Resolution-invariant Person Re-Identification

Shunan Mao[1] , **Shiliang Zhang**[1] and **Ming Yang**[2]
[1]Peking University
[2]Horizon Robotics, Inc
{snmao, slzhang.jdl}@pku.edu.cn, ming.yang@horizon-robotics.com

Abstract

Exploiting resolution invariant representation is critical for person Re-Identification (ReID) in real applications, where the resolutions of captured person images may vary dramatically. This paper learns person representations robust to resolution variance through jointly training a Foreground-Focus Super-Resolution (FFSR) module and a Resolution-Invariant Feature Extractor (RIFE) by end-to-end CNN learning. FFSR upscales the person foreground using a fully convolutional auto-encoder with skip connections learned with a foreground focus training loss. RIFE adopts two feature extraction streams weighted by a dual-attention block to learn features for low and high resolution images, respectively. These two complementary modules are jointly trained, leading to a strong resolution invariant representation. We evaluate our methods on five datasets containing person images at a large range of resolutions, where our methods show substantial superiority to existing solutions. For instance, we achieve Rank-1 accuracy of 36.4% and 73.3% on *CAVIAR* and *MLR-CUHK03*, outperforming the state-of-the art by 2.9% and 2.6%, respectively.

1 Introduction

Person Re-identification (ReID) aims to find a probe person from a large-scale person image gallery collected by a camera network. [Li *et al.*, 2019] Person ReID is challenging since it is confronted by many appearance variations due to camera viewpoint, person pose, illumination, background, *etc*. Thanks to the introduction of many benchmark datasets like *VIPeR* [Gray and Tao, 2008], *CUHK03* [Li *et al.*, 2014], *Market1501* [Zheng *et al.*, 2015] and *MSMT17* [Wei *et al.*, 2018], most of these challenges are covered in these datasets, leading to a significant progress in person ReID performance.

Among the above challenges, varying resolutions of person images are probably the most common one, due to the distance to a camera, or camera focus and resolution. Matching persons at different resolutions requires the ReID algorithms to attend to distinct visual cues. For example, Fig. 1 illustrates two instances of three persons sampled from *CAVIAR* [Cheng *et al.*, 2011]. With high resolution image samples, those three persons can be distinguished by their hair styles or strips on the pants. As these details are not available in low resolution images, a ReID method needs to resort to silhouettes or global textures for a reliable matching. Moreover, the high and low resolution samples of the same person may even present a larger discrepancy than to those samples from different persons at a similar resolution. Therefore, dedicated treatments are desired for ReID methods to cope with large resolution variations of person images.

Figure 1: Illustration of 6 images from 3 persons in *CAVIAR* [Cheng *et al.*, 2011]. Person ReID needs to match the same person and discern different persons across different resolutions.

Matching persons at dramatically different resolutions has not been extensively studied, partly because of the limitation of current ReID benchmark datasets. Most widely used benchmark datasets usually consist of person images with limited resolution variations. *CAVIAR* [Cheng *et al.*, 2011] is particularly collected to consider two levels of resolutions. *MLR-VIPeR* and *MLR-CUHK03* [Jiao *et al.*, 2018] are adapted from *VIPeR* [Gray and Tao, 2008] and *CUHK03* [Li *et al.*, 2014] by including three levels of resolutions, respectively. These datasets have inspired many works on low-resolution person ReID [Li *et al.*, 2015; Jing *et al.*, 2015; Wang *et al.*, 2016; Jiao *et al.*, 2018; Wang *et al.*, 2018b], yet not many efforts on how to handle person images with a large range of resolution variance.

Traditional methods [Li *et al.*, 2015; Jing *et al.*, 2015; Wang *et al.*, 2016] address person ReID with varying person resolutions mainly by learning a shared feature space between low and high resolutions. Recent approaches focus on deep learning based Super-Resolution (SR) [Jiao *et al.*, 2018; Wang *et al.*, 2018b]. Although SR methods can recover some visual details, they do not differentiate person foregrounds and backgrounds and are not optimized for person ReID, *i.e.*, their goal is to minimize the pixel-level L2 loss, rather than the person ReID errors. In practice, SR methods are not capable to fully recover the missing details in low resolution images. We argue that the person feature extractor shall be explicitly designed and optimized to combat against challenging resolution variance in real-world scenarios.

