

Article

# A Domain Adaptive Person Re-Identification Based on Dual Attention Mechanism and Camstyle Transfer

Chengyan Zhong<sup>1</sup>, Guanqiu Qi<sup>2,\*</sup>, Neal Mazur<sup>2</sup>, Sarbani Banerjee<sup>2</sup>, Devanshi Malaviya<sup>2</sup> and Gang Hu<sup>2</sup>

<sup>1</sup> Key Laboratory of Industrial Internet of Things and Networked Control, Ministry of Education, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; S190301022@stu.cqupt.edu.cn

<sup>2</sup> Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA; mazurnm@buffalostate.edu (N.M.); banerjs@buffalostate.edu (S.B.); malavidd01@mail.buffalostate.edu (D.M.); hug@buffalostate.edu (G.H.)

\* Correspondence: qig@buffalostate.edu

**Abstract:** Due to the variation in the image capturing process, the difference between source and target sets causes a challenge in unsupervised domain adaptation (UDA) on person re-identification (re-ID). Given a labeled source training set and an unlabeled target training set, this paper focuses on improving the generalization ability of the re-ID model on the target testing set. The proposed method enforces two properties at the same time: (1) camera invariance is achieved through the positive learning formed by unlabeled target images and their camera style transfer counterparts; and (2) the robustness of the backbone network feature extraction is improved, and the accuracy of feature extraction is enhanced by adding a position-channel dual attention mechanism. The proposed network model uses a classic dual-stream network. Comparative experimental results on three public benchmarks prove the superiority of the proposed method.

**Keywords:** person re-identification; unsupervised domain adaptation; position-channel dual attention mechanism



**Citation:** Zhong, C.; Qi, G.; Mazur, N.; Banerjee, S.; Malaviya, D.; Hu, G. A Domain Adaptive Person Re-Identification Based on Dual Attention Mechanism and Camstyle Transfer. *Algorithms* **2021**, *14*, 361. <https://doi.org/10.3390/a14120361>

Academic Editor: Jörg Rothe

Received: 5 November 2021

Accepted: 7 December 2021

Published: 13 December 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Person re-identification (re-ID) is not only a hot research topic, but also has considerable practical value in computer vision. Given an interested query pedestrian, person re-ID focuses on matching the same pedestrian in a camera network without any overlapping area [1–4]. Following the development of deep learning, supervised person re-ID methods have already achieved high rank-1 accuracy and mAP accuracy on widely used datasets [5–7]. However, these methods heavily rely on a large amount of labeled information in the target domain dataset. It is often difficult to obtain labeled information. Various unsupervised person re-ID methods have been proposed to overcome the lack of labeled information [8–11]. For example, the labeled source domain dataset and the unlabeled target domain dataset are used to train the model, or the unlabeled target domain dataset is directly used to train the model.

Existing unsupervised person re-ID methods can be divided into three categories: (a) Domain adaptation is used to align the feature distribution between source domain and target domain [12–14]. (b) Generative adversarial network (GAN) is used to realize the transformation of pedestrian image style, while the identity annotations of source domain images are retained [15–17]. (c) Generated pseudo-labels on the target domain are used for training, and pseudo labels are assigned to similar images by different methods, such as clustering and KNN search [18–21]. The methods of the third category are restricted by pseudo labels and has poor accuracy in actual application scenarios. Therefore, this article focuses on the first two categories. The corresponding methods of the first two categories use labeled data on the source domain. So, they can be regarded as transfer learning.

Given labeled source data and unlabeled target data, cross-domain re-ID methods are dedicated to learning a model that adapts to the target domain. For cross-domain person re-ID methods, there is no overlap between labeled source domain data and unlabeled target domain data. Both the source domain and target domain have completely different classes (pedestrian identities). In addition, since the imaging styles of different cameras are different, each domain can be further divided into different sub-domains. So, poor migration performance is attributed to two main factors: intra-domain difference and inter-domain conversion. In order to achieve good migration performance, both factors should be considered in the design process. The intra-domain difference is mainly caused by the inconsistency of camera parameters. However, a lot of existing work ignores the first factor.

In fact, in the same target domain, the images captured by different cameras still have obvious style differences. The images captured by one camera can be regarded as a sub-domain of the target domain. The commonly used datasets DukeMTMC-ReID and Market1501 have eight and six sub-domains, respectively, which can be derived from the corresponding camera numbers. In a real-world scene, the distribution of a sub-domain may be considerably different from the distribution of other sub-domains, due to different camera types and image collection scenes. In this case, it is not appropriate to treat any target domain as a whole. It is better to reduce the deviation between each sub-domain in the source and target domains to achieve domain adaptation.

In order to alleviate the above issue, Zhong [17] applied CycleGAN (CamStyle) to generate camera style conversion images to achieve data enhancement for person re-ID. CycleGAN is used to train an image-to-image conversion model for each pair of cameras. Subsequently, the obtained model can generate new image samples that are converted from the source domain to target domain. Since CycleGAN can only be modeled by a one-to-one domain mapping, this method can only learn the mapping between a pair of cameras used in one model. Therefore, multiple models need to be trained to build a complete camera style conversion network by using the CycleGAN method. For example, there are six different cameras in the Market1501 dataset, so  $C_6^2 = 15$  different models need to be trained separately. Similarly, DukeMTMC-ReID needs an additional 28 different models. As the number of cameras increases, the time complexity and the number of parameters increase dramatically. In addition, the cross-camera relationship is ignored in this model. In the proposed method, StarGAN [22] is used to remove the above limitations, and a similarity preservation term is applied to the loss function to achieve image-to-image conversion of camera perception.

Some recently published solutions [23–26] confirm that neighborhood invariance is effective in dealing with changes in the target domain. These methods set up a memory bank to search for the nearest neighbor of each probe in the entire dataset and impose consistent constraints. Since the target domain is unlabeled and lacks the corresponding strong constraints, these models cannot well suppress the impacts of changes between different cameras (including viewing angle and background). In this case, proximity search tends to select candidates captured by the same camera as the probe, but these candidates are not actually correct. In order to solve this issue, this paper uses two loss functions to impose constraints on both inter-camera and intra-camera matching for the enhancement of neighborhood invariance.

