IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 17, 2020, accepted August 27, 2020, date of publication September 14, 2020, date of current version October 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023594

Multi Scale-Adaptive Super-Resolution Person Re-Identification Using GAN

MUHAMMAD ADIL[1], SAQIB MAMOON[1], ALI ZAKIR[1], MUHAMMAD ARSLAN MANZOOR[2], AND ZHICHAO LIAN[1]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]School of Information Science and Technology, East China University of Science and Technology, Shanghai 200237, China

Corresponding author: Zhichao Lian (lzcts@163.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 30919011401 and Grant 30919011231, in part by the China Postdoctoral Foundation under Grant 2015M581800, in part by the National Key Research and Development Program of China under Grant 2016YFF0103604, in part by the Visiting Scholar Foundation of Key Laboratory of Biorheological Science and Technology, Ministry of Education, Chongqing University, under Grant CQKLBST-2018-011, and in part by the Foundation of Shandong Provincial Key Laboratory of Digital Medicine and Computer Assisted Surgery under Grant SDKL-DMCAS-2018-04.

ABSTRACT In real-world surveillance systems, the person images captured by the camera network consists of various low-resolution (LR) images. It creates a resolution mismatching problem when compared against high-resolution images of a targeted person. It significantly affects the performance of person re-Identification. This problem is known as Low-Resolution Person re-identification (LR PREID). An efficient strategy would be to exploit image super-resolution (SR) with person re-identification as a mutual learning approach. In this paper, we propose a novel method MSA-SR-PREID to solve this problem. The model takes low-resolution images on different resolutions and resized them to pre-defined fixed resolution. The design of the super-resolution network consists of ESRGAN and the de-Noising module to generate super-resolution images. The SR images are later passed to the re-identification network to learn the unique descriptors to recognize a person identity. The performance of this model is evaluated on four competitive benchmarks, MLR-VIPeR, MLR-DukeMTMC-reID, VR-MSMT17, and VR-Market1501. The comparison with similar state-of-the-art demonstrates the superiority of our model.

INDEX TERMS Person re-identification, low-resolution person re-identification, super-resolution, image de-noising.

I. INTRODUCTION

Person re-identification (PREID) aims to identify all the occurrences of the person of interest in the surveillance network or from different timestamps of a single camera. It facilitates a wide range of applications varying from camera surveillance [1] to computational forensics [2].

The problems: pose and viewpoint variations, light intensity and background changes, low-resolution (LR) and scale variations (SV), and other challenges like partial occlusion make person re-identification a non-trivial problem. Significant efforts have been made to eradicate or minimize these challenges [3]–[8]. However, in person re-identification, Low-Resolution (LR) is still considered a major challenge for large scale surveillance systems.

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

Recently, deep learning techniques are being used for PREID and achieved significant alleviation in performance on popular benchmarks [9]–[15]. However, regardless of various LR and SV, most of these approaches are based on the premise that both query and gallery images are of the same or sufficiently high resolutions. This assumption may not hold in real-world re-identification problems as there happen to be many variations in the resolution of the person images under surveillance network [16]. These variations occur due to the natural changes in the distance between person and cameras, e.g., pedestrian movement relative to the camera. This also creates resolution inconsistency and significant loss in visual information. Consequently, standard re-identification models may fail to identify the multiple instances of the respective persons, which are undesirable in security and surveillance.

Figure 1 demonstrates low-resolution and cross-resolution person re-identification. Resolution discrepancy between the

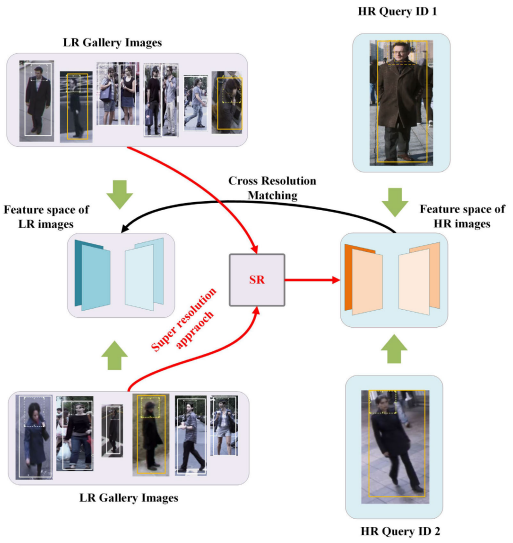

FIGURE 1. Illustration of low-resolution person images on a different resolutions, representing the real-world person re-Identification problem.

high-resolution (HR) query images and the Low-resolution (LR) gallery images or vice versa creates an unaligned feature distributions that affects the pedestrian matching performance. For example, a standard PREID method [17] can experience up to 19.2% Rank-1 performance drop when applied to cross-resolution PREID [18]. This gives rise to a more difficult task, which requires the algorithm to match LR gallery images with HR probe images. It needs addressing cross-resolution matching because LR surveillance images contain much less discriminative information, as the image acquisition process significantly loses the details of the images [19]. This problem is known as Low-Resolution Person re-identification (LR-PREID). To cope with this problem, an adequate strategy is to use an image super-resolution (SR) method to improve the resolution of LR query images to minimize the distribution discrepancy with HR gallery images.

Several methods [16], [18]–[24] have been introduced to address the LR PREID problem. However, there are some common drawbacks in these methods: (1) Instead of trying to recover the misplaced discriminative appearance information, they perform a transformation of the cross resolution representation in pre-defined feature space [20]–[22], [24], [25]. This does not solve the information amount disparity issue since the pixel-to-pixel high-resolution supervision lacks in cross-view pedestrian images. (2) They focus exclusively on hand-crafted visual features instead of taking advantage of a deep neural network's capability to learn and optimize the discriminative features automatically [20], [21]. Naturally, image (SR) would provide an effective way to minimize the dilemma of resolution discrepancy because of its ability to produce high-frequency details [26]. (3) Utilization of multi-branch super-resolution networks to generate high-resolution person images like CSR-GAN, employed three consecutive GANs unit for image SR [16], [19]. It makes the overall network complex, requires more memory and power.

This study addresses the LR person re-identification problem by employing the image Super-Resolution technique with the PREID network in a unified framework named multi-scale adaptive super-resolution person re-Identification (MSA-SR PREID). We introduce a scale-adaptive module for resolution mismatching problem by resizing them to pre-defined fixed resolution. The focus of this research is to investigate the LR PREID, for that, we consider all the images as LR images-(varies in the range of width within [8,32). LR images have mostly blurry and coarse edges, which make it difficult to extract the low-level features (such as edges, color, pixel intensity, pixel gradient, and orientation, etc.) that provide the baseline for high-level semantics features, especially after performing downsampling operation on the already low-resolution images.