This paper proposes to jointly optimize person image resolution and feature extraction for person ReID. Specifically, we propose a deep network consisting of two modules. The Foreground-Focus Super-Resolution (FFSR) module upscales the resolution of an input image using a fully convolutional auto-encoder with skip connections. Different from general SR modules, FFSR is jointly trained with the person ReID loss and a foreground focus loss, which recovers the details on the person body and suppresses the cluttered backgrounds. The subsequent Resolution-Invariant Feature Extractor (RIFE) extracts person representations for person ReID. RIFE consists of several feature learning blocks, each of which adopts two CNN branches to learn features from low and high-resolution images, respectively. This design learns more dedicated feature extractors for low-resolution inputs. In other words, RIFE explicitly differentiates high and low resolution inputs during feature learning to ensure its robustness to resolution variance. Features from those two branches are fused with the weights predicted by a Dual-Stream Block (DSB) as the resolution invariant feature.

By jointly training FFSR and RIFE, our approach achieves consistent improvements on the three multi-resolution ReID datasets, *i.e.*, *CAVIAR* [Cheng *et al.*, 2011], *MLR-VIPeR*, and *MLR-CUHK03* [Jiao *et al.*, 2018]. Besides those three datasets, we also construct two large datasets with large variations of person resolutions, *i.e.*, *VR-Market1501* and *VR-MSMT17* by modifying *Market1501* [Zheng *et al.*, 2015] and *MSMT17* [Wei *et al.*, 2018], respectively. On these two datasets, our method also achieves promising performance. To our best knowledge, this is an original work that jointly considers foreground focus super resolution and multiple CNN branches for resolution invariant representations in person ReID. Extensive ablation studies as well as comparisons on five datasets have shown the competitive performance of the proposed approach.

2 Related Work

This section briefly reviews low-resolution person ReID and image super-resolution, which are closely related to our work.

Low-Resolution Person Re-ID. Some works use matric learning methods to address low-resolution person ReID mainly by learning a shared feature space between low and high resolutions. For example, JUDEA [Li *et al.*, 2015] optimizes the distance between images of different resolutions by requiring features on the same person to be close to each other. SLD^{2}L [Jing *et al.*, 2015] uses the Semi-Coupled Low-Rank dictionary learning to build the mapping between features from low and high-resolution images. SDF [Wang *et al.*, 2016] learns a discriminating surface to separate feasible and infeasible functions in the scale distance function space. [Chen *et al.*, 2019] uses adversarial loss and reconstruction loss to decrease distance between deep features from different resolution. Other works use deep learning based Super-Resolution (SR). CSR-GAN [Wang *et al.*, 2018b] focuses on the super resolution part, and uses a deep Cascaded SR-GAN as well as several handcraft restrictions to enhance the image resolution. SING [Jiao *et al.*, 2018] adds a Super Resolution network before the feature extraction and trains two networks jointly.

Super-Resolution benefits from the advance of deep models. SRCNN [Dong *et al.*, 2014] first introduces a Fully Convolutional Network for image Super-Resolution. Many works [Kim *et al.*, 2016; Tai *et al.*, 2017] have been proposed by designing deeper, wider, and denser network architectures. SRGAN [Ledig *et al.*, 2017] designs additional loss functions to recover more semantic cues. Those works are general SR models and do not concern with image contents. SFT-GAN [Wang *et al.*, 2018a] uses image segmentation to help the texture super resolution. [Yu *et al.*, 2018] use an encoder-decoder structure to leverage attributes and use GAN [Goodfellow *et al.*, 2014] and STN [Jaderberg *et al.*, 2015] to make the generated faces appear realistic.

Different from SING and CSR-GAN, our FFSR focuses on person foreground and RIFE learns different feature extractors for high and low-resolution images. This further enhances the robustness to resolution variance.

3 Problem Formulation

In surveillance videos, a person image can be regarded as a sample of one person captured by a camera, where the resolution is decided by shooting parameters like sensor resolution, shooting distance, camera focus, imaging processor, *etc.* *i.e.*,

$$I_i^r = \text{camera-sample}(P_i, \theta), \tag{1}$$

where I_i^r is a person image with resolution r and image index i. P_i denotes the person ID label of I_i^r, θ denotes the shooting parameters.

It is hard to precisely define the resolution r, because the parameters θ could be complicated. For simplicity, for I_i^r in a dataset $\mathbb{D}$, we use a scalar $r \in [0, 1]$, computed with $width(I_i^r)/width_{max}$ as its resolution, where $width_{max}$ is the maximum width of images on $\mathbb{D}$. For example with $width_{max} = 96$, resizing an original 128×48 sized image to 64×24 degrades its resolution from 0.5 to 0.25. To simplify the definition of resolution, we note that enlarging an image with interpolation does not enhance its resolution.