In order to obtain more discriminative feature embedding, attention mechanism is introduced. DANet [27,28] introduced a self-attention mechanism to capture the dependency of features in the spatial and channel dimensions, respectively. Specifically, two parallel attention modules, position attention module and channel attention module, are attached to the top of the expanded full convolutional network. In the position attention module, a self-attention mechanism is introduced to capture the spatial dependency between any two positions in the feature map. For the features of a certain location, the weighted sum method is used to aggregate and update the features of all locations. Any weight is determined by the feature similarity of the corresponding two locations. In other words, any two

positions with similar characteristics can promote each other, regardless of their distance in the spatial dimension. In the channel attention module, a similar self-attention mechanism is used to capture the channel dependency between any two channel graphs, and the weighted sum of all the channel graphs is used to update each channel graph. Finally, the outputs of two attention modules are merged to further enhance the feature representation.

Inspired by the weaknesses of existing solutions, this paper proposes a novel unsupervised domain adaptive framework for person re-ID. According to the number of cameras, the target domain is divided into the corresponding sub-domains to perform image style conversion between domains. In addition, due to the poor robustness of image features, a dual attention mechanism is added to the learning framework to analyze feature dependencies.

The main contributions of this paper are summarized as follows.

- StarGAN is introduced into pedestrian image processing to reduce the distribution deviation between different sub-domains in the target dataset. Fast style conversion is applied to multi-domain images. The dataset is expanded while generating high-quality images.
- A dual-channel attention network is integrated to the feature extraction network. More discriminative features are obtained without affecting domain style. The feature dependence from both spatial and channel dimensions is obtained to further enhance feature representation.
- The effectiveness of the proposed method is verified by comparing with state-of-the-art methods on both Market-1501 and DukeMTMC-reID datasets.

The rest of this paper is structured as follows: Section 2 discusses related work; Section 3 presents the proposed method; Section 4 compares the proposed method with state-of-the-art methods and analyzes the related experimental results; and Section 5 concludes this paper.

## 2. Related Work

### 2.1. Unsupervised Domain Adaptation

The source domain is labeled and the target domain is not labeled. The UDA methods focus on solving the target domain without any labels [29–32]. In order to learn the discriminative features in the target domain, early-stage methods focus on the feature/sample mapping between the source domain and target domain. Some existing solutions [16,33–35] reduce the deviation between the source domain and target domain at the image level. The image-to-image conversion from the source domain to the target domain is first performed [22,36], and then the converted images are used to train the model. In addition, some methods [14,37–39] associate two domains with common auxiliary tasks. Wang et al. [14] proposed an attribute consistent framework to achieve UDA on unlabeled target domains by learning the semantics of pedestrian attributes. Huang et al. [37] performed human body segmentation and posed estimation on two domains at the same time to align local features and improve the generalization performance of the corresponding model. Some recently published solutions [9,23–25] recognize the importance of mining discriminative features in the target domain. Yu et al. [9] first used the features of source domain as references to measure whether the image features of the target domain are similar. Then, soft labels were assigned to unlabeled pedestrian identities to explore potential pairwise relationships. Finally, contrast loss was used to reinforce the relationships. Zhong et al. [23] introduced the concept of memory to store the intermediate features of the target data during the training process. Three invariance constraints (named as sample invariance, camera invariance and neighborhood invariance) were imposed on target domain samples to reduce the influence of changes in the corresponding domain. Ding et al. [25] proposed an adaptive exploration (AE) method. In the target domain, the distance between all pedestrians is maximized, and the distance between similar pedestrians is minimized. Yang et al. [24] proposed a patch-based unsupervised framework to learn discriminative features from image patches instead of the entire image. At present, many

methods [40–42] first adopt pseudo-label estimation schemes, then use some clustering algorithms to label target samples, and finally train the corresponding model accordingly. The above operations are repeated until the trained model converges, which results in high computation costs.

## 2.2. Generative Adversarial Networks

In recent years, generative adversarial networks have shown significant improvements in various computer vision tasks, especially image-to-image translation. Radford et al. [43] introduced a deep convolutional generative confrontation network (DCGAN) by proposing some constraints on the network structure. The training process is stable and high resolution images are generated. Pix2Pix [44] as an extension of GANs uses conditional GANs to achieve the mapping from an input image to the output image by combining both adversarial loss and L1 loss. Additionally, this method requires paired data in the training process. Some existing methods [22,36,45] overcome the above limitation. Liu et al. [45] proposed a coupling generative adversarial network (CoGAN) to learn the joint distribution of multi-domain images without any paired image tuples. It can learn the joint distribution by only drawing samples from the marginal distribution, which is achieved by implementing weight sharing constraints. Cyclic consistency adversarial network (CycleGAN) [36] uses cyclic consistency in the image-to-image conversion process without any paired samples to retain key attributes. However, CycleGAN has limited scalability. It can only learn the mapping between two domains. When images are translated between multiple domains, multiple models need to be trained. So, Choi et al. [22] proposed a unified generative confrontation network (StarGAN), which allows one model to learn the mapping between multiple domains.