To enhance the feature extraction capability of the network, we employed a GAN based enhanced super-resolution generative adversarial network [27]to recreate HR counterparts of LR images. This allows us to effectively extract the distinct visual appearance information. It composes of Residual in Residual dense block for effective image regeneration. Our proposed framework utilizes only one super-resolution network as compared to the [16], [19], decreasing the complexity and computation time. Besides, a de-Noising module is introduced to remove the noise from re-generated images. Moreover, to keep the network lightweight, ResNet50 [28] is employed for feature extraction. Besides, two parameters, Random erasing and Linear Warm-Up, are employed to enhance the feature extraction capability of ResNet50. Our model has achieved state-of-the-art results for Rank-1 accuracies on MLR-VIPeR 62.00%, MLR-DUKEMTMC-REID 79.06%, VR-MSMT17 60.65%, and VR-Market1501 68.26%.

The main contribution of our work is summarized as the following points:

- 1) The multi-scale-adaptive super-resolution person re-Identification network (MSA-SR PREID) adequately addresses the resolution-mismatching problem as it takes arbitrary input size and resizes it to a pre-defined resolution using a scale-adaptive module. (Figure-2).
- 2) An enhanced super-resolution generative adversarial network is utilized for effective image re-generation. It uses the Residual-in-Residual Dense Block to generate more realistic and natural images. The relativistic discriminator network is utilized to enhance the identification information effectively. The Batch normalization layer is removed for stable training and consistent performance [27]. The perceptual loss is improved by extracting the feature map before applying the activation function to avoid the feature sparsity [27]. To support the in-depth training, the network employs residual scaling-(scaling down the residuals before being added to the main path or other residuals) and smaller initialization strategy.

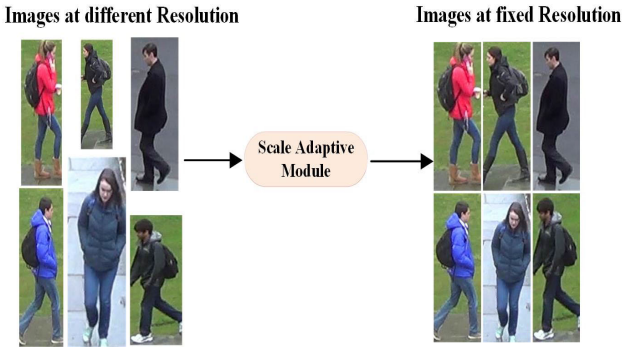

FIGURE 2. Resizing images on a fixed resolution. The images are taken from DukeMTMC-REID.

- 3) The de-Noising module employed to eliminate the noisy artifacts created by GAN due to the decompressing nature of the JPEG format images, which are utilized for evaluation.
- 4) As of its applications, the evaluation of the proposed network is made under different settings with following ablation studies:1) evaluation on MLR datasets, 2) evaluation on VR datasets, 3) evaluation of super-resolution network (SRNet), 4) evaluation of image de-noising module and 5) evaluation of re-identification parameters.

II. RELATED WORK

This section covers a brief overview of the current work of low-resolution person re-identification and super-resolution.

A. PERSON RE-IDENTIFICATION

Person re-identification has attracted extensive research in the past decade [29]–[36]. Many existing approaches [35], [37]–[40] addresses the uncontrolled variables in person re-identification, such as pose and viewpoint variations, light intensity, and other difficulties like partial occlusion. For instance, Liu *et al.* [41]] developed a pose-transferable network based on the generative adversarial (GAN) framework [42]; it precisely predicts the pose variations. Zhong *et al.* [43] addressed the camera-invariant subspace to cope with the permutations in the theme caused by multiple cameras. Yang [4] proposed a patch-based unsupervised learning framework that learns discriminative features from the image patches, instead of the whole images. Meng *et al.* [9] learn the identity labels annotated at the untrimmed video level. Munjal *et al.* [6] fused the query information into a Siamese network as a guide for global context information, proposal generation and similarity calculation. Another research trend is domain adaptation [44], [45] for PREID [5], [46], [47], where networks trained on the source domain can have a significant performance drop in the evaluation on the target domain [5]. Several methods [15], [34], [48] addresses background variations by leveraging attention frameworks [25], [49], [50] to emphasize the discriminative parts. However, most of the methods, as mentioned earlier, designed on the assumption that all person images are of similar or sufficiently high resolutions, which might not be practical for real-world applications.

B. LOW-RESOLUTION PERSON RE-IDENTIFICATION (LR PREID)

Recently, various methods [16], [18]–[24], [51] have been proposed for the LR PREID problem. Jing *et al.* [20] introduced a semi-coupled low-rank discriminant dictionary learning hand-crafted method to uncover the relationship between features of LR and HR images. However, they used the LR problem images and considered all gallery images as HR, which is usually the ideal case. In contrast, the real-world REID environment has various LR gallery images. Li *et al.* [21] used the same approach in which they used heterogeneous class mean discrepancy (HCMD) criterion for cross-scale image domain alignment to match the LR problem image against HR gallery images. However, this approach does not extract discriminative appearance information that is lost in the acquisition of images. Jiao *et al.* [19] proposed super-resolution and identity joint learning (SING) to optimize image SR and PREID process simultaneously. Wang *et al.* [22] learn a discriminative surface for re-identifying the persons using feasible and infeasible functions. Instead of recovering the missing discriminative appearance information of LR images, they transformed multiple resolution representation to a pre-defined feature space. Mao *et al.* [24] proposed two modules: a) foreground-focus super-resolution (FFSR) model to recover the resolution loss in LR input images, b) and a resolution-invariant person re-ID module to extract features. Given their promising results, it relied on the annotation of the foreground mask to direct the learning of image recovery for each training image.

Chen *et al.* [23] proposed resolution adaptation and re-identification Network (RAIN). They used adversarial loss and reconstruction loss to reduce the difference between different resolution deep features by aligning the feature distributions of HR and LR images. Li *et al.* [18] proposed a framework to study both learning resolution-invariant representation and exploiting image super-resolution for improving cross-resolution PREID performance. Huang *et al.* [51] developed a degradation invariance learning framework for real-world PREID. Their proposed network consists of two stages: a degradation invariance learning to remove real-world degradations (like low-resolution, weak illumination, and blurring) by a Degradation Decomposition Generative Adversarial Network (DDGAN) and a robust identity representation learning by a Dual Feature Extraction Network (DFEN). Zheng Wang proposed cascaded SRGAN [16] network for LR PREID and utilized the GAN network to generate HR counterparts of LR images. They used multiple super-resolution generative adversarial networks (SRGAN) to generate high-resolution gallery images, i.e., each SRGAN worked on a specific resolution. It made the overall network more complex and increase computation time and power.

