. He is currently a professor with Hefei University of Technology, China. His research interests include object recognition and tracking, computer vision, and machine learning.
E-mail: sunrui@hfut.edu.cn

YANG Zi was born in 1997. She received her B.S. degree from West Anhui University, in 2015. She is currently pursuing her M.S. degree with Hefei University of Technology. Her research interests include object recognition and tracking, computer vision, and image processing.
E-mail: yangzi@mail.hfut.edu.cn

ZHAO Zhenghui was born in 1994. She received her B.S. degree from Hefei University of Technology, in 2015. She is currently pursuing her M.S. degree with Hefei University of Technology. Her research interests include object recognition and tracking, computer vision, and visible-infrared image processing.
E-mail: 904150289@qq.com

ZHANG Xudong was born in 1966. He received his B.S. degree from Hefei University of Technology in 1989, M.S. degree from Hefei University of Technology in 1992, and Ph.D. degree from University of Science and Technology of China in 2005. He spent three months in collaborative research at Heilbronn University in Germany in 2006. Currently, he is a professor at Hefei University of Technology, China. His main research interests include image processing, pattern recognition, and intelligent information processing.
E-mail: xudong@hfut.edu.cn