

Recover and Identify: A Generative Dual Model for Cross-Resolution Person Re-Identification

Yu-Jhe Li^{1,3*} Yun-Chun Chen^{1,2,3*} Yen-Yu Lin² Xiaofei Du⁵ Yu-Chiang Frank Wang^{1,3,4}

¹National Taiwan University ²Academia Sinica

³MOST Joint Research Center for AI Technology and All Vista Healthcare

⁴Asus Intelligent Cloud Services ⁵Umbo Computer Vision

{yujheli, ycchen, ycwang}@ntu.edu.tw, yylin@citi.sinica.edu.tw, xiaofei.du@umbocv.com

Abstract

Person re-identification (*re-ID*) aims at matching images of the same identity across camera views. Due to varying distances between cameras and persons of interest, resolution mismatch can be expected, which would degrade person re-ID performance in real-world scenarios. To overcome this problem, we propose a novel generative adversarial network to address cross-resolution person re-ID, allowing query images with varying resolutions. By advancing adversarial learning techniques, our proposed model learns resolution-invariant image representations while being able to recover the missing details in low-resolution input images. The resulting features can be jointly applied for improving person re-ID performance due to preserving resolution invariance and recovering re-ID oriented discriminative details. Our experiments on five benchmark datasets confirm the effectiveness of our approach and its superiority over the state-of-the-art methods, especially when the input resolutions are unseen during training.

1. Introduction

Person re-identification (*re-ID*) [49] aims at recognizing the same person across images taken by different cameras, and is an active research topic in computer vision. A variety of applications ranging from person tracking [1], video surveillance system [25], to computational forensics [42] are highly correlated this research topic. Nevertheless, due to the presence of background clutter, occlusion, illumination or viewpoint changes, person re-ID remains a challenging task for practical applications.

Driven by the recent success of convolutional neural networks (CNNs), several learning-based methods [19, 34, 39, 51] have been proposed. Despite promising performances, these methods are typically developed under the assumption

* indicates equal contributions.



Figure 1: **Illustration and challenges of cross-resolution person re-ID.** Note that existing approaches typically leverage SR models with pre-selected resolutions followed by person re-ID modules. This cannot not be easily applied to query images with varying or unseen resolutions.

that both query and gallery images are of *similar* or *sufficiently high* resolutions. This assumption, however, may not hold in practice since image resolutions would vary drastically. For instance, query images captured by surveillance cameras are often of low resolution (LR) whereas those in the gallery set are carefully selected beforehand and are of high resolution (HR). As a result, direct matching of LR query images and HR gallery ones would lead to non-trivial *resolution mismatch* problems.

To address cross-resolution person re-ID, most existing methods [22, 44] employ super-resolution (SR) models to convert LR inputs into their HR versions followed by person re-ID. However, these methods suffer from two limitations. First, each employed SR model is designed to upscale image resolutions by a particular factor. Thus, these methods need to *pre-determine* the resolutions of LR queries so that

the corresponding SR models can be applied. However, designing SR models for each possible resolution input makes these methods hard to scale. Second, in the real-world scenario, queries can be with *various* resolutions even with the resolutions that are *unseen* during training. As illustrated in Figure 1, queries with varying or unseen resolutions would restrict the applicability of the person re-ID methods that employ SR models since one cannot assume the resolutions of the input images will be known in advance.

In this paper, we propose *Cross-resolution Adversarial Dual Network* (CAD-Net) for cross-resolution person re-ID. The key characteristics of CAD-Net are two-fold. First, to address the resolution variations, CAD-Net derives the *resolution-invariant representations* via adversarial learning. This allows our model to handle images of *varying* and even *unseen* resolutions. Second, CAD-Net learns to recover the missing details in LR input images. Together with the resolution-invariant features, our model generates HR images *preferable for person re-ID*, achieving the state-of-the-art performance on cross-resolution person re-ID. It is worth noting that the above image resolution recovery and cross-resolution person re-ID are realized by a *single* model learned in an *end-to-end* fashion.

The contributions of this paper are highlighted below:

- We propose an end-to-end trainable network which advances adversarial learning strategies for cross-resolution person re-ID.
- Our model learns resolution-invariant representations while recovering the missing details in LR input images, resulting in improved cross-resolution person re-ID performance.
- Our model is able to handle query images with varying or even unseen resolutions without the need to pre-determine the input resolutions.
- Extensive experimental results on five challenging datasets confirm that our method performs favorably against the state-of-the-art person re-ID approaches.

2. Related Work

Person re-ID. A variety of existing methods [2, 3, 10, 24, 34, 37, 38] are developed to address various challenges in person re-ID, such as background clutter, viewpoint changes, and pose variations. For instance, Yang *et al.* [51] learn a camera-invariant subspace to deal with the style variations caused by different cameras. Liu *et al.* [35] develop a pose-transferable framework based on the generative adversarial network (GAN) [16] to yield pose-specific images for tackling the pose variations. Several methods [30, 39, 40] addressing background clutter leverage attention mechanisms [4, 5, 9, 33] to emphasize the discriminative parts. Another research trend focuses on domain adaptation [8, 20] for person re-ID [13, 45]. By viewing

image-to-image translation methods as a data augmentation technique, these methods employ image translation modules, e.g., CycleGAN [52], to generate viewpoint specific images with labels. However, the above approaches typically assume that both query and gallery images are of similar or sufficiently high resolutions, which might not be practical for real-world applications.