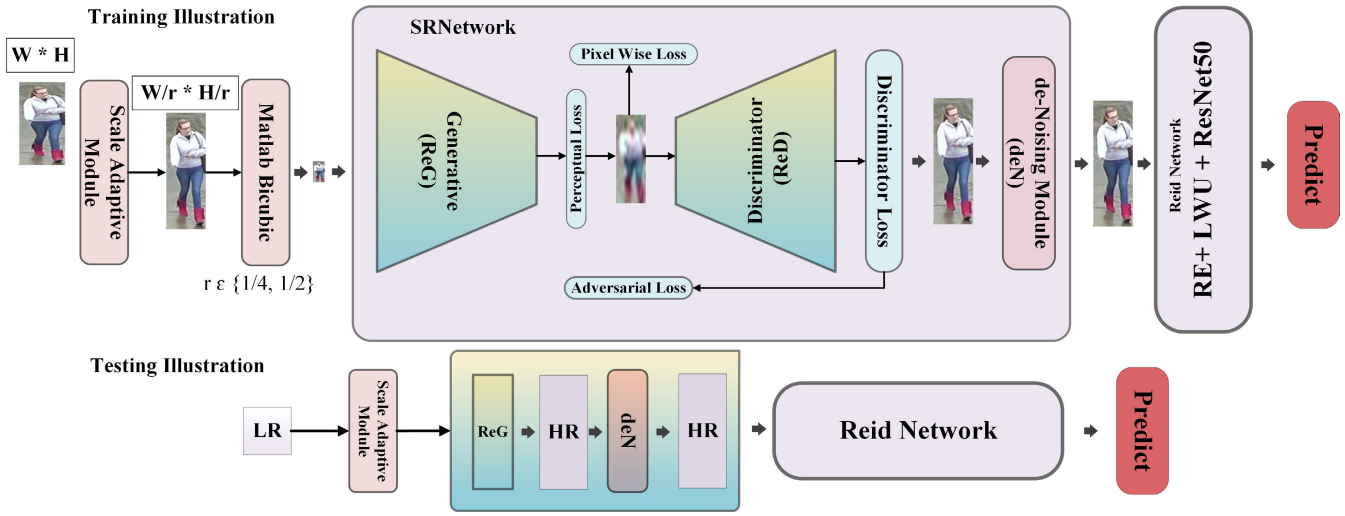

FIGURE 3. The overall architecture of the proposed network. It consists of scale adaptive module, SRNet and Re-identification network. The scale-adaptive module is used to resize the images to a pre-defined resolution. The SRNet takes fixed low-resolution images and re-creates them in the HR form and passed them to the re-identification network. The upper part of the diagram depicts the training of the network, and the lower part illustrate the testing of the network. In testing, the generated SR images are directly passed to the de-noising module.

C. SUPER-RESOLUTION ADVANCEMENTS

The extremely challenging task of creating an HR counterpart of its LR image is known as super-resolution (SR). It has attracted significant attention within the computer vision community and has a wide range of practical applications [52]. Recently, Haris *et al.* [53] proposed a recurrent framework for super-resolution, where they extracted context from continuous frames and combines these contexts to produce recurrent output frames by a back-projection module. Similarly, Li [54] developed a feedback network for image super-resolution that employs an iterative up-and-down sampling feedback block with more dense skip connections to learn better representations. Shamsolmoali *et al.* [55] utilized the least square function as a discriminant loss function for stable training and introduced a gradual GAN to use all the image details. Therefore, the proposed model can effectively create SR results, even up to large scaling factors. Xu *et al.* [56] developed a pipeline to re-generate realistic training data by simulating the imaging process and designed a dual CNN to capture the extracted radiance information, initially in raw images. Zhang *et al.* [57] introduced local and non-local attention blocks to extract features that capture the long-range dependencies between pixels. Dai *et al.* [58] employed an attention mechanism to capture long-distance spatial contextual information for single image super-resolution, which generates a better performance of PSNR. Several frameworks are developed based on a densely connected network, residual blocks [52], [59], deep back projection [60], and dense residual network [61]. Besides, unsupervised learning and reinforcement learning methods [62], [63] also have been exploited to solve SR problems. However, hallucinated details of images that have been processed by the SRGAN [52] are often accompanied by over smooth textures. The ESRGAN [27] has been introduced to address this problem, which can effectively enhance the restoration and perception quality of the image by using Residual-in-Residual Dense Block.

This study addresses the LR-PREID problem and introduces a novel multi-scale-adaptive super-resolution person re-identification (MSA-SR PREID). As compared to most of the previous studies, the proposed network recovers the visual details lost in the image acquisition process. The de-noising module further refines the visual descriptors, which in turn directly assist the re-identification network for feature extraction. Re-identification parameters (Random erasing + Linear Warm UP) further enhance the feature extraction capability. The ablation studies report the effectiveness of the proposed network. The designed framework achieved the state-of-the-art results on all four benchmarks.

III. MSA-SR-PREID ARCHITECTURE

The complete illustration of MSA-S-PREID is shown in Figure-3. It is consists of three parts. First, the person images of arbitrary resolution are passed to the scale-adaptive module to resize them to a pre-defined fixed resolution, as shown in Table-1. Second, the SRNet, comprises of ESRGAN and de-noising module, is used to re-create the LR images into HR and removes the de-compressing artifacts, respectively. Third, the final HR image is passed to the re-identification network (REID) to extract and learn discriminative features.

A. SCALE-ADAPTIVE MODULE

The image height and width is represented by W & H, respectively, as shown in Figure-3. In the training phase, to allow the designed framework to handle images at different resolutions, we embedded the scale-adaptive module. It is a high-quality convolution based filter named Alias,

TABLE 1. Statistical analysis of benchmarks.

Dataset	Identities			Images			LR Size
	Total	Training	Testing	Total	Training	Testing	
VIPeR	632	316	316	1264	632	632	32 × 12
DukeMTMC-rID	1812	702	702+408	36441	16522	17661	70 × 30
MSMT17	4101	1042	2246+807	126411	32964	68239	Width = [32,128)
Market1501	1501	751	751	32217	12937	19733	Width = [8,32)

from IMAGE library. It takes images on different resolutions and resize them on a fixed resolution, which varies for each dataset. We employed Scale adaptive module for VR-Market1501, and VR-MSMT17 benchmarks, to resize them on a single resolution before passing them to the SRNet.

