Feature-Level Camera Style Transfer for Person Re-Identification

Yang Liu 1,2, Hao Sheng 1,2,3,* Shuai Wang 1,2, Yubin Wu 1,2 and Zhang Xiong 1

1 State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China; liu.yang@buaa.edu.cn (Y.L.); shuawang@buaa.edu.cn (S.W.); yubin.wu@buaa.edu.cn (Y.W.); xiongz@buaa.edu.cn (Z.X.)
2 Beihang Hangzhou Innovation Institute Yuhang, Xixi Octagon City, Yuhang District, Hangzhou 310023, China
3 Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, China
* Correspondence: shenghao@buaa.edu.cn

Abstract: The person re-identification (re-ID) problem has attracted growing interest in the computer vision community. Most public re-ID datasets are captured by multiple non-overlapping cameras, and the same person may appear dissimilar in different camera views due to variances of illuminations, viewpoints and postures. These differences, collectively referred to as camera style variance, make person re-ID still a challenging problem. Recently, researchers have attempted to solve this problem using generative models. The generative adversarial network (GAN) is widely used for the pose transfer or data augmentation to bridge the camera style gap. However, these methods, mostly based on image-level GAN, require huge computational power during the training of generative models. Furthermore, the training process of GAN is separated from the re-ID model, which makes it hard to achieve a global optimal for both models simultaneously. In this paper, the authors propose to alleviate camera style variance in the re-ID problem by adopting a feature-level Camera Style Transfer (CST) model, which can serve as an intra-class augmentation method and enhance the model robustness against camera style variance. Specifically, the proposed CST method transfers the camera style-related information of input features while preserving the corresponding identity information. Moreover, the training process can be embedded into the re-ID model in an end-to-end manner, which means the proposed approach can be deployed with much less time and memory cost. The proposed approach is verified on several different person re-ID baselines. Extensive experiments show the validity of the proposed CST model and its benefits for re-ID performance on the Market-1501 dataset.

Keywords: person re-identification; camera style transfer; feature generation; generative adversarial network; data augmentation

1. Introduction

Person re-ID [1] aims to match images of the same person appearing in different camera views. Due to the camera style variance in the surveillance system, images of the same pedestrian often show significant differences in pose, illumination and background, which increase the intra-class variations in the model and harm its retrieval performance. Thus, camera style variance is a major difficulty in the person re-ID task.

With the rapid development of deep neural networks in the past decade, researchers have explored various deep learning-based methods for person re-ID [2]. Many researchers have tried to solve camera style variance as a part of the optimization problem in a common classification model. Some researchers focus on the innovation of deep learning loss functions. Wang et al. [3,4] study the extension of classification losses based on softmax, while Deng et al. [5] and Hermans et al. [6] investigate the usage of contrastive losses based on hard sample mining. Sun et al. [7] and Chen et al. [8] pay attention to the modification of the base model, which extracts pedestrians’ features. However, these methods sidestep the camera style variance that innately exists in re-ID datasets. Therefore, they cannot solve the camera style variance problem fundamentally.
Another prevalent choice is using generative models to make camera style-guided image-level data augmentation. Benefitting from the augmented dataset, the re-ID model can directly perceive these variations. With the rapid progress of the GAN theory [9–12], many GAN-based practices [5,13,14] have been proven as helpful alliances in re-ID tasks. However, these methods are coupled with off-the-backbone generative networks which cost tremendous time and memory in the training process. Taking CamStyle [14] as an example, the training process of the person re-ID model on the extended dataset only requires about one hour. However, it takes a couple of days to train the generative model, which is indicated as a bunch of CycleGANs [15] in this task. Furthermore, the introduction of a complex generation network splits the training process of the entire model, and it makes it hard to achieve an optimal for both models simultaneously.

