Cascaded SR-GAN for Scale-Adaptive Low Resolution Person Re-identification

Zheng Wang[1], **Mang Ye**[2], **Fan Yang**[3], **Xiang Bai**[4,*], **Shin'ichi Satoh**[1,3,*]
[1] National Institute of Informatics, Japan
[2] Hong Kong Baptist University, China
[3] The University of Tokyo, Japan
[4] Huazhong University of Science of Technology, China
{wangz, yang, satoh}@nii.ac.jp, mangye@comp.hkbu.edu.hk, xbai@hust.edu.cn

Abstract

Person re-identification (REID) is an important task in video surveillance and forensics applications. Most of previous approaches are based on a key assumption that all person images have uniform and sufficiently high resolutions. Actually, various low-resolutions and scale mismatching always exist in open world REID. We name this kind of problem as Scale-Adaptive Low Resolution Person Re-identification (SALR-REID). The most intuitive way to address this problem is to increase various low-resolutions (*not only low, but also with different scales*) to a uniform high-resolution. SR-GAN is one of the most competitive image super-resolution deep networks, designed with a fixed upscaling factor. However, it is still not suitable for SALR-REID task, which requires a network not only synthesizing high-resolution images with different upscaling factors, but also extracting discriminative image feature for judging person's identity. (1) To promote the ability of scale-adaptive upscaling, we cascade multiple SR-GANs in series. (2) To supplement the ability of image feature representation, we plug-in a re-identification network. With a unified formulation, a Cascaded Super-Resolution GAN (CSR-GAN) framework is proposed. Extensive evaluations on two simulated datasets and one public dataset demonstrate the advantages of our method over related state-of-the-art methods.

1 Introduction

Person re-identification (REID) is the task of visually matching images of the same person, extracted from non-overlapping camera views in open surveillance spaces [Wang *et al.*, 2016a]. Since biometric cues, such as face and gait, are usually unreliable or even infeasible in the uncontrolled surveillance environment, the appearance of individuals is mainly exploited [Ye *et al.*, 2017; 2016]. In order to overcome the variations in illumination, occlusion and alignment, existing methods typically address the REID task by designing feature representation [Liao *et al.*, 2015; An *et al.*, 2015; Zhang *et al.*, 2016; Zhao *et al.*, 2017] or learning distance metrics [Chen *et al.*, 2015; Wang *et al.*, 2017b; 2017a; Yu *et al.*, 2017].

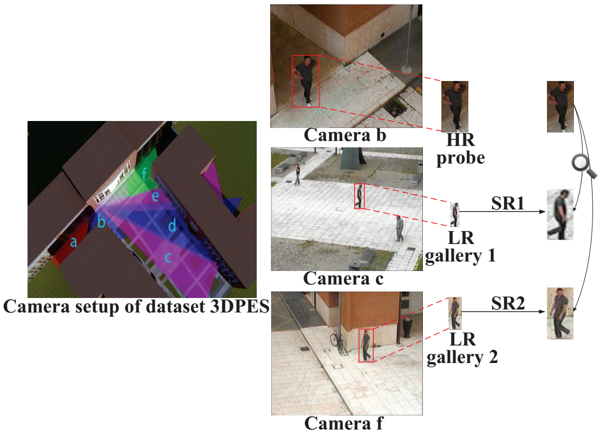


Figure 1: Illustration of person image resolutions with different scales in the open-space person re-identification. Three images of the same person are captured in three different camera views in the 3DPES dataset. The resolutions of these images are significantly different. The person image captured by camera b is relatively HR, but the person images captured by camera c and camera f are relatively LR. Meanwhile, the resolutions of any pair of these images are different. Intuitively, the LR gallery images should be enlarged to HR, and then the re-identification process carries on. Due to two different LRs, super-resolution (SR) modules with two upscaling factors are needed.

Generally, without regard to *various low-resolutions and scale mismatching*, most of these methods make an assumption that all person images have sufficiently high resolutions (HR), and they resize the images to a uniform scale before re-identification. However, it is common that the resolution of surveillance person image varies a lot, due to variations in the person-camera distance and camera deployment settings. Taking 3DPES dataset [Baltieri *et al.*, 2011] as an example,

[*] corresponding author

Figure 1 shows a usual situation in open space REID, that person images are not only low resolution (LR), but also holding different scales. This gives rise to a more challenging task that given a HR probe image, the algorithm is expected to match against LR gallery images with different scales. It requires to address cross-resolution matching since LR images contain much less information with discriminative appearance details largely lost in the image acquisition process. We name this kind of problem as Scale-Adaptive Low Resolution Person Re-identification (SALR-REID).