Cross-resolution person re-ID. A number of methods [6, 22, 23, 31, 43, 44] have been proposed to address the problem of resolution mismatch in person re-ID. Li *et al.* [31] jointly perform multi-scale distance metric learning and cross-scale image domain alignment. Jing *et al.* [23] develop a semi-coupled low-rank dictionary learning framework to seek a mapping between HR and LR images. Wang *et al.* [43] learn a discriminating scale-distance function space by varying the image scale of LR images when matching with the HR ones. Nevertheless, these methods adopt hand-crafted descriptors, which cannot easily adapt the developed models to the tasks of interest, and thus may lead to sub-optimal person re-ID performance.

Recently, three CNN-based methods [6, 22, 44] are presented for cross-resolution person re-ID. The network of SING [22] is composed of several SR sub-networks and a person re-ID module to carry out LR person re-ID. On the other hand, CSR-GAN [44] cascades multiple SR-GANs [28] and progressively recovers the details of LR images to address the resolution mismatch problem. In spite of their promising results, such methods require the training of pre-defined SR models. As mentioned earlier, the degree of resolution mismatch, i.e., the resolution difference between the query and gallery images, is typically *unknown beforehand*. Moreover, if the resolution of the input LR query is unseen during training, the above methods cannot be easily applied or might not lead to satisfactory performance. Apart from these methods, RAIN [6] aligns the feature distributions of HR and LR images, showing some performance improvement over existing algorithms.

Similar to RAIN [6], our method also performs feature distribution alignment between HR and LR images. Our model differs from RAIN [6] in two aspects. First, our model derives resolution-invariant representations and recovers the missing details in LR input images. By jointly considering features of both modalities, our algorithm further improves the performance. Second, the HR image recovery is learned in an end-to-end fashion, allowing our model to recover HR images preferable for person re-ID. Experimental results demonstrate that our approach can be applied to input images of varying and even unseen resolutions using only a single model.

Cross-resolution vision applications. The issues regarding cross-resolution handling have been studied in the literature. For face recognition, existing approaches typically

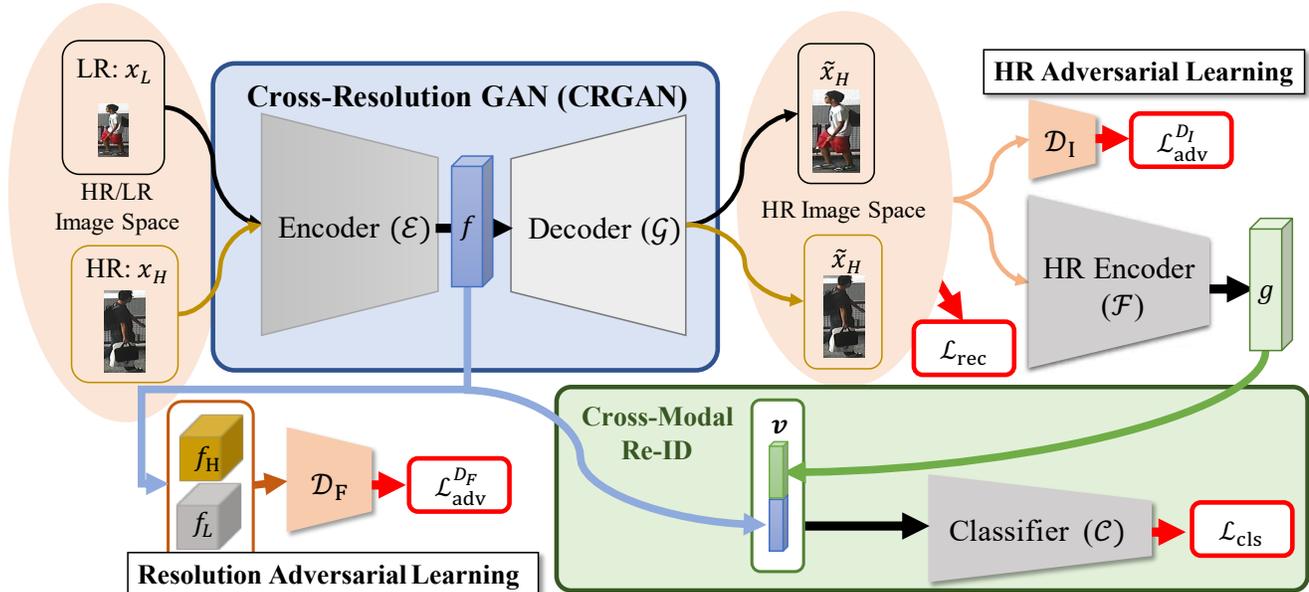


Figure 2: **Overview of Cross-resolution Adversarial Dual Network (CAD-Net).** CAD-Net comprises Cross-Resolution GAN (CRGAN) and Cross-Modal Re-ID network. The former learns resolution-invariant representations and recovers the missing details in LR input images, while the latter considers both feature modalities for cross-resolution person re-ID.

rely on face hallucination algorithms [46, 53] or SR mechanisms [12, 14, 26] to super-resolve the facial details. Unlike the above existing methods that focus on synthesizing the facial details, our model learns to recover re-ID oriented discriminative details. Together with the derived resolution-invariant features, our model would considerably boost the person re-ID performance while allowing query images with varying and even unseen resolutions.

3. Proposed Method

In this section, we first provide an overview of our proposed approach. We then describe the details of each network component as well as the loss functions.

3.1. Algorithmic Overview

We define the notations to be used in this paper. In the training stage, we have access to a set of N HR images $X_H = \{x_i^H\}_{i=1}^N$ and its corresponding label set $Y_H = \{y_i^H\}_{i=1}^N$, where $x_i^H \in \mathbb{R}^{H \times W \times 3}$ and $y_i^H \in \mathbb{R}$ are the i^{th} HR image and its label, respectively. To allow our model to handle images of different resolutions, we generate a *synthetic* LR image set $X_L = \{x_i^L\}_{i=1}^N$ by down-sampling each image in X_H , followed by resizing them back to the original image size via bilinear up-sampling (i.e., $x_i^L \in \mathbb{R}^{H \times W \times 3}$), where x_i^L is the synthetic LR image of x_i^H . Obviously, the label set Y_L for X_L is identical to Y_H .