In order to overcome camera style variance in the person re-ID problem, a feature-level Camera Style Transfer (CST) approach is proposed to serve as data augmentation with low time and memory cost. The authors propose utilizing an additional camera style feature extractor to learn camera-related information. The proposed CST approach then takes source identity features and target camera style features as input, and it generates features that transfer their camera information while maintaining their identity information. An Adaptive Batch Normalization (ABN) layer is designed to inject camera information into the original identity feature. Compared with image-level GAN-based methods, the proposed CST model can produce high-quality features precisely and speedily. Moreover, extensive experiments show that the proposed CST model can adapt to different baselines and metric learning methods, proving the universality of the proposed method. Finally, a lightweight framework is presented, where the generative model is embedded into a person re-ID model in an end-to-end manner. In summary, this paper makes the following main contributions.

- To overcome camera variations, a feature-level Camera Style Transfer (CST) model for person re-ID is proposed, acting as a data augmentation approach which benefits the training process of re-ID model.
- To achieve camera style transfer while preserving the identity information, an Adaptive Batch Normalization (ABN) layer is designed to inject camera information and produce high-quality features.
- To optimize the generative model and re-ID model simultaneously, a lightweight framework is proposed to train the generative model and re-ID model in an end-to-end manner.

The rest of this paper is organized as follows: Related work is discussed in Section 2. The proposed feature-level CST model is described in Section 3. Section 4 introduces the end-to-end training process when implementing the CST model in re-ID baselines. Experimental results are reported and analyzed in Section 5, which are followed by conclusions in Section 6.

2. Related Works

2.1. Generative Adversarial Network (GAN)

GAN [9] aims at learning a mapping function from latent space to a target distribution. The training objective function of GAN has been widely discussed and modified since its establishment. Mao et al. use LSGAN [10] to minimize Pearson divergence. Zhao et al. [11] treat the discriminator as an energy function that attributes lower energies to the regions near the data manifold and higher energies to other regions. Martin Arjovsky et al. propose Wasserstein GAN [16] to stabilize the training process of GAN. Lars Mescheder et al. [12] discuss regularization strategies to stabilize GAN’s training.

2.2. Image Generation

At the same time, remarkable progress has been made in image generation with the development of GAN theory. Alec et al. [17] firstly bridge the gap between convolutional network and GAN. Zhu et al. introduce CycleGAN [15] to translate images from two
different domains circularly. Choi et al. propose StarGAN [18] to perform image-to-image translation for multiple domains using only a single model. Karras et al. [19] describe an unique training strategy which can generate images from low resolution to high resolution gradually. Image generation can be used in image processing [20,21], and also can be used for data augmentation in computer vision tasks.

2.3. Camera Style Transfer

Another mentionable topic is the usage of adversarial training in camera style transfer. Sohn et al. [22] arrange a GAN-based network in handling unsupervised domain adaptation problem for face recognition. Yin et al. [23] propose a center-based feature transfer framework to enhance the face recognition model. Gao et al. [24] introduce a covariance preserving GAN into the feature space in low-shot learning research. These research studies on camera style transfer mainly focus on the face recognition task.

2.4. Person Re-Identification

A large family of approaches treat person re-ID as a metric learning problem. Many efforts focus on the improvement of loss functions and the basic structure of feature encoders. Alexander Hermans et al. [6] introduce triplet loss based on hard sample mining into the re-ID theory. Chen et al. [25] use a novel scheme to make instance-guided context rendering. Sun et al. [7] discuss the advantage of local features through a unique partition strategy. Chen et al. [8] propose the High-Order Attention (HOA) module to model and utilize the complex and high-order statistics information in an attention mechanism. These methods mainly regard person re-ID as a classification problem yet neglect camera style variance, which innately exists in the re-ID problem.

Due to the rapid growth of GAN theory in recent years, several researchers have exploited using GAN in image-level data augmentation for person re-ID. CycleGAN [15] is applied in some works [5,26] to make cross-domain image style conversion to solve domain adaption problems. StarGAN [18] is adopted to generate pedestrian images with different camera styles due to its flexibility. Liu et al. [27] manage to improve the quality of generated images by decomposing the process of transformation into a set of sub-transformation. Zheng et al. [13] design a generative network by exploiting and reorganizing the structure information and the appearance information in the dataset. Chen et al. [25] formulate a dual condition GAN which can enrich person images with contextual variations.

