remote sensing

MDPI

Technical Note

Multi-Scale Image- and Feature-Level Alignment for Cross-Resolution Person Re-Identification

Guoqing Zhang [1,2,3,4], Zhun Wang [1], Jiangmei Zhang [2,*], Zhiyuan Luo [1] and Zhihao Zhao [2]

[1] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; guoqingzhang@nuist.edu.cn (G.Z.); zhunwang@nuist.edu.cn (Z.W.); luozhiyuan@nuist.edu.cn (Z.L.)
[2] Fundamental Science on Nuclear Wastes and Environment Safety Laboratory, Southwest University of Science and Technology, Mianyang 621010, China; haolin05002@swust.edu
[3] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China
[4] Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: zjm@swust.edu.cn

Abstract: Cross-Resolution Person Re-Identification (re-ID) aims to match images with disparate resolutions arising from variations in camera hardware and shooting distances. Most conventional works utilize Super-Resolution (SR) models to recover Low Resolution (LR) images to High Resolution (HR) images. However, because the SR models cannot completely compensate for the missing information in the LR images, there is still a large gap between the HR image recovered from the LR images and the real HR images. To tackle this challenge, we propose a novel Multi-Scale Image- and Feature-Level Alignment (MSIFLA) framework to align the images on multiple resolution scales at both the image and feature level. Specifically, (i) we design a Cascaded Multi-Scale Resolution Reconstruction (CMSR^2) module, which is composed of three cascaded Image Reconstruction (IR) networks, and can continuously reconstruct multiple variables of different resolution scales from low to high for each image, regardless of image resolution. The reconstructed images with specific resolution scales are of similar distribution; therefore, the images are aligned on multiple resolution scales at the image level. (ii) We propose a Multi-Resolution Representation Learning (MR^{2}L) module which consists of three-person re-ID networks to encourage the IR models to preserve the ID-discriminative information during training separately. Each re-ID network focuses on mining discriminative information from a specific scale without the disturbance from various resolutions. By matching the extracted features on three resolution scales, the images with different resolutions are also aligned at the feature-level. We conduct extensive experiments on multiple public cross-resolution person re-ID datasets to demonstrate the superiority of the proposed method. In addition, the generalization of MSIFLA in handling cross-resolution retrieval tasks is verified on the UAV vehicle dataset.

Keywords: cross-resolution; super-resolution; multi-branch network; person re-identification

Citation: Zhang, C.; Wang, Z.; Zhang, J.; Luo, Z.; Zhao, Z. Multi-Scale Image- and Feature-Level Alignment for Cross-Resolution Person Re-Identification. *Remote Sens.* **2024**, *16*, 278. https://doi.org/10.3390/rs16020278

Academic Editors: Chiman Kwan, Eugene Levin, Roman Shults and Surya Prakash Tiwari

Received: 1 December 2023
Revised: 2 January 2024
Accepted: 8 January 2024
Published: 10 January 2024

1. Introduction

The goal of Person Re-Identification [1–3] is to match the identity of the target image across non-overlapping surveillance cameras. Recently, person re-ID has attracted attention from both the community and the industry due to its practical value in public safety and private property security. With the development of deep neural networks [4], great progress has been made in person re-ID and most of the existing studies focus on challenges from camera settings [5–7], occlusions [8,9], viewpoints [10–12], illumination-adaptive [13–15], modalities [16–18] and cloth-changing [19,20]. Although these methods have achieved inspiring matching accuracy on public benchmarks [21,22], they have a common limitation

in that they are designed based on a prerequisite that the images captured by different cameras are of the same resolution.

However, due to the difference in hardware and shooting distance between cameras, this prerequisite is hard to meet in practical applications. Therefore, the application of these methods is limited in real-world scenes because they cannot adapt to images with various resolutions. The practical task of matching images of the same individual but with different resolution scales is referred to as Cross-Resolution Person Re-Identification. As shown in Figure 1, the challenge of cross-resolution person re-ID is two-fold: First, the query and gallery sets are composed of low-resolution (LR) and high-resolution (HR) images, respectively, and the matching accuracy of re-ID models will decrease if their features are matched directly because the LR and HR images are of different distribution. Second, resolution disparity also exists between LR query images, resulting that matching LR query images with various resolutions and HR gallery images with unified resolution becomes harder. From the analyses above, we can conclude that the main difficulty lies in cross-resolution person re-ID because the distribution of images with different resolutions is not aligned, so the model is disturbed to learn discriminative information and is forced to establish the connection between the distribution of different resolutions.

Figure 1. The challenge of cross-resolution person re-ID. (**a**) The resolution misalignment problem not only exists between LR query and HR gallery images, (**b**) but it also exists between LR query images with different resolution scales.

To tackle the resolution mismatch challenge in person re-ID, some approaches have been made in recent years and they can be broadly classified into two categories. The methods of the first category [23,24] aim to learn resolution-irrelevant feature representations in order to map features to a common space. However, these methods force the images with higher resolutions which contain more appearance details to align with the images with lowest resolution which contain the least identity clues, and the results show that the additional information that lies in images with higher resolutions is lost. The re-ID models suffer from very limited discriminative semantic, so their performance is not optimal. The methods of the second category [25,26] are based on super-resolution algorithms. The primary idea of these methods is that the SR models can recover part of the person description which is missing in LR images, so the disparity between the amount of identity semantic contained in the images with different resolutions is reduced. However, because the SR model cannot fully complete the missing clues in LR images, and the amount of information in reconstructed LR images also varies, cross-resolution challenges still exist. The re-ID models are forced to construct the connection between reconstructed LR images and HR images, which distracts the attention of the model in learning discriminative clues.