As shown in Figure 2, our network comprises two components: Cross-Resolution Generative Adversarial Network (CRGAN) and Cross-Modal Re-ID network. To achieve

cross-resolution person re-ID, our CRGAN simultaneously learns a resolution-invariant representation $f \in \mathbb{R}^{h \times w \times d}$ ($h \times w$ is the spatial size of f whereas d denotes the number of channels) from the input cross-resolution images, while producing the associated HR images as the decoder outputs. The recovered HR output image will be encoded as an HR representation $g \in \mathbb{R}^{h \times w \times d}$ by the HR encoder. For person re-ID, we first concatenate f and g to form a joint representation $v = [f, g] \in \mathbb{R}^{h \times w \times 2d}$. The classifier then takes the joint representation v as input to perform person identity classification. The details of each component are elaborated in the following subsections.

As for testing, our network takes a query image resized to $H \times W \times 3$ as the input, and computes the joint representation $v = [f, g] \in \mathbb{R}^{h \times w \times 2d}$. We then apply global average pooling (GAP) to v for deriving a joint feature vector $u = \text{GAP}(v) \in \mathbb{R}^{2d}$, which is applied to match the gallery images via nearest neighbor search with Euclidean distance. It is worth repeating that, the query image during testing can be with varying resolutions or with unseen ones during training (verified in experiments).

3.2. Cross-Resolution GAN (CRGAN)

In CRGAN, we have a cross-resolution encoder \mathcal{E} which converts input images across different resolutions into resolution-invariant representations, followed by a high-resolution decoder \mathcal{G} recovering the associated HR versions.

Cross-resolution encoder \mathcal{E} . Since our goal is to perform cross-resolution person re-ID, we encourage the cross-

resolution encoder \mathcal{E} to extract resolution-invariant features for input images across resolutions (e.g., HR images X_H and LR ones X_L). To achieve this, we advance adversarial learning strategies and deploy a resolution discriminator \mathcal{D}_F in the latent *feature space*. This discriminator \mathcal{D}_F takes the feature maps f_H and f_L as inputs to determine whether the input feature maps are from X_H or X_L . To be more precise, we define the feature-level adversarial loss $\mathcal{L}_{\text{adv}}^{\mathcal{D}_F}$ as

$$\mathcal{L}_{\text{adv}}^{\mathcal{D}_F} = \mathbb{E}_{x_H \sim X_H} [\log(\mathcal{D}_F(f_H))] + \mathbb{E}_{x_L \sim X_L} [\log(1 - \mathcal{D}_F(f_L))], \quad (1)$$

where $f_H = \mathcal{E}(x_H)$ and $f_L = \mathcal{E}(x_L) \in \mathbb{R}^{h \times w \times d}$ denote the encoded HR and LR image features, respectively.¹

With loss $\mathcal{L}_{\text{adv}}^{\mathcal{D}_F}$, our resolution discriminator \mathcal{D}_F aligns the feature distributions across resolutions, carrying out the learning of resolution-invariant representations.

High-resolution decoder \mathcal{G} . In addition to learning the resolution-invariant representation f , our CRGAN further synthesizes the associated HR images. This is to recover the missing details in LR input images, together with the person re-ID task to be performed later in the cross-modal re-ID network.

To achieve this goal, we have an HR decoder \mathcal{G} in our CRGAN which reconstructs (or recovers) the HR images as the outputs. To accomplish this, we apply an HR reconstruction loss \mathcal{L}_{rec} between the reconstructed HR images and their corresponding HR ground-truth images. Specifically, the HR reconstruction loss \mathcal{L}_{rec} is defined as

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x_H \sim X_H} [\| \mathcal{G}(f_H) - x_H \|_1] + \mathbb{E}_{x_L \sim X_L} [\| \mathcal{G}(f_L) - x_H \|_1], \quad (2)$$

where the HR ground-truth image associated with x_L is x_H . Following Huang *et al.* [21], we adopt the ℓ_1 norm in the loss \mathcal{L}_{rec} as it preserves image sharpness. We note that both X_H and X_L will be shuffled during training. That is, images of the same identity but different resolutions will not necessarily be observed by CRGAN at the same time.

It is worth noting that, while the aforementioned HR reconstruction loss \mathcal{L}_{rec} could reduce information loss in the latent feature space, we follow Ledig *et al.* [28] and introduce skip connections between the cross-resolution encoder \mathcal{E} and the HR decoder \mathcal{G} . This would facilitate the learning process of image reconstruction, as well as allowing more efficient gradient propagation.

To encourage the HR decoder \mathcal{G} to produce more perceptually realistic HR outputs and associate with the task of person re-ID, we further adopt adversarial learning in the *image space* and introduce an HR image discriminator \mathcal{D}_I which takes the recovered HR images (i.e., $\mathcal{G}(f_L)$ and

¹For simplicity, we omit the subscript i , denote HR and LR images as x_H and x_L , and represent their corresponding labels as y_H and y_L .

