Real-world Person Re-Identification via Degradation Invariance Learning

Yukun Huang[1], Zheng-Jun Zha[1],[*] Xueyang Fu[1], Richang Hong[2], Liang Li[3]
[1]University of Science and Technology of China, China
[2]Hefei University of Technology, China
[3]Institute of Computing Technology, Chinese Academy of Sciences, China
kevinh@mail.ustc.edu.cn, {zhazj, xyfu}@ustc.edu.cn, hongrc.hfut@gmail.com, liang.li@ict.ac.cn

Abstract

Person re-identification (Re-ID) in real-world scenarios usually suffers from various degradation factors, e.g., low-resolution, weak illumination, blurring and adverse weather. On the one hand, these degradations lead to severe discriminative information loss, which significantly obstructs identity representation learning; on the other hand, the feature mismatch problem caused by low-level visual variations greatly reduces retrieval performance. An intuitive solution to this problem is to utilize low-level image restoration methods to improve the image quality. However, existing restoration methods cannot directly serve to real-world Re-ID due to various limitations, e.g., the requirements of reference samples, domain gap between synthesis and reality, and incompatibility between low-level and high-level methods. In this paper, to solve the above problem, we propose a degradation invariance learning framework for real-world person Re-ID. By introducing a self-supervised disentangled representation learning strategy, our method is able to simultaneously extract identity-related robust features and remove real-world degradations without extra supervision. We use low-resolution images as the main demonstration, and experiments show that our approach is able to achieve state-of-the-art performance on several Re-ID benchmarks. In addition, our framework can be easily extended to other real-world degradation factors, such as weak illumination, with only a few modifications.

1. Introduction

Person re-identification (Re-ID) is a pedestrian retrieval task for non-overlapping camera networks. It is very challenging since the same identity captured by different cameras usually have significant variations in human pose, view, illumination conditions, resolution and so on. To withstand the interference of identity-independent variations, the ma-
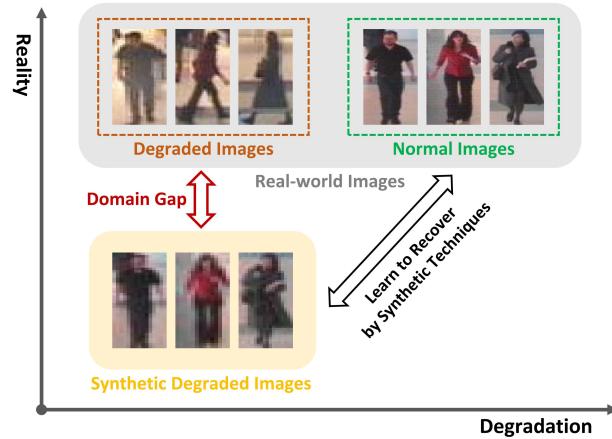

Figure 1. Existing methods [28, 20, 18] only use simple synthetic techniques, such as down-sampling for low resolution or gamma correction for low illumination, to alleviate the image degradation issue in Re-ID. This is far from the complex degradation in real-world scenarios, which leads to the domain gap.

jor target of person Re-ID is to extract robust identity representations. With the powerful representation learning capability, deep convolutional neural networks-based methods have achieved remarkable performance on publicly available benchmarks. For example, the rank-1 accuracy on Market-1501 [57] has reached 94.8% [58], which is very close to the human-level performance.

However, there are still some practical issues that need to be solved for real-world surveillance scenarios, and low quality images caused by various degradation factors are one of them. Several previous works [18, 38] have demonstrated that such degradations have a serious negative impact on the person Re-ID task. On the one hand, these degradations lead to pool visual appearances and discriminative information loss, making representation learning more difficult; on the other hand, it brings the feature mismatch problem and greatly reduces the retrieval performance.

---

*Corresponding author.

Existing methods, which focus on alleviating the low-level degradation issue, can be classified into three types:

1) Data augmentation. This kind of methods [1] synthesize more training samples under different low-level visual conditions to improve the generalization performance of the model. However, there is a domain gap between synthetic data and real-world data. For example, most of cross-resolution person Re-ID works use the simple down-sampling operator to generate low-resolution images. While the real-world low-resolution images captured usually contain more degradations, such as noise and blurring.

2) Combination with low-level vision tasks. This type of methods [20, 49, 38, 18], which usually consists of a two-stage pipeline, *i.e.*, combine Re-ID backbone with existing image restoration or enhancement modules to eliminate the effects of degradations. Nevertheless, most existing low-level vision algorithms require aligned training data, which is impossible to collect in real-world surveillance scenarios.

3) Disentangled representation learning. In recent years, some studies attempt to utilize generative adversarial networks (GANs) to learn disentangled representations, which is invariant to some certain interference factors, *e.g.*, human pose [8] or resolution [5, 28]. Ge *et al.* [8] propose FD-GAN for pose-invariant feature learning without additional pose annotations of the training set. To guide the extraction of disentangled features, auxiliary information usually needs to be introduced, which inevitably leads to additional estimation errors or domain bias.