The intrinsic discrepancy between images with different resolutions lies in the amount of discriminative clues contained in the images, and the LR and HR images cannot be

aligned well at the HR scale because the SR model cannot fully recover the missing information in LR images. HR images contain more person descriptions, making alignment at the LR scale easier to achieve. However, when only LR scale alignment is performed, the model cannot benefit from the additional information in HR images. Based on these analyses, we propose a novel Multi-Scale Image- and Feature-Level Alignment framework (MSIFLA) for cross-resolution person re-ID, as exhibited in the conceptual diagram in Figure 2. Specifically, we first reconstruct all images into LR images, regardless of their resolution, so the amount of information contained in images with different resolutions is aligned. Then, we reconstruct LR images to higher resolution scales and eventually align all images at the HR scale, so the model can also benefit from the discriminative clues that lie in HR images. The proposed framework consists of two modules: (i) a Cascaded Multi-Scale Resolution Reconstruction (CMSR^2) module which gradually reconstructs each image from the IR scale to the HR scale without distinguishing the resolution scale, so the images with different resolutions are aligned first on the LR scale and then on higher resolution scales. (ii) a Multi-Resolution Representation Learning (MR^{2}L) module which consists of three-person re-ID networks to extract feature representations from the reconstructed images with different resolutions separately. Each re-ID network focuses on the feature learning of specific resolution scale images to extract identity-relevant features. By utilizing the discriminative semantic lies in each resolution scale for supervised training, the images are further aligned at the feature level.

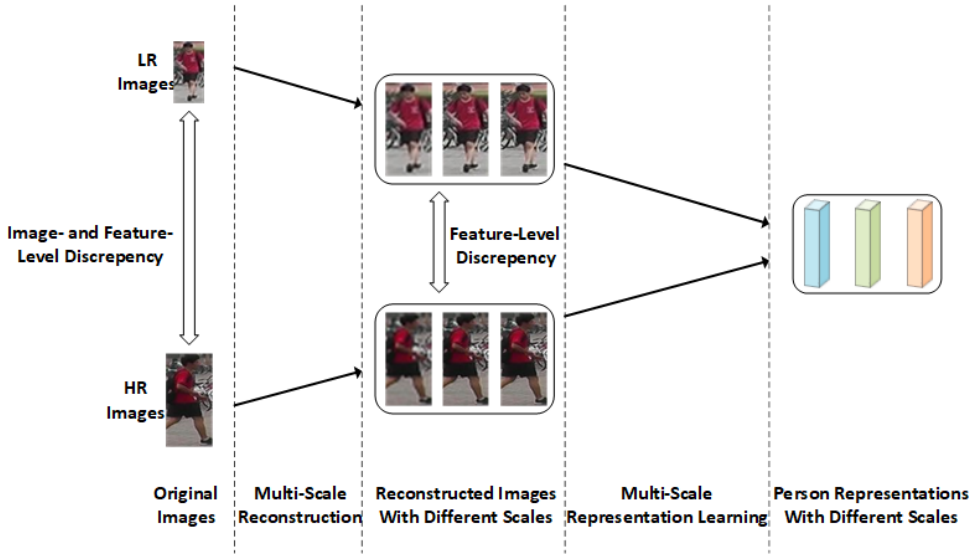

Figure 2. The illustration of image- and feature-level alignment.

The contribution of this paper can be summarized as follows:

- We propose a Cascaded Multi-Scale Resolution Reconstruction module (CMSR^2) to align the images with different resolutions at the image level. Specifically, we first reconstruct all images into LR images, regardless of their resolution, so all images are aligned on the LR scale. Then, we reconstruct and align these images on higher resolution scales, so the model can also benefit from the discriminative clues that lie in HR images.
- We design a Multi-Resolution Representation Learning module (MR^{2}L) to align the images with different resolutions at the feature level. Specifically, we utilize the image-level aligned person images for supervised training to encourage the features of the reconstructed images to be aligned on each resolution scale.
- Experimental results on five cross-resolution person re-ID datasets demonstrate the superiority of the proposed method compared to other state-of-the-art methods. In addition, the generalization of the proposed method is verified on a UAV simulation cross-resolution vehicle dataset.

2. Related Work

2.1. Deep Learning Person Re-Identification

Person re-ID aims to retrieve the person of interest from surveillance systems consisting of multiple disjoint cameras. The key challenge of person re-ID lies in the style disparities between cameras due to different camera settings, viewpoints, backgrounds, illuminations and resolutions [27,28]. Recently, thanks to the development of advanced network structure and deep learning [29–32], person re-ID has achieved remarkable progress. Zhu et al. [10] addressed the challenge of viewpoint variation by projecting the features of people with different viewpoints into a unified space and modeling the representations on identity- and viewpoint-level. Zhong et al. [5] proposed to transfer each image in the training set to the styles of other cameras by CycleGAN-based image translation functions, the transferred images smoothed the domain gap between different cameras. Tian et al. [33] studied the interference of backgrounds to the accuracy of re-ID models and proposed a person segmentation-guided model to learn representations that are robust to background variation. Zeng et al. [14] tacked the illumination variation challenge by disentangling the illumination-invariant information from identity-relevant information, so the model can learn illumination-irrelevant representations. Miao et al. [34] deal with the Visible Infrared re-ID task by bridging the modality gap via two intermediate modalities. However, these works ignored the problem of resolution variation between cameras caused by hardware and shooting distance, so their effectiveness in real-world scenarios is limited.

2.2. Cross-Resolution Person Re-Identification

Cross-resolution person re-ID attempts to tackle the resolution mismatch issue between images with different resolutions. Recently, remarkable achievements have been made [25,26,35–37] for addressing this challenging task. These methods can be broadly categorized into two groups: (1) learning resolution-invariant features [23,24] and (2) utilizing super-resolution models [25,26,35]. In the first group, Li et al. [24] minimized the heterogeneous class mean discrepancy to align the distribution of person images with different resolutions. Chen et al. [23] proposed aligning the feature representations across resolutions via adversarial learning. However, these approaches face a limitation in that the fine-grained information lies in the HR images are lost due to aligning the features of HR and LR images. In the second group, Jiao et al. [25] proposed exploiting person identity signals to supervise the training of the super-resolution model and train the person re-ID and super-resolution model in an end-to-end manner, so the SR model learned to retain the discriminative semantic essential for person re-ID tasks. Wang et al. [26] cascaded three SR-GANs to recover LR images with three different resolution scales to HR images, so the model can adapt to LR images with various resolutions. Li et al. [36] employed adversarial learning to extract both resolution-invariant and high-resolution features; therefore, the model can adapt to varying resolutions. However, these methods ignore the fact that the images with different resolutions contain different levels of information and the information learned from different resolutions is complementary.