$\mathcal{G}(f_H)$) and their corresponding HR ground-truth images as inputs to distinguish whether the input images are real or fake [28, 44]. Specifically, we define the image-level adversarial loss $\mathcal{L}_{\text{adv}}^{\mathcal{D}_I}$ as

$$\mathcal{L}_{\text{adv}}^{\mathcal{D}_I} = \mathbb{E}_{x_H \sim X_H} [\log(\mathcal{D}_I(x_H))] + \mathbb{E}_{x_L \sim X_L} [\log(1 - \mathcal{D}_I(\mathcal{G}(f_L)))] + \mathbb{E}_{x_H \sim X_H} [\log(\mathcal{D}_I(x_H))] + \mathbb{E}_{x_H \sim X_H} [\log(1 - \mathcal{D}_I(\mathcal{G}(f_H)))]. \quad (3)$$

It is also worth repeating that the goal of this HR decoder \mathcal{G} is not simply to recover the missing details in LR input images, but also to have such recovered HR images aligned with the learning task of interest (i.e., person re-ID). Namely, we encourage the HR decoder \mathcal{G} to perform *re-ID oriented* HR recovery, which is further realized by the following cross-modal re-ID network.

3.3. Cross-Modal Re-ID

As shown in Figure 2, the cross-modal re-ID network first applies an HR encoder \mathcal{F} , which takes the reconstructed HR image from CRGAN as input, to derive the HR feature representation $g \in \mathbb{R}^{h \times w \times d}$. Then, a classifier \mathcal{C} is learned to complete person re-ID.

As for the input to the classifier \mathcal{C} , we jointly consider the feature representations of two different modalities for person identity classification, i.e., the resolution-invariant representation f and the HR representation g . The former preserves content information, while the latter observes the recovered HR details for person re-ID. Thus, we have the classifier \mathcal{C} take the concatenated feature representation $v = [f, g] \in \mathbb{R}^{h \times w \times 2d}$ as the input. In this work, the adopted classification loss \mathcal{L}_{cls} is the integration of the identity loss \mathcal{L}_{id} and the triplet loss \mathcal{L}_{tri} [19], and is defined as

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}}, \quad (4)$$

where the identity loss \mathcal{L}_{id} computes the softmax cross entropy between the classification prediction and the corresponding ground-truth one-hot vector, while the triplet loss \mathcal{L}_{tri} is introduced to enhance the discrimination ability during person re-ID process and is defined as

$$\mathcal{L}_{\text{tri}} = \mathbb{E}_{(x_H, y_H) \sim (X_H, Y_H)} \max(0, \phi + d_{\text{pos}}^H - d_{\text{neg}}^H) + \mathbb{E}_{(x_L, y_L) \sim (X_L, Y_L)} \max(0, \phi + d_{\text{pos}}^L - d_{\text{neg}}^L), \quad (5)$$

where d_{pos} and d_{neg} are the distances between the positive (same label) and the negative (different labels) image pairs, respectively, and $\phi > 0$ serves as the margin. We note that weighted identity classification loss [7] can also be adopted to improve person identity classification.

It can be seen that the above cross-resolution person re-ID framework is very different from existing one like CSR-GAN [44], which addresses SR and person re-ID *separately*. More importantly, the aforementioned identity loss \mathcal{L}_{id} not only updates the classifier \mathcal{C} , but also refines the HR

decoder \mathcal{G} in our CRGAN. This is the reason why our CRGAN is able to produce *re-ID oriented* HR outputs, i.e., the recovered HR details preferable for person re-ID.

Full objective. The total loss function \mathcal{L} for training our proposed CAD-Net is summarized as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{adv}}^{\mathcal{D}_F} \cdot \mathcal{L}_{\text{adv}}^{\mathcal{D}_F} + \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}^{\mathcal{D}_I} \cdot \mathcal{L}_{\text{adv}}^{\mathcal{D}_I}, \quad (6)$$

where $\lambda_{\text{adv}}^{\mathcal{D}_F}$, λ_{rec} , and $\lambda_{\text{adv}}^{\mathcal{D}_I}$ are the hyper-parameters used to control the relative importance of the corresponding losses. We note that losses $\mathcal{L}_{\text{adv}}^{\mathcal{D}_F}$, \mathcal{L}_{rec} , and $\mathcal{L}_{\text{adv}}^{\mathcal{D}_I}$ are developed to learn CRGAN, while loss \mathcal{L}_{cls} is designed to update both CRGAN and cross-modal re-ID network.

To train our network using training HR images and their down-sampled LR ones, we minimize the HR reconstruction loss \mathcal{L}_{rec} for updating our CRGAN, and the classification loss \mathcal{L}_{cls} for jointly updating CRGAN and cross-modal re-ID network. The image-level adversarial loss $\mathcal{L}_{\text{adv}}^{\mathcal{D}_I}$ is computed for producing perceptually realistic HR images while the feature-level adversarial loss $\mathcal{L}_{\text{adv}}^{\mathcal{D}_F}$ is optimized for learning resolution-invariant representations.

4. Experiments

We first provide the implementation details, followed by dataset descriptions and settings. Both quantitative and qualitative results are presented, including ablation studies.

4.1. Implementation Details

We implement our model using PyTorch. ResNet-50 [18] pretrained on ImageNet is used to build the cross-resolution encoder \mathcal{E} and the HR encoder \mathcal{F} . Note that since \mathcal{E} and \mathcal{F} work for different tasks, these two components do not share weights. The classifier \mathcal{C} is composed of a global average pooling layer and a fully connected layer followed a softmax activation. The architecture of the resolution discriminator \mathcal{D}_F is the same as that adopted by Tsai *et al.* [41]. The structure of the HR image discriminator \mathcal{D}_I is similar to ResNet-18 [18]. Our HR decoder \mathcal{G} is similar to that proposed by Miyato *et al.* [36]. Components \mathcal{D}_F , \mathcal{D}_I , \mathcal{G} , and \mathcal{C} are all randomly initialized. We use stochastic gradient descent to train the proposed model. For components \mathcal{E} , \mathcal{G} , \mathcal{F} , and \mathcal{C} , the learning rate, momentum, and weight decay are 1×10^{-3} , 0.9, and 5×10^{-4} , respectively. For the two discriminators \mathcal{D}_F and \mathcal{D}_I , the learning rate is set to 1×10^{-4} . The batch size is 32. The margin ϕ in the triplet loss \mathcal{L}_{tri} is set to 2. We set the hyper-parameters in all the experiments as follows: $\lambda_{\text{adv}}^{\mathcal{D}_F} = 1$, $\lambda_{\text{rec}} = 1$, and $\lambda_{\text{adv}}^{\mathcal{D}_I} = 1$. All images of various resolutions are resized to $256 \times 128 \times 3$ in advance. We train our model on a single NVIDIA GeForce GTX 1080 GPU with 12 GB memory.