B. SUPER-RESOLUTION NETWORK(SRNet)

SRNet is composed of ESRGAN and the de-Noising module. It takes the output of the scale-adaptive module and generates the high-resolution images by a enlarging factor of 2, and 4. The generator network generates an image sample $I^{z+1} = G_z(I^z)$. However, before passing it to the final activation function, it extracts the feature map to calculate the perceptual loss. The pixel-wise loss is calculated, and the generated image is passed to the discriminator network to distinguish between generated image I^{z+1} and real image $\hat{I}^{z+1}$. The real image $\hat{I}^{z+1}$ is fed to the discriminator network during the training process, which has the same resolution as of the super-resolution image I^{z+1}. After that, the generator network updates its adversarial loss function and creates another sample of the same image. This whole process continues until the discriminator network could not differentiate between real and generated images. We train the generator function G_z to estimate the corresponding HR counterpart of $\hat{I}^{z+1}$ for a given LR input image.

Total loss of SRNet is:

$$L_{Totalloss} = L_{Gen}(L_{PerceptualLoss} + \lambda L_G^{Ra} + \gamma L1) + L_{Dis}^{Ra} \quad (1)$$

In equation (1), L_G^{Ra} represents the adversarial loss of relativistic generator network, $L1$ is pixel-wise loss, and λ & γ are trade-off parameters.

1) PERCEPTUAL LOSS

The perceptual loss works on a feature-level loss to enhance the perceived quality and texture details of the generated images [64]. The feature map of the original image $\hat{I}^{z+1}$ and the generated image I^{z+1} is compared using Euclidean distance. Following the concept of [27], the feature map was extracted before applying the final activation function in the generator network. This approach solved the following problems:

- In the person re-identification, the illumination variation exists in most of the benchmarks. The extraction of feature maps after the activation function further cause illumination inconsistency, which directly affects the model performance.
- It provides strong supervision between feature maps in reconstructing the LR into HR. As most of the person images are not sufficiently HR, this factor significantly improves the model re-generation capability.

Feature map α_{ij} obtained by after j^{th}-convolution layer and before the i^{th} max-pooing layer. The perceptual loss is calculated as the Euclidean distance between the feature representations of a super-resolution image $G_z(I^z)$ and its corresponding real image $\hat{I}^{z+1}$. Euclidean computation between feature maps is given in equation (2)

$$L_{Perceploss} = \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\alpha_{ij} \left(\hat{I}^{z+1} \right)_{xy} - \alpha_{ij} \left(G_z \left(I^z \right) \right)_{xy} \right)^2 \quad (2)$$

Rather than encouraging the pixels of the output image I^{z+1} to exactly match the pixels of the target image $\hat{I}^{z+1}$, perceptual loss encourages them to have similar feature representations as computed by the loss network.

2) PIXEL-WISE LOSS

The network utilizes the pixel-wise loss to improve the pixel-level accuracy of the generated image. It forces the HR image I^{z+1} to be similar to the ground truth $\hat{I}^{z+1}$ on the pixel values. The L1 loss is employed for better performance and convergence as compared to the L2 loss that often results in oversmooth results.

$$L1 = \sum_{x}^{W} \sum_{y}^{H} \left\| G_z \left(I^z \right)_{xy} - \left(\hat{I}^{z+1} \right)_{xy} \right\|_1 \quad (3)$$

From equation (3), the L1-Norm distance between SR image $G_z(I^z)_{xy}$ and ground truth image$(\hat{I}^{z+1})_{xy}$ is calculated.

3) RELATIVISTIC GAN LOSSES

Most of the preliminary studies used standard GAN in image SR, however, we employ a realistic discriminatory loss in our SR network to ensure the generated HR images are more natural and realistic than gallery images. The standard discriminator D_{st} in generative adversarial network aims to classify the images, according to the equation (4):

$$\begin{aligned} D_{st} &= \sigma(f_d(\hat{I}^{z+1}) \rightarrow 1 \\ D_{st} &= \sigma(f_d\left(I^{\mathbf{z}}\right))) \rightarrow 0 \end{aligned} \quad (4)$$

Equation (4), represents the working of standard GAN. Here, D_{st} is the output of the discriminator to classify whether the input images are real or generated, $f_d(.)$ is the discriminator feature vector, and σ represents the sigmoid function. Here,

adversarial loss works as a binary classifier to check either it is real or fake. We employ the relativistic GAN D_{st} [65], which computes the probability to differentiate between the actual $\hat{I}^{z+1}$ and the generated data $G_z(I^z)$ by computing the distance:

$$D_{Ra}(\hat{I}^{z+1}, G_z\left(I^z\right)) \tag{5}$$

Relativistic GAN (RGAN) outputs the person images with sharp edges and contains more visual and frequency details as compared with standard GAN. This is explained in the equation (6):

$$\begin{aligned} &D_{Ra}(Real, Fake) = C(Real) - E(C(Fake)) \rightarrow 1 \\ &Calculating\ how\ much\ more\ realistic\ than\ the\ fake\ one \\ &D_{Ra}\left(Fake, Real\right) = C\left(Fake\right) - E\left(C\left(Real\right)\right) \rightarrow 0 \\ &Calculating\ how\ much\ less\ realistic\ than\ the\ real\ one \end{aligned} \tag{6}$$

Here $E\left(.\right)$ is the average of all (Fake or Real) data in the mini-batch and $f_d\left(.\right)$ represents the discriminator output value. This slight improvement makes the model more effective than the classic discriminator network.

Equation 7 represents the discriminator network loss:

$$\begin{aligned} L_G^{Ra} = &-E_{\hat{I}^{z+1}}\left[log\left(D_{Ra}\left(\hat{I}^{z+1}, G_z\left(I^z\right)\right)\right)\right] \\ &- E_{G_z(I^z)}\left[\log\left(1 - D_{Ra}\left(G_z\left(I^z\right), \hat{I}^{z+1}\right)\right)\right] \end{aligned} \tag{7}$$

Contrary to it, Equation 8 represents the adversarial loss for the relativistic generative network.

$$\begin{aligned} L_G^{Ra} = &-E_{\hat{I}^{z+1}}\left[\log\left(1 - D_{Ra}\left(\hat{I}^{z+1}, G_z\left(I^z\right)\right)\right)\right] \\ &- E_{G_z(I^z)}\left[\log\left(D_{Ra}\left(G_z\left(I^z\right), \hat{I}^{z+1}\right)\right)\right] \end{aligned} \tag{8}$$

The network is trained in a mutual training strategy for both real image $\hat{I}^{z+1}$ and generated images $G_z(I^z)$, by simultaneously reducing the loss of the discriminator and generator networks. In the traditional GAN, when the gradient of the discriminator reaches the optimal value, i.e., $(1 - D_{\hat{I}^{z+1}}) \rightarrow 0$ is optimal; it stops learning from real images $\hat{I}^{z+1}$ and focuses mostly on the generated $G_z(I^z)$ images. At this stage, traditional GAN is not learning how to generate images more realistic. In contrast, RGAN learns from both as its gradients depend on both $\hat{I}^{z+1}$ and $G_z(I^z)$.