3. The Proposed Method

3.1. Overview

In this paper, we propose a Multi-Scale Image- and Feature-Level Alignment (MSIFLA) framework for cross-resolution person re-ID. Different from previous methods which aim to align images on the LR scale by learning resolution-irrelevant representations or align them on the HR scale by recovering LR images to their HR versions, we take advantage of both ideas and align the images not only at the LR scale but also at the HR scale. The framework consists of two modules, a Cascaded Multi-Scale Resolution Reconstruction ($CMSR^2$) model which consists of three IR networks to reconstruct each image to three variants with different resolution scales and a Multi-Resolution Representation Learning (MR^2L) module which consists of three-person re-ID networks, each dedicated to learning

discriminative information from reconstructed images at specific resolution scales. Figure 3 illustrates the overall framework of our proposed method.

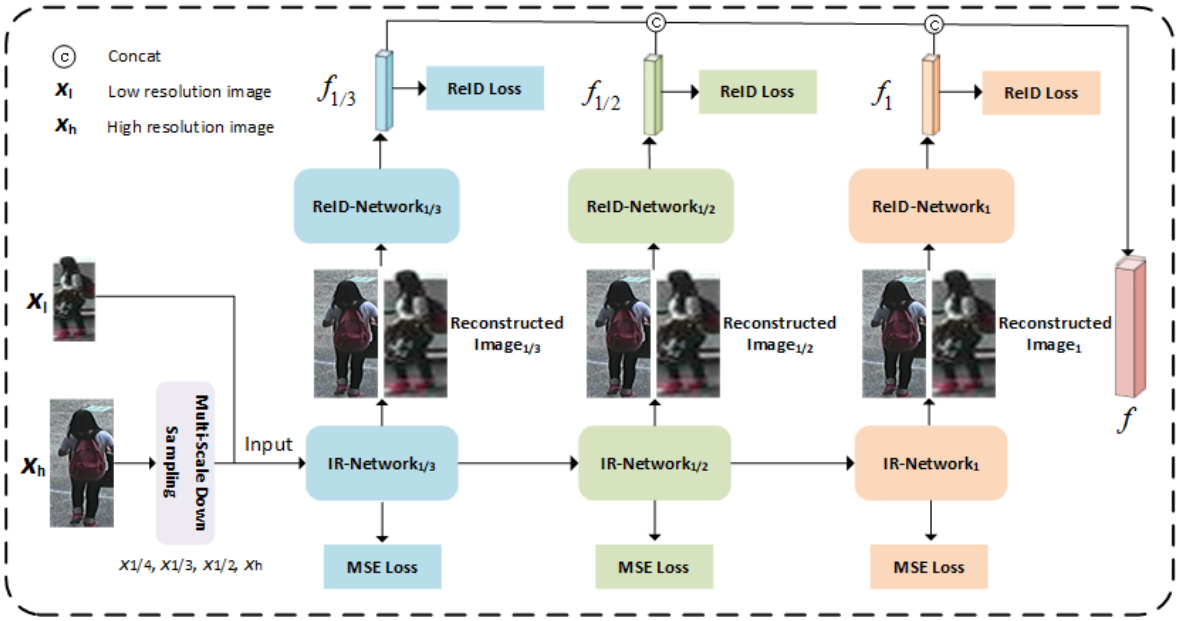

Figure 3. The overall architecture of our proposed method.

Assuming that $X_H = \{x_h, l\}$ is an HR image set, where x_h denotes a HR image and l is corresponding identity label. In order to train the IR networks, we down sample each x_h in the image set X_H to three variants with different resolution scales using down-sampling rate randomly picked from $\{4, 3, 2\}$, and we denote the resolution scales as $LR_{1/4}$, $LR_{1/3}$ and $LR_{1/2}$, respectively. Then the down-sampled images are resized back to their original size via bilinear up-sampling and denoted as $x_{1/4}$, $x_{1/3}$, $x_{1/2}$. The identity label information remains unchanged in the process of down-sampling and resizing. The down-sampled images are only used in the training phase of the proposed framework. IR networks are trained using images with different resolutions generated by HR image subsampling. In this way, each IR network focuses on image reconstruction with specific resolutions and can adapt images with different resolutions as input.

3.2. Cascaded Multi-Scale Resolution Reconstruction Module

Extracting and matching the features of LR and HR images directly will yield poor results because the images are of different distribution. In order to achieve higher retrieval accuracy, the images needed to be aligned at the image level. However, if the images are aligned on the LR scale, the additional information lies in HR images that are abandoned. If we recover LR images to the HR scale by adopting SR models, the reconstructed images cannot fully recover the missing information, so the recovered LR images are still different from genuine HR images.

Motivated by the analysis above, instead of solely aligning them on the LR or HR scale, we propose a Cascaded Multi-Scale Resolution Reconstruction (CMSR^2) module which consists of three IR networks to align the images on three resolution scales. Specifically, each input image, regardless of its resolution, goes forward through all image reconstruction networks consecutively and generates three reconstructed images with three specific resolution scales. Because each original image is reconstructed into three different resolution scales, alignment is achieved at the image level.

Each IR network consists of two components, an encoder and a decoder. The encoder is used to extract feature representations from the images, and it consists of two parts. Each part contains four convolutional layers. The decoder consists of two deconvolutional layers followed by a convolutional layer for image reconstruction. Three IR networks are of the

same architecture but without weight sharing. To alleviate the data loss caused by image reconstruction, residual connections are used between the two components of the encoder, and between the encoder and the decoder.

Note that CMSR^2 is different from previous cascaded SR model-based methods [26] on two major aspects: First, we aim to reconstruct images to three variants with different resolution scales while the idea of [26] is to recover images to HR scale. Second, the input images go forward through all IR networks in CMSR^2, while they go forward through different numbers of SR networks according to their resolution in [26].