4.2. Datasets

We evaluate the proposed method on five datasets, each of which is described as follows.

CUHK03 [29]. The CUHK03 dataset comprises 14,097 images of 1,467 identities with 5 different camera views. Following CSR-GAN [44], we use the 1,367/100 training/test identity split.

VIPeR [17]. The VIPeR dataset contains 632 person-image pairs captured by 2 cameras. Following SING [22], we randomly divide this dataset into two non-overlapping halves based on the identity labels. Namely, images of a subject belong to either the training set or the test set.

CAVIAR [11]. The CAVIAR dataset is composed of 1,220 images of 72 person identities captured by 2 cameras. Following SING [22], we discard 22 people who only appear in the closer camera, and split this dataset into two non-overlapping halves according to the identity labels.

Market-1501 [48]. The Market-1501 dataset consists of 32,668 images of 1,501 identities with 6 camera views. We use the widely adopted 751/750 training/test identity split.

DukeMTMC-reID [50]. The DukeMTMC-reID dataset contains 36,411 images of 1,404 identities captured by 8 cameras. We adopt the benchmarking 702/702 training/test identity split.

4.3. Experimental Settings and Evaluation Metrics

We evaluate the proposed method using *cross-resolution person re-ID* setting [22] where the test (query) set is composed of LR images while the gallery set contains HR images only. In all of the experiments, we adopt the standard single-shot person re-ID setting [22, 32] and use the average cumulative match characteristic as the evaluation metric.

4.4. Evaluation and Comparisons

Following SING [22], we consider multiple low-resolution (MLR) person re-ID and evaluate the proposed method on *four synthetic* and *one real-world* benchmarks. To construct the synthetic MLR datasets (i.e., MLR-CUHK03, MLR-VIPeR, MLR-Market-1501, and MLR-DukeMTMC-reID), we follow SING [22] and down-sample images taken by one camera by a randomly selected down-sampling rate $r \in \{2, 3, 4\}$ (i.e., the size of the down-sampled image becomes $\frac{H}{r} \times \frac{W}{r} \times 3$), while the images taken by the other camera(s) remain unchanged. The CAVIAR dataset inherently contains realistic images of multiple resolutions, and is a *genuine* and more challenging dataset for evaluating MLR person re-ID.

We compare our approach with methods developed for cross-resolution person re-ID, including JUDEA [31],

Table 1: **Results of cross-resolution re-ID (%)**. Bold and underlined numbers indicate top two results, respectively.

Method	MLR-CUHK03			MLR-VIPeR			CAVIAR			MLR-Market-1501			MLR-DukeMTMC-reID		
	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10
JUDEA [31]	26.2	58.0	73.4	26.0	55.1	69.2	22.0	60.1	80.8	-	-	-	-	-	-
SLD ² L [23]	-	-	-	20.3	44.0	62.0	18.4	44.8	61.2	-	-	-	-	-	-
SDF [43]	22.2	48.0	64.0	9.3	38.1	52.4	14.3	37.5	62.5	-	-	-	-	-	-
SING [22]	67.7	90.7	94.7	33.5	57.0	66.5	33.5	72.7	89.0	74.4	87.8	91.6	65.2	80.1	84.8
CSR-GAN [44]	71.3	92.1	97.4	37.2	62.3	71.6	34.7	72.5	87.4	76.4	88.5	91.9	67.6	81.4	85.1
CamStyle [51]	69.1	89.6	93.9	34.4	56.8	66.6	32.1	72.3	85.9	74.5	88.6	93.0	64.0	78.1	84.4
FD-GAN [15]	73.4	93.8	97.9	39.1	62.1	72.5	33.5	71.4	86.5	79.6	<u>91.6</u>	93.5	67.5	82.0	85.3
Ours (<i>f</i> only)	77.6	<u>96.2</u>	98.5	41.2	66.3	75.6	<u>41.5</u>	<u>75.3</u>	85.6	80.1	90.6	93.2	73.4	84.4	86.8
Ours (<i>g</i> only)	<u>79.7</u>	97.4	<u>98.7</u>	<u>41.7</u>	<u>66.4</u>	<u>76.1</u>	38.9	73.1	<u>90.6</u>	<u>82.2</u>	91.3	<u>94.5</u>	<u>74.1</u>	<u>85.1</u>	<u>88.2</u>
Ours	82.1	97.4	98.8	43.1	68.2	77.5	42.8	76.2	91.5	83.7	92.7	95.8	75.6	86.7	89.6

SLD²L [23], SDF [43], SING [22], and CSR-GAN [44], and methods developed for standard person re-ID, including CamStyle [51] and FD-GAN [15]. For methods developed for cross-resolution person re-ID, the training set contains HR images and LR ones with all three down-sampling rates $r \in \{2, 3, 4\}$ for each person. For methods developed for standard person re-ID, the training set contains HR images for each identity only.