The task of person ReID can be described as matching a query person against the collected person image dataset using a feature representation f, with the goal of minimizing the distance between images of the same person, meanwhile maintaining larger distances between images of different persons. Considering the variance of image resolution, we denote the objective function $\mathcal{O}$ of person ReID as,

$$\begin{aligned}
&\min_f \ \mathcal{O}(r_1, r_2) = \mathrm{D}_{sim}(r_1, r_2)/\mathrm{D}_{dif}(r_1, r_2), \\
&\mathrm{D}_{sim}(r_1, r_2) \triangleq \sum_{P_i = P_{i'}} \|f_i^{r_1} - f_{i'}^{r_2}\|_2^2, \\
&\mathrm{D}_{dif}(r_1, r_2) \triangleq \sum_{P_i \neq P_{i'}} \|f_i^{r_1} - f_{i'}^{r_2}\|_2^2,
\end{aligned} \tag{2}$$

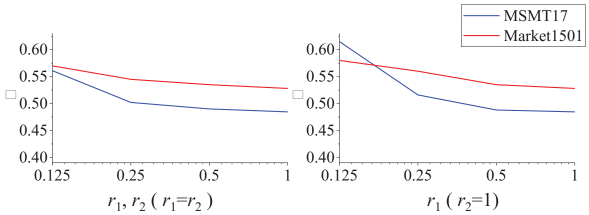

Figure 2: Values of object function $\mathcal{O}$ in Eq. (2) computed with variations of resolution on *MSMT17* and *Market1501*. (a) fixes $r_1 = r_2$ and increase r_1 and r_2 from 0.125 to 1. (b) fixes $r_2 = 1$ and increase r_1 from 0.125 to 1. It verifies that, both low resolution and varied resolution increase the difficulty of person ReID.

where $\|\cdot\|_2^2$ computes the distance between feature vectors. $\mathrm{D}_{dif}(\cdot)$ and $\mathrm{D}_{sim}(\cdot)$ compute the distance between two images of the same person and different persons, respectively. We use superscripts r_1 and r_2 to denote resolutions of two images considered in distance computation.

Before proceeding to the formulation of our algorithm, we first illustrate the effects of resolution variance to person ReID performance on two large ReID datasets *Market1501* [Zheng *et al.*, 2015] and *MSMT17* [Wei *et al.*, 2018], respectively. We first train a ResNet50 baseline [He *et al.*, 2016] as the feature extractor, then compute $\mathcal{O}(r_1, r_2)$ on two datasets with different combinations of r_1 and r_2. Fig. 2 (a) fixes $r_1 = r_2$ and increases their values from 0.125 to 1. We observe that, lower resolution leads to larger $\mathcal{O}$, resulting in a lower person ReID accuracy. Fig. 2 (b) fixes $r_2 = 1$ and increases r_1 from 0.125 to 1. It is clear that, larger variance of resolution corresponds to increased person ReID difficulty. We also observe that, the curves in Fig. 2 (b) are more abrupt than the ones in Fig. 2 (a), indicating that varied-resolution ReID could be more challenging than the low-resolution case.

Our solution is inspired by the above observations, *i.e.*, to improve person ReID accuracy, two compared images should present at 1) high resolution and 2) similar resolution. The person image resolution should be enhanced to recover visual details. To facilitate feature extraction, the SR model is expected to focus on the person foreground and suppress the cluttered backgrounds. Meanwhile, the feature extractor should be able to alleviate the resolution variances. Those two intuitions correspond to two modules in our network, *i.e.*, the Foreground Focus Super Resolution (FFSR) and Resolution Invariant Feature Extractor (RIFE), respectively.

For an input person image I_i^r, FFSR first enhances its resolution to r', $r \leq r'$, then it is processed by RIFT for resolution invariant feature extraction. The forward computation of our network can be denoted as,

$$I_i^{r'} = \mathcal{M}^{FFSR}(I_i^r), \;\; f_i = \mathcal{M}^{RIFE}(I_i^{r'}), \qquad (3)$$

where f_i is the final feature, $\mathcal{M}^{FFSR}$ and $\mathcal{M}^{RIFE}$ denote the two modules, respectively.