In the training phase, the down-sampled images (i.e., $x_{1/4}$, $x_{1/3}$, $x_{1/2}$) and their HR version x_h are fed to the IR-Network$_{1/3}$, and the outputs are denoted as $x^1_{1/4}$, $x^1_{1/3}$, $x^1_{1/2}$ and x^1_h. We expect IR-Network$_{1/3}$ to reconstruct images on $LR_{1/3}$ scale and the generated images are similar to $x_{1/3}$, so we adopt MSE loss to narrow the distance between the outputs of IR-Network$_{1/3}$ and the down-sampled images with $LR_{1/3}$ scale, which is formulated as:

$$L_{IR_{1/3}} = \left\| x^1_{1/4} - x_{1/3} \right\|_2^2 + \left\| x^1_{1/3} - x_{1/3} \right\|_2^2 + \left\| x^1_{1/2} - x_{1/3} \right\|_2^2 + \left\| x^1_h - x_{1/3} \right\|_2^2 \quad (1)$$

After that, $x^1_{1/4}$, $x^1_{1/3}$, $x^1_{1/2}$ and x^1_h go forward through the IR-Network$_{1/2}$ and the IR-Network$_1$ consecutively. We denote the outputs of the IR-Network$_{1/2}$ as $x^2_{1/4}$, $x^2_{1/3}$, $x^2_{1/2}$ and x^2_h and the outputs of the IR-Network$_1$ as $x^3_{1/4}$, $x^3_{1/3}$, $x^3_{1/2}$ and x^3_h. Similarly, the loss function for the IR-Network$_{1/2}$ can be formulated as:

$$L_{IR_{1/2}} = \left\| x^2_{1/4} - x_{1/2} \right\|_2^2 + \left\| x^2_{1/3} - x_{1/2} \right\|_2^2 + \left\| x^2_{1/2} - x_{1/2} \right\|_2^2 + \left\| x^2_h - x_{1/2} \right\|_2^2 \quad (2)$$

Then, the loss function for the IR-Network$_1$ can be formulated as:

$$L_{IR_1} = \left\| x^3_{1/4} - x_h \right\|_2^2 + \left\| x^3_{1/3} - x_h \right\|_2^2 + \left\| x^3_{1/2} - x_h \right\|_2^2 + \left\| x^3_h - x_h \right\|_2^2 \quad (3)$$

The overall loss function for the IR networks is written as:

$$L_{IR} = L_{IR_{1/3}} + L_{IR_{1/2}} + L_{IR_1} \quad (4)$$

3.3. Multi-Resolution Representation Learning Module

Although the CMSR^2 module aligns the person to be matched at the image level by reconstructing each original image to three variants with different resolution scales, they are still not aligned at the feature level. The main reason is that by adopting a single-person re-ID network, the discriminative information lies in reconstructed images with different resolution scales mixed together, so the person to be matched cannot be aligned on different scales separately. We not only expect the reconstructed images with each resolution scale to be of similar distribution, but we also expect the extracted features to contain discriminative information related to the corresponding resolution scales.

To this end, we propose a Multi-Resolution Representation Learning (MR^{2}L) module which utilizes three-person re-ID networks to learn feature representations from the reconstructed images with three resolution scales separately. By extracting and matching the features on each scale, the person to be matched is aligned at the feature level. All person re-ID networks share the same architecture, but there is no weight sharing between them. Thus, each network can focus on extracting discriminative clues from images with a specific resolution scale, avoiding interference from other scales.

The architecture of the person re-ID network is depicted in Figure 4. We adopt ResNet-50 [38] as the backbone network and the extracted tensor T of an input image is partitioned into 4 horizontal parts [39] through an average pooling layer. The features of these parts are concatenated to form the output of the re-ID network.

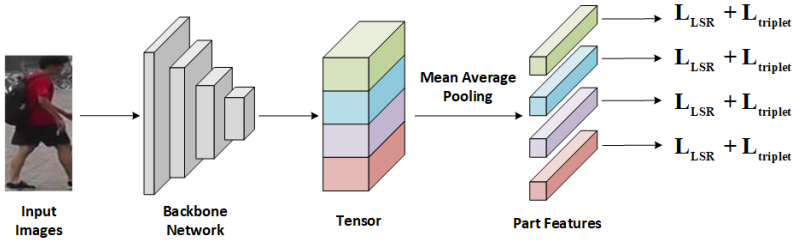

Figure 4. The architecture of person re-ID network.

For an input image (denoted as y), three variants of reconstructed images (denoted as $y_{1/3}$, $y_{1/2}$ and y_1) generated by MR^{2}L module go forward through three re-ID networks, and the extracted features are denoted as $f_{1/3}$, $f_{1/2}$ and f_1, respectively. Thus, the final feature representation of an input image is written as:

$$f = [f_{1/3}, f_{1/2}, f_1] \tag{5}$$

where the lower scripts represent the corresponding resolution scales.

For a reconstructed image y_i, $i \in \{1/3, 1/2, 1\}$, the extracted feature of each person re-ID network can be expressed as:

$$f_i = [f_i^1, f_i^2, f_i^3, f_i^4] \tag{6}$$

where the upper scripts represent the indexes of horizontal parts. Then, Formula (5) can be re-written as:

$$f = [f_{1/3}^1, f_{1/3}^2, f_{1/3}^3, f_{1/3}^4, f_{1/2}^1, f_{1/2}^2, f_{1/2}^3, f_{1/2}^4, f_1^1, f_1^2, f_1^3, f_1^4] \tag{7}$$

Then, we define the loss function of person re-ID network1/3. We adopt cross-entropy loss with label smoothing regularization (LSR) [18] and triplet loss for supervised raining. For a given reconstructed image $y_{1/3}$, the cross-entropy loss is formulated as:

$$L_{Cross_{1/3}} = -\sum_{h=1}^{4}\sum_{c=1}^{C} q_{1/3}^h(c) log(p_{1/3}^h(c)) \tag{8}$$

where $p_{1/3}^h(c)$ is the logits of class which is predicted using feature representation $f_{1/3}^h$. We denote C as the total number of identities in the training set. The ground-truth distribution is recorded as $q_{1/3}^h(c)$, and it can be expressed as:

$$q_{1/3}^h(c) = \begin{cases} 1, l = c \\ 0, l \neq c \end{cases} \tag{9}$$

where l is the ground truth person identity label of the given image. We adopt label smoothing regularization to assign less confidence on the ground truth label, and thus the label distribution can be written as:

$$qLSR_{1/3}^r(c) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{C}, l = c \\ \frac{\varepsilon}{C}, l \neq c \end{cases} \tag{10}$$

where $\varepsilon \in [0, 1]$. Then, we define the cross-entropy loss with label smooth regularization as:

$$L_{LSR_{1/3}} = -\sum_{h=1}^{4}(1-\varepsilon) log(p_{1/3}^h(l)) + \frac{\varepsilon}{C}\sum_{c=1}^{C} log(p_{1/3}^h(c)) \tag{11}$$