C. de-Noising (deN) IN GAN

The presence of noise disrupt not only the human perception of images but also the efficiency of networks. Therefore, to have well-performing de-noising strategies at our disposal is crucial to make sure the optimal functionality of image processing pipelines. The utilization of the Joint Photographic Group (.jpg) format to train the GAN network creates noisy artifacts due to the decompressing nature of.JPG images. To solve this problem, we incorporate a non-local means de-noising [66] module. It is based on the principle of replacing a pixel color with an average of identical pixel colors. It takes a noisy input image and selects a pixel in the noise; takes the search patch around the selected pixel and search for similar patches in the image; calculates the Euclidean distance and computes the average of all the Euclidean distances. The selected noisy pixel is finally replaced by the resultant pixel, which enhances the overall performance of the network. We can also balance the perceptual quality and fidelity by controlling the de-Noising hyper-parameters without re-training the model. Removal of unwanted artifacts assists the re-identification model in extracting information. The ablation studies suggests that it is effective to eliminate unwanted decompressing factors.

D. PERSON RE-IDENTIFICATION NETWORK

For REID, we employ the Resnet-50 [28] with dropout and Kaiming weights initialization strategy. The pre-trained weights on Image-Net are used as an initialization. We employ two hyper-parameters: a) random erasing along with the standard augmentation techniques and b) linear warm-up to improve the generalization of the PREID network. They are described below.

1) RANDOM ERASING

Random erasing introduced by Zhong *et al.* [67] is analogous to dropout in the case of input data space embedded into the network architecture. It is designed, specifically, for image recognition challenges. It is a promising technique and assures that a network pays attention to the entire image. It works on the random selection of a patch in an image and replaces it with a mean of 0s to 255s pixel values or random values. We employ image and object aware random erasing that focuses on both the image background and the object (Or ground truth of the object). Figure-4

FIGURE 4. The illustration of image and object aware random erasing. It works by selecting a random patch from both the Image and bounding box of the object to replace it with a mean of 0s to 255s pixel values or random values.

By replacing specific patches, it forces the model to learn other characteristics of the image.

2) LINEAR WARM UP

Linear Warm-Up uses the linear scaling rule where the learning rate is increased by 'k' times (the value of 'k' depends upon the batch size) [68]. It has two strategies: constant warm-up and gradual warm-up. The gradual warm-up is applied in the proposed model because constant warm-up

causes a spike in training when the learning rate changes abruptly. It starts with a small learning rate and then gradually increases by a constant for each epoch until it reaches k times.

The final HR image goes directly to the re-identification network, which extracts the features to use for person re-identification.

IV. EXPERIMENTS

This section aims to explain the experimental setup, implementation settings, results, and evaluation details.

A. EXPERIMENTAL SETUP

We evaluate the proposed method on four large-scale person re-identification benchmarks: VIPeR, DukeMTMC-reID, VR-MSMT17, and VR-Market1501. The results are compared with the previous state of the art on all four benchmarks. The Table-1 represents the statistical analysis of the benchmarks.

VIPeR (View Point Invariant Pedestrian recognition) [69] dataset was introduced in 2007, one of the most challenging datasets due to one to one corresponding. Following SING [19], every image is down-sampled with a ratio $r \in \frac{1}{4}$, such that $(\frac{H}{4} \times \frac{W}{4} \times 3)$, as shown in Figure-5. MLR-VIPeR has 316 identities for training and testing.

(a) MLR-DukeMTMC-reID (b) MLR-VIPeR

FIGURE 5. Example image pairs from two datasets. Each column shows two images of the same identity with fixed resolutions and different scales, where images in the bottom row are LR.

DukeMTMC-REID (Duke Multi-Target Multi-Camera REID) [70] is a subset of the DukeMTMC dataset for image-base re-identification in the format of Market1501 dataset. It contains 16, 522 training images of 702 identities, 2, 228 query images of the other 702 identities and 17, 661 gallery images (702 ID + 408 distractor IDs).

The transformation of DukeMTMC-ReID into MLR is carried out the same way as for the VIPeR dataset.

MSMT17 [46] is developed by the 15 cameras network and contains many real environment-challenging factors. It is not designed for the LR-PREID problem; therefore, it is reconstructed into VR-MSMT17 (Various Resolution MSMT17) by [24]. Images are downsampled to make the width within the range of [32, 128), consisting of 96 different resolutions. Original settings of 1,041 and 3,060 for training and testing, respectively on VR-MSMT17 are kept.

Market1501 [71] is collected in front of a supermarket at Tsinghua University. Similarly, as MSMT17, Market1501 is reconstructed into VR- Market1501 (Various Resolution Market1501) by [24]. All images are down-sampled to make the width within the range of [8, 32), consisting of 24 different resolutions separately. We kept the original settings of 751 and 710 identities for training and testing, respectively.

Compared with existing datasets, VR-Market1501 and VR-MSMT17 are considerably more substantial in size and are more challenging due to the extensive range of resolution variance in both query and gallery images.

B. IMPLEMENTATION DETAILS

We employed a frequent performance metrics, Cumulative Matching Characteristic (CMC top-K) for evaluation. CMC addresses the probability of reacquiring a minimum one correct identity within the top-K predictions (CMC top-1, top-5 and top-10 are adopted here). The Standard training-testing ratio is used and is summarized in Table-1.

The training process includes three steps.

1− The scale adaptive module is used to transform the images on a pre-defined fixed resolutions.