Table 1 reports the quantitative results recorded at ranks 1, 5, and 10 on all five adopted datasets. For CSR-GAN [44] on MLR-CUHK03, CAVIAR, MLR-Market-1501, and MLR-DukeMTMC-reID, and CamStyle [51] and FD-GAN [15] on all five adopted datasets, their results are obtained by running the released code with the default implementation setup. For SING [22], we reproduce their results on MLR-Market-1501 and MLR-DukeMTMC-reID.

We note that the performance of our method can be further improved by applying pre-/post-processing methods, attention mechanisms, or re-ranking. For fair comparisons, no such techniques are used in all of our experiments.

Results. In Table 1, our method performs favorably against all competing methods on all five datasets. We observe that our method consistently outperforms the best competitors [15, 44] by 4% ~ 8% at rank 1. The performance gains can be ascribed to three main factors. First, unlike most existing person re-ID methods, our model performs cross-resolution person re-ID in an end-to-end learning fashion. Second, our method learns resolution-invariant representations, allowing our model to recognize persons in images of different resolutions. Third, our model learns to recover the missing details in LR input images, thus providing additional discriminative evidence for person re-ID.

The advantage of deriving joint representation $v = [f, g]$ can be assessed by comparing with two of our variant methods, i.e., Ours (*f* only) and Ours (*g* only). In method “Ours (*f* only)”, the classifier \mathcal{C} only takes the resolution-invariant representation f as input. In method “Ours (*g* only)”, the classifier \mathcal{C} only takes the HR representation g as input. We

observe that deriving joint representation v consistently improves the performance over these two baseline methods. We note that method “Ours (*f* only)” achieves a better performance than method “Ours (*g* only)” on the CAVIAR dataset. We attribute the results to the higher resolution variations exhibited in the CAVIAR dataset.

4.5. Evaluation of the Recovered HR Images

To demonstrate that our CRGAN is capable of recovering the missing details in LR images of varying and even unseen resolutions, we evaluate the quality of the recovered HR images on the MLR-CUHK03 *test set* using SSIM, PSNR, and LPIPS [47] metrics. We employ the ImageNet-pretrained AlexNet [27] when computing LPIPS. We compare our CRGAN with CycleGAN [52], SING [22], and CSR-GAN [44]. For CycleGAN [52], we train the model to learn a mapping between LR and HR images. We report the quantitative results of the recovered image quality and person re-ID in Table 2 with two different settings: (1) LR images of resolutions seen during training, i.e., $r \in \{2, 3, 4\}$, and (2) LR images of unseen resolution, i.e., $r = 8$.

For seen resolutions (i.e., left block), we observe that our results using SSIM and PSNR metrics are slightly worse than CSR-GAN [44] while compares favorably against SING [22] and CycleGAN [52]. However, our method performs favorably against these three methods using LPIPS metric and achieves the state-of-the-art performance when evaluating on cross-resolution person re-ID task. These results indicate that (1) SSIM and PSNR metrics are low-level pixel-wise metrics, which do not reflect high-level perceptual tasks and (2) the end-to-end learning of cross-resolution person re-ID would result in better person re-ID performance and recover more perceptually realistic HR images as reflected by LPIPS.

For unseen resolution (i.e., right block), our method performs favorably against all three competing methods on all the adopted evaluation metrics. These results suggest that our method is capable of handling unseen resolution (i.e., $r = 8$) with favorable performance in terms of both image

Table 2: **Quantitative results of cross-resolution person re-ID on the MLR-CUHK03 test set.** *Left block:* resolutions are seen during training. *Right block:* resolution is not seen during training.

Method	Down-sampling rate $r \in \{2, 3, 4\}$ (seen)				Down-sampling rate $r = 8$ (unseen)			
	SSIM \uparrow	PSNR \uparrow	LPIPS [47] \downarrow	Rank 1 (%) \uparrow	SSIM \uparrow	PSNR \uparrow	LPIPS [47] \downarrow	Rank 1 (%) \uparrow
CycleGAN [52]	0.55	14.1	0.31	62.1	0.42	12.7	0.37	40.5
SING [22]	0.65	18.1	0.18	67.7	0.52	14.5	0.34	54.2
CSR-GAN [44]	0.76	21.5	0.13	71.3	0.67	17.2	0.25	62.1
Ours	0.73	20.2	0.07	82.1	0.71	19.8	0.11	78.6