For a batch with M identities, and N images for each identity, given a feature representation $f^a_{1/3} = [f^{a,1}_{1/3}, f^{a,2}_{1/3}, f^{a,3}_{1/3}, f^{a,4}_{1/3}]$ of an anchor image, the triplet loss is formulated as:

$$L_{triplet_{1/3}} = \sum_{h=1}^{4} [m + \left\| f^{a,h}_{1/3} - f^{p,h}_{1/3} \right\| - \left\| f^{a,h}_{1/3} - f^{n,h}_{1/3} \right\|]_+ \quad (12)$$

where the feature of the hardest positive sample and the feature of the hardest negative sample are denoted as $f^{p,h}_{1/3}$ and $f^{n,h}_{1/3}$, respectively. The margin is denoted as m, $[z]_+ = max(z, 0)$, and $||f_1 - f_2||$ indicates the Euclidean distance between two features.

The loss function for person re-ID network1/3 can be formulated as:

$$L_{\mathrm{ReID}_{1/3}} = L_{LSR_{1/3}} + L_{triplet_{1/3}} \quad (13)$$

The overall loss function for person re-ID networks can be expressed as:

$$L_{\mathrm{ReID}} = L_{\mathrm{ReID}_{1/3}} + L_{\mathrm{ReID}_{1/2}} + L_{\mathrm{ReID}_1} \quad (14)$$

where $L_{\mathrm{ReID}_{1/2}}$ and L_{ReID_1} are the loss function for re-ID network1/2 and re-ID network1, respectively. Similar to some multi-task learning methods, we do not add L_{IR} and L_{ReID} together and backward them separately.

4. Experiments

4.1. Datasets

Our experiments are based on five public cross-resolution person re-ID datasets, comprising one real-world dataset (i.e., CAVIAR) and four synthetic datasets (i.e., MLR-Market-1501, MLR-CUHK03, MLR-VIPeR and MLR-SYSU).

(1) The CAVIAR dataset [40] consists of images captured by 2 cameras, which contains 1220 images with 72 identities. The person images captured by the two cameras are of different resolutions because the shooting distance of the two cameras is different. Following the experiment setting in [25], we exclude 22 identities that appear in only one camera. For the remaining 50 identities, we randomly select 10 HR and 10 LR images for each identity to construct the dataset.

(2) The MLR-Market-1501 dataset is constructed based on Market-1501 [21] which contains images captured by 6 cameras. The dataset includes 3561 training images with 751 identities and 15,913 testing images with 750 identities. We adopt the same strategy as described in [25] to generate LR images with 3 resolution scales. Specifically, we randomly pick one camera and down-sample each image in the picked camera by randomly picking a down-sampling rate, and the images of other cameras remain unchanged.

(3) The MLR-CUHK03 dataset is constructed based on CUHK03 [41] which contains 14,097 images with 1467 identities captured by 5 cameras. The training set includes images with 1367 identities, and the other 100 identities are used for testing. The down-sampling strategy is the same as for MLR-Market-1501.

(4) The MLR-VIPeR dataset is constructed based on VIPeR [42] which contains 1264 images with 632 identities captured by 2 cameras. We randomly partition the dataset into two non-overlapping parts for training and testing according to the identity label. The down-sampling strategy is the same as for MLR-Market-1501.

(5) The MLR-SYSU dataset is constructed based on SYSU [43] which contains 24,446 images with 502 identities captured by 2 cameras. For each identity, we randomly choose 3 images from each camera. We separate the dataset into two non-overlapping parts for training and testing according to the identity label. The down-sampling strategy is the same as for MLR-Market-1501.

As shown in Figure 5, unlike images captured by conventional cameras, images taken by UAVs exhibit more variations in viewpoints and are characterized by more pronounced background noise.

(a) Surveillance cameras (b) UAV cameras

Figure 5. Comparison of vehicle images captured by surveillance cameras and UAVs.

To validate the generalization of our method, we constructed a synthetic cross-resolution dataset (MLR-VRU) based on the UAV-captured vehicle re-ID dataset VRU [44]. We select 1415 vehicle IDs, totaling 14,611 images for the training set, and each image undergoes downsampling with an equal probability or remains unchanged to simulate resolution changes. The test set is composed of 486 vehicle IDs, totaling 5549 images. For each identity ID, one image is randomly chosen to form the gallery set, while the other images are downsampled to create the query set. The down-sampling strategy is the same as for MLR-Market-1501.

4.2. Experiment Settings

Our code is implemented based on PyTorch, and all experiments are performed on a single RTX 3090 GPU. All images are resized to 256×128, and a total of 60 epochs are trained. The batch consists of a total of 20 images for 5 identities, we pick 2 HR images and 2 LR images for each identity ID. Note that only HR images are used to train the IR networks. For the CMSR^2 module, the kernel size for the encoder is set to 3. The starting learning rate is settled at 3×10^{-3} for the first 30 epochs and decreased to 3×10^{-4} for the remaining epochs. For the MR^{2}L module, we adopt ResNet-50 [38], which is pre-trained on ImageNet [45] as the backbone architecture of each person re-ID network. The initial learning rate is set to 3×10^{-4} and decreased to 3×10^{-5} for the remaining epochs. The margin for the triplet loss is configured as 0.5.

The Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP) are used to evaluate our method.

4.3. Comparison with State-of-the-Art

We compare our method with some state-of-the-art methods, including JUDEA [24], SLD^{2}L [46], SDF [47], SING [25], CSR-GAN [26], CAD-Net [36], INTACT [37], PRI [35] and APSR [48]. The comparison results are reported in Table 1. In the real-world dataset (i.e., CAVIAR), we can see that our method achieves 62.4% in Rank-1 and outperforms all other methods by a very large margin. In particular, we achieve an 18.1% performance boost in Rank-1 over the best competitor (i.e., PCB + PRI). From Table 1, we can also see that the proposed method achieves the best results in Rank-1 on four synthetic datasets. Our method outperforms the compared methods by 0.9%, 6%, 10.6% and 9.5% in Rank-1 on MLR-Market-1501, MLR-CUHK03, MLR-VIPeR and MLR-SYSU, respectively. The experiments demonstrate the superiority of our proposed method.