Actually, a couple of works [Jing *et al.*, 2015; Li *et al.*, 2015; Jiao *et al.*, 2018] have paid attention to LR and resolution mismatching problem. They set the probe image as a LR one, while all the gallery images are typically HR with a uniform scale. To address the mismatching of two resolutions (LR and HR), they constructed a cross-resolution relationship by a large number of training samples. However, their assumption is in a relatively ideal condition. In practice, gallery images are always not only LR, but also holding different scales. They have neglected the variation of LR scales. Multiple resolution scales require more training samples to construct relationships, and it cannot be guaranteed that those relationships work perfectly in matching. The common weakness makes these works inappropriate to address the SALR-REID problem. So far, to our best knowledge, only one pioneer research [Wang *et al.*, 2016b] investigated the SALR-REID problem. Instead of recovering the missing discriminative appearance information of LR images, the method performed multiple resolution representation transformation in a pre-defined feature space. Hence, the pioneer research does not inherently solve the information mis-matching challenge, whose effectiveness still needs promotion.

Intuitively, we employ super-resolution (SR) techniques to alleviate the resolution mismatch problem. During the process of super-resolving, SR methods [Dong *et al.*, 2016; Johnson *et al.*, 2016; Jiang *et al.*, 2017] supplement and recover the missing appearance information (Figure 1), so that LR gallery images and the HR probe image can be treated equally. The generative adversarial network (GAN) for image SR [Ledig *et al.*, 2016], i.e., SR-GAN, offers an effective solution to image SR, and is also one of the most competitive SR approaches. Nevertheless, (1) a fixed SR module is not suitable for the scale-adaptive LR scenario. SR-GAN can only increase the resolution with a fixed upscaling factor, while the SALR-REID problem requires to synthesize HR images with multiple upscaling factors. (2) Generic-purpose SR methods are designed to improve image visual sense rather than the re-identification performance. The ability of SR-GAN to capture inter-image similarity is limited as it is defined based on perceptually intra-image differences. A direct connection between super resolution and re-identification may suffer from suboptimal compatibility.

In this paper, we address the SALR-REID problem by exploring re-identification and multiple cascaded SR-GANs in series with a unified framework. We name the new framework as Cascaded Super-Resolution GAN (CSR-GAN). The contributions are as follows: (1) We cascade multiple SR-GANs in series, which is capable of super-resolving LR images with scale-adaptive upscaling. As far as we know, we are the first to propose a cascaded super-resolution deep network for scale-adaptive low resolutions. (2) The proposed CSR-GAN improves the integration compatibility between scale-adaptive super-resolution and re-identification, and consequently enhances the similarity of LR-HR pair during SR process. (3) For LR-HR intra-image pair, we design a common-human loss to make the super-resolved image look more like human. For LR-HR inter-image pair, we introduce the unique-human loss to make person image representation discriminative. Together with generator and discriminator losses, a joint loss function is optimized in a hybrid network architecture. Experimental results on SALR-REID datasets show the superiority of our CSR-GAN approach.

2 Revisit SR-GAN

As we know, GAN [Goodfellow *et al.*, 2014] provides a powerful framework for generating plausible-looking natural images with high perceptual quality. SR-GAN [Ledig *et al.*, 2016] proposes a very deep ResNet [He *et al.*, 2016] architecture using the concept of GAN for photo-realistic image super-resolution. It sets a new state-of-the-art for image SR with a fixed upscaling factor (4×). Following [Goodfellow *et al.*, 2014], it tries to solve the adversarial min-max problem as Eq. 1. The general idea behind this formulation is that it allows one to train a generator model G_{θ_G} with the goal of fooling a differentiable discriminator D_{θ_D} that is trained to distinguish super-resolved images from real images [Ledig *et al.*, 2016].

$$\begin{aligned}\min_{\theta_G}\max_{\theta_D} & \mathbb{E}_{\hat{I}\sim p_{train}(\hat{I})}[\log D_{\theta_D}(\hat{I})]+ \\ & \mathbb{E}_{I\sim p_G(I)}[\log(1-D_{\theta_D}(G_{\theta_G}(I)))].\end{aligned} \tag{1}$$

In this formulation, $\hat{I}$ stands for the real HR image, while I stands for the LR image to be super-resolved [1]. The generator function G_{θ_G} is parametrized by θ_G, and the discriminator function D_{θ_D} is parametrized by θ_D. Generally, the target of previous supervised SR algorithms is commonly the minimization of the mean squared error (MSE) [Wang and Bovik, 2009] between the recovered HR image and the ground truth. Besides the MSE loss, SR-GAN also defines a perceptual loss using high-level feature maps of the VGG network [Simonyan and Zisserman, 2014], which makes the super-resolved image and HR reference image perceptually similar. However, to address the SALR-REID task, two more functionalities need to be annexed.