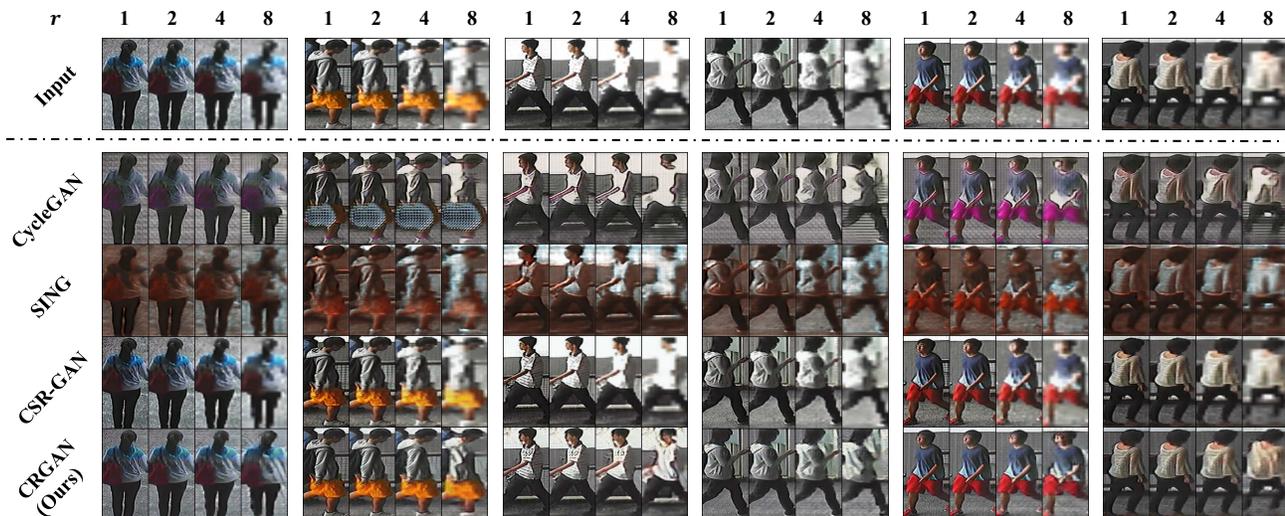


Figure 3: **Visual results of the recovered HR images on the MLR-CUHK03 test set.** We present the visual comparison among CycleGAN [52], SING [22], CSR-GAN [44], and the proposed CRGAN.

quality and person re-ID. Note that we only train our model with HR images and LR ones with $r \in \{2, 3, 4\}$.

Figure 3 presents six examples. For each person, there are four different resolutions (i.e., $r \in \{1, 2, 4, 8\}$). Note that images with down-sampling rate $r = 1$ indicate that the images remain their original sizes and are the corresponding HR images of the LR ones. We observe that when LR images with down-sampling rate $r = 8$ are given, our model recovers the HR details with the highest visual quality among all competing methods. Both quantitative and qualitative results above confirm that our model can handle a range of seen resolutions and generalize well to unseen resolutions using just one single model, i.e., CRGAN.

4.6. Ablation Study

To analyze the importance of each developed loss function, we conduct an ablation study on the MLR-CUHK03 dataset. Table 3 reports the quality of the recovered HR images and the performance of cross-resolution person re-ID recorded at rank 1.

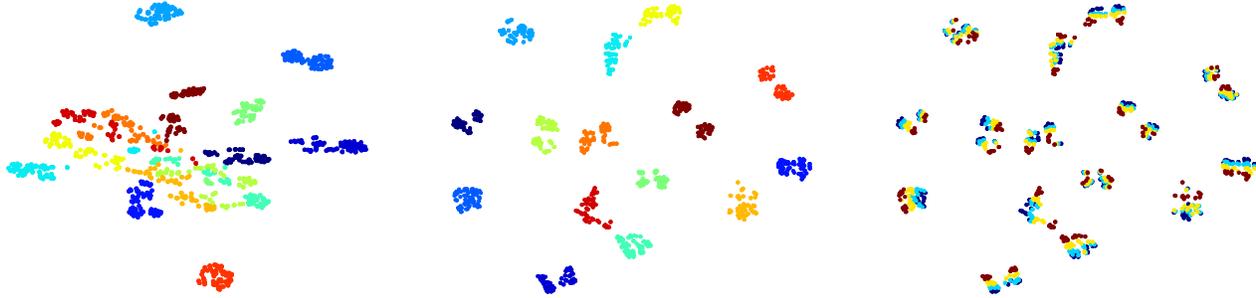
Image-level adversarial loss $\mathcal{L}_{adv}^{\mathcal{D}_I}$. When loss $\mathcal{L}_{adv}^{\mathcal{D}_I}$ is turned off, our model is not encouraged to produce percep-

tually realistic HR images as reflected by LPIPS, resulting in a performance drop of 2.3% at rank 1.

Feature-level adversarial loss $\mathcal{L}_{adv}^{\mathcal{D}_F}$. Without loss $\mathcal{L}_{adv}^{\mathcal{D}_F}$, our model does not learn resolution-invariant representations and thus suffers from the resolution mismatch issue. Significant performance drops in the recovered image quality and person re-ID performance occur, indicating the importance of our method for learning resolution-invariant representations to address the resolution mismatch issue.

HR reconstruction loss \mathcal{L}_{rec} . Once loss \mathcal{L}_{rec} is excluded, there is no explicit supervision to guide the CRGAN to perform image recovery, and the model implicitly suffers from information loss in compressing visual images into semantic feature maps. Severe performance drops in terms of the recovered image quality and person re-ID performance are hence caused.

Classification loss \mathcal{L}_{cls} . Although our model is still able to perform image recovery without loss \mathcal{L}_{cls} , our model cannot perform discriminative learning for person re-ID since data labels are not used during training. Thus, significant performance drop in person re-ID occurs.



(a) Ours w/o \mathcal{L}_{adv}^{DF} : colored w.r.t identity.

(b) Ours: colored w.r.t identity.

(c) Ours: colored w.r.t resolution.

Figure 4: **2D visualization of the resolution-invariant feature vector w on the MLR-CUHK03 test set via t-SNE.** Data of different identities (each in a unique color) derived by our model *without* and *with* observing the feature-level adversarial loss \mathcal{L}_{adv}^{DF} are shown in (a) and (b), respectively. The same data but with resolution-specific colorization, i.e., one color for each down-sampling rate $r \in \{1, 2, 4, 8\}$, are depicted in (c). Note that images with $r = 8$ are not seen during training.

The ablation study demonstrates that the losses \mathcal{L}_{adv}^{DF} , \mathcal{L}_{rec} , and \mathcal{L}_{cls} are crucial to our method, while the loss \mathcal{L}_{adv}^{DI} is helpful for improving the performance of cross-resolution person re-ID as well as the quality of the recovered images.

4.7. Resolution-Invariant Representation f

To demonstrate the effectiveness of our model in deriving the resolution-invariant representations, we first apply global average pooling to f to obtain the resolution-invariant feature vector $w = \text{GAP}(f) \in \mathbb{R}^d$. We then visualize w on the MLR-CUHK03 *test set* in Figure 4.