With a training set $\mathbb{T} = \{(I_i^r, I_i^h, P_i)\}, i = 1, ..., N$, where I_i^h is the groundtruth high-resolution image and P_i is the person ID label, the network is optimized with two losses computed on two modules, *i.e.*,

$$\mathcal{L} = \sum_{i=1:N} \mathcal{L}^{FFSR}(I_i^r) + \alpha \mathcal{L}^{RIFE}(I_i^r), \qquad (4)$$

where α balances the two losses. The following section introduces our network architecture and the implementations of those two loss functions.

4 Proposed Methods

Our network architecture is illustrated in Fig. 3. This section introduces the FFSR and RIFT modules, respectively.

4.1 Foreground-Focus Super-Resolution

As the initial stage before feature extraction, FFSR model should be compact and efficient to compute. Additionally, FFSR is expected to work with varied resolutions, *e.g.*, perform super-resolution to low resolution inputs, and preserve original details of high resolution inputs.

Instead of following existing SR models [Kim *et al.*, 2016; Tai *et al.*, 2017], we use a light-weight FFSR module illustrated in Fig. 3. FFSR is implemented based on the auto-encoder architecture. The first several convolutional layers down-sample the input with stride width 2. Then, small convolutional kernels with stride width 1 are applied for feature extraction. Following the RED-net [Mao *et al.*, 2016] and U-net [Ronneberger *et al.*, 2015], we add symmetric skip connections between low and high layers. Skip connects could preserve the visual cues in original images, hence help to enhance the quality of reconstructed images.

Pixel-wised distance Mean Square Error (MSE) is commonly applied for SR model training. Simply minimizing the MSE may not be optimal for person ReID task, because it does not differentiate person foregrounds and backgrounds. Person foregrounds generally provide more valuable cues for person ReID. To recover more visual cues on person foregrounds and depress cluttered backgrounds, we propose the foreground-focus SR loss $\mathcal{L}_{FFSR}$, *i.e.*,

$$\mathcal{L}^{FFSR}(I_i^r) = \|M \odot (I_i^{r'} - I_i^h)\|_2^2, \qquad (5)$$

where $\odot$ denotes element-wise multiply and M is a mask with the same size of I_i^r.

Our method is compatible with different mask generation strategies. Image segmentation algorithms like [Insafutdinov *et al.*, 2016] can be applied to generate binary foreground masks. With a well-trained person bounding box detector, person foregrounds are more likely to appear in the center of bounding boxes. For simplicity, Gaussian kernels can be applied as foreground masks, as illustrated in Fig. 3.

4.2 Resolution-Invariant Feature Extractor

Since super-resolution is an ill-posed problem, solely applying FFSR is not strong enough to achieve resolution invariance. We further design RIFE to generate resolution invariant features. As illustrated in Fig. 1, high and low-resolution images convey substantially different amount of visual cues, they should be treated with different feature extractors. RIFE explicitly differentiates high and low resolution images into two feature extraction streams. As shown in Fig. 3, RIFE consists of several Dual-Stream Blocks (DSB), each introduces two feature extraction streams with an identical architecture but different training objectives. The following part first introduces the forward procedure of RIFE, then discusses its training objectives.