Based on the above observations, we argue that the lack of supervised information about real degradations is the main difficulty in solving real-world Re-ID. This inspires us to think about how to adaptively capture the real-world degradations with limited low-level supervision information. In this work, we propose a Degradation-Invariant representation learning framework for real-world person Re-ID, named **DI-REID**. With self-supervised and adversarial training strategies, our approach is able to preserve identity-related features and remove degradation-related features. The DI-REID consists of: (a) a content encoder and a degradation encoder to extract **content** and **degradation** features from each pedestrian image; (b) a decoder to generate images from previous features; (c) a reality discriminator and a degradation discriminator to provide domain constraints.

To effectively capture the real-world degradations, we generate images by switching the content or degradation features of self-degraded image pairs and real image pairs. The reality discriminator is employed to reduce the domain gap between the synthesis and reality, while the degradation discriminator aims to estimate the degree of degradation of inputs. Utilizing these two discriminators is beneficial for degradation-invariant representation learning. Since the degradation degree may not have a certain discrete division, we use rankGAN [29] as the degradation discriminator

to solve this problem.

In summary, our contribution is three-fold:

- We introduce a new direction to improve the performance of person re-identification affected by various image degradations in real-world scenarios. Our method can alleviate the need for large amounts of labeled data in existing image restoration methods.

- We propose a degradation invariance learning framework to extract robust identity representations for real-world person Re-ID. With the self-supervised and disentangled representation learning, our method is able to capture and remove the real-world degradations without extra labeled data.

- Experiments on several challenging Re-ID benchmarks demonstrate that our approach favorably performs against the state-of-the-art methods. With a few modifications, our method is able to cope with different kinds of degraded images in real-world scenarios.

2. Related Work

Since our work is related with feature representation learning and GANs, we first briefly summarize these two aspects of works.

2.1. Feature Representation Learning

Person re-identification, including image-based Re-ID [24, 57] and video-based Re-ID [54, 31], is a very challenging task due to dramatic variations of human pose, camera view, occlusion, illumination, resolution and so on. An important objective of Re-ID is to learn identity representations, which are robust enough for the interference factors mentioned above. These interference factors can be roughly divided into high-level variations and low-level variations.

Feature learning against high-level variations. Such variations include pose, view, occlusions, *etc*. Since these variations tend to be spatially sensitive, one typical solution is to leverage local features, *i.e.*, pre-defined regional partition [39, 42, 43, 34], multi-scale feature fusion [33, 4, 60], attention-based models [2, 3, 25, 35, 56] and semantic parts extraction [11, 39, 22, 55]. These methods usually require auxiliary tasks, such as pose estimation or human parsing. The research line described above has been fully explored and will not be discussed in detail. In this work, we focus on the low-level variation problem.

Feature learning against low-level variations. Such variations include illumination, resolution, weather, *etc*. Low-level variations tend to have global consistency and can be alleviated by image restoration methods, such as super-resolution or low-light enhancement.

Most existing Re-ID methods, which are developed for low-level variations, focus on the cross-resolution issue.

Jiao *et al.* [20] propose to optimize SRCNN and Re-ID network simultaneously in an end-to-end fashion. It is the first work to introduce super-resolution methods to deal with low-resolution Re-ID. To improve the scale adaptability of SR methods, Wang *et al.* [49] adopt the cascaded SRGAN structure to progressively recover lost details. Mao *et al.* [38] propose Foreground-Focus Super-Resolution module to force the SR network to focus on the human foreground, then a dual stream module is used to extract resolution-invariant features. On the other hand, several methods [5, 28] utilize adversarial learning to extract the resolution-invariant representations.

Similar to the resolution issue, illumination is another common problem in real-world scenarios. The main impact of illumination variations is the change in color distribution, which has been studied in [45, 23]. In view of the lack of illumination diversity in the current re-identification datasets, Bak *et al.* [1] introduce SyRI dataset, which provides 100 virtual humans rendered with different illumination maps. Based on the Retinex theory, Huang *et al.* [18] propose a joint framework of Retinex decomposition and person Re-ID to extract illumination-invariant features.

2.2. Generative Adversarial Networks

Generative Adversarial Network is first proposed by Goodfellow *et al.* [9] to estimate generative models, and then spawn a large number of variants and applications. For person Re-ID, the usage of GANs can be roughly divided into three categories: domain transfer [59, 50, 7, 1, 30, 48], data augmentation [37, 26, 40, 58, 16] and feature representation learning [8, 32, 5, 28, 52]. Liu *et al.* [30] utilize multiple GANs to perform factor-wise sub-transfers and achieves superior performance over other unsupervised domain adaptation methods. Zheng *et al.* [58] integrate the discriminative model and the generative model into a unified framework to mutually benefit the two tasks. Hou *et al.* [16] propose STCnet to explicitly recover the appearance of the occluded areas based on the temporal context information. Similar to STCnet, Li *et al.* [28] propose Cross-resolution Adversarial Dual Network to simultaneously reconstruct the missing details and extract resolution-invariant features.