- To promote the ability of *scale-adaptive upscaling*, it requires combining multiple SR-GANs, so that scale-adaptive LR images can be enlarged to a uniform HR.
- To supplement the ability of *discriminative person representation extracting*, it requires plugging in the re-identification network, so that identity appearance information can be captured during SR.

[1] In this paper, if image $\hat{I}$ has a hat, it means that the image is a real image. Otherwise, image I without a hat means that it is a super-resolved image.

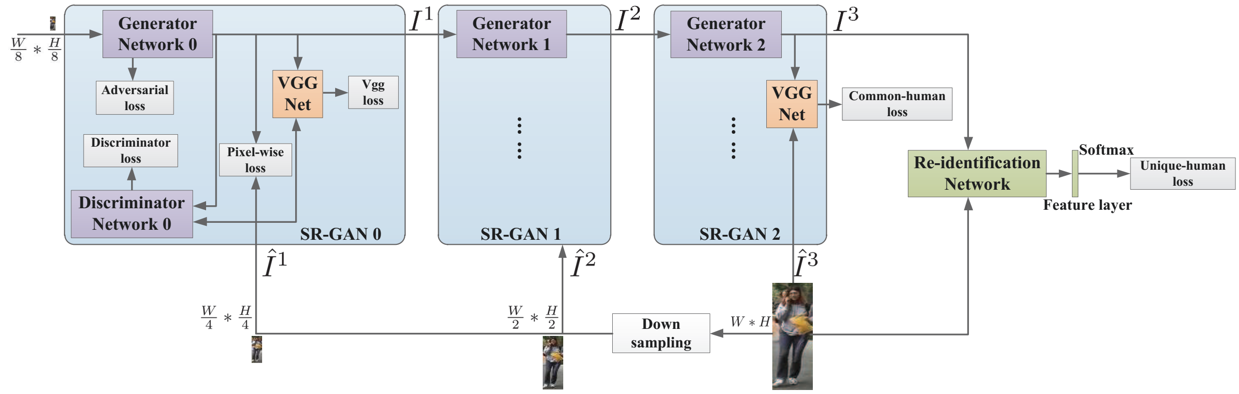


Figure 2: The architecture of of the proposed CSR-GAN. The CSR-GAN consists of three cascaded SR-GANs, and a re-identification network. Each SR-GAN includes a generator network, which enlarges the input image with a double upscaling factor (2×), and a corresponding discriminator network. The pixel-wise and adversarial loss are designed to make the output image of generator network to be a real image. The VGG loss is designed for perceptual similarity. These three kinds of losses form the generator network losses. The discriminator losses are designed to distinguish super-resolved images from real images. At the last SR-GAN, a common-human loss is designed by the output of VGG network, which tries to make the super-resolved image look more like human. Finally, a unique-human loss is for the plugged-in re-identification network, which makes the extracted features of the same person similar. (For simplicity, we do not draw the details in SR-GAN 1 and SR-GAN 2.)

3 Our Approach

We propose the Cascaded Super-Resolution GAN (CSR-GAN) framework in a unified architecture. Figure 2 shows the architecture. Suppose that all person images have the same weight-to-height ratio. W and H respectively denote the weight and height of the original HR image. To make the architecture easy to follow, we use a three-cascaded SR-GANs to enlarge the LR person image, whose resolution is lower than $\frac{W}{2} * \frac{W}{2}$. Each generator network can enlarge the input image with a double upscaling factor. That is to say, $\frac{W}{8} * \frac{W}{8}$, $\frac{W}{4} * \frac{W}{4}$ and $\frac{W}{2} * \frac{W}{2}$ are respectively image shapes input to three generator networks. Here, we denote I^k as the input image of each generator G_k, and $I^{k+1} = G_{\theta_{G_k}}(I^k)$ as the output image of the generator, where $k \in [0, 2]$ is the scale index of input image and the ID of sub SR-GAN. $\hat{I}^{k+1}$ denotes the real image, which has the same scale as the super-resolved image $G_{\theta_{G_k}}(I^k)$ or I^{k+1}. We set r_k standing for the scale ratio of LR image to HR image, then the resolution of image I^k is described as $r_kH * r_kW$. The scale ratios are respectively $r_0 = 1/8$, $r_1 = 1/4$, $r_2 = 1/2$, and $r_3 = 1$.