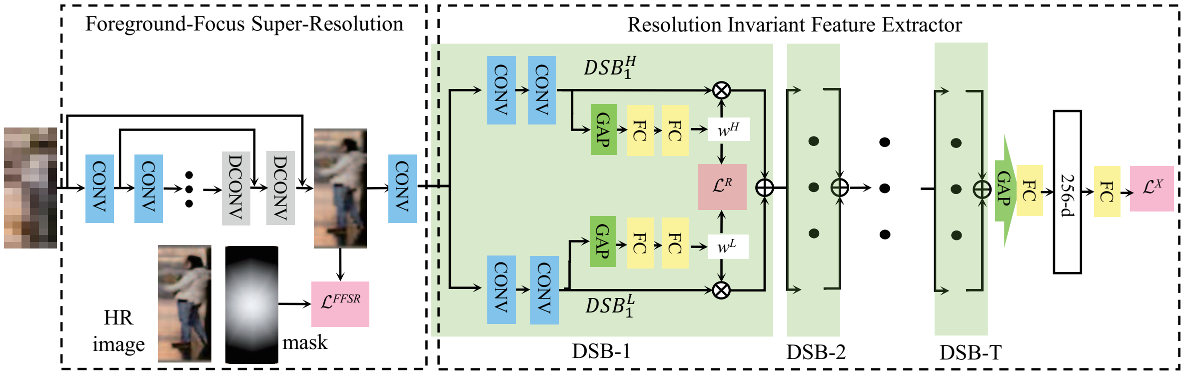

Figure 3: The architecture of our network, which consists of two modules: Foreground-Focus Super-Resolution (FFSR) and Resolution-Invariant Feature Extractor (RIFE). FFSR is an auto-encoder with skip connections trained with both person ReID loss and a foreground-focus super resolution loss $\mathcal{L}^{FFSR}$. RIFE consists of several Dual-Stream Blocks (DSB), each progressively learns resolution invariant features through two CNN streams. Features from two streams are fused with weights learned by a resolution weighting loss $\mathcal{L}^R$. RIFE finally outputs a 256-d feature, which is hence used to compute the cross-entropy loss $\mathcal{L}^X$.

In RIFE, each DSB applies two streams of convolutional layers to extract feature maps for high and low-resolution inputs, respectively. For the t-th DSB, we denote its two streams as DSB_t^L and DSB_t^H, and their generated feature maps as $\mathbf{m}_t^L$ and $\mathbf{m}_t^H$, where superscripts L and H denote the low and high-resolution streams, respectively. $\mathbf{m}_t^L$ and $\mathbf{m}_t^H$ are adaptively fused as the output of the DSB to achieve better robustness to resolution variance. For example, $\mathbf{m}_t^L$ is fused with larger weights for low-resolution images, because DSB_t^L is more suited for feature extraction on low-resolution images. We denote the computation of output feature $\mathbf{m}_t$ of t-th DSB as,

$$\mathbf{m}_t = w_t^L \times \mathbf{m}_t^L + w_t^H \times \mathbf{m}_t^H, \tag{6}$$

where w_t^L and w_t^H are related to the resolution of input image. For high-resolution images, w^L would be smaller than w^H, and vice versa. As shown in Fig. 3, those two weights are predicted with two FC layers based on $\mathbf{m}_t^L$ and $\mathbf{m}_t^H$, respectively.

In order to learn w_t^L and w_t^H, we introduce the resolution weighting loss $\mathcal{L}^R$ into each DSB. With a training image I_i^r, the $\mathcal{L}_t^R$ for the t-th DSB is defined as,

$$\mathcal{L}_t^R(I_i^r) = \|w_t^L - (1-r)\|_2^2 + \|w_t^H - r\|_2^2, \tag{7}$$

where r denotes the resolution of I_i^r.

The fused feature map $\mathbf{m}_t$ is propagated to the next DSB. Stacking multiple DSBs leads to a deep neural network with strong feature learning capability. The output of final DSB is processed with a Global Average Pooling (GAP) layer and a Fully Connected (FC) layer as the final feature f. A FC layer is trained on f to predict person ID labels. A cross entropy loss can be computed as the person ReID loss, *i.e.*,

$$\mathcal{L}^X(I_i^r) = CrossEntropy(\text{FC}(f_i), P_i), \tag{8}$$

where P_i denotes the person ID label of a training image I_i^r.

With T DSBs in total, RIFE is trained with one cross entropy loss and T resolution weighting losses. The RIFE loss on training image I_i^r can be represented as

$$\mathcal{L}^{RIFE}(I_i^r) = \mathcal{L}^X(I_i^r) + \beta \sum_{t=1:T} \mathcal{L}_t^R(I_i^r), \tag{9}$$

where parameter β weights the two losses.

Fusing features with Eq. (6) enforces DSB^L and DSB^H to focus on low and high-resolution images during training. For low-resolution images, the back propagated person ReID loss makes more modifications to DSB^L than to DSB^H because of larger w^L. This mechanism finally learns different parameters for DSB^L and DSB^H, respectively and leads to a strong resolution invariant representation. Implementation details of RIFE will be presented in Sec. 5.2.

5 Experiment

5.1 Datasets

We evaluate our methods on five datasets, including three existing datasets and two datasets we constructed.

CAVIAR [Cheng *et al.*, 2011] contains 1220 images of 72 identities. Images are captured by one High-Resolution (HR) camera and one Low-Resolution (LR) camera. Among 72 identities, 50 have images from two cameras. Those 50 identities are divided into a LR query set and a HR gallery set.