## 2.3. Self-Attention Modules

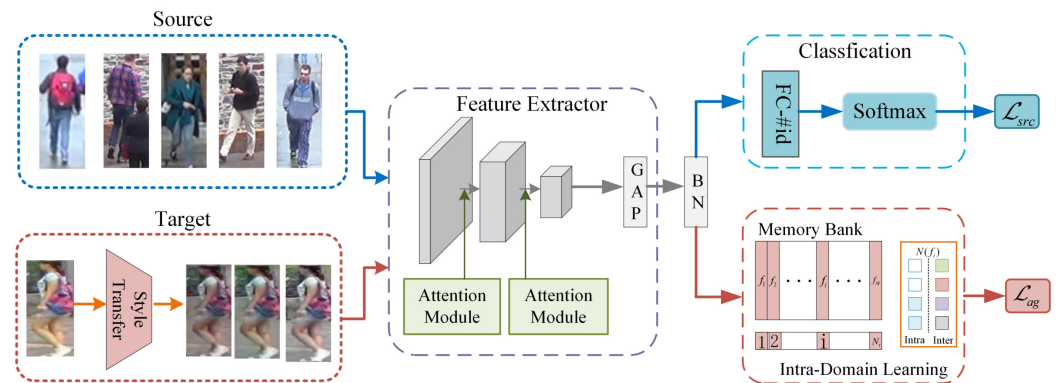
The attention module can establish a long-range dependency model and has been widely used in many tasks [46,47]. The attention mechanism in pedestrian re-ID models aims to highlight key image features to avoid misalignment caused by posture changes, occlusion, or lack of body parts in the bounding boxes [47]. Existing attention mechanisms can be divided into two categories: hard attention and soft attention. The former solutions usually use a pose estimation model to locate rough areas, and then apply the corresponding local features to person re-ID [48]. These hard region-level attentions heavily rely on pose estimation, which is usually inaccurate. Additionally, they do not consider pixel-level information that is important for person re-ID in the selected areas. The soft-attention mechanism usually inserts a trainable layer in the baseline to mask the convolutional feature map to highlight the area rich in information [49]. There are usually two soft-attention mechanisms: spatial attention and channel attention. The former solutions enable the model to focus on valuable features in different spatial locations. The latter solutions enable the model to perform channel re-calibration to improve its characterization capability. On this basis, Fu et al. [27] integrated both types of solutions to propose a dual-channel attention network (DANet), which models semantic relevance in spatial and channel dimensions, respectively. In this paper, this network is added to the backbone network of person re-ID feature extraction.

## 3. The Proposed Method

### 3.1. Overview of the Proposed Framework

As shown in Figure 1, the overall network structure of the proposed method refers to hetero-homogeneous learning (HHL), proposed by Zhong et al. [50]. Both source and target domain samples are forwarded to the network after intra-domain style conversion. The first branch as a classification process is used to learn with labeled source domain samples. With the aid of the memory bank, the learning of the second branch is supervised by the consistency of both intra-camera and inter-camera neighborhoods. The backbone network of the proposed method uses IBN-net [51], which integrates instance normalization (IN)

and batch normalization (BN). IN mainly learns the correlation of visual changes, such as color, style, true, and false, etc. BN mainly learns the content-related information, which can accelerate training and learn more distinguishing features.



**Figure 1.** The framework of the proposed method.

For UDA person re-ID, a labeled source dataset  $S = \{X_s, Y_s\}$  that contains  $N_s$  pedestrian images is obtained. The source domain contains  $N_s$  images of  $P$  pedestrians. Each image corresponds to an identity label  $y_i^s$ . There are  $N_t$  unlabeled target images  $\{x_i^t\}_{i=1}^{N_t}$  from the unlabeled target dataset  $T = \{X_t\}$ . The identity of each target image  $x_t$  in  $\{X_t\}$  is unknown. In addition, the camera index of an image (for example,  $C_s = \{c_i^s\}_{i=1}^{N_s}$  and  $C_t = \{c_i^t\}_{i=1}^{N_t}$ ) is available in both domains. Given the above information, learning a model that can be well generalized to the target domain is achieved.

### 3.2. Supervised Learning for Source Domain

Since the identity labels of source domain images are available, the training process of source domain images can be reduced to a classification problem. Cross entropy loss is used to optimize the network by the following equation.

$$\mathcal{L}_{src} = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log p(y_{s,i} | x_{s,i}) \quad (1)$$

where  $n_s$  is the number of source images in the training batch, and  $p(y_{s,i} | x_{s,i})$  is the predicted probability that the source image  $x_{s,i}$  belongs to the identity  $y_{s,i}$ , which is obtained by the classification module.

A model obtained by training on the labeled source data can produce high accuracy on the same distribution testing dataset. However, when both the testing set and source domain have different distributions, the performance is severely degraded. Next, a method based on example storage is introduced to overcome this problem. This method considers the intra-domain changes of the target domain during the network training.

### 3.3. Intra-Domain Learning

In order to improve the generalization ability of the network on the target testing set, this paper considers the invariance learning in the network by estimating the similarity between target images. A sample memory is first constructed to store the latest features of all target images. The sample memory is a key–value structure [51], which has a key memory ( $\mathcal{K}$ ) and a value memory ( $\mathcal{V}$ ). In the sample memory, each slot stores features in the key part, and labels in the value part. Given a target dataset containing  $N_t$  unlabeled images, each image instance is treated as a separate category. Therefore, the sample memory contains slots, and each slot stores the features and labels of target images. In the initialization process, the values of all the features in the key memory are initialized to zero. For simplicity, the corresponding index as the label of a target sample is specified and stored in value memory. For example, the category of the first target image in value

memory is assigned as  $\mathcal{V}[i] = i$ . During the training process, the labels in value memory are fixed. In each training iteration, each target training sample is input into the deep re-ID network for forward propagation, and the  $L2$  normalized feature  $f(x_{t,i})$  of the output of the fully connected layer is obtained. During the back propagation process, the characteristics of the training samples in key memory are updated in the following way.

$$\mathcal{K}[i] \leftarrow \alpha \mathcal{K}[i] + (1 - \alpha) f(x_{t,i}) \quad (2)$$

where  $\mathcal{K}[i]$  is the key memory of  $x_{t,i}$  image in the  $i$ -th slot.  $\alpha \in [0, 1]$  controls the update rate. Then  $L2$ -normalization  $\mathcal{K}[i]$  is performed by  $\mathcal{K}[i] \leftarrow \|\mathcal{K}[i]\|_2$ .