Table 1. Comparisons of our method with state-of-the-art methods on CAVIAR, MLR-Market-1501, MLR-CUHK03, MLR-VIPeR and MLR-SYSU. Rank-1 (%) and Rank-5 (%) are reported.

Methods	CAVIAR		MLR-Market-1501		MLR-CUHK03		MLR-VIPeR		MLR-SYSU	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
JUDEA [24]	22.0	60.1	-	-	26.2	58.0	26.0	55.1	18.3	41.9
SLD^{2}L [46]	18.4	44.8	-	-	-	-	20.3	44.0	20.3	34.8
SDF [47]	14.3	37.5	-	-	22.2	48.0	9.3	38.1	13.3	26.7
SING [25]	33.5	72.7	74.4	87.8	67.7	90.7	33.5	57.0	50.7	75.4
CSR-GAN [26]	32.3	70.9	76.4	88.5	70.7	92.1	37.2	62.3	-	-
CAD-Net [36]	42.8	76.2	83.7	92.7	82.1	97.4	43.1	68.2	-	-
INTACT [37]	44.0	81.8	88.1	95.0	86.4	97.4	46.2	73.1	-	-
PRI [35]	43.2	78.5	84.9	93.5	85.2	97.5	-	-	-	-
PCB + PRI [35]	44.3	**83.7**	88.1	94.2	86.2	**97.9**	-	-	-	-
APSR [48]	44.0	77.6	-	-	84.1	97.5	48.8	73.2	63.7	83.5
Ours	**62.4**	81.2	**89.6**	**95.6**	**92.4**	96.2	**60.3**	**85.7**	**75.4**	**88.1**

4.4. Ablation Study

We conduct a series of experiments to evaluate the effectiveness of each component of our proposed method, and the results are listed in Table 2. The baseline is similar to the traditional cross-resolution re-ID method which reconstructs the LR images to HR images for comparison. The CMSR^2 module constructs each image to three resolution scales and all reconstructed images go forward through the same re-ID network for feature extraction. From Table 2, we can see that the CMSR^2 module slightly improves the performance over baseline in Rank-1 on all five datasets. However, the performance decreases in Rank-5 on CAVIAR, MLR-VIPeR and MLR-SYSU. The main reason is that three IR networks aim to reconstruct images of different resolution scales, so the images constructed by different IR networks can be considered as different domains. The person re-ID network is disturbed to learn discriminative information from learning to reduce the domain gap between the reconstructed images. When combined with the MR^{2}L module, the performance is improved by a large margin in both Rank-1 and Rank-5 in all five datasets. By feeding the images reconstructed by different IR networks to different person re-ID networks, each network can focus on extracting feature representations of a specific resolution scale, so the network will not be disturbed from trying to reduce the domain gap between different variants of reconstructed images.

Table 2. Comparisons of different variants of the proposed method on CAVIAR, MLR-Market-1501, MLR-CUHK03, MLR-VIPeR and MLR-SYSU. Rank-1 (%) and Rank-5 (%) are reported.

Methods	CAVIAR		MLR-Market-1501		MLR-CUHK03		MLR-VIPeR		MLR-SYSU	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Baseline	50.4	76.0	87.7	94.3	88.8	94.7	55.2	83.2	68.7	86.1
Baseline + CMSR^2	51.6	71.2	88.0	94.9	90.5	94.8	55.9	81.3	70.0	85.3
Baseline + CMSR^2 + MR^{2}L	62.4	81.2	89.6	95.6	92.4	96.2	60.3	85.7	75.4	88.1

We also conduct experiments to evaluate the performance of utilizing recovered images of different resolution scales, and the results are shown in Table 3. The notation IR_1 denotes that all images are reconstructed to HR images for comparison which is similar to traditional cross-resolution person re-ID methods. Specifically, we remove IR-Network-1/3, IR-Network-1/2 from CMSR^2 module and remove ReID-Network-1/3, ReID-Network-1/2 from MR^{2}L module. Similarly, the notation $IR1 + IR\frac{1}{2}$ denote that each image is reconstructed by IR-Network-1/2 and IR-Network-1. Then, the features extracted by ReID-Network-1/2 and ReID-Network-1 are concatenated for comparison. The notation $IR1 + IR\frac{1}{2} + IR\frac{1}{3}$ denotes the full version of the proposed method. From Table 3, we can see that the performance of $IR1 + IR\frac{1}{2}$ improves over IR_1 in Rank-1 because the images

are aligned on two resolution scales instead of solely aligned on HR scale. We can also see that $IR1 + IR\frac{1}{2} + IR\frac{1}{3}$ further improves the performance and achieves the best results in both Rank-1 and Rank-5 on all five datasets compared to IR_1 and $IR1 + IR\frac{1}{2}$. The notation $IR1 + IR\frac{1}{2} + IR\frac{1}{3} + IR\frac{1}{4}$ denotes that all images are first reconstructed to the resolution scale of $IR\frac{1}{4}$, then gradually recover to the scale of HR. Specifically, we add an additional IR network named IR-Network-1/4 and a re-ID network named ReID-Network-1/4 to the framework. We can see that the results are worse than $IR1 + IR\frac{1}{2} + IR\frac{1}{3}$ on CAVIAR, MLR-Market-1501, MLR-CUHK03 and MLR-SYSU, which cannot demonstrate the superiority of $IR1 + IR\frac{1}{2} + IR\frac{1}{3} + IR\frac{1}{4}$. The main reason is that the images of $IR\frac{1}{4}$ contain too little discriminative information which is not enough for person matching. Aligning all images to such a low-resolution scale causes the images with higher resolutions to lose too much information, so adding $IR\frac{1}{4}$ to the framework does not yield better re-ID performance. Above all, we adopt $IR1 + IR\frac{1}{2} + IR\frac{1}{3}$ as our final method from the perspective of both performance and simplicity.

Table 3. Comparisons of using different configurations of resolution scales on CAVIAR, MLR-Market-1501, MLR-CUHK03, MLR-VIPeR and MLR-SYSU. Rank-1 (%) and Rank-5 (%) are reported.