MLR-VIPeR and *MLR-CUHK03* are constructed on *VIPeR* [Gray and Tao, 2008] and *CUHK03* [Li *et al.*, 2014] datasets, where both were captured by two cameras. Following S-ING [Jiao *et al.*, 2018], every image from one camera is down-sampled with a ratio evenly selected from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$ as the query set. Original images from the other camera are used as the test set. *MLR-VIPeR* has 316 identities for training and testing. *MLR-CUHK03* has 100 identities for testing and 1367 for training, respectively.

VR-Market1501 and *VR-MSMT17* are constructed by us based on *Market1501* [Zheng *et al.*, 2015] and *MSMT* [Wei *et al.*, 2018], respectively. *VR-Market1501* contains 32,217 images of 1,501 people captured by 6 cameras. *VR-MSMT17* consists of 126,441 images of 4,101 persons from 15 cameras. All images are down-sampled to make the width within the range of [8, 32) in *VR-Market1501* and [32,128) in *VR-MSMT17*. Hence these two datasets present 24 and 96 different resolutions separately. We keep original divisions of training and testing sets, *i.e.*, 751 and 710 identities for

SR model	Mask	JL	Rank-1	Rank-5	FLOPs(G)
Bilinear	-	-	47.6	66.4	-
SRCNN	-	-	46.7	65.0	2.53
VDSR	-	-	48.5	67.7	32.79
AE	-	-	48.1	67.1	3.71
FFSR	Gaussian	-	49.4	67.8	3.71
FFSR	Gaussian	√	52.8	69.0	3.71
FFSR	Deepercut	√	**52.9**	**69.2**	3.71

Table 1: ReID Performance of different SR methods on *VR-MSMT17*. FLOPs shows the SR complexity. JL denotes joint learning with ReID loss.

structure	weight learning	Rank-1	Rank-5
ResNet50	-	47.6	66.4
two ResNet50	-	49.1	68.2
two ResNet50	√	50.4	67.4
RIFE	√	**53.3**	**70.1**

Table 2: Performance of different feature extractors on *VR-MSMT17*

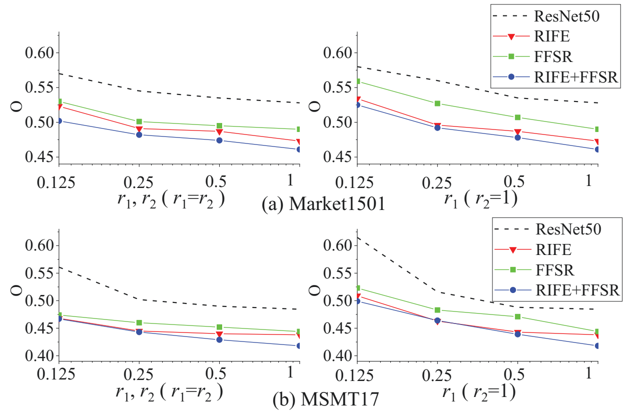

Figure 4: Effects of FFSR and RIFT to the object function $\mathcal{O}$ in Eq. (2). This figure follows the configurations of Fig. 2. It is clear that FFSR and RIFE boost the robustness to resolution variance.

training and testing on *VR-Market1501*, 1,041 and 3,060 for training and testing on *VR-MSMT17*. Compared with existing datasets, *VR-Market1501* and *VR-MSMT17* are substantially larger in size and are more challenging, because both query and gallery images show a large range of resolution variance.

5.2 Implementation Details

Our FFSR module is a 12-layer fully convolutional network. Two convolutional layers with a stride of 2 and two transposed convolutional layers are applied to down-sample and up-sample the feature maps, respectively. We use ResNet50 [He *et al.*, 2016] as the backbone of RIFE module. Each main block of ResNet50 is modified to be a DSB by duplicating its convolutional layers as DSB^L and DSB^H, respectively. Following ResNet50, our RIFE has 4 DSBs. Two FC layers with the output channels of 64 and 1 are used for w^L and w^H prediction.

Our network is trained on PyTorch by Stochastic Gradient Descent (SGD). Training is finished with three steps. 1) We initialize and pre-train the FFSR model on *ImageNet* [Russakovsky *et al.*, 2015] with the MSE loss. Then it is fine-tuned on person ReID training datasets with $\mathcal{L}^{FFSR}$. 2) We initialize and fine-tune the RIFE module on target dataset with $\mathcal{L}^{RIFE}$. 3) FFSR and RIFE modules are jointly trained with the loss function in Eq. (4). We fix hyperparameters as $\alpha = 1, \beta = 0.1$ for all datasets. Each step has 60 epoches and the batch size is set as 32. The initial learning rate is set as 0.01 at the first two steps and 0.001 at the final step. The learning rate is reduced ten times after 30 epoches. Input images are resized to 256×128 in *VR-market1501* and 384×128 in other datasets. The final 256-D feature is used for ReID with Euclidean distance. All of our experiments are implemented with GTX 1080Ti GPU, Intel i7 CPU, and 128GB memory.