3. Proposed Method

3.1. Overview

As shown in Figures 2 and 3, our proposed DI-REID consists of two stages: a **degradation invariance learning** by a Degradation Decomposition Generative Adversarial Network (DDGAN) and a robust **identity representation learning** by a Dual Feature Extraction Network (DFEN).

To learn degradation-invariant representations, we attempt to capture and separate the real-world degradation component from a single image. This is an ill-posed problem and extremely difficult since there are no degradation annotations or reference images in the real-world scenarios. Therefore, we synthesize self-degraded images to provide prior knowledge and guidance with self-supervised methods, such as down-sampling, gamma correction and so on. During the degradation invariance learning stage, the aligned self-degraded image pairs and the non-aligned real image pairs are used to train DDGAN in turn, which helps to narrow the domain gap between synthesis and reality.

For identity representation learning, we find that using only degradation-invariant representations does not lead to superior Re-ID performance. This is because degradation invariance forces the network to abandon those discriminative but degradation-sensitive features, *e.g.*, color cues to illumination invariance. Therefore, we design a dual feature extraction network to simultaneously extract both types of features. Besides, an attention mechanism is introduced for degradation-guided feature selection.

3.2. Network Architecture

Content Encoder E_c. The content encoder E_c is used to extract content features for image generation as well as degradation-invariant identity representation, and DDGAN and DFEN share the same content encoder. In particular, a multi-scale structure is employed for E_c to facilitate gradient back propagation.

Degradation Encoders E_d and E_d'. Due to the domain gap between real-world images and self-degraded images, we design a degradation encoder E_d and a self-degradation encoder E_d' to capture the degradation information, respectively. Note that the weights of E_d and E_d' are not shared, and E_d' is encouraged to convert synthetic degradation features into real-world degradation features.

Decoder G. Similar to [58], we utilize the adaptive instance normalization (AdaIN) layers [17] to fuse content and degradation features for image generation.

Reality discriminator D_r. The reality discriminator D_r forces the decoder to generate images that are close to the realistic distribution. This can indirectly facilitate the self-degradation encoder E_d' to produce real-world degradation features.

Degradation discriminator D_d. The degradation discriminator resolves the degree of degradation of the input, encouraging the encoders to learn disentangled content and degradation representations.

Identity Encoder E_{id}. As a pre-trained Re-ID backbone network, the identity encoder provides identity preserving constraints for degradation invariance learning. This encoder is used to extract discriminative but degradation-sensitive features during the identity representation learning phase.
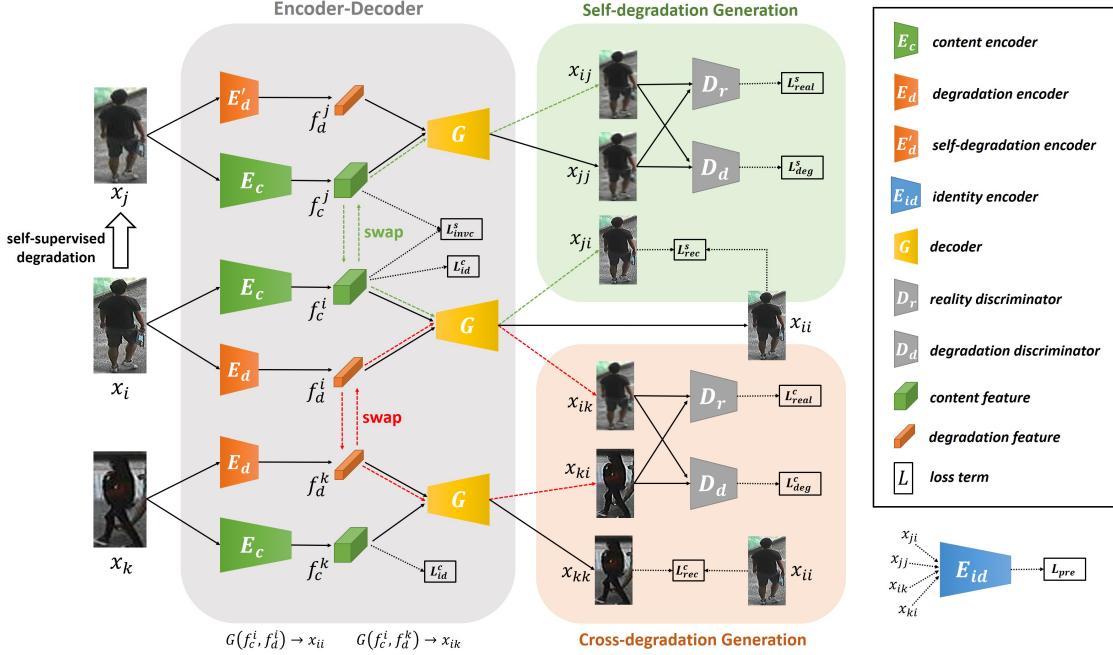
Figure 2. Overview of the proposed Degradation Decomposition Generative Adversarial Network, DDGAN. A self-degraded image pair $\{x_i, x_j\}$ and a real image pair $\{x_i, x_k\}$ are alternately used to train the DDGAN. For each pair, input images are decomposed into content features f_c and degradation features f_d, which are then swapped and combined to generate four reconstructed images, *e.g.*, $\{x_{ii}, x_{jj}, x_{ij}, x_{ji}\}$.