GAN-based methods can be helpful to alleviate camera style variance for the person re-ID task. However, due to the huge computing resources required by the training of GAN, image-level generative models are usually used as a pre-training or data augmentation process in re-ID. The training process of GAN is usually separated from the training of the re-ID model. In order to reduce the time and memory consumption of model training, the authors establish the generative model on a feature level and embed the generative model into the re-ID model in an end-to-end manner. Experimental results show that the proposed feature-level Camera Style Transfer (CST) approach can significantly improve the performance of multiple re-ID baselines with less time and memory cost.

3. Feature-Level Camera Style Transfer

In the person re-ID problem, camera style variance often leads to differences in viewpoint, illumination, and background, making it hard to determine whether two images captured by different cameras are of the same person or not. Some researchers apply pose-guided image generation as data augmentation. However, the image-level generative model takes a huge amount of time and memory. In addition, the training process of the image generation model is separated from the re-ID model, which makes it hard to achieve optimal performance for both models. To overcome these shortcomings, a GAN-based feature-level Camera Style Transfer (CST) approach is proposed to alleviate the camera style variance in the person re-ID task.
3.1. Problem Modeling

The person re-ID problem is related to determining whether the two input images belong to the same person. Traditional methods often consist of feature extraction and metric learning. Camera style variance can harm the representational ability of traditional identity features. This paper proposes extracting an additional camera style feature to provide camera-related information. Furthermore, a generative model is proposed to make camera style transfer to alleviate camera style variance for the re-ID problem.

Identity Feature and Camera Style Feature. Generally speaking, a deep learning-based person re-ID approach consists of two main parts: representative feature extraction and distance metric learning. In a typical approach, a feature extractor $E_I$ takes a pair of person images $(x_i, x_j)$ with labels $(y_i, y_j)$ as input and outputs the extracted identity features $(f_i^I, f_j^I)$. Then, the distance of the features is measured to determine the similarity of images $x_i$ and $x_j$. If the two images have the same label, i.e., $y_i = y_j$, the approach should determine that $x_i$ and $x_j$ are of the same person. However, if $x_i$ and $x_j$ were captured by different cameras and thus have different camera labels $c_i \neq c_j$. The camera style information contained in the extracted features makes it hard to make the right decision. An extra feature extractor $E_C$ is used to capture features related to camera style $(f_i^C, f_j^C)$. Then, the extracted identity feature $(f_i^I, f_j^I)$ and camera style feature $(f_i^C, f_j^C)$ are used in a feature-level generative model to complete the camera style transfer of a given feature. The necessity of the extra independent feature extractor $E_C$ is discussed in Section 5.6.1.

The Generative Model for Camera Style Transfer. The generative model aims to make camera style transfer at the feature level. Thus, it takes a pair of features $(f_i^I|y_i, f_j^C|c_j)$ as input where $f_i^I|y_i$ denotes the source identity feature with identity label $y_i$, which is extracted from source image $x_i$ by extractor $E_I$. Similarly, $f_j^C|c_j$ denotes the target camera style feature with camera label $c_j$, which is obtained from target image $x_j$ by an extra camera style feature extractor $E_C$. Then, the generator $G$ generates feature $f'^I$ conditioned on identity label $y_i$ from $f_j^I$ and camera label $c_j$ from $f_j^C$. The whole generation process of the generative model can be expressed as:

$$
\begin{align*}
    f_i^I|y_i &= E_I(x_i) \\
    f_j^C|c_j &= E_C(x_j) \\
    f'^I|y_i, c_j &= G(f_i^I, f_j^C)
\end{align*}
$$

The generated feature $f'^I|y_i, c_j$ has the same identity label $y_i$ as the source feature $f_i^I|y_i$ and the same camera label $c_j$ as the target feature $f_j^C|c_j$.

3.2. Adversarial Camera Style Feature-to-Feature Translation

The generative model aims to make camera style transfer while preserving input features’ identity information. Thus, the goal of the generative model can be divided into two parts: judging the authenticity of generated features and determining whether generated features complete the camera style transfer. The objective functions of the generative model are designed based on these two goals.