2− The resized images are then downsampled by using Matlab bi-cubic kernel function to obtain the LR images. Super-Resolution network is trained with the LR images with a mini-batch size eight. The training of the SR network is divided into two steps. Firstly, the network is trained with the $L1-$ loss to improve the PSNR value. The learning rate (lr) is initialized with 2×10^{-4} and lr is set to reduce by factor of 2 after every 1×10^4 iterations. Secondly, the PSNR oriented model is initialized for the generator network. Total loss of GAN is calculated with the function:

$$L_{TotalLoss} = L_{Gen}\left(L_{PercepLoss} + \lambda L_G^{Ra} + \gamma L1\right) + L_{Dis}^{Ra} \quad (9)$$

where $\lambda = 5 \times 10^{-3}$ & $\gamma = 1 \times 10^{-2}$. The learning rate is set to 1×10^{-4}, and reduces at every $10K$, $20K$, $30K$, and $40K$ iterations with a total number of $50K$. Training is carried out on RGB channels. Augmentation is performed on training dataset with random horizontal flip and 90−degree rotation.

3− A pre-trained re-identification network on ImageNet is trained with HR images obtained from super-resolution network. Training is conducted on RTX2080Ti with a total time of 18 hours. The MSA-SR PREID network is developed on PyTorch framework.

V. EVALUATION OF MSA-SR-PREID

This section consists of two parts. In first, we show the effectiveness of SRNet in the image generation task. The second part deals with the evaluation of the Re-Identification network on four competitive benchmarks.

A. SRNet EVALUATION

In this section, the aim is to demonstrate the effectiveness of SRNet in high-resolution image generation tasks. Figure-6 illustrates low-resolution images (LR), HR images (HR) generated by SRNet, and HR images with de-Noising (HR+deN). The HR images generated by SRNet, contains unwanted decompressing artifacts, which can be observed in the Figure-6. By integrating the de-Noising module, the Super

FIGURE 6. Super-resolution results of our modified version of ESRGAN. From left to right, low-resolution image (LR), high-resolution image (HR), and high-resolution image after passing through image de-Noising module (HR+deN).

Resolution network generates more adequate images, and it eliminates those artifacts and provides better visibility for feature extraction. The HR images generated by SRNet preserve information of a person's body, and discriminate background environment from persons with prominent edges.

B. PERSON RE-IDENTIFICATION EVALUATION

1) COMPARE METHODS

We compare the findings of our network with previous state-of-the-art LR-PREID methods, which includes, SDF [22], DAMA [76], JUDEA [21], SLDL [20], SING [19], CSR-GAN [16], RAIN [24], CR-GAN [73], Densenet121 [77], SE-resnet50 [72], DSPDL [78] and methods developed for standard PREID, such as FDGAN [17], CamStyle [43], DSMIN [75], and PL-Net [74]. For standard person re-identification, the training set contains HR images for each identity.

2) RESULTS

The experiments are performed on four benchmarks, MLR-VIPeR, MLR-Duke-MTMC-reID, VR-MSMT17, and VR-Market1501. The quantitative results of MLR-VIPeR and MLR-Duke are illustrated in Table-2. Our network yields Rank-1 accuracy of 62.00% and Rank-5 accuracy of 74.48% on MLR-VIPeR. The accuracy improvement is 14% and 2.08% than the PL-NET [74] in Rank-1 and Rank-5, respectively. The increase in the performance is 5.36% in Rank 1, as compared to PL-Net+LOMO [74]. However, PL-Net+LOMO [74] performed better in Rank-5, as it employs LOMO, which is a method based on hand-crafted feature extraction machine learning approach. As the VIPeR dataset consists of fewer training samples for each identity, it is challenging to learn discriminative features, especially when it comes to end-to-end training of deep neural networks from scratch. For MLR-Duke, MSA-SR-PREID achieves state-of-the-art results i.e., 79.62% Rank-1, and 90.03% Rank-5 accuracy. In comparison with CR-GAN [73], improvement in the results are 1.86%, and 1.9%, on Rank-1 and Rank-5, respectively.

TABLE 2. The experimental results of the proposed network on MLR-VIPeR and MLR-Duke.

Methods	MLR-VIPeR		MLR-DukeMTMC-reID	
	Rank 1%	Rank 5%	Rank 1%	Rank 5%
Densenet121 [72]	31.4	63.1	-	-
SE-resnet50 [72]	33.5	63.6	-	-
JUDEA [21]	26	55.1	-	-
SLD2L [20]	20.3	44	-	-
SDF [22]	9.52	38.1	-	-
SING [19]	33.5	57	65.2	80.1
CSR-GAN [16]	37.2	62.3	67.6	81.4
CR-GAN [73]	43.1	68.2	75.6	86.7
RIPR [24]	41.6	64.9	-	-
CamStyle [43]	34.4	56.8	64	78.1
FD-GAN [17]	39.1	62.1	67.5	82
PL-Net-ResNet50 [74]	47.47	72.47	-	-
DSMIN [75]	49.1	-	-	-
PL-NET + LOMO [74]	56.65	**82.59**	-	-
MSA-SR-PREID	**62.00%**	74.48%	**79.06%**	**90.00%**

On VR-MSMT17 and VR-Market1501, the experimental results and comparison are summarized in Table-3. We adopted the standard settings of [24] for experimentation. The experiments are performed with three different configurations settings on both datasets. First, the experiment is conducted without re-creating HR images i.e., without SRNet. The PREID network is trained on LR images. On VRMSMT17, the network achieves 52.85% at Rank-1 and 68.91% at Rank-5, outperform the [19], [72], [77], and [16]. On VR-Market1501, it yields 60.68% at Rank-1,

TABLE 3. The experimental results on VR-Market1501 and VR-MSMT17 benchmarks.

Methods	VR-MARKET1501		VR-MSMT17	
	Rank 1%	Rank 5%	Rank 1%	Rank 5%
Densenet121 [77]	60	78.8	51.2	67.4
SE-resnet50 [72]	58.2	78.6	52.3	68.9
SING [19]	60.5	81.8	52.1	68.3
CSR-GAN [16]	59.8	81.3	51.9	67.5
RIPR [24]	66.9	84.7	55.5	72.4
ResNet50 (RE + WU)	60.68	79.83	52.85	68.91
MSA-SR-PREID (r $\epsilon \frac{1}{4}$)	67.1	84.67	54.78%	70.90%
MSA-SR-PREID (r $\epsilon \frac{1}{2}$)	**68.26%**	**85.71%**	**60.65%**	**75.20%**

79.83% at Rank-5, better [16], [19], [72], [77], and RIPR-ResNet50 [24]. Second, the LR images are passed to the SRNet to recreate them in HR with a downsampling factor of $r \in \frac{1}{4}$. In comparison with [16], [19], [72], [74], the proposed network achieves 8.9%, 6.6%, 7.3% and 0.2% improvement at Rank-1 on VR-Market1501, respectively. The model outperformed the CSR-GAN with 2.8%, SING by 2.68%, and [74] by 2.48% on Rank-1 on VR-MSMT17. Third, to further elaborate the effectiveness of MSA-SR-PREID, the experiments using downsampling factor $r \in \frac{1}{2}$ are conducted. MSA-SR-PREID yield state of the art results on both benchmarks. It surpasses the previous state of the art [74] by 5.15% at Rank 1, 2.8% at Rank 5, on VR-MSMT17. When trained on VR-Market1501, 1.36%, and 1.01% increase in the performance is reported on Rank-1 and Rank-5, respectively.