### 3.4. Camera-Aware Neighborhood Invariance

When the label space (given by the identity annotation and the number of identities) is unknown, it is not feasible to directly analyze the categories of target samples. In this case, the pairwise relationship is a potential clue to guide the feature learning in target domain. In characteristics learning, it is usually assumed that each sample is likely to share the same basic label with its nearest neighbors. The probability that  $x_i^t$  and  $x_j^t$  share the same identity can be instantly obtained by using the memory bank mentioned above as follows.

$$p_{ij} = \frac{\exp(s \times \mathcal{K}_j^T f(x_i^t))}{\sum_{n=1}^{N_i} \exp(s \times \mathcal{K}_n^T f(x_i^t))} \quad (3)$$

where  $s$  is the scale factor, which can adjust the sharpness of the probability distribution. Based on the above assumptions, ECN [23] was proposed to maximize the probability of each detection image and its nearest neighbor in the entire dataset as follows.

$$\mathcal{L}_{ag} = - \sum_j w_{i,j} \log p_{ij}, w_{i,j} = \begin{cases} \frac{1}{|\Omega_i|}, j \neq i \\ 1, j = i \end{cases}, \forall j \in \Omega_i \quad (4)$$

where  $\Omega_i$  represents the nearest neighbor of  $x_i^t$  in the entire target domain.  $|\Omega(x_i^t)|$  indicates the size of the neighbor set. For convenience, this loss function is called the camera-independent neighborhood loss because it treats all candidates equally, regardless of their camera indexes, when searching for neighbors.

Due to the scene changes between cameras, there is a significant difference in the distribution of similarity between inter-camera matching and intra-camera matching [52]. The average pairwise similarity of inter-camera matches is smaller than that of intra-camera matches. In this case, Equation (3) pushes the positive match between cameras away from the probe, which causes issues. As an intuitive solution, a larger neighborhood is selected. However, such an approach inevitably involves more negative matches, which is not conducive to feature learning.

To solve this problem, this paper proposes to enforce neighborhood invariance for intra-camera matching and inter-camera matching, respectively. It has the following two assumptions.  $O_i^{intra}$  represents an instance set that shares the same camera with  $x_i^t$ .  $O_i^{inter}$  represents an instance set that has a different camera index from  $x_i^t$ . For both intra-camera matching and inter-camera matching of samples, only the instances in  $O_i^{intra}$  and  $O_i^{inter}$  can be accessed respectively. Therefore, the probability that  $x_i^t$  shares the same identity with candidate  $x_j^t$  in the camera is formalized as follows.

$$p_{i,j}^{intra} = \frac{\exp(s \times \mathcal{K}_j^T f(x_i^t))}{\sum_{n \in O_i^{intra}} \exp(s \times \mathcal{K}_n^T f(x_i^t))} \quad (5)$$

As shown in Equation (6), the definition of the probability that the candidate has the same identity between  $x_i^t$  and the camera is similar to Equation (5).

$$p_{i,j}^{inter} = \frac{\exp\left(s \times \mathcal{K}_j^T f(x_i^t)\right)}{\sum_{n \in O_i^{inter}} \exp\left(s \times \mathcal{K}_n^T f(x_i^t)\right)} \quad (6)$$

Therefore, the original camera-agnostic loss function shown in Equation (3) is replaced by the following two camera-perceived loss functions.

$$\begin{aligned} \mathcal{L}_{intra} &= - \sum_j w_{i,j} \log p_{i,j}^{intra}, \forall j \in \Omega_i^{intra} \\ \mathcal{L}_{inter} &= - \sum_j w_{i,j} \log p_{i,j}^{inter}, \forall j \in \Omega_i^{inter} \end{aligned} \quad (7)$$

where  $\Omega_i^{intra}$  and  $\Omega_i^{inter}$  represent the neighborhood set of  $x_i^t$  between  $O_i^{intra}$  and  $O_i^{inter}$ , respectively. This paper defines the neighborhood based on the relative similarity with the top-1 neighbor.

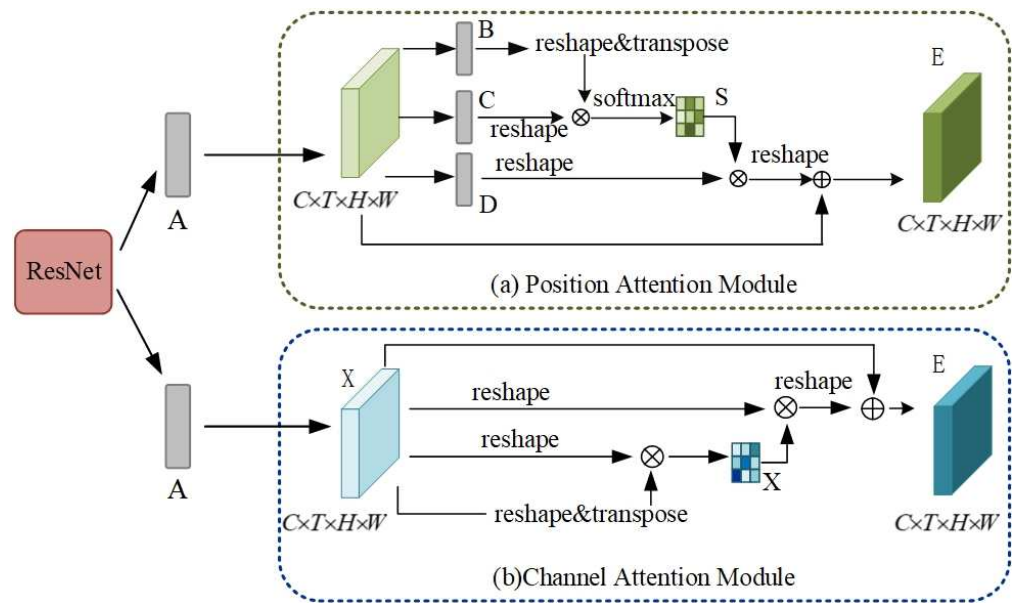
In addition, the mining neighborhood used in intra-camera matching is more reliable than the mining neighborhood used in inter-camera matching without interference from cross-camera changes. Therefore, it is much easier to learn the discriminative intra-camera representation first, which can be conducive to accurate inter-camera matching. Therefore, the proposed method uses  $\mathcal{L}_{intra}$  before participation in  $\mathcal{L}_{inter}$ .