GAN Designing. To determine whether generated features are real or not, a discriminator $D$ is introduced into the generative model to calculate adversarial losses. Paired with the generator $G$, the whole generative model plays a min–max game to improve the authenticity of generated features:

$$
\min_{\theta_G} \max_{\theta_D} \mathcal{L}_{adv} = \mathbb{E}_{f_i^I \sim P_G} [\log(1 - D(G(f_i^I, f_j^C)|y_i, c_j))] + \mathbb{E}_{f_j^I \sim P_I} [\log D(f_j^I|y_i, c_i)]
$$
where $\theta_G, \theta_D$ represents the parameters of generator $G$ and discriminator $D$. $f_i^I$ is the real source feature with identity label $y_i$ and camera label $c_i$, while $f' = G(f_i^I, f_i^C)$ is the generated feature with the same identity label $y_i$ and different camera label $c_j$. $\mathcal{P}_G$ and $\mathcal{P}_I$ denote the distributions of generated features and real features, respectively. Furthermore, WGAN-GP [28] is selected to play the role of adversarial loss, since the authors found its relatively stable performance in the generation objectives. Therefore, a gradient penalty loss is added to confirm the Lipschitz constraint of the entire loss function:

$$L_{gp} = \mathbb{E}_{f \sim \mathcal{P}_I}[\|\nabla_f D(f)\|_2^2 - 1]$$

(3)

where $\mathcal{P}_I$ corresponds to the distribution that samples uniformly along the straight line between $\mathcal{P}_G$ and $\mathcal{P}_I$, $\| \cdot \|_2$ indicates L2-norm.

**Camera Classification Loss.** To ensure that the generator makes accurate camera style transfer toward target features, camera labels can be utilized as supervisory information in the GAN training. To be more specific, the real feature $f_i^I$ should preserve its camera label $c_i$ and the generated feature $f'$ has transferred camera label $c_j$. Following the practice of StarGAN [18], this objective can be decomposed into two parts. One is camera domain classification loss for real features $\mathcal{L}^r_{cam}$, and the other is camera domain classification loss for generated features $\mathcal{L}^f_{cam}$:

$$\mathcal{L}^r_{cam} = \mathbb{E}_{f_i^I \sim \mathcal{P}_I}[- \log D_{cam}(f_i^I|c_i)]$$

(4)

$$\mathcal{L}^f_{cam} = \mathbb{E}_{f' \sim \mathcal{P}_G}[- \log D_{cam}(G(f_i^I, f_i^C)|c_j)]$$

(5)

where $D_{cam}$ is the camera discriminator, and the definitions of the other symbols are the same as shown in Equation (2).

**Overall Objective Function.** The loss functions mentioned before are combined to formulate total loss function for $D$ and $G$, respectively:

$$L_D = -L_{adv} + \lambda_{cam}(\mathcal{L}^r_{cam} + \mathcal{L}^f_{cam}) + \lambda_{gp}L_{gp}$$

(6)

$$L_G = L_{adv} + \lambda_{cam}\mathcal{L}^f_{cam} + \lambda_{gp}L_{gp}$$

(7)

where $\lambda_{cam}$ and $\lambda_{gp}$ are hyper-parameters to control the contribution of different parts to the overall loss function. The impact of these hyper-parameters is studied in Section 5.3.

### 3.3. Adaptive Batch Normalization

In order to make camera style transfer while preserving identity information, this paper proposes a novel Adaptive Batch Normalization (ABN) layer. When trying to add camera style information into feature generation, the authors first tried to simply add a one-hot camera label to the original identity feature. However, they found that the one-hot camera label is merely a label, and it does not contain enough camera style information for the network to learn from. Then, the authors tried to use the concatenation of identity feature and camera feature, but the result is still not ideal. The authors believe the reason is that the network sees a feature as a whole, the identity part and the camera style part are not clearly separated, leading to semantic confusion. Inspired by Adaptive Instance Normalization [29], in which the style code can be injected through linear affine parts in image style transformation, the authors proposed an Adaptive Batch Normalization (ABN) layer that can effectively embed camera information into an identity feature and accomplish camera style transfer in a feature level. ABN layers are applied to replace the corresponding normalization layers in generator $G$ by a similar yet more concise way:

$$ABN(x, y) = m_1(y) \frac{x - \mu(x)}{\sigma(x)} + m_2(y),$$

(8)
where \( \mu(\cdot) \) and \( \sigma(\cdot) \) denote the mean calculation and the standard deviation calculation, \( x \) represents features that normally pass through the layer, and \( y \) represents style features that are mapped into the appropriate dimensions through mapping functions \( m_1(\cdot) \) and \( m_2(\cdot) \). Note that the normalization layers are placed into the feature level, so the related statistics are calculated only along the batch dimension. The ablation study shows the effectiveness of the proposed ABN layer, and the detailed experimental results are shown in Section 5.6.2.

### 3.4. Network Architecture

The main architectures of generator \( G \) and discriminator \( D \) are illustrated in Figure 1. Since the whole camera style transfer process took place at the feature level, a convolutional layer is not necessary, and only fully connected (FC) layers are used in the network. The dimensions of input identity feature and camera-style feature are 2048 and 512, respectively. For generator \( G \), a U-Net based architecture is adopted as the backbone. The identity branch consists of three parts: the downsampling part, the transition part and the upsampling part. The downsampling part acts as the role of an encoder, reducing the dimension of features, while the upsampling part acts as a decoder and has the opposite effect. Batch normalization layers are applied after each FC layer in these two parts. The transition part, namely a serial of ABN blocks, plays the key role of fusing two branches of features and makes a style transfer, where the proposed ABN layers are applied after each FC layer in this part. All normalization layers in \( G \) are followed with a Leaky ReLU layer (the negative slope is set to be 0.2, which is the same as below). The camera style branch is a Multi-Layer Perception (MLP) for mapping camera style information into ABN layers. For \( D \), a series of FC layers reduce the dimension of input features from 2048 to 1 with a reduction rate of 0.5 per layer. An additional classifier is connected to the 512-dimensional layer for camera classification loss. All FC layers in \( D \) are followed with a Leaky ReLU layer except the last one, while normalization layers are not used in \( D \), since they would do harm to Conditional GAN training. The image in the red dotted box in Figure 1 is not generated; it is placed here only to demonstrate that the generated feature \( f' \) maintains identity information while transferring camera style.

**Figure 1.** The structure of generator \( G \) and discriminator \( D \) of the proposed CST approach. \( E_i \) and \( E_C \) are the identity feature extractor and camera style feature extractor, respectively.

Through the proposed CST approach, given an input feature, multiple features that maintain the same identity but vary in camera style can be generated. These augmented features can be used to alleviate camera style variance during the training process of the re-ID model.
4. Using Transferred Feature for Person Re-ID

In this section, the proposed CST approach is implemented in different re-ID models in an end-to-end manner. Through a multi-stage joint training process, the augmented features produced by the CST approach can help to overcome camera style variance and improve the performances of baseline re-ID models.

In order to optimize the feature generation model with re-ID model simultaneously, the authors design an end-to-end framework which can easily embed the CST approach into any baseline re-ID model. The training process including three stages: pre-training stage, generative model training stage, and joint training stage.

4.1. Pre-Training Stage

The pre-training stage is used to train two feature extractors, namely identity feature extractor $E_I$ and camera style feature extractor $E_C$. The training process is illustrated in Figure 2. ResNet50 [30] and ResNet18 pre-trained on ImageNet are used as the backbone network for $E_I$ and $E_C$, respectively. For both extractors, the last fully-connected (FC) layer in ResNet is replaced by a batch normalization (BN) layer with corresponding dimensions.

![Figure 2](image.png)

Figure 2. In pre-training stage, feature extractors $E_I$ and $E_C$ are trained independently with identity classifier $C_I$ and camera style classifier $C_C$. Cross Entropy (CE) loss is chosen to be the loss function.