To be more precise, we select 15 different identities, each of which is indicated by a unique color, as shown in Figure 4a and Figure 4b. In Figure 4a, we observe that without the feature-level adversarial loss \mathcal{L}_{adv}^{DF} , our model cannot establish a well-separated feature space. When loss \mathcal{L}_{adv}^{DF} is imposed, as shown in Figure 4b, the projected feature vectors are well separated. These two figures indicate that without loss \mathcal{L}_{adv}^{DF} , our model does not learn resolution-invariant representations, thus implicitly suffering from the negative impact induced by the resolution mismatch issue.

We note that the projected feature vectors in Figure 4b are well separated, suggesting that sufficient person re-ID ability can be exhibited by our model. On the other hand, for Figure 4c, we colorize each image resolution with a unique color in each identity cluster (four different down-sampling rates $r \in \{1, 2, 4, 8\}$). We observe that the projected feature vectors of the same identity but different down-sampling rates are all well clustered. We note that images with down-sampling rate $r = 8$ are not present in the training set (i.e., unseen resolution).

The above visualizations demonstrate that our model learns resolution-invariant representations and generalizes well to unseen image resolution (e.g., $r = 8$) for cross-resolution person re-ID.

Table 3: **Ablation study on the MLR-CUHK03 dataset.** Bold and underlined numbers indicate top two results, respectively.

Method	SSIM \uparrow	PSNR \uparrow	LPIPS [47] \downarrow	Rank 1 (%) \uparrow
Ours	0.73	<u>20.2</u>	0.07	82.1
Ours w/o \mathcal{L}_{adv}^{DI}	0.67	18.5	0.17	<u>79.8</u>
Ours w/o \mathcal{L}_{adv}^{DF}	0.54	14.2	0.34	67.6
Ours w/o \mathcal{L}_{rec}	0.45	12.9	0.40	66.7
Ours w/o \mathcal{L}_{cls}	<u>0.72</u>	21.4	<u>0.11</u>	1.7

5. Conclusions

We have presented an *end-to-end trainable* generative adversarial network, CAD-Net, for addressing the resolution mismatch issue in person re-ID. The core technical novelty lies in the unique design of the proposed CRGAN which learns the *resolution-invariant* representations while being able to recover *re-ID oriented* HR details. Our cross-modal re-ID network jointly considers the information from two feature modalities, leading to better person re-ID capability. Extensive experimental results demonstrate that our approach performs favorably against existing cross-resolution and standard person re-ID methods on five challenging benchmarks, and produces perceptually higher quality HR images using only a *single* model. Visualization of the resolution-invariant representations further verifies our ability in handling query images with *varying* or even *unseen* resolutions. Thus, the use of our model for practical person re-ID applications can be strongly supported.

Acknowledgements. This work is supported in part by Ministry of Science and Technology (MOST) under grants 107-2628-E-001-005-MY3, 108-2634-F-007-009, and 108-2634-F-002-018, and Umbo Computer Vision.

References

- [1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008. 1
- [2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018. 2
- [3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, 2018. 2
- [4] Yun-Chun Chen and Winston H Hsu. Saliency aware: Weakly supervised object localization. In *ICASSP*, 2019. 2
- [5] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin. Deep semantic matching with foreground detection and cycle-consistency. In *ACCV*, 2018. 2
- [6] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *AAAI*, 2019. 2
- [7] Yun-Chun Chen, Yu-Jhe Li, Aragorn Tseng, and Tsungnan Lin. Deep learning for malicious flow detection. In *PIMRC*, 2017. 4
- [8] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *CVPR*, 2019. 2
- [9] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint learning of semantic matching and object co-segmentation. *arXiv*, 2019. 2
- [10] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 2
- [11] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 5
- [12] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *ICCV*, 2017. 3
- [13] 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 reidentification. In *CVPR*, 2018. 2
- [14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2016. 3
- [15] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018. 6
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2
- [17] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 5
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5
- [19] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 1, 4
- [20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2
- [21] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 4
- [22] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *AAAI*, 2018. 1, 2, 5, 6, 7
- [23] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, 2015. 2, 6
- [24] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 2
- [25] Furqan M Khan and François Brémont. Person re-identification for real-world surveillance systems. *arXiv*, 2016. 1
- [26] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 3
- [27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 6
- [28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2, 4
- [29] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 5
- [30] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 2
- [31] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015. 2, 5, 6
- [32] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 5
- [33] Jih-Yuan Lin, Min-Sheng Wu, Yu-Cheng Chang, Yun-Chun Chen, Chao-Te Chou, Chun-Ting Wu, and Winston H Hsu. Learning volumetric segmentation for lung tumor. *IEEE ICIP VIP Cup Tech. Report*, 2018. 2
- [34] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv*, 2017. 1, 2
- [35] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, 2018. 2
- [36] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *ICLR*, 2018. 5
- [37] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng

- Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, 2018. 2
- [38] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 2018. 2
- [39] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 1, 2
- [40] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 2
- [41] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 5
- [42] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 2013. 1
- [43] Zheng Wang, Ruimin Hu, Yi Yu, Junjun Jiang, Chao Liang, and Jinqiao Wang. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, 2016. 2, 6
- [44] 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*, 2018. 1, 2, 4, 5, 6, 7
- [45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 2
- [46] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*, 2017. 3
- [47] Richard Zhang, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6, 7, 8
- [48] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5
- [49] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv*, 2016. 1
- [50] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 5
- [51] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 1, 2, 6
- [52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 6, 7
- [53] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, 2016. 3