### 3.5. Style Transfer

The image style change caused by cameras is a key factor that affects the testing process of person re-ID. In order to realize the invariance property of cameras in the target domain, the proposed method learns the images of the same pedestrian with different camera styles in an unlabeled target domain. These images retain different amounts of personal identity information and reflect other camera styles, which can be used to generate new target images. The proposed method adopts the CamStyle method to learn the camera style transfer model in the target dataset. Unlike the image-to-image translation methods using CycleGAN [44], CamStyle based on StarGAN [51] was proposed. StarGAN can train multi-camera image-to-image conversion with a single model, while CycleGAN needs to train a conversion model for each pair of cameras. Suppose the images in the target dataset are captured by  $C$  cameras. A StarGAN model is first trained, which can perform image-to-image conversion between each camera pair. For the real target image  $x_{t,j}$  collected by camera  $j$  ( $j \in 1, 2, \dots, C$ ) in the target dataset,  $C$  fake images (camera style transfer)  $x_{t^*,1}, x_{t^*,2}, \dots, x_{t^*,C}$  that more or less contain the same number of people as  $x_{t,j}$  are generated by using the learned StarGAN model. However, their styles are similar to cameras  $1, 2, \dots, C$ . These  $C$  images contain the style transferred from camera  $j$  that has the style of the real image  $x_{t,j}$ .

### 3.6. Dual Attention Network

For the attention mechanism, the proposed method adopts the dual attention network proposed by Fu et al. [27]. The specific structure is shown in Figure 2. The long-range contextual information in both the space and channel dimensions is captured, respectively.



**Figure 2.** The details of the position attention module and channel attention module.

### 3.6.1. Position Attention Module

The location attention module is introduced to build a rich context model based on local features. The location attention module encodes broad context information into local features, thereby enhancing their representation capabilities. As shown in Figure 2a, the given local feature  $A \in R^{C \times H \times W}$  is first input into a convolutional layer to generate two new feature maps  $B$  and  $C$ , where  $\{B, C\} \in R^{C \times H \times W}$ . Then they are reshaped into  $R^{C \times N}$ , where  $N = H \times W$  is the number of pixels. Next, matrix multiplication is performed between the transpose of  $C$  and  $B$ , and the softmax layer is applied to calculate the space. Note that the figure  $S \in R^{N \times N}$ .

$$s_{ji} = \frac{\exp(B_i \times C_j)}{\sum_{i=1}^N \exp(B_i \times C_j)} \quad (8)$$

where  $s_{ji}$  represents the influence of the  $i$ -th position on the  $j$ -th position. As the similarity of the feature representations of two locations increases, the correlation between them gets close.

Additionally, feature  $A$  is sent into the convolutional layer to generate a new feature map  $D \in R^{C \times N}$ , and then the obtained feature map is reshaped into  $R^{C \times N}$ . Next, matrix multiplication is performed between the transpose of  $D$  and  $S$ , and the result is reshaped to  $R^{C \times H \times W}$ . Finally, the reshaped result is multiplied by the scale parameter  $\alpha$  and an element-wise summation is performed by using feature  $A$  to obtain the final output  $E \in R^{C \times H \times W}$  as follows.

$$E_j = \alpha \sum_{i=1}^N (s_{ji} D_i) + A_j \quad (9)$$

where  $\alpha$  is initialized to 0, and then more weights are gradually assigned by learning. According to Equation (9), the resulting feature  $E$  of each location is the weighted sum of the features of all locations and the original feature. Therefore, there is a global context view, and the contexts based on the spatial attention map are selectively aggregated. Similar semantic features are mutually improved, thereby improving intra-class compactness and semantic consistency.



### 3.6.2. Channel Attention Module

Each channel graph of high-level features can be regarded as a class-specific response, and different semantic responses are related to each other. By taking advantage of the interdependence between channel mappings, the inter-dependent feature mapping is emphasized, and the feature representation of specific semantics is improved. Therefore, a channel attention module is used to explicitly model the inter-dependence between channels. The structure of the channel attention module is shown in Figure 2b. Different from the position attention module, this module directly calculates the channel attention map  $X \in R^{C \times C}$  from the original feature  $A \in R^{C \times H \times W}$ . Specifically,  $A$  is reshaped into  $R^{C \times N}$ , and then matrix multiplication is performed between  $A$  and  $A$ 's transpose. Finally, the softmax layer is applied to obtain the channel attention map  $X \in R^{C \times C}$  as follows.

$$x_{ji} = \frac{\exp(A_i \times A_j)}{\sum_{i=1}^C \exp(A_i \times A_j)} \quad (10)$$

where  $x_{ji}$  represents the influence of the  $i$ -th channel on the  $j$ -th channel. In addition, a matrix multiplication is performed between the transpose of  $X$  and  $A$ , and the result is reshaped to  $R^{C \times H \times W}$ . Then the result is multiplied by a scale parameter  $\beta$  and an element-wise summation operation is performed on  $A$  to obtain the final output  $E \in R^{C \times H \times W}$  as follows.

$$E_j = \beta \sum_{i=1}^C (x_{ji} A_i) + A_j \quad (11)$$

The weights are gradually learned from 0. Equation (11) shows that the final feature of each channel is the weighted sum of all channel features and the original features, thereby establishing a long-term semantic dependency between feature maps. It is conducive to improving the distinguish ability of features.

In order to make full use of remote context information, the characteristics of these two attention modules are integrated. Specifically, the outputs of the two attention modules are transformed through the convolutional layer, and the element summation is performed to complete the feature fusion. Lastly, the final prediction map is generated through the convolutional layer.

## 4. Experiments

### 4.1. Dataset and Evaluation Metrics

The proposed method is applied to the widely used person re-ID datasets Market1501 and DukeMTMC-ReID for verification.

Market1501 was collected on the campus of Tsinghua University, constructed and published in 2015. It includes 1501 pedestrians captured by 6 cameras and a total of 32,668 images. Among them, 12,936 images of 751 pedestrians (identities) are used for training, and 19,732 images of 750 pedestrians (identities) are used for testing.