Two extractors $E_I$ and $E_C$ are trained independently as two classification tasks, while person identity classifier $C_I$ and camera style classifier $C_C$ are introduced to calculate identity loss $L_{id}$ and camera style loss $L_{cam}$ as follows:

$$L_{id} = \mathbb{E}_{x_i \sim P_I}[- \log C_I(E_I(x_i|y_i))]$$
$$L_{cam} = \mathbb{E}_{x_j \sim P_C}[- \log C_C(E_C(x_j|c_j))]$$

where $x_i$ is the input source image with identity label $y_i$, and $x_j$ is the input target image with camera label $c_j$. $f^I_i = E_I(x_i|y_i)$ is the extracted identity feature, and $f^C_j = E_C(x_j|c_j)$ is the extracted camera style feature, while $P_I$ and $P_C$ are the distributions of corresponding features.

After finishing the pre-training stage, extractors $E_I$ and $E_C$ are trained to extract high-quality person identity features $f^I$ and camera style features $f^C$; these features are used as input in the generative model training stage to train the GAN for camera style transfer.
4.2. Generative Model Training Stage

The generative model training stage is to train a GAN to accomplish camera style transfer. The network structure is introduced in Section 3.4. The training process is illustrated in Figure 3. The dotted boxes and lines in the figure mean that during the generative model training stage, all parameters in feature extractors $E_I$ and $E_C$ stop updating, and gradient back-propagation is stopped before it reaches input features $f^I$ and $f^C$. $E_I$ and $E_C$ only provide input data for GAN in this stage.

Figure 3. In generative model training stage, identity feature $f^I$ and camera style feature $f^C$ are input into GAN, the generated feature $f'$ maintains the same identity label as $f^I$ and the same camera style label as $f^C$.

The identity feature $f^I$ and camera style feature $f^C$ are input into GAN in pairs through a feature reassemble algorithm. To adequately train the CST model, each identity feature should be paired with camera style features of different camera labels. $f^I$ and $f^C$ share the same batch size $N$, and the number of different camera labels $M$ is much less than $N$. The input image batch can be properly arranged so that each $N$ of camera style feature $f^C$ in a same batch cover all $M$ different camera labels. Then, for each of the $N$ identity features $f^I_k$ in a batch, one camera style feature $f^C_j$ of each camera label is chosen to form a feature pair $(f^I_k | y_i = 1, f^C_j | c_j = 1)$. Thus, the batch size of feature pairs is $N \times M$.

The loss functions use in this stage are introduced in Section 3.2. Through this training stage, the generator $G$ is trained to produce feature $f' | y_i, c_j = G(f^I_i | y_i, f^C_j | c_j)$. The generated feature $f'$ maintains the same identity label $y_i$ as the input source feature $f^I_i$ while the camera label is transferred to $c_j$, which is the same as target feature $f^C_j$.

4.3. Joint Training Stage

The joint training stage jointly trains the baseline re-ID model and the generative model. The training process is shown in Figure 4. During this stage, the parameters of identity feature extractor $E_I$ are unfrozen, so that $E_I$ can be further trained with help of augmented data. Because camera classification is a relatively easy task, camera style feature extractor $E_C$ remains frozen in this stage. The generator $G$ is stable after the previous training stage, so discriminator $D$ is removed from this training stage, and $G$ continues to update with a small learning rate.
Figure 4. In the joint training stage, the re-ID baseline and the CST model are trained jointly. The generated features act as feature-level data augmentation to improve the performance of the re-ID model.

Identity classifier \( C_I \) is used for both real input features and generated features, cross-entropy losses \( L_{id}^r \) (for real features) and \( L_{id}^f \) (for generated features) are calculated as follows:

\[
L_{id}^r = \mathbb{E}_{f_i \sim P_i}[- \log C_I(E_I(x_i|y_i))] \tag{11}
\]

\[
L_{id}^f = \mathbb{E}_{f' \sim P_G}[- \log C_I(G(E_I(x_i|y_i), f'C_j))] \tag{12}
\]

where \( f_i = E_I(x_i|y_i) \) is the real identity feature, and \( f' = G(f_i, f'C_j) \) is the generated feature. \( P_i \) and \( P_G \) denote the distributions of real features and generated features, respectively.