Methods	CAVIAR		MLR-Market-1501		MLR-CUHK03		MLR-VIPeR		MLR-SYSU	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
IR_1	50.4	76.0	87.7	94.3	88.8	94.7	55.2	83.2	68.7	86.1
$IR1 + IR\frac{1}{2}$	52.0	74.4	88.9	95.4	89.0	94.4	57.5	84.0	72.0	86.3
$IR1 + IR\frac{1}{2} + IR\frac{1}{3}$	62.4	81.2	89.6	95.6	92.4	96.2	60.3	85.7	75.4	88.1
$IR1 + IR\frac{1}{2} + IR\frac{1}{3} + IR\frac{1}{4}$	61.6	77.6	89.1	95.5	90.5	95.2	61.3	88.0	74.2	87.1

As shown in Table 4, with only 14611 training samples, we achieve a rank-1 accuracy of 65.7% and an mAP score of 75.7% on the MLR-VRU dataset. We achieve excellent performance even with a small number of training samples, significant background noise, complex visual variations, and a large number of similar vehicles. Under the same experimental settings, increasing the number of training samples significantly improves the performance of cross-resolution retrieval. This indicates that increasing sample diversity can promote the model to acquire more discriminative semantics but also lead to a notable escalation in training time. As the number of training samples reaches a certain extent, the performance presented in the same training epochs is lower than in the case of fewer samples. This is partly because handling larger datasets requires more training epochs for the model to converge. Additionally, IR is a lightweight image reconstruction network that converges quickly and might not adapt well to significant variations in the dataset. For easier deployment, cross-resolution matching models should excel with limited data and remain relatively lightweight. Therefore, we chose to construct the training set with only 1415 IDs for reference in future work.

Table 4. The performance of the proposed method on the MLR-VRU dataset and the sensitivity to variations in training data volume.

Identities/Images	Rank-1	Rank-5	Rank-10	mAP
1415/14,611	65.7	88.4	92.9	75.7
2418/27,918	74.1	92.0	94.5	82.1
4413/52,172	79.6	97.7	99.5	87.4
7086/80,532	74.6	95.8	98.5	83.8

4.5. Visualization

We are the first to evaluate cross-resolution matching performance on the MLR-VRU dataset. Additionally, since there is limited open-source code available for cross-resolution

re-ID and the absence of fair comparative strategies, we validate the effectiveness of our proposed method through visualization.

During the visualization experiments, we apply rotation and size scaling to the retrieval results to ensure easy comparison with the query image. Specifically, when the aspect ratio is more than 1, the image is rotated 90 degrees clockwise; otherwise, it remains unchanged. Then, we resize the image to 288 × 144 pixels.

Figure 6 presents some retrieval results, which can accurately match the correct samples in the gallery set even when the resolution of the query images is low. The first and fourth rows show results that are highly similar in appearance to the query images. This indicates that MSIFLA can extract the overall structure of vehicles and generate sufficiently discriminative features, enabling accurate matching in categories with minor visual disparity. The second and third rows of query images are significantly degraded, lacking detailed vehicle information, making it quite challenging for humans to accurately discern the category to which these images belong. However, the proposed method can still accurately match solely based on the overall vehicle contour.

Figure 6. Visualizations of partial retrieval results of the proposed method on the MLR-VRU dataset. The first column shows low-resolution query images, while the following five columns display the top five retrieval results from high-resolution gallery images. The bounding boxes are used to indicate correct (green) or incorrect (red) retrieval results.

5. Conclusions

In this paper, we have proposed the Multi-Scale Image- and Feature-Level Alignment (MSIFLA) framework for cross-resolution person re-ID by learning feature representations from reconstructed images with three resolution scales. Specifically, we first propose a Cascaded Multi-Scale Resolution Reconstruction ($CMSR^2$) module which consists of three IR networks to reconstruct each image regardless of its resolution to three variants with different resolution scales from low to high consecutively. Thereby, the distribution gap between images with different resolutions is reduced and the images are aligned on three resolution scales at the image level. We further propose a Multi-Resolution Representation Learning (MR^2L) module which consists of three-person re-ID networks to supervise the training of IR networks and learn feature representations from the reconstructed images. Each person re-ID network focuses on learning representations from images with a specific resolution scale, avoiding the disturbance from resolution variation. By matching the features on three resolution scales, the images with different resolutions are also aligned at the feature level. Extensive experiments affirm the superiority of our proposed method. In addition, performing multi-scale down-sampling on high-resolution images to simulate various low-resolution images collected in real-world scenarios enhances the robustness of the model to resolution changes, enabling it to generalize to other scenarios for improved performance on tasks related to resolution variations.

This study has some limitations. The design of the image reconstruction network is relatively simple, and it tends to converge prematurely when the dataset is large, which is not conducive to improving the quality of reconstructed images. In future work, we will refine the structure of the image reconstruction network.

Author Contributions: Conceptualization, methodology, writing, funding acquisition, and supervision, G.Z. and J.Z.; software, validation, and data curation, Z.W., Z.L., Z.Z. and G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant number: 62172231); Natural Science Foundation of Jiangsu Province of China (Grant number: BK20220107); Nuclear energy development project (Grant number: 23zg6106).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hauptmann, A.; Yang, Y.; Zheng, L. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
2. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
3. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [CrossRef] [PubMed]
4. Zhang, G.; Fang, W.; Zheng, Y.; Wang, R. SDBAD-Net: A Spatial Dual-Branch Attention Dehazing Network based on Meta-Former Paradigm. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 60–70. [CrossRef]
5. Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camera style adaptation for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5157–5166.
6. Zhang, G.; Zhang, H.; Lin, W.; Chandran, A.K.; Jing, X. Camera contrast learning for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4096–4107. [CrossRef]
7. Chung, D.; Delp, E.J. Camera-aware image-to-image translation using similarity preserving StarGAN for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
8. Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 542–551.
9. Chen, P.; Liu, W.; Dai, P.; Liu, J.; Ye, Q.; Xu, M.; Chen, Q.; Ji, R. Occlude them all: Occlusion-aware attention network for occluded person re-id. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11833–11842.