3.3. Degradation Invariance Learning

We aim to propose a general degradation invariance learning network against various real-world degradations under limited supervised information. In this section, we only describe the most common unsupervised DDGAN. More details about the semi-supervised DDGAN for unpaired data are given in the supplement.

Formulation. Our proposed DDGAN is alternately trained by a self-degraded image pair $p^s = \{x_i, x_j\}$ and a real image pair $p^c = \{x_i, x_k\}$, which are referred to as Self-degradation Generation and Cross-degradation Generation. For example, as shown in Figure 2, during the self-degradation generation phase, the input pair $\{x_i, x_j\}$ is decomposed into content features $\{f_c^i, f_c^j\}$ and degradation features $\{f_d^i, f_d^j\}$ by the encoders E_c, E_d and E_d'. After that, all features are combined in pairs to generate new images $\{x_{ii}, x_{ij}, x_{jj}, x_{ji}\}$ by the decoder G, where x_{ij} is generated from $G(f_c^i, f_d^j)$.

3.3.1 Self-degradation Generation

Given a self-degraded image pair $p^s = \{x_i, x_j\}$, where $x_j = F_{deg}(x_i)$, the type of the self-supervised degradation function F_{deg} depends on the specific real-world degradation factors. Since x_j and x_j are pixel-wise aligned, their content features should be consistent. We provide this con-

straint using a invariable content loss:

$$L_{invc}^s = ||E_c(x_i) - E_c(x_j)||_1. \quad (1)$$

Further, we can reconstruct the images x_{ii} and x_{ji} with a pixel-wise reconstruction loss:

$$L_{recon}^s = ||G(f_c^i, f_d^i) - x_i||_1 + ||G(f_c^j, f_d^i) - x_i||_1. \quad (2)$$

Note that L_{recon}^s should not be applied to the reconstructed images x_{ij} and x_{jj} due to the adaptive effects of the self-degradation encoder E_d'.

To ensure that the appearance of the reconstructed pedestrian images does not change significantly, an identity feature preserving loss is adopted:

$$L_{pre}^s = ||E_{id}(G(f_c^i, f_d^j)) - E_{id}(x_i)||_1 \\ + ||E_{id}(G(f_c^j, f_d^i)) - E_{id}(x_j)||_1. \quad (3)$$

As mentioned earlier, the self-supervised degradation function F_{deg} tends to introduce undesired domain bias between reality and synthesis, which leads to the learned features to deviate from the real-world distribution. To alleviate this issue, we introduce a reality adversarial loss:

$$L_{real}^s = \mathbb{E}[log(D_r(x_i)) + log(1 - D_r(G(f_c^i, f_d^j)))] \\ + \mathbb{E}[log(D_r(x_k)) + log(1 - D_r(G(f_c^j, f_d^j)))], \quad (4)$$

where both x_i and x_k are real-world images.

At last, our main objective is to learn a degradation in-dependent representation. In other words, after switching the content features of the input image pair, the degradation score ranking of reconstructed images should be consistent with the original ranking. To provide such a constraint, we introduce a degradation ranking loss:

$$L_{deg}^s = max(0, (D_d(x_i) - D_d(G(f_c^i, f_d^j))) * \gamma + \epsilon)$$
$$+ max(0, (D_d(G(f_c^j, f_d^i)) - D_d(x_j)) * \gamma + \epsilon), \quad (5)$$

where $\gamma = 1$ is the rank label of the input image pair, and the margin ϵ controls the difference of degradation scores. A higher degradation score means lower image quality.