Our goal is not only to train generator functions $\{G_k\}$ that estimate the corresponding HR counterpart $\hat{I}^{k+1}$ for a given LR input image I^k, but also to train re-identification function F that extracts feature for the final HR image. Following [Ledig *et al.*, 2016], we train the generator networks with pixel-wise loss, VGG loss, and adversarial loss, which together form the generator network loss l^{SR}_{Gen}. For super resolution, we also train the discriminator networks with loss l^{SR}_{Dis} from the perspectives of game theory, and boost the generator networks indirectly. In addition, a common-human loss l_{Com} is designed to make person image looks better. The re-identification network is trained with unique-human loss l_{Uni}. Finally, all losses are combined together as Eq. 2.

$$l_{total} = l^{SR}_{Gen} + l^{SR}_{Dis} + l_{Com} + l_{Uni}. \tag{2}$$

For simplify, we balance the weights of these losses equally. The details of losses are described in the following subsections.

3.1 Generator Network Loss

The generator network loss is critical for the performance of our generator network. In general, super-resolution is commonly modeled based on the pixel-wise MSE [Dong *et al.*, 2016]. In particular, we introduce the adversarial loss for each generator network k, and exploit the VGG loss for perceptual similarity as [Ledig *et al.*, 2016] did. The pixel-wise loss $l^{SR_k}_{MSE}$, the VGG loss $l^{SR_k}_{VGG}$, and the adversarial loss $l^{SR_k}_{Adv}$ are demonstrated as follows.

Pixel-wise loss. The pixel-wise MSE loss is calculated as:

$$l^{SR_k}_{MSE} = \frac{1}{r^2_{k+1}WH} \sum_{x=1}^{r_{k+1}W} \sum_{y=1}^{r_{k+1}H} (\hat{I}^{k+1}_{x,y} - G_{\theta_{G_k}}(I^k)_{x,y})^2. \tag{3}$$

This is the most widely used optimization target for image super-resolution on which many state-of-the-art approaches rely [Shi *et al.*, 2016].

VGG loss. Furthermore, a VGG loss based on the ReLU activation layers of the pre-trained 19 layer VGG network, described in [Simonyan and Zisserman, 2014], is exploited for perceptual similarity, measuring on higher semantical level which a naive MSE loss is unable to handle. It should be mentioned that the VGG losses are adopted in all the VGG networks, who share common parameters. We denote $\phi_{i,j}$ as the feature map obtained by the j-th convolution before the i-th maxpooling layer within the VGG19 network. The VGG loss is calculated as the Euclidean distance between the feature representations of a super-resolved image $G_{\theta_{G_k}}(I^k)$ and its corresponding HR image $\hat{I}^{k+1}$. Here, $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps with the VGG network.

$$l_{VGG}^{SR_k} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(\hat{I}^{k+1})_{x,y} - \phi_{i,j}(G_{\theta_{G_k}}(I^k))_{x,y})^2. \tag{4}$$

Adversarial loss. Besides the losses for image content described above, we add a generative component. Along with the losses for deceiving the discriminator networks, the adversarial loss is defined based on the probabilities of the discriminator over all training samples. It is denoted as:

$$l_{Adv}^{SR_k} = -\log D_{\theta_{D_k}}(G_{\theta_{G_k}}(I^k)), \tag{5}$$

where $D_{\theta_{D_k}}(G_{\theta_{G_k}}(I^k))$ is the probability that the super-resolved image $G_{\theta_{G_k}}(I^k)$ is a real image. For better gradient behavior we minimize $-\log D_{\theta_{D_k}}(G_{\theta_{G_k}}(I^k))$ instead of $\log[1 - D_{\theta_{D_k}}(G_{\theta_{G_k}}(I^k))]$ [Goodfellow *et al.*, 2014].

At last, we formulate the generator network loss l_{Gen}^{SR} as the weighted sum of afore-mentioned three kinds of losses [2]:

$$l_{Gen}^{SR} = \sum_{k=0}^{2} l_{MSE}^{SR_k} + \alpha \sum_{k=0}^{2} l_{VGG}^{SR_k} + \beta \sum_{k=0}^{2} l_{Adv}^{SR_k}, \tag{6}$$

where α and β respectively denote the weights for the VGG loss and the adversarial loss.