5.3 Ablation Study

Validity of FFSR: To show the validity of our FFSR model, we fix the feature extraction module as ResNet50 and test different super resolution methods including Bilinear interpolation, SRCNN [Dong *et al.*, 2014], VDSR [Kim *et al.*, 2016], as well as variants of our module, *i.e.*, baseline Auto Encoder (AE), FFSR trained with the Gaussian mask and segmented mask by deepercut [Insafutdinov *et al.*, 2016], as well as training with/without person ReID loss. The experiments are conducted on the large *VR-MSMT17*. We illustrate experimental results in Table 1 In Table 1, with the Gaussian mask our method outperforms the baseline AE, indicating the validity of emphasizing the person foreground in SR for person ReID. It is also clear that, jointly training with person ReID loss substantially boosts the ReID accuracy. Conveying more accurate foreground locations, segmented mask further outperforms the Gaussian mask. We also compare the computational complexity of FFSR with other super resolution methods. It can be observed that, FFSR introduces marginal computational overhead to the compact SRCNN [Dong *et al.*, 2014], but shows substantially better performance, *e.g.*, outperforms SRCNN by 6.2% in Rank-1 Accuracy. FFSR also substantially outperforms VDSR in the aspects of both accuracy and complexity.

Validity of RIFE: To show the validity RIFE module, we fix the super resolution module as the Bilinear interpolation and compare RIFE with three feature extractors, *i.e.*, a) ResNet50 baseline, b) two ResNet50 with their features fused with equal weight, and c) two ResNet50 with their features fused with learned weights learned with Eq. (6). We summarize the experimental results in Table 2. Table 2 shows that increasing the amount of network parameters by fusing two ResNet50 only brings marginal improvements over the one branch ResNet50 baseline. Fusing two ResNet50 with learned weights brings 1.3% improvements to the Rank-1 Accuracy. Among compared methods, RIFE module achieves the best performance, substantially outperforming baseline by 5.7% in Rank-1 Accuracy.

Effect to the Objective Function: We further show the effect of FFSR and RIFE to the person ReID objective function defined in Eq. (2). Referring to Fig. 2, we illustrate the effect

Dataset	*CAVIAR*		*MLR-VIPeR*		*MLR-CUHK03*		*VR-Market1501*		*VR-MSMT17*	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Densenet121 [Huang *et al.*, 2017]	31.1	65.5	31.4	63.1	70.8	91.3	60.0	78.8	51.2	67.4
SE-resnet50 [Hu *et al.*, 2018]	30.8	65.1	33.5	63.6	70.8	92.3	58.2	78.6	52.3	68.9
JUDEA [Li *et al.*, 2015]	20.0	60.1	26.0	55.1	26.2	58.0	-	-	-	-
SLD^{2}L [Jing *et al.*, 2015]	18.4	44.8	20.3	44.0	-	-	-	-	-	-
SDF [Wang *et al.*, 2016]	14.3	37.5	9.52	38.1	22.2	48.0	-	-	-	-
SING [Jiao *et al.*, 2018]	33.5	72.7	33.5	57.0	67.7	90.7	60.5	81.8	52.1	68.3
CSR-GAN [Wang *et al.*, 2018b]	32.3	70.9	37.2	62.3	70.7	92.1	59.8	81.3	51.9	67.5
ResNet50	29.6	64.0	29.9	62.2	67.4	91.7	57.0	78.7	47.6	66.4
FFSR	31.1	68.7	40.3	**65.3**	70.5	92.3	59.2	80.1	52.8	69.0
RIFE	35.7	**74.9**	33.9	63.6	69.7	91.5	62.6	82.4	53.3	70.1
FFSR+RIFE	**36.4**	72.0	**41.6**	64.9	**73.3**	**92.6**	**66.9**	**84.7**	**55.5**	**72.4**

Table 3: Comparison with recent works on five datasets.

of FFSR and RIFE in Fig. 4. It is clear that, both FFSR and RIFE decreases the objective $\mathcal{O}$, indicating improved ReID accuracy. It is also clear that, either FFSR and RIFE decrease the slope of the original curves, implying they improve the robustness of a ReID system to resolution variants. Combining FFSR and RIFE brings the best performance. Following part compares our methods with recent works.