3.3.2 Cross-degradation Generation

For the cross-degradation generation, we also perform image encoding and decoding on the input real image pair $p^c = \{x_i, x_k\}$, where x_i and x_k are directly sampled from the real-world data. To provide the regularization constraint, we also introduce a self-reconstruction loss:

$$L_{recon}^c = ||G(f_c^i, f_d^i) - x_i||_1 + ||G(f_c^k, f_d^k) - x_k||_1, \quad (6)$$

a reality adversarial loss:

$$L_{real}^c = \mathbb{E}[log(D_r(x_i)) + log(1 - D_r(G(f_c^i, f_d^k)))]$$
$$+ \mathbb{E}[log(D_r(x_k)) + log(1 - D_r(G(f_c^k, f_d^i)))], \quad (7)$$

and an identity feature preserving loss:

$$L_{pre}^c = ||E_{id}(G(f_c^i, f_d^k)) - E_{id}(x_i)||_1$$
$$+ ||E_{id}(G(f_c^k, f_d^i)) - E_{id}(x_k)||_1. \quad (8)$$

Different from self-degradation generation, x_i and x_k here have completely inconsistent content information, which means the invariable content loss is no longer applicable.

Since the purpose of degradation invariance learning is to improve the real-world person Re-ID, we use a standard identification loss to provide task-driven constraints:

$$L_{id}^c = \mathbb{E}[-log(p(y_i|x_i))] + \mathbb{E}[-log(p(y_k|x_k))], \quad (9)$$

where the predicted probability $p(y_i|x_i)$ and $p(y_k|x_k)$ are based on the content features f_c^i and f_c^k, respectively.

For unsupervised degradation invariance learning, the real-world training data does not have any degradation-related supervised information. To take advantages of real data to model the real-world degradation distribution, we also introduce a degradation ranking loss:

$$L_{deg}^c = max(0, (D_d(x_i) - D_d(G(f_c^i, f_d^k))) * \gamma + \epsilon)$$
$$+ max(0, (D_d(G(f_c^k, f_d^i)) - D_d(x_k)) * \gamma + \epsilon),$$
$$\begin{cases} \gamma = -1 & if \quad D_d(x_i) > D_d(x_k) \\ \gamma = 1 & if \quad D_d(x_i) < D_d(x_k) \end{cases},$$
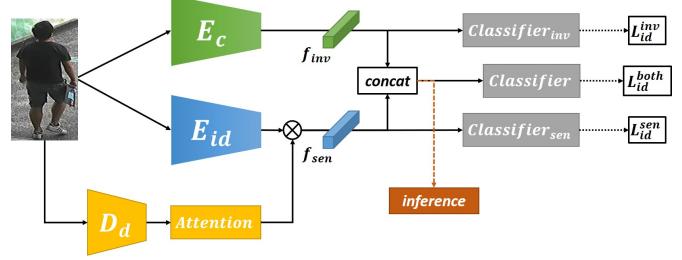$$(10)$$

Figure 3. Overview of proposed Dual Feature Extraction Network (DFEN) for robust identity representation learning.

where the rank label γ depends on the predicted degradation scores of the real-world images x_i and x_k. In this way, the disentangled content and degradation features can be learned to approximate the real-world distribution without extra supervised information.

3.3.3 Optimization

For self-degradation generation, the total objective is:

$$L_{total}^s = \lambda_{invc}L_{invc}^s + \lambda_{recon}L_{recon}^s + \lambda_{pre}L_{pre}^s$$
$$+ \lambda_{real}L_{real}^s + \lambda_{deg}L_{deg}^s. \quad (11)$$

For cross-degradation generation, the total objective is:

$$L_{total}^c = \lambda_{id}L_{id}^c + \lambda_{recon}L_{recon}^c + \lambda_{pre}L_{pre}^c$$
$$+ \lambda_{real}L_{real}^c + \lambda_{deg}L_{deg}^c. \quad (12)$$

These two optimization phases are performed alternately.

3.4. Identity Representation Learning

As described in 3.1, the DFEN extracts the degradation-invariant features f_{inv} and the degradation-sensitive features f_{sen} as identity representations, where the degradation-invariant features are the content features without dimension reduction.

Given a normal image, both the f_{inv} and f_{sen} should be kept; while for a degraded image, it should keep f_{inv} and suppress f_{sen} for Re-ID. To achieve this goal, we introduce a degradation-guided attention module, which inputs the degradation cues and outputs the attentive weights of f_{sen}. Although both E_d and D_d can provide the degradation information, we choose D_d for better interpretability. Given an input image x_i, the final identity representation is formulated as:

$$f_{id}^i = concat(f_{inv}^i, f_{sen}^i \odot Att(D_d(x_i))), \quad (13)$$

where $\odot$ denotes element-wise product.

In addition, we use multiple classifiers to better coordinate these two types of features. The total objective is:

$$L_{total}^{id} = \lambda_{inv}L_{inv}^{id} + \lambda_{sen}L_{sen}^{id} + \lambda_{both}L_{both}^{id}, \quad (14)$$

Table 1. Cross-resolution Re-ID performance (%) compared to the state-of-the-art methods on the MLR-CUHK03, MLR-VIPeR and CAVIAR datasets, respectively.