3.2 Discriminator Network Loss

The discriminator network loss is defined based on the probabilities of the discriminator over all training images and super-resolved images. The discriminator should distinguish super-resolved images from real ones. Hence, the loss of D_k needs to be minimized when judging a real image $\hat{I}^{k+1}$ to be positive or a super-resolved image $G_{\theta_{G_k}}(I^k)$ to be negative. Then, the total loss is calculated as:

$$l_{Dis}^{SR} = -\sum_{k=0}^{2} \log D_{\theta_{D_k}}(\hat{I}^{k+1}) + \sum_{k=0}^{2} \log D_{\theta_{D_k}}(G_{\theta_{G_k}}(I^k)), \tag{7}$$

where $D_{\theta_{D_k}}(G_{\theta_{G_k}}(I^k))$ is the probability that the super-resolved image $G_{\theta_{G_k}}(I^k)$ is a real image, and $D_{\theta_{D_k}}(\hat{I}^{k+1})$ is the probability that the original image $\hat{I}^{k+1}$ is a real image. For better gradient behavior we minimize $-\log D_{\theta_{D_k}}(\hat{I}^{k+1})$ instead of $\log[1 - D_{\theta_{D_k}}(\hat{I}^{k+1})]$ [Goodfellow *et al.*, 2014].

3.3 Common-human Loss

Generally, the output of VGG network [Simonyan and Zisserman, 2014] is the category of input image. Suppose that the VGG network is able to judge each real image $\hat{I}^3$ as a human category. To make the super-resolved I^3 look like a human, we design a common-human loss to constrain its category results to be the same as that of $\hat{I}^3$. The common-human loss is calculated as the Euclidean distance between category results of $\hat{I}^3$ and I^3. It is defined as below:

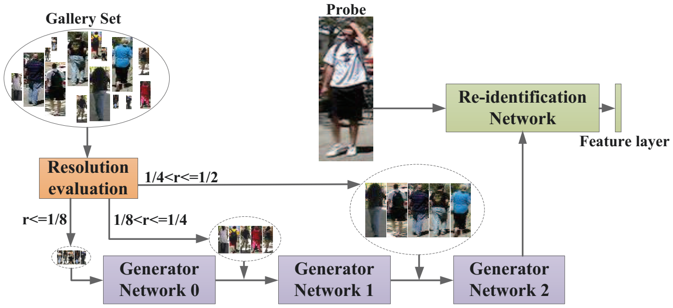


Figure 3: The evaluation process. The re-identification network is used to extracted feature. Different from the HR probe image, scale-adaptive LR gallery images are assigned to different stages of the cascaded generator network, based on their resolution scale ratio. After the cascaded generator networks, all gallery images are enlarged to the uniform HR.

$$l_{Com} = \frac{1}{1000} \sum_{c=1}^{1000} (\psi_c(\hat{I}^3) - \psi_c(I^3))^2, \tag{8}$$

where c denotes the dimension index, and ψ_c is possibility that an image is assigned to the c-th category. The total number of categories is 1000.

3.4 Unique-human Loss

Given an original image $\hat{I}^3$ or an super-resolved image I^3, the output of re-identification network is $z = [z_1, z_2, ..., z_M] \in \mathbb{R}^M$, where M is the number of person IDs. So the predicted probability of each ID label m is calculated as: $p(m|I^3) = \frac{exp(z_m)}{\sum_{i=1}^{M} exp(z_m)}$. To simplify the equation, we omit the correlation between m and I^3. The cross entropy of identification loss is formulated as below:

$$l_{Uni} = -\sum_{m=1}^{M} log(p(m))q(m). \tag{9}$$

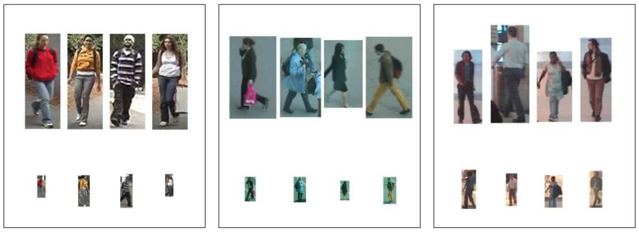
Let y be the ground-truth ID label, so that $q(y) = 1$, and $q(m) = 0$ for all $m \neq y$. In this case, minimizing the identification loss is equivalent to maximizing the possibility of being assigned to the ground-truth class.

4 Implementation Details

The training process includes the following three steps: (1) We first initialize the re-identification network separately. We choose ResNet-50 [He *et al.*, 2016] as the base. The ResNet-50 is pre-trained with ImageNet [Russakovsky *et al.*, 2015], and then fine-tuned with the Market-1501 [Zheng *et al.*, 2015] dataset. (2) The cascaded generator networks are initialized with MSE losses. (3) The whole network is trained simultaneously with all the losses. For each sub SR-GAN, the settings of strides and feature maps are the same as [Ledig *et al.*, 2016].

During the evaluation process (as Figure 3 shows), a resolution detection module is used to assign the LR gallery images to different stages of cascaded generator networks. The

[2] In this paper, if no otherwise specified, following [Ledig *et al.*, 2016], we set $\alpha = 2 * 10^{-6}$ and $\beta = 1 * 10^{-3}$.