10. Zhu, Z.; Jiang, X.; Zheng, F. Viewpoint-aware loss with angular regularization for person re-identification [J/OL]. *arXiv* **2019**. [CrossRef]
11. Zhang, H.; Zhang, G.; Chen, Y.; Zheng, Y. Global relation-aware contrast learning for unsupervised person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8599–8610.
12. Sun, X.; Zheng, L. Dissecting person re-identification from the viewpoint of viewpoint. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 608–617.
13. Zhang, Z.; Da Xu, R.Y.; Jiang, S.; Li, Y.; Huang, C.; Deng, C. Illumination adaptive person reid based on teacher-student model and adversarial training. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Virtual, 25–28 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 2321–2325.
14. Zeng, Z.; Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.Y.; Satoh, S. Illumination-adaptive person re-identification. *IEEE Trans. Multimed.* **2020**, *22*, 3064–3074. [CrossRef]
15. Zhang, G.; Luo, Z.; Chen, Y.; Zheng, Y.; Lin, W. Illumination unification for person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6766–6777. [CrossRef]
16. Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; Hou, Z. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3623–3632.
17. Liu, H.; Tan, X.; Zhou, X. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Trans. Multimed.* **2020**, *23*, 4414–4425. [CrossRef]
18. Chen, Y.; Zhang, G.; Zhang, H.; Zheng, Y.; Lin, W. Multi-level Part-aware Feature Disentangling for Text-based Person Search. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 2801–2806.
19. Qian, X.; Wang, W.; Zhang, L.; Zhu, F.; Fu, Y.; Xiang, T.; Jiang, Y.G.; Xue, X. Long-term cloth-changing person re-identification. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
20. Zhang, G.; Liu, J.; Chen, Y.; Zheng, Y.; Zhang, H. Multi-biometric unified network for cloth-changing person re-identification. *IEEE Trans. Image Process.* **2023**, *32*, 4555–4566. [CrossRef] [PubMed]
21. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
22. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3754–3762.
23. Chen, Y.C.; Li, Y.J.; Du, X.; Wang, Y.C.F. Learning resolution-invariant deep representations for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8215–8222.
24. Li, X.; Zheng, W.S.; Wang, X.; Xiang, T.; Gong, S. Multi-scale learning for low-resolution person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3765–3773.
25. Jiao, J.; Zheng, W.S.; Wu, A.; Zhu, X.; Gong, S. Deep low-resolution person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
26. Wang, Z.; Ye, M.; Yang, F.; Bai, X.; Satoh, S. Cascaded SR-GAN for scale-adaptive low resolution person re-identification. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; Volume 1, p. 4.
27. Zhang, G.; Ge, Y.; Dong, Z.; Wang, H.; Zheng, Y.; Chen, S. Deep high-resolution representation learning for cross-resolution person re-identification. *IEEE Trans. Image Process.* **2021**, *30*, 8913–8925. [CrossRef] [PubMed]
28. Zhang, G.; Chen, Y.; Lin, W.; Chandran, A.; Jing, X. Low resolution information also matters: Learning multi-resolution representations for person re-identification. In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021; pp. 1295–1301.
29. Yan, C.; Fan, X.; Fan, J.; Wang, N. Improved U-Net remote sensing classification algorithm based on Multi-Feature Fusion Perception. *Remote Sens.* **2022**, *14*, 1118. [CrossRef]
30. Chen, W.; Ouyang, S.; Tong, W.; Li, X.; Zheng, X.; Wang, L. GCSANet: A global context spatial attention deep learning network for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1150–1162. [CrossRef]
31. Wang, X.; Tan, L.; Fan, J. Performance Evaluation of Mangrove Species Classification Based on Multi-Source Remote Sensing Data Using Extremely Randomized Trees in Fucheng Town, Leizhou City, Guangdong Province. *Remote Sens.* **2023**, *15*, 1386. [CrossRef]
32. Ma, M.; Ma, W.; Jiao, L.; Liu, X.; Li, L.; Feng, Z.; Yang, S. A multimodal hyper-fusion transformer for remote sensing image classification. *Inf. Fusion* **2023**, *96*, 66–79. [CrossRef]
33. Tian, M.; Yi, S.; Li, H.; Li, S.; Zhang, X.; Shi, J.; Yan, J.; Wang, X. Eliminating background-bias for robust person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5794–5803.
34. Miao, Z.; Liu, H.; Shi, W.; Xu, W.; Ye, H. Modality-aware Style Adaptation for RGB-Infrared Person Re-Identification. In Proceedings of the IJCAI, Montreal, QC, Canada, 19–27 August 2021; pp. 916–922.

35. Han, K.; Huang, Y.; Chen, Z.; Wang, L.; Tan, T. Prediction and recovery for adaptive low-resolution person re-identification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 193–209.
36. Li, Y.J.; Chen, Y.C.; Lin, Y.Y.; Du, X.; Wang, Y.C.F. Recover and identify: A generative dual model for cross-resolution person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8090–8099.
37. Cheng, Z.; Dong, Q.; Gong, S.; Zhu, X. Inter-task association critic for cross-resolution person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2605–2615.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
40. Dong, S.C.; Cristani, M.; Stoppa, M.; Bazzani, L.; Murino, V. Custom pictorial structures for re-identification. In Proceedings of the British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011; Volume 6.
41. Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
42. Gray, D.; Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In Proceedings of the Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Proceedings, Part I 10; Springer: Berlin/Heidelberg, Germany, 2008; pp. 262–275.
43. Chen, Y.C.; Zheng, W.S.; Lai, J.H.; Yuen, P.C. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 1661–1675. [CrossRef]
44. Lu, M.; Xu, Y.; Li, H. Vehicle Re-Identification Based on UAV Viewpoint: Dataset and Method. *Remote Sens.* **2022**, *14*, 4603. [CrossRef]
45. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
46. Jing, X.Y.; Zhu, X.; Wu, F.; You, X.; Liu, Q.; Yue, D.; Hu, R.; Xu, B. 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, Boston, MA, USA, 7–12 June 2015; pp. 695–704.
47. Wang, Z.; Hu, R.; Yu, Y.; Jiang, J.; Liang, C.; Wang, J. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In Proceedings of the IJCAI, New York, NY, USA, 9–15 July 2016; Volume 2, p. 6.
48. Han, K.; Huang, Y.; Song, C.; Wang, L.; Tan, T. Adaptive super-resolution for person re-identification with low-resolution images. *Pattern Recognit.* **2021**, *114*, 107682. [CrossRef]