The proposed framework achieves the state-of-the-art results in the meantime utilizing less number of parameters as compared to the CSR-GAN [16]. Our designed approach is cost-effective, lightweight, yet deep architecture maintains the optimal network depth. Table 4 reports the number of parameters comparison against CSR-GAN, a super-resolution based LR-PREID approach.

TABLE 4. Comparisons of parameter numbers.

Method	Number of parameters
CSR-GAN	172,157,836
MSA-SR-PREID	54,787,980

VI. ABLATION STUDIES

This section analyzes the benefit of three major hyper-parameters tested during extensive experimentation. Moreover, experiments are also conducted to evaluate the performance of SRNet and losses used in the network.

A. EFFECTIVENESS OF SRNet IN PREID

The experimentation process is carried out to quantify the capability of SRNet, against nearest and Bi-cubic interpolation functions. Nearest and bi-cubic interpolation function is used for upsampling the images. The PREID network is trained on the upsample images. Table-5 reports the re-identification results on MLR-VIPeR. The nearest interpolation works on the principle of translating the nearby pixel values as compared to the bicubic interpolation, which works on averaging the four translated pixel values for each output pixel value. Both interpolation functions are not an optimal approach for upsampling the LR images as it produces the blurry images and causes the jaggies effect. The SRNet utilizes the advanced architecture for re-generating the High-resolution images with better perceptual details, sharp edges, and contains more spatial and context information. This, in turn, assists the re-identification network to extract the more robust and distinctive identity features.

B. CHOICE OF LOSSES IN SRNet

The experiments are performed to show the importance of Pixel Wise loss and Perceptual Loss in SRNet. Table 7 summarizes the evaluation results with Perceptual loss (PR Loss), Pixel Wise loss (PW Loss), and both PWLoss + PR Loss. Note that the experimental results mentioned above are carried with the same settings of adversarial loss and PREID network. In PR Loss, perceptual loss and GAN loss are used with all other hyperparameters, achieves 61.60% at Rank-1, which is 0.40% performance drop. In PWLoss, pixel-wise loss and GAN loss are employed with all other hyperparameters, achieves 62.08% at Rank-l. Although the Rank-1 performance is improved by 0.08%, the proposed framework exhibits strong generalization capability by the incorporation of both losses. It can be seen with a 3.23% enhancement in Rank-5.

C. CONTRIBUTION OF IMAGE DE-NOISING

The experiments are conducted with and without de-Noising module on MLR-VIPeR and MLR-Duke. As reported in Table-6, network reports the boost in performance with DN module in all ranks on both benchmarks, as Rank-1 accuracy of MLR-VIPeR increased from 59.80% to 62.00%. Similarly, the results enhancement can also be observed on MLR-Duke proves the importance of the image de-Noising module.

D. CONTRIBUTION OF PREID PARAMETERS

Initially, experiments were performed with the standard baseline model. Table-8 reports the Rank-1 results of MLR-VIPeR and MLR-Duke. First, we show the effect of random erasing (RE) by adding it to the baseline model. As shown in Table-8, the addition of random erasing improves the results over baseline on both benchmarks. It demonstrates that random-erasing is an efficient way to improve the discrimination capability of the target domain. Second, the experiment is performed with the addition of linear warm-up (WU) in the baseline. The results improvements on both benchmarks can be observed in the Table-8. Finally, the best performance is achieved by the model on the incorporation of both hyper-parameters.i.e. 5%, 3.82% improvement to the baseline model can be observed on Rank 1 when tested on MLR-VIPeR and MLR-Duke, respectively. It exhibits the complimentary benefit of these two components.

TABLE 5. SRNet evaluation against nearest and bi-cubic interpolation.

Interpolation/SRNet	Nearest		bicubic		**MSA-SR-PREID**	
Dataset	Rank -1	Rank 5	Rank 1	Rank 5	Rank 1	Rank 5
MLR-VIPeR	55.79%	64.21%	57.59	65.85	**62.00%**	**74.48%**

TABLE 6. Evaluation of with and without the utilization of the de-noising module.

Image DeNoising/Dataset	MLR-VIPeR			MLR-DukeMTMC-reID		
	Rank -1	Rank-5	Rank 10	Rank -1	Rank-5	Rank 10
Without deN	59.80%	73.10%	83.40%	78.30%	89.20%	92.10%
With deN	**62.00%**	**74.48%**	**84.80%**	**79.60%**	**90.00%**	**92.70%**

TABLE 7. The contribution of pixel wise loss (PWLoss) and perceptual loss (PRLoss).

	MLR-VIPeR	
Loss	Rank 1%	Rank 5%
MSA-SR-PREID (PRLoss)	61.60%	70.88%
MSA-SR-PREID (PWLoss)	62.08%	71.25%
MSA-SR-PREID (PWLoss + PRLoss)	**62.00%**	**74.48%**

TABLE 8. Contribution analysis of re-identification network hyperparameters.

Model/Dataset	MLR-Duke	MLR-VIPeR
ResNet50	75.80%	56.96%
ResNet50 + RE	76.79%	58.86%
ResNet50 + WU	78.63%	59.17%
ResNet50 +WU+RE	**79.62%**	**62.00%**

VII. CONCLUSION

In this paper, we have proposed multi-scale-adaptive super-resolution person re-identification framework, intending to address the LR-PREID. We have introduced a scale-adaptive module to deal with the resolutions variations problem and a de-Noising module to solve the noising effect in the image generation process by GAN during the utilization of.JPG format images. We have shown that our model can generate more realistic and natural images and able to extracts the deep features to use for person re-identification. Our proposed method outperformed the previous state of the art LR-PREID methods by a large margin.