(a) SALR-VIPeR (b) SALR-PRID (c) CAVIAR

Figure 4: Example image pairs from three datasets. Each column shows two images of the same identity from two different cameras with different resolutions, where images in the bottom row are LR. (a) the SALR-VIPeR dataset; (b) the SALR-PRID dataset; (c) the CAVIAR dataset.

assignment depends on the scale ratio r of LR to HR. When a LR image is assigned to the generator network k, it will be resized to the input shape of corresponding generator network $r_kH * r_kW$. After cascaded SR, the super-resolved gallery images and the HR probe image are put into the re-identification network to extract features. Finally, features are used for person re-identification.

5 Experiments

5.1 Experimental Datasets and Settings

Following [Wang *et al.*, 2016b], the evaluation is run on two simulated person datasets SALR-VIPeR and SALR-PRID, which are based on the VIPeR dataset [Gray *et al.*, 2007] and the PRID450S dataset [Roth *et al.*, 2014] respectively, and the public CAVIAR dataset [Cheng *et al.*, 2011].

SALR-VIPeR. The widely used VIPeR dataset [Gray *et al.*, 2007] contains 1264 outdoor images obtained from two views of 632 persons. All images of individuals are normalized to a size of $128 * 48$ pixels. To construct the SALR-VIPeR dataset, we set images from camera A as the HR probe set, whose resolution remains unchanged. While images from camera B are set as the LR gallery set, which are down-sampled randomly to different scales. The scale ratios range from 0.1 to 0.25. Some example images are shown in Figure 4(a).

SALR-PRID. The PRID450S [Roth *et al.*, 2014] is a challenge dataset, particularly there is camera characteristics variation. It contains 450 single shot image pairs captured over two spatially disjoint camera views. All images are normalized to $168 * 80$ pixels. We construct SALR-PRID dataset following the way of constructing SALR-VIPeR. Some example images are shown in Figure 4(b).

CAVIAR. The CAVIAR dataset [Cheng *et al.*, 2011] contains images of 72 individuals captured from 2 cameras in a shopping mall. This dataset is suitable for testing SALR-REID, as the resolution of images captured from the second camera is much lower than that in the first camera (Figure 4(c)). Among the 72 people, 18 were only captured in a single camera view with no low resolution images, and they were thus removed. The remaining persons were used in our experiments, where a HR image of each person is selected to form the probe set, and a LR image of each person is selected to form the gallery set.

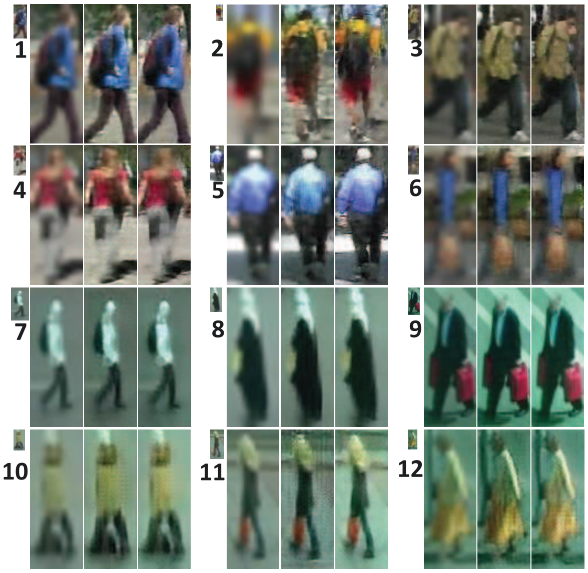


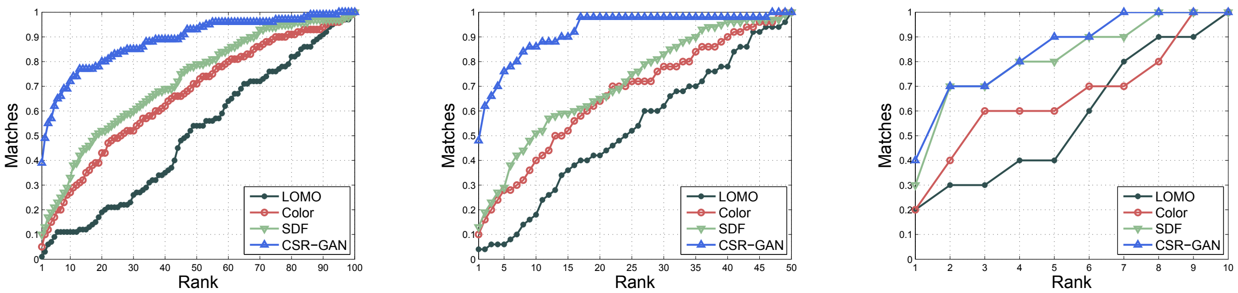
Figure 5: Some subjective results for scale-adaptive SR. We show 12 groups of results for 12 LR images. Image 1-6 are selected from the SALR-VIPeR dataset, and image 7-12 are selected from the SALR-PRID dataset. For each group, from left to right, four images are respectively a LR image, a SR image by bicubic, a SR image by CSR-GAN without common-human loss, and a SR image by CSR-GAN with common-human loss.