DukeMTMC-ReID is a person re-ID subset of the DukeMTMC dataset. It contains 16,522 training images from 702 people, and 2228 query images from the other 702 people. The search gallery consists of 17,661 images. During the training process, the images and camera labels are only used in the training set of each dataset. No other annotation information is used. The two datasets are used as the source domain and target domain, respectively. During the testing process, the cumulative matching features (CMC) of rank-1, rank-5, rank-10, and average accuracy (mAP) are evaluated in the testing dataset of the target domain.

### 4.2. Deep Re-ID Model

The proposed method uses the ResNet-based IBN-net as the backbone. The last downsampling work of ResNet is discarded, resulting in a total stride of 16.  $\mathcal{L}_{intra}$  and

$\mathcal{L}_{inter}$  participated in the second and fourth rounds of training, respectively. In terms of the optimizer, stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of  $1 \times 10^{-5}$  is used. The learning rates of the backbone layer and the new layer are set to 0.0013 and 0.004, respectively. They are divided by 10 in the sixth period. The entire training process lasts 15 epochs. Each mini-batch of data contains 32 source images and 32 target images. The size of all input images is adjusted to  $256 \times 128$ . The setting of scale factor  $s = 10$ , neighborhood range = 0.85, memory update momentum  $\sigma = 0.7$ , and beta distribution parameter  $\alpha = 0.6$  are used in experiments. During the testing process, the output of the final batch normalization layer is used as the image embedding. The cosine similarity is used as a measure of retrieval. All experiments are performed on two TESLA P100 using the PyTorch platform.

#### 4.3. Parameter Analysis

**Beta distribution parameter  $\alpha$ :** The parameter  $\alpha$  determines the distribution of the interpolation coefficient  $\lambda$ . Assigning a large value to  $\alpha$  results in a strong regularization. The value of parameter  $\alpha$  is changed to five different values and the corresponding performance is evaluated under the above settings. As shown in Figure 3, both the rank-1 accuracy and mAP fluctuate little with the change in  $\alpha$ . It confirms that the proposed method is relatively robust to cross-domain hybrid settings.

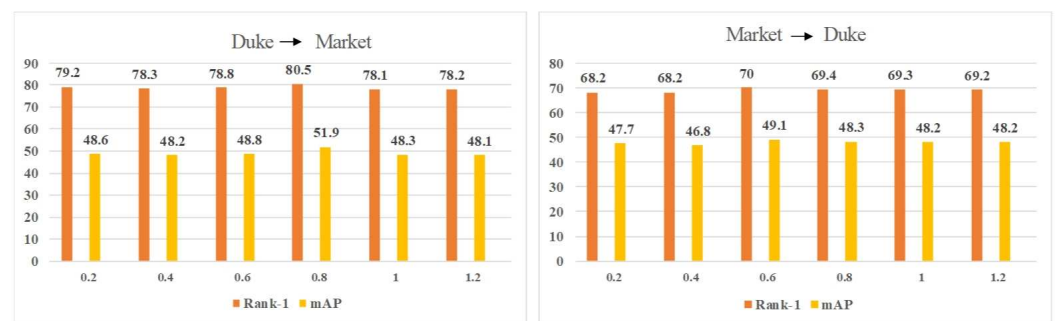


Figure 3. Evaluation with different values of the beta distribution parameter  $\alpha$ .

**Scaling factor  $s$ :** The scale factor  $s$  in Equation (3) is critical to the final performance. A large  $s$  can sharpen the probability distribution and simplify optimization. However, assigning too a large value to  $s$  may cause the task to be too trivial to learn discriminative features. The model under five different values of  $s$  is trained, and the corresponding results are shown in Table 1. When  $s = 10$  and  $s = 12$ , the proposed method achieves the best performance on Market-1501 and DukeMTMC-reID, respectively. When  $s$  becomes too large or too small, the corresponding performance of the proposed method drops considerably.

Table 1. Evaluation with different scale factor  $s$  values.

$s$	Duke to Market				Market to Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
6	68.5	82.3	86.6	38.3	61.8	72.6	76.4	39.3
8	75.1	86.4	89.7	45.6	65.9	75.9	80.4	44.2
10	<b>80.1</b>	<b>89.9</b>	<b>93.2</b>	<b>60.1</b>	68.1	79.1	82.3	46.9
12	78.8	88.9	91.9	54.3	<b>69.5</b>	80.4	83.4	48.5
14	76.6	87.4	90.6	53.5	68.9	<b>80.7</b>	<b>84.3</b>	<b>53.0</b>

#### 4.4. Ablation Study

In order to verify the effectiveness of the attention mechanism, DukeMTMC-reID is used as the source domain to evaluate the performance of the attention mechanism on Market-1501. In the main body network, the proposed method adds location attention and

channel attention to the first and fifth layers of the backbone network. The experiments prove that adding two channels at the same time by controlling variables helps improve the accuracy of the model. As shown in Table 2, when no attention mechanism is added, the accuracy of both mAP and Rank-1 is the lowest, which is about 3% lower than the corresponding values. However, only the positional attention mechanism or the channel attention mechanism is still less effective than the best accuracy. In addition, more experiments are implemented. The channel attention mechanism is added to the first layer and the position attention mechanism is added to the fifth layer. The position attention mechanism and channel attention mechanism are added to the first and fifth layers, respectively. The final accuracy shows that the performance of the proposed model is optimal only when both the position attention mechanism and channel attention mechanism are added at the same time.

**Table 2.** An ablation study on DukeMTMC-reID to Market1501. The best accuracy is reached when channel and location attention are added.