5.4 Comparison with Recent Works

We compare our method with five recent low-resolution ReID methods including three traditional methods, *i.e.*, JUDEA [Li *et al.*, 2015], SLD^{2}L [Jing *et al.*, 2015], SDF [Wang *et al.*, 2016], and two deep learning based methods, *i.e.*, SING [Jiao *et al.*, 2018], CSR-GAN [Wang *et al.*, 2018b]. Three deep neural networks including ResNet50, Densenet121 [Huang *et al.*, 2017], and SE-resnet50 [Hu *et al.*, 2018] are also implemented and compared. We summarize the experimental results on five datasets in Table 3, which show the reported performance of JUDEA, SING, SDF, SING, and CSR-GAN on *CAVIAR*, *MLR-VIPeR*, and *MLR-CUHK03*. Performance of compared methods on *VR-Market1501* and *VR-MSMT17* are implemented with the code provided by their authors.

From the comparison we observe that, deep learning based methods substantially outperform the traditional ones. Our method shows promising performance on the first three datasets. On *CAVIAR*, our RIFE module outperforms the recent CSR-GAN by 3.4% in Rank-1 Accuracy. Combining FFSR and RIFE further boosts the performance, and outperforms CSR-GAN by 4.1%. On *MLR-VIPeR* and *MLR-CUHK03*, our method outperforms CSR-GAN by 4.4% and 2.6% in Rank-1 Accuracy, respectively.

Our method also shows promising performance on *VR-Market1501* and *VR-MSMT17*. Among existing methods that are compared, SING shows the best performance on *VR-Market1501*. FFSR+RIFE outperforms SING by 6.4%. FFSR and RIFE also achieves the best performance on *VR-MSMT17*, outperforming SE-resnet50 by 3.5%. It can be observed that, combining FFSR and RIFE commonly leads to the best performance on those five datasets. We show some results of image super resolution and person ReID in Fig. 5.

Figure 5: Sample results of person ReID and super resolution on *VR-MSMT17*. Our method (second row) outperforms the ResNet50 (first row). It is interesting to observe that, without joint learning with person ReID loss, FFSR module gets a lower ReID accuracy, but produces images with better visual quality.

6 Conclusion

This paper proposes a deep neural network composed of FFSR and RIFE modules for resolution invariant person re-identification. FFSR upscales the person foreground using a fully convolutional auto-encoder with skip connections learned with a foreground focus loss. RIFE adopts two feature extraction streams weighted by a dual-attention block to learn features for low and high resolution images, respectively. These two complementary modules are jointly trained to optimize the person ReID objective, leading to a strong resolution invariant representation. Extensive experiments on five datasets have shown the validity of introduced components and the promising performance of our methods.

Acknowledgments

This work is supported in part by Beijing Natural Science Foundation under Grant No. JQ18012, in part by Natural Science Foundation of China under Grant No. 61620106009, 61572050, 91538111, in part by Peng Cheng Laboratory.

References

[Chen *et al.*, 2019] Yun-Chun Chen, Yu-Jhe Li, Xiao-fei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *AAAI*, 2019.

[Cheng *et al.*, 2011] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*. Citeseer, 2011.

[Dong *et al.*, 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Gray and Tao, 2008] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017.

[Insafutdinov *et al.*, 2016] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, pages 34–50. Springer, 2016.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

[Jiao *et al.*, 2018] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. *AAAI*, 2018.

[Jing *et al.*, 2015] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, pages 695–704, 2015.

[Kim *et al.*, 2016] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[Li *et al.*, 2015] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, pages 3765–3773, 2015.

[Li *et al.*, 2019] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, 2019.

[Mao *et al.*, 2016] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NIPS*, pages 2802–2810, 2016.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[Tai *et al.*, 2017] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, volume 1, page 5, 2017.

[Wang *et al.*, 2016] Zheng Wang, Ruimin Hu, Yi Yu, Junjun Jiang, Chao Liang, and Jinqiao Wang. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, pages 2669–2675, 2016.

[Wang *et al.*, 2018a] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018.

[Wang *et al.*, 2018b] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, pages 3891–3897, 2018.

[Wei *et al.*, 2018] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[Yu *et al.*, 2018] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *CVPR*, pages 908–917, 2018.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.