Settings. Following [Wang *et al.*, 2016b], all datasets are randomly divided into training set and testing set. Persons for training and testing are respectively 532 and 100 (SALR-VIPeR), 400 and 50 (SALR-PRID), and 44 and 10 (CAVIAR). The probe set consists of all HR images per person. LR images are randomly downsampled and selected to construct the gallery set. Cumulative Matching Characteristic (CMC) curves [Wang *et al.*, 2007] were used to calculate the average performance, and the value of CMC@k indicates the percentage of the real match ranked in the top k.

5.2 Evaluation on Scale-Adaptive SR

In this subsection, we prove that the proposed method is applicable for scale-adaptive SR. In Figure 5, we show some subjective results on scale-adaptive LR images. The SR results are generated respectively by bicubic, CSR-GAN without common-human loss, and CSR-GAN with common-human loss. Here, we choose images from SALR-VIPeR and SALR-PRID as examples. From the figure, we can find that CSR-GAN is effective for scale-adaptive SR, and the effectiveness varies with different LR scales. We can also find that the generated result by CSR-GAN with common-human loss looks better than without common-human loss. For example, for image 2, the SR image without common-human loss drops out its human shape a little. For the SR images 7, 10, 11 and 12, there are less noises in the image background with common-human loss, where we guess that CSR-GAN pays more attention to the human body and additional information for the background are smoothed. Hence, the common-human loss is every useful for person image SR.

Meanwhile, we performed a Mean Opinion Score (MOS) test to quantify the ability of scale-adaptive super resolution.



(a) SALR-VIPeR (b) SALR-PRID (c) CAVIAR

Figure 6: Experimental results on three datasets for SALR-REID problem. For each dataset, we compare the proposed method (CSR-GAN) with a popular person re-identification method (LOMO), a basic person image descriptor method (Color) and the state-of-the-art method for SALR-REID problem (SDF). (a) the SALR-VIPeR dataset; (b) the SALR-PRID dataset; (c) the CAVIAR dataset.

Dataset	r	nearest	bicubic	CSR-GAN
SALR-VIPeR	$(0, \frac{1}{8}]$	1.05	1.12	**1.98**
	$(\frac{1}{8}, \frac{1}{4}]$	2.14	2.25	**3.78**
SALR-PRID	$(0, \frac{1}{8}]$	1.05	1.20	**2.05**
	$(\frac{1}{8}, \frac{1}{4}]$	2.30	2.55	**3.83**
CAVIAR	$(\frac{1}{4}, \frac{1}{2}]$	3.10	3.25	**4.20**

Table 1: The MOS test results on the testing images of three different datasets. We compared the proposed CSR-GAN method with the nearest and the bicubic methods.

	$rank@1$	$rank@5$	$rank@10$	$rank@20$
JUDEA	26.0	55.1	69.2	82.3
SLD^{2}L	20.3	44.0	62.0	78.2
SDF	9.52	38.1	52.4	68.0
SING	33.5	57.0	66.5	76.6
CSR-GAN	**37.2**	**62.3**	**71.6**	**83.7**

Table 2: Comparing with state-of-the-art LR person re-identification methods on MLR-VIPER. The $1^{st}/2^{nd}$ best results are indicated in red/blue.

Specifically, we asked 5 raters to assign an integral score from 1 (bad quality) to 5 (excellent quality) to the super-resolved images. Each rater rated all the testing images of three datasets. We found no significant differences between the ratings of the identical images. As images with different resolution scales will go to different process of CSR-GAN, all the images are divided into different groups depending on their scale ratio. The experimental results of the conducted MOS tests are summarized in Table 1. As can be seen, the proposed method outperforms the general SR methods (nearest and bicubic) in all the scale ratio groups. It should be mentioned that we did not compared with the other supervised learning SR methods, because most of those methods just enlarge the image with a fixed upscaling factor.