For the IDE baseline, triplet loss is also used to enhance the performance of the re-ID model. The generative model can provide abundant positive and negative samples for hard sample mining. With the help of the generated feature, the triplet loss used in this stage can be written as:

\[
L_{trip}^{r+f} = \frac{1}{N} \sum_{i=1}^{N} \text{ReLU}([\min(d(f_i, f_p)) - \max(d(f_i, f_n)) + m]) \tag{13}
\]

where \( N \) is the batch size, positive sample \( f_p \) and negative sample \( f_n \) are both chosen from the union set of real features and generated features, and \( m \) is the margin for triplet loss.

The total loss function is as follows:

\[
L_{total} = L_{id}^r + \lambda_{gen} L_{id}^f + L_{trip}^{r+f} \tag{14}
\]

where \( \lambda_{gen} \) controls the contribution of generated features; the impact of this weight is studied in Section 5.3.

Through the above multi-stage training process, the proposed CST approach can help improve the performance of the baseline re-ID model by two means: First, the camera style transferred features can alleviate the difficulties brought by camera style variance. Second, the generated features also act as data augmentation, providing more training samples, which is especially helpful for hard sample mining.

5. Experiments
5.1. Baseline Methods

Three re-ID models are chosen as baselines to verify the effectiveness of the proposed CST approach. **ID-discriminative Embedding (IDE)** [1] is a famous person re-ID baseline proposed by Zheng et al. in 2016; it has been widely studied and used in the re-ID task due to its simplicity and effectiveness. The IDE baseline used in this paper is a slightly improved version (using a batch normalization layer instead of the original fully connected block); it outperforms the original IDE baseline, but it is still referred to as
IDE in this paper for simplicity. **Part-based Convolutional Baseline (PCB)** [7] is a re-ID baseline proposed by Sun et al. in 2018; the authors proposed dividing the input image horizontally and output part-level features. **Multiple Granularity Network (MGN)** [31] proposed combining global and partial features obtained by a multi-branch deep network. These three baselines represent models based on global features, partial features, and the combination of global and partial features; thus, they are chosen to evaluate the effectiveness of the proposed CST approach.

### 5.2. Implementation Details

The proposed method is implemented on the PyTorch Framework. All experiments are held on two NVIDIA 2080 Ti GPUs. For the IDE baseline, input images are resized to $256 \times 128 \times 3$. Random horizontal flip and random erasing (with probabilities 0.5) are used for pre-processing, and the batch size is set to be 64. The network is optimized within 450 epochs (1–60 for pre-training, 61–360 for generative model training and 361–450 for joint training). Adam optimizer is adopted for all networks. For $G$ and $D$, the learning rate starts with 0.001 and is divided by 10 every 100 epochs. For $E_I$ and $E_C$, the initial learning rate is set to be 0.00035, and it decays at the 360th epoch and 400th epoch with the same decay rate, 0.1. For the PCB baseline, input images are resized to $384 \times 128 \times 3$ with horizontal flip, and the batch size is 256. SGD optimizer with a learning rate of 0.1, and a weight decay of $5 \times 10^{-4}$ is used for $E_I$ in the pre-training stage. For the joint training stage, Adam optimizers with a learning rate of $1 \times 10^{-3}$ are used for $G$ and $D$. The whole network is trained for 150 epochs (1–20 for pre-training, 21–100 for generative model training and 101–150 for joint training). The learning rates of $G$ and $D$ decay by 0.1 every 30 epochs. For MGN baseline, input images are resized to $384 \times 128 \times 3$ with random horizontal flip, and the batch size is 256. The SGD optimizer with a learning rate of 0.1 and weight decay of $5 \times 10^{-4}$ is used for feature extractors in the pre-training stage. For the joint training stage, Adam optimizers with a learning rate of $1 \times 10^{-3}$ are applied for both $G$ and $D$. The network is trained for 520 epochs (1–320 for pre-training, 321–440 for generative model training and 441–520 for joint training). The learning rates of $G$ and $D$ decay by 0.1 every 60 epochs.

The performance of the proposed method is evaluated on the Market-1501 [32] dataset. **Market-1501** contains 32,668 labeled images with 1501 identities from six camera views. The training set has 12,936 images with 751 identities, from which 1182 images of 80 identities are chosen as the validation set. Images in the validation set are not used in the training stage, and they are used for parameter