Method	MLR-CUHK03			MLR-VIPeR			CAVIAR		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
CamStyle [59]	69.1	89.6	93.9	34.4	56.8	66.6	32.1	72.3	85.9
FD-GAN [8]	73.4	93.8	97.9	39.1	62.1	72.5	33.5	71.4	86.5
JUDEA [27]	26.2	58.0	73.4	26.0	55.1	69.2	22.0	60.1	80.8
SLD^{2}L [21]	-	-	-	20.3	44.0	62.0	18.4	44.8	61.2
SDF [47]	22.2	48.0	64.0	9.3	38.1	52.4	14.3	37.5	62.5
SING [20]	67.7	90.7	94.7	33.5	57.0	66.5	33.5	72.7	89.0
CSR-GAN [49]	71.3	92.1	97.4	37.2	62.3	71.6	34.7	72.5	87.4
FFSR+RIFE [38]	73.3	92.6	-	41.6	64.9	-	36.4	72.0	-
RAIN [5]	78.9	97.3	98.7	42.5	68.3	79.6	42.0	77.3	89.6
CAD-Net [28]	82.1	**97.4**	**98.8**	43.1	68.2	77.5	42.8	76.2	91.5
ResNet50	60.2	86.6	93.2	28.5	53.8	65.2	20.2	61.0	79.8
ResNet50 (tricks†)	75.1	91.3	95.7	42.1	63.9	71.5	40.6	76.2	91.0
Ours	**85.7**	97.1	98.6	**50.3**	**77.9**	**87.3**	**51.2**	**83.6**	**94.4**

† Here all tricks we used include RandomHorizontalFlip, RandomCrop, BNNeck [36] and triplet loss.

where each loss term consists of a cross-entropy loss and a triplet loss with hard sample mining [14].

4. Experiments

To evaluate our approach on person Re-ID task against various real-world degradations, we focus on two major degradation factors, *i.e.*, resolution and illumination.

4.1. Datasets

We conduct experiments on four benchmarks: CAVIAR [6], MLR-CUHK03 and MLR-VIPER for cross-resolution Re-ID, and MSMT17 [50] for cross-illumination Re-ID.

The **CAVIAR** dataset comprises 1,220 images of 72 identities captured by two different cameras in an indoor shopping center in Lisbon. Due to the resolution of one camera is much lower than that of the other, it is very suitable for evaluating genuine cross-resolution person Re-ID.

The **MLR-CUHK03** and **MLR-VIPeR** datasets are based on the CUHK03 [24] and VIPeR [10], respectively. MLR-CUHK03 includes 14,097 images of 1,467 identities, while MLR-VIPeR contains 632 person image pairs captured from two camera views. Following SING [20], each image from one camera is down-sampled with a ratio randomly picked from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$ to construct cross-resolution settings, where the query set consists of LR images while the gallery set is only composed of HR images.

The **MSMT17** dataset, which contains 32,621/93,820 bounding boxes for training/testing, is collected by 15 surveillance cameras on the campus, including both outdoor and indoor scenes. To cover as many time periods as

possible, four days with different weather conditions in one month were selected for collecting the raw video.

4.2. Implementation Details

The proposed approach is implemented in PyTorch with two NVIDIA 1080Ti GPUs. All the used images are resized to $256 \times 128 \times 3$. We employ a multi-scale ResNet50 [13] structure for the content encoder, and both discriminators D_{real} and D_{deg} follow the popular multi-scale Patch-GAN structure [19]. More details about the optimizations and structures can be found in the supplement.

4.3. Experimental Settings and Evaluation Metrics

We employ the single-shot Re-ID settings and use the average Cumulative Match Characteristic (CMC) [20, 38, 5, 28] for evaluating cross-resolution Re-ID. In addition, we choose downsampling with a ratio that obeys uniform distribution $U[2, 4]$ as the self-degradation function.

For cross-illumination Re-ID, we follow the standard protocols of corresponding datasets. The mean Average Precision (mAP) and CMC are adopted to evaluate the retrieval performance. Gamma correction is used as the self-degradation function, where the gamma value obeys a uniform distribution $U[2, 3.5]$.

4.4. Re-ID Evaluation and Comparisons

Cross-Resolution. We compare our DI-REID with the state-of-the-art cross-resolution Re-ID methods as well as standard Re-ID methods. As shown in Table 1, our approach achieves superior performance on all three adopted datasets

Table 2. Cross-illumination Re-ID performance (%) compared to the state-of-the-art methods on the MSMT17 dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
GoogLeNet [44]	47.6	65.0	71.8	23.0
PDC [41]	58.0	73.6	79.4	29.7
GLAD [51]	61.4	76.8	81.6	85.9
PCB [43]	68.2	81.2	85.5	40.4
IANet [15]	75.5	85.5	88.7	46.8
ResNet50	57.4	72.9	78.4	29.2
ResNet50 (tricks)	68.8	80.9	84.7	35.8
Ours	**75.5**	**86.2**	**89.5**	**47.1**

Table 3. Ablation Study on the CAVIAR dataset.