5.3 Evaluation on SALR-REID Datasets

In this subsection, we prove that the proposed method is suitable for the SALR-REID problem. We evaluated the effectiveness of the proposed method by comparing with one of the most popular person re-identification method LOMO [Liao *et al.*, 2015], a basic person image descriptor method with RGB color, and the state-of-the-art method SDF [Wang *et al.*, 2016b], on the SALR-VIPeR, SALR-PRID and the CAVIAR datasets, respectively. In particular, when we used LOMO and Color to extract features, all the scale-adaptive LR gallery images were resized to the uniform HR with bicubic. The obtained results are shown in Figure 6. As can be seen, the general person re-identification method LOMO is almost no effect, even performs worse than the basic color descriptor. Consequently, general descriptors are not suitable for the SALR-REID problem. Meanwhile, we can also find that our approach has improvements on all the three datasets, compared with LOMO, Color and SDF. The promotions are significant on the SALR-VIPeR and SALR-PRID dataset, where the scale-adaptive LR problem is serious.

5.4 Comparison with State-of-the-art LR Methods

Recently, some approaches have paid attention to LR problem, and proposed algorithms to address the resolution mismatching problem, such as JUDEA [Li *et al.*, 2015], SLD^{2}L [Jing *et al.*, 2015] and SING [Jiao *et al.*, 2018]. These methods assume that all gallery images are HR, and probe images are LR with a fixed down-sampling factor. For example, MLR-VIPeR [Jiao *et al.*, 2018] was constructed by down-sampling images with a ratio picked from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$. To evaluate the ability of addressing resolution mismatching, we compared our method with these methods on the MLR-VIPeR dataset. Table 2 shows the results. From the table, we can see that although our method is proposed for SALR-REID problem, it outperforms related state-of-the-art methods on fixed the resolution mismatching problem with a big margin.

6 Conclusion

This paper focuses on a new issue, i.e., scale-adaptive low resolution person re-identification. To promote the ability of scale-adaptive upscaling and image feature extracting, we propose a new framework CSR-GAN. Besides generator network and discriminator network losses, a common-human loss and a unique-human loss are introduced, and optimized in a hybrid network architecture. Using extensive experiments, we have confirmed that CSR-GAN is capable of super-resolving person images with adaptive upscaling, achieves a considerable promotion on SALR-REID datasets, and outperforms state-of-the-art LR person re-identification methods.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR1686, Japan.

References

[An *et al.*, 2015] Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu. Person re-identification with reference descriptor. *IEEE Trans. Circuits Syst. Video Technol.*, 2015.

[Baltieri *et al.*, 2011] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Joint ACM workshop on Human gesture and behavior understanding*, 2011.

[Chen *et al.*, 2015] Jiaxin Chen, Zhaoxiang Zhang, and Yunhong Wang. Relevance metric learning for person re-identification by exploiting listwise similarities. *IEEE Trans. Image Proc.*, 2015.

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

[Dong *et al.*, 2016] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[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*, 2014.

[Gray *et al.*, 2007] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007.

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

[Jiang *et al.*, 2017] Junjun Jiang, Jiayi Ma, Chen Chen, Xinwei Jiang, and Zheng Wang. Noise robust face image super-resolution through smooth sparse representation. *IEEE Trans. Cybern.*, 2017.

[Jiao *et al.*, 2018] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *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*, 2015.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[Ledig *et al.*, 2016] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[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*, 2015.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[Roth *et al.*, 2014] Peter M Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*. 2014.

[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. *Int. J. Comput. Vision*, 2015.

[Shi *et al.*, 2016] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Wang and Bovik, 2009] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Proc. Mag.*, 2009.

[Wang *et al.*, 2007] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[Wang *et al.*, 2016a] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng. Zero-shot person re-identification via cross-view consistency. *IEEE Trans. Multimedia*, 2016.

[Wang *et al.*, 2016b] 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*, 2016.

[Wang *et al.*, 2017a] Zheng Wang, Ruimin Hu, Chen Chen, Yi Yu, Junjun Jiang, Chao Liang, and Shin'ichi Satoh. Person reidentification via discrepancy matrix and matrix metric. *IEEE Trans. Cybern.*, 2017.

[Wang *et al.*, 2017b] Zheng Wang, Ruimin Hu, Yi Yu, Junjun Jiang, Jiayi Ma, and Shin'ichi Satoh. Statistical inference of gaussian-laplace distribution for person verification. In *ACM Multimedia*, 2017.

[Ye *et al.*, 2016] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Trans. Multimedia*, 2016.

[Ye *et al.*, 2017] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, 2017.

[Yu *et al.*, 2017] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

[Zhang *et al.*, 2016] Yaqing Zhang, Xi Li, Liming Zhao, and Zhongfei Zhang. Semantics-aware deep correspondence structure learning for robust person re-identification. In *IJCAI*, 2016.

[Zhao *et al.*, 2017] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[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*, 2015.