Method	Duke to Market			
	Rank-1	Rank-5	Rank-10	mAP
None	77.8	88.1	91.7	48.4
Channel	78.8	88.3	91.7	48.6
Position	78.7	88.9	92	48.5
Position-Channel	77.9	88	91.5	50.5
Channel-Position	79.1	88.7	92.5	53.5
Channel+Position	<b>80.5</b>	<b>89.5</b>	<b>93.2</b>	<b>60.1</b>

#### 4.5. Comparison with State-of-the-Art Methods

**Results on Market-1501 dataset.** DukeMTMC-reID is used as the source domain to evaluate the performance of the proposed method on Market-1501. The results are compared with the representative works in different directions, including methods based on style transfer [16,34,52], methods based on pseudo-label estimation [42,51], and methods in mining domains [9,23–25]. As shown in Table 3, the proposed method is superior to the current leading methods in terms of Rank-1 accuracy and mAP.

**Table 3.** Comparison with state-of-the-art cross-domain methods on Market-1501 and DukeMTMC-reID. The proposed method shows better accuracy on both datasets.

Method	Market1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PTGAN [16]	38.6	-	66.1	-	27.4	-	50.7	-
SPGAN [52]	51.5	70.1	76.8	22.8	41.1	56.6	63	22.3
CamStyle [53]	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
HHL [50]	62.2	78.8	84	31.4	46.9	61	66.7	27.2
MAR [9]	67.7	81.9	-	40	67.1	79.8	-	48
PAUL [24]	68.5	82.4	87.4	40.1	72	82.7	86	53.2
ARN [54]	70.3	80.4	86.3	39.4	60.2	73.9	79.5	33.4
ECN [23]	75.1	87.6	91.6	43	63.3	75.8	80.4	40.4
UDA [55]	75.8	89.5	93.2	53.7	68.4	80.1	83.5	49
PAST [41]	78.4	-	-	54.6	72.4	-	-	54.3
SSG [42]	80	90	92.4	58.3	73	80.6	83.2	53.4
CV-DA [51]	79.7	89	91.4	59.8	71.1	81.2	84.2	52.6
Ours	80.5	89.9	93.2	60.1	71.2	81.7	84.3	53.0

**Results of DukeMTMC-reID dataset.** Market-1501 is used as the source domain, and the performance of the proposed method is evaluated on DukeMTMC-reID. As shown on

the right side of Table 3, the performance of the proposed method is competitive with other state-of-the-art methods.

## 5. Conclusions

In this paper, a cross-domain person re-ID model is proposed, which considers both intra-domain changes and inter-domain transfers. The neighborhood invariance method is used to supervise feature learning in the target domain. However, due to the huge differences between cameras, the neighbor search often has a relatively large deviation. So, the proposed method applies StarGan to transform the image style. In addition, this paper uses a channel attention mechanism and a position attention mechanism to improve the robustness of feature extraction of the backbone network. An ablation study verifies the effectiveness of each proposed module. Comparative experiments verify that the proposed method outperforms state-of-the-art methods. The proposed method achieves higher accuracy of two objective evaluation indicators, rank-1 and mAP, on two general datasets. In future research, unsupervised learning of person re-ID will be further explored to improve the generalization capability of the proposed model in practical applications. Additionally, data security enhancement in research will also be considered to prevent the related data from being used by illegal organizations.

The General Data Protection Regulation (GDPR) was issued by the European Union in 2018. When cameras are deployed in public areas for capturing images used in further analysis, data privacy protection should indeed be considered. In addition, China has also issued relevant regulations, such as the Personal Information Protection Law and the Data Security Law. Security cameras currently focus on capturing the overall appearance information of pedestrians. The resolution of current security cameras is not high enough. Generally, they do not obtain clear facial information. So, it is difficult to achieve the related analysis of pedestrian information by facial recognition. As a basic principle of research, when any research involves obtaining and processing pedestrian images, the related research must strictly abide by the laws and regulations of various countries. Any personal identity information involved in the related research is not allowed to be obtained, kept, or sold.

**Author Contributions:** Conceptualization, C.Z. and G.Q.; methodology, G.Q., N.M. and G.H.; software, C.Z. and G.Q.; validation, S.B. and D.M.; formal analysis, C.Z., G.Q. and S.B.; investigation, D.M. and G.H.; resources, C.Z. and G.Q.; data curation, S.B. and D.M.; writing—original draft preparation, C.Z. and G.Q.; writing—review and editing, G.Q., N.M., S.B. and G.H.; visualization, D.M.; supervision, G.Q. and N.M.; project administration, G.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. 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.
2. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 17–35.
3. Qi, G.; Hu, G.; Wang, X.; Mazur, N.; Zhu, Z.; Haner, M. EXAM: A Framework of Learning Extreme and Moderate Embeddings for Person Re-ID. *J. Imaging* **2021**, *7*, 6. [[CrossRef](#)] [[PubMed](#)]
4. Li, Y.; Chen, S.; Qi, G.; Zhu, Z.; Haner, M.; Cai, R. A GAN-Based Self-Training Framework for Unsupervised Domain Adaptive Person Re-Identification. *J. Imaging* **2021**, *7*, 62. [[CrossRef](#)]

5. 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.
6. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **2019**, *22*, 2597–2609. [[CrossRef](#)]
7. Li, Y.; Wang, X.; Zhu, Z.; Huang, X.; Li, P.; Qi, G.; Rong, Y. A Novel Person Re-ID Method based on Multi-Scale Feature Fusion. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 7154–7159.
8. Jin, X.; Lan, C.; Zeng, W.; Chen, Z. Global distance-distributions separation for unsupervised person re-identification. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 735–751.
9. Yu, H.X.; Zheng, W.S.; Wu, A.; Guo, X.; Gong, S.; Lai, J.H. Unsupervised person re-identification by soft multilabel learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2148–2157.
10. Li, H.; Chen, Y.; Tao, D.; Yu, Z.; Qi, G. Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1480–1494. [[CrossRef](#)]
11. Li, H.; Dong, N.; Yu, Z.; Tao, D.; Qi, G. Triple Adversarial Learning and Multi-View Imaginative Reasoning for Unsupervised Domain Adaptation Person Re-Identification. Available online: <https://ieeexplore.ieee.org/abstract/document/9495801> (accessed on 5 November 2021).
12. Wu, A.; Zheng, W.S.; Lai, J.H. Unsupervised person re-identification by camera-aware similarity consistency learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 6922–6931.
13. Lin, S.; Li, H.; Li, C.T.; Kot, A.C. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv* **2018**, arXiv:1807.01440.
14. Wang, J.; Zhu, X.; Gong, S.; Li, W. Transferable joint attribute-identity deep learning for unsupervised 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. 2275–2284.
15. Zou, Y.; Yang, X.; Yu, Z.; Kumar, B.V.; Kautz, J. Joint disentangling and adaptation for cross-domain person re-identification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 87–104.
16. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap 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. 79–88.
17. 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.
18. Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; Yang, Y. A bottom-up clustering approach to unsupervised person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8738–8745.
19. Fan, H.; Zheng, L.; Yan, C.; Yang, Y. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–18. [[CrossRef](#)]
20. Zhai, Y.; Lu, S.; Ye, Q.; Shan, X.; Chen, J.; Ji, R.; Tian, Y. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9021–9030.
21. Zhu, Z.; Luo, Y.; Chen, S.; Qi, G.; Mazur, N.; Zhong, C.; Li, Q. Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification. *J. Vis. Commun. Image Represent.* **2021**, *80*, 103303. [[CrossRef](#)]
22. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
23. Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; Yang, Y. Invariance matters: Exemplar memory for domain adaptive person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 598–607.
24. Yang, Q.; Yu, H.X.; Wu, A.; Zheng, W.S. Patch-based discriminative feature learning for unsupervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3633–3642.
25. Ding, Y.; Fan, H.; Xu, M.; Yang, Y. Adaptive exploration for unsupervised person re-identification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *16*, 1–19. [[CrossRef](#)]
26. Zhong, C.; Jiang, X.; Qi, G. Video-based Person Re-identification Based on Distributed Cloud Computing. *J. Artif. Intell. Technol.* **2021**, *1*, 110–120. [[CrossRef](#)]
27. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
28. Zhu, Z.; Luo, Y.; Wei, H.; Li, Y.; Qi, G.; Mazur, N.; Li, Y.; Li, P. Atmospheric Light Estimation Based Remote Sensing Image Dehazing. *Remote Sens.* **2021**, *13*, 2432. [[CrossRef](#)]

29. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
30. Chen, Y.; Zhu, X.; Gong, S. Person re-identification by deep learning multi-scale representations. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2590–2600.
31. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning transferable features with deep adaptation networks. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015.
32. Zhu, Z.; Luo, Y.; Qi, G.; Meng, J.; Li, Y.; Mazur, N. Remote Sensing Image Defogging Networks Based on Dual Self-Attention Boost Residual Octave Convolution. *Remote Sens.* **2021**, *13*, 3104. [[CrossRef](#)]
33. Huang, Y.; Wu, Q.; Xu, J.; Zhong, Y. SBSGAN: Suppression of inter-domain background shift for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 9527–9536.
34. Liu, J.; Zha, Z.J.; Chen, D.; Hong, R.; Wang, M. Adaptive transfer network for cross-domain person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7202–7211.
35. Sun, J.; Qi, G.; Mazur, N.; Zhu, Z. Structural Scheduling of Transient Control Under Energy Storage Systems by Sparse-Promoting Reinforcement Learning. *IEEE Trans. Ind. Inform.* **2022**, *18*, 744–756. [[CrossRef](#)]
36. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
37. Huang, H.; Yang, W.; Chen, X.; Zhao, X.; Huang, K.; Lin, J.; Huang, G.; Du, D. Eanet: Enhancing alignment for cross-domain person re-identification. *arXiv* **2018**, arXiv:1812.11369.
38. Zhu, Z.; Wei, H.; Hu, G.; Li, Y.; Qi, G.; Mazur, N. A Novel Fast Single Image Dehazing Algorithm Based on Artificial Multiexposure Image Fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–23. [[CrossRef](#)]
39. Zheng, M.; Qi, G.; Zhu, Z.; Li, Y.; Wei, H.; Liu, Y. Image Dehazing by an Artificial Image Fusion Method Based on Adaptive Structure Decomposition. *IEEE Sens. J.* **2020**, *20*, 8062–8072. [[CrossRef](#)]
40. Wu, J.; Liao, S.; Wang, X.; Yang, Y.; Li, S.Z. Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 886–891.
41. Zhang, X.; Cao, J.; Shen, C.; You, M. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 8222–8231.
42. Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; Huang, T.S. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 6112–6121.
43. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
44. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
45. Liu, M.Y.; Tuzel, O. Coupled generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 469–477.
46. Ning, X.; Gong, K.; Li, W.; Zhang, L.; Bai, X.; Tian, S. Feature Refinement and Filter Network for Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3391–3402. [[CrossRef](#)]
47. Xia, B.N.; Gong, Y.; Zhang, Y.; Poellabauer, C. Second-order non-local attention networks for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3760–3769.
48. Zheng, L.; Huang, Y.; Lu, H.; Yang, Y. Pose-invariant embedding for deep person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 4500–4509. [[CrossRef](#)]
49. Chen, B.; Deng, W.; Hu, J. Mixed high-order attention network for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October 2019; pp. 371–381.
50. Zhong, Z.; Zheng, L.; Li, S.; Yang, Y. Generalizing a person retrieval model hetero-and homogeneously. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–188.
51. Jia, X.; Wang, X.; Mi, Q. An unsupervised person re-identification approach based on cross-view distribution alignment. *IET Image Process.* **2021**, *15*, 2693–2704. [[CrossRef](#)]
52. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 994–1003.
53. Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camstyle: A novel data augmentation method for person re-identification. *IEEE Trans. Image Process.* **2018**, *28*, 1176–1190. [[CrossRef](#)] [[PubMed](#)]
54. Li, Y.J.; Yang, F.E.; Liu, Y.C.; Yeh, Y.Y.; Du, X.; Frank Wang, Y.C. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–178.
55. Song, L.; Wang, C.; Zhang, L.; Du, B.; Zhang, Q.; Huang, C.; Wang, X. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognit.* **2020**, *102*, 107173. [[CrossRef](#)]