Methods	Rank-1	Rank-5	Rank-10
Ours w/o DIL[1]	44.6	82.2	93.8
Ours w/o multi-scale	47.2	82.4	95.2
Ours w/o attention	48.0	82.6	93.8
Ours (f_{inv} only)	41.0	80.8	92.2
Ours (f_{sen} only)	45.4	80.0	91.2
Ours	51.2	83.6	94.4

and consistently outperform all competing methods at rank-1. Note that our approach outperforms the best competitor [28] by 8.4% at rank-1 on the only real-world cross-resolution dataset CAVIAR. It proves the effectiveness of our approach to the real-world resolution degradation.

Cross-Illumination. To demonstrate that our DI-REID is capable of dealing with various real-world degradation, extended evaluation on the real-world MSMT17 dataset is also performed for cross-illumination Re-ID. As reported in Table 2, competitive Re-ID performance is also achieved by our approach compared with existing state-of-the-art methods. It is worth mentioning that we only use the illumination degradation prior without introducing extra structural or semantic priors of human body parts.

4.5. Feature Analysis and Visualizations

Degradation-invariant identity features. We provide the comparison on learned degradation-invariant identity features with a ResNet-50 baseline model. The features are generated by the content encoder and visualizations are shown in Figure 4. All the feature maps are produced after three downsampling layers for a balance between high-level semantics and fine-grained details. It is clear that de-

---

[1]For the w/o DIL configuration, we skip the stage of Degradation Invariance Learning (directly assigning ImageNet pretrained weights to the content encoder), and the degradation-guided attention module is disabled.
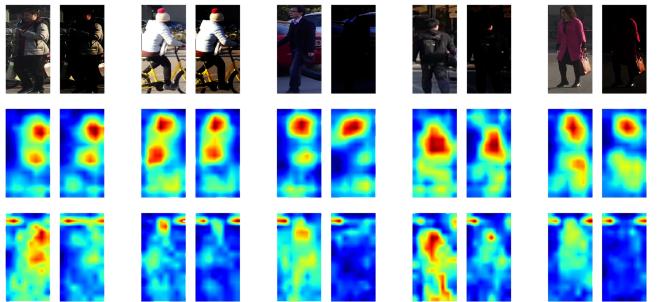

Figure 4. Visualizations of degradation-invariant features. **Top**: input images, **middle**: features produced by our DI-REID, **bottom**: features produced by a ResNet-50 baseline.

(a) Low-Resolution (LR) to High-Resolution (HR).
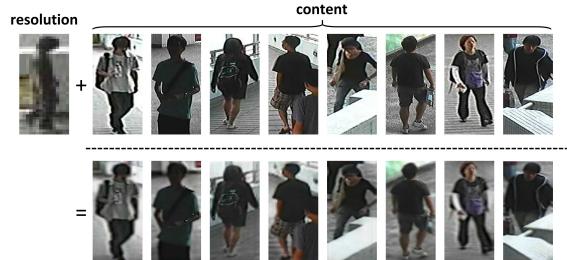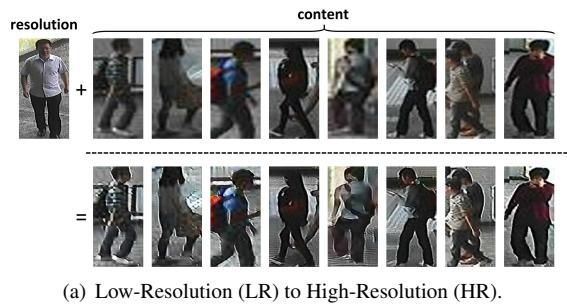

(b) High-Resolution to Low-Resolution.

Figure 5. Examples of cross-resolution image generation. (a) Images generated by HR degradation + LR content; (b) Images generated by LR degradation + HR content.

spite the degradation of illumination, the attentive regions of degradation-invariant features are basically consistent. Besides, even in extremely low-light conditions (*e.g.*, 6th and 10th columns in Figure 4), our method still can extract effective discriminative features. We also find that the learned degradation-invariant features are more focused on local areas, although no such guidance and constraints are used.

Analysis of representation disentanglement. Since the proposed DI-REID framework extracts degradation-invariant features via disentangled representation learning, it is necessary to analyze the disentangled representations for more insights and interpretability.

As shown in Figures 5 and 6, we provide the cross-degradation generation results under cross-resolution and cross-illumination settings, respectively. By reorganizing

Figure 6. Examples of cross-illumination image generation. The generated results are compared to two state-of-the-art low-light image enhancement methods: LIME [12] and DeepUPE [46].