REFERENCES

[1] F. M. Khan and F. Bremond, "Person re-identification for real-world surveillance systems," 2016, *arXiv:1607.05975*. [Online]. Available: http://arxiv.org/abs/1607.05975
[2] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 1–37, Nov. 2013.
[3] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 393–402.
[4] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3633–3642.
[5] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.
[6] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 811–820.
[7] A. Abbas, A. Muhammad, Z. Lian, and C. Huang, "Synchronized ReID with expanded cross neighborhood re ranking," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2018, pp. 144–148.
[8] Z. Wang, Z. Wang, Y. Wu, J. Wang, and S. Satoh, "Beyond intramodality discrepancy: A comprehensive survey of heterogeneous person re-identification," *CoRR*, vol. abs/1905.10048, 2019. [Online]. Available: http://arxiv.org/abs/1905.10048
[9] J. Meng, S. Wu, and W.-S. Zheng, "Weakly supervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 760–769.
[10] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2148–2157.
[11] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5735–5744.
[12] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 667–676.
[13] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.
[14] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.
[15] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5363–5372.
[16] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded SR-GAN for scale-adaptive low resolution person re-identification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, vol. 1, no. 2, p. 4.
[17] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, and X. Wang, "Fd-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1222–1233.
[18] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C.-F. Wang, "Recover and identify: A generative dual model for cross-resolution person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8090–8099.
[19] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[20] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 695–704.

[21] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3765–3773.

[22] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proc. IJCAI*, vol. 2, 2016, p. 6.

[23] Y.-C. Chen, Y.-J. Li, X. Du, and Y.-C. F. Wang, "Learning resolution-invariant deep representations for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8215–8222.

[24] S. Mao, S. Zhang, and M. Yang, "Resolution-invariant person re-identification," 2019, *arXiv:1906.09748*. [Online]. Available: http://arxiv.org/abs/1906.09748

[25] Y.-C. Chen, P.-H. Huang, L.-Y. Yu, J.-B. Huang, M.-H. Yang, and Y.-Y. Lin, "Deep semantic matching with foreground detection and cycle-consistency," in *Computer Vision—ACCV*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham, Switzerland: Springer, 2019, pp. 347–362.

[26] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[27] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 11133, L. Leal-Taixé and S. Roth, Eds. Munich, Germany: Springer, Sep. 2018, pp. 63–79.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.

[30] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[31] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.

[32] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: http://arxiv.org/abs/1711.08184

[33] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

[34] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[35] H. Fang, J. Chen, and Q. Tian, "Multi-branch body region alignment network for person re-identification," in *MultiMedia Modeling*. Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, Eds. Cham, Switzerland: Springer, 2020, pp. 341–352.

[36] Y. Huang, Z.-J. Zha, X. Fu, and W. Zhang, "Illumination-invariant person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 365–373.

[37] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[38] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 542–551.

[39] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4610–4617.

[40] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107036.

[41] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4099–4108.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[43] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.

[44] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "CrDoCo: Pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1791–1800.

[45] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," 2017, *arXiv:1711.03213*. [Online]. Available: http://arxiv.org/abs/1711.03213

[46] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[47] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–19, Apr. 2020.

[48] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.

[49] Y.-C. Chen and W. H. Hsu, "Saliency aware: Weakly supervised object localization," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1907–1911.

[50] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation," 2019, *arXiv:1906.05857*. [Online]. Available: http://arxiv.org/abs/1906.05857

[51] Y. Huang, Z.-J. Zha, X. Fu, R. Hong, and L. Li, "Real-world person re-identification via degradation invariance learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14084–14094.

[52] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[53] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3897–3906.

[54] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3867–3876.

[55] P. Shamsolmoali, M. Zareapoor, R. Wang, D. K. Jain, and J. Yang, "G-GANISR: Gradual generative adversarial network for image super resolution," *Neurocomputing*, vol. 366, pp. 140–153, Nov. 2019.

[56] X. Xu, Y. Ma, and W. Sun, "Towards real scene super-resolution with raw images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1723–1731.

[57] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," 2019, *arXiv:1903.10082*. [Online]. Available: http://arxiv.org/abs/1903.10082

[58] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.

[59] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[60] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[61] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," 2018, *arXiv:1812.10477*. [Online]. Available: http://arxiv.org/abs/1812.10477

[62] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 701–710.

[63] K. Yu, C. Dong, L. Lin, and C. C. Loy, "Crafting a toolchain for image restoration by deep reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2443–2452.

[64] J. Johnson, A. Alahi, and L. Fei-Fei, *Perceptual Losses for Real-Time Style Transfer and Super-Resolution* (Lecture Notes in Computer Science), vol. 9906, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, 2016, pp. 694–711.

[65] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*. [Online]. Available: http://arxiv.org/abs/1807.00734
[66] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Process. Line*, vol. 1, pp. 208–212, Sep. 2011.
[67] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896
[68] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: Training imagenet in 1 hour," 2017, *arXiv:1706.02677*. [Online]. Available: http://arxiv.org/abs/1706.02677
[69] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. VS-PETS Workshop*, 2007, vol. 3, no. 5, pp. 1–7.
[70] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, *Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking* (Lecture Notes in Computer Science), vol. 9914, G. Hua and H. Jégou, Eds. Springer, 2016, pp. 17–35.
[71] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[72] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[73] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, and Y.-C. Frank Wang, "Cross-resolution adversarial dual network for person re-identification and beyond," 2020, *arXiv:2002.09274*. [Online]. Available: http://arxiv.org/abs/2002.09274
[74] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[75] L. Wu, Y. Wang, X. Li, and J. Gao, "What-and-where to match: Deep spatially multiplicative integration networks for person re-identification," *Pattern Recognit.*, vol. 76, pp. 727–738, Apr. 2018.
[76] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, p. 1541.
[77] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[78] K. Li, Z. Ding, S. Li, and Y. Fu, "Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

MUHAMMAD ADIL received the M.S. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, where he is currently pursuing the Ph.D. degree in computer science. His research interests include deep learning, person re-identification, and object detection.

SAQIB MAMOON received the B.S. degree in computer science from the University of Punjab, Pakistan, and the M.S. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, where he is currently pursuing the Ph.D. degree. His research interests include machine learning and computer vision.

ALI ZAKIR received the B.S. and M.S. degrees in computer science from the University of Peshawar, Pakistan. He is currently pursuing the Ph.D. degree with the Nanjing University of Science and Technology. His research interests include multi-object detection and tracking.

MUHAMMAD ARSLAN MANZOOR received the M.S. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China. He is currently pursuing the Ph.D. degree with the East China University of Science and Technology. His research interests include deep learning, natural language processing, and sentiment analysis.

ZHICHAO LIAN received the B.S. and M.S. degrees in computer science from Jilin University, Changchun, China, in 2005 and 2008, respectively, and the Ph.D. degree in electrical and electronics engineering from Nanyang Technological University, Singapore, in 2013. From 2012 to 2014, he was a Postdoctoral Associate with the Department of Statistics, Yale University. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include image processing, pattern recognition, and neuroimaging.