Figure 7. t-SNE visualization of degradation features on a 1000-sample split, which randomly selected from the MSMT17 dataset.

content and degradation features, our DDGAN is able to generate new samples with degradation characteristics of the degradation provider and content characteristics of the content provider. In other words, our framework is capable of extracting degradation-independent features as identity representations for the person Re-ID task. In addition, although high-quality image generation is not our purpose, these additional generated samples are expected to be utilized for data augmentation for further performance improvement.

Analysis of Authenticity. Since this work focuses on the real-world degradations, we also analyze the authenticity of degradation features, that is, whether the real-world degradation information is captured. As illustrated in Figure 6, our approach achieves very consistent illumination adjustments without causing over-enhancement or under-enhancement. Compared to the existing state-of-the-art low-light image enhancement methods: LIME [12] and DeepUPE [46], our results are more natural and close to the real-world illumination distribution. We emphasize that our approach does not utilize any of the illumination supervised information of the original dataset, but only with the self-supervised guidance of gamma correction. The t-SNE visualization of learned degradation features of the real-world MSMT17 dataset are shown in Figure 7, and a significant illumination distribution along the manifold can be observed.

4.6. Ablation Study

We study the contribution of each component of our approach on the CAVIAR dataset. As shown in Table 3, all the components consistently achieve performance improvements, where the contribution of degradation invariance learning is most significant, resulting in a performance rise of 6.6% at Rank-1. We believe the reason is that the learned features is able to simultaneously take account of both identity discriminability and degradation invariance.

We also provide the analysis of degradation-invariant features and degradation-sensitive features, *i.e.*, f_{inv} and f_{sen}. It can be observed that f_{sen} performs better at Rank-1, while f_{inv} performs better at Rank-5 and Rank-10.

5. Conclusion

In this paper, we propose a degradation-invariance feature learning framework for real-world person Re-ID. With the capability of disentangled representation and the self-supervised learning, our method is able to capture and remove real-world degradation factors without extra labeled data. In future work, we consider integrating other semi-supervised feature representation methods, *e.g.*, graph embedding [53], to better extract pedestrian features from noisy real-world data.

6. Acknowledgement

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 189–205, 2018.

[2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2590–2600, 2017.

[5] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8215–8222, 2019.

[6] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Bmvc*, volume 1, page 6, 2011.

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2018.

[8] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 1222–1233, 2018.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[11] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[12] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019.

[16] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019.

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[18] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, and Wei Zhang. Illumination-invariant person re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 365–373, New York, NY, USA, 2019. ACM.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[20] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 695–704, 2015.

[22] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[23] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color invariants for person reidentification. *IEEE Transactions on pattern analysis and machine intelligence*, 35(7):1622–1634, 2012.

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[25] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[26] Xiang Li, Ancong Wu, and Wei-Shi Zheng. Adversarial open-world person re-identification. In *Proceedings of the*

European Conference on Computer Vision (ECCV), pages 280–296, 2018.

[27] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3765–3773, 2015.

[28] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[29] Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. Adversarial ranking for language generation. In *Advances in Neural Information Processing Systems*, pages 3155–3165, 2017.

[30] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7202–7211, 2019.

[31] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. Dense 3d-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1s):1–19, 2019.

[32] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. Deep adversarial graph attention convolution network for text-based person search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 665–673, 2019.

[33] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 192–196, 2016.

[34] Jiawei Liu, Zheng-Jun Zha, Hongtao Xie, Zhiwei Xiong, and Yongdong Zhang. Ca3net: Contextual-attentional attribute-appearance network for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 737–745, 2018.

[35] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[36] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[37] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[38] Shunan Mao, Shiliang Zhang, and Ming Yang. Resolution-invariant person re-identification. *arXiv preprint arXiv:1906.09748*, 2019.

[39] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[40] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–667, 2018.

[41] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3960–3969, 2017.

[42] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2019.

[43] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[45] Rahul Rama Varior, Gang Wang, Jiwen Lu, and Ting Liu. Learning invariant color features for person reidentification. *IEEE Transactions on Image Processing*, 25(7):3395–3410, 2016.

[46] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019.

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

[48] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019.

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

[50] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[51] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedes-

trian retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 420–428. ACM, 2017.

[52] Zheng-Jun Zha, Jiawei Liu, Di Chen, and Feng Wu. Adversarial attribute-text embedding for person search with natural language query. *IEEE Transactions on Multimedia*, 2020.

[53] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, and Tat-Seng Chua. Robust (semi) nonnegative graph embedding. *IEEE transactions on image processing*, 23(7):2996–3012, 2014.

[54] Wei Zhang, Shengnan Hu, Kan Liu, and Zhengjun Zha. Learning compact appearance representation for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2442–2452, 2018.

[55] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017.

[56] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017.

[57] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[58] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[59] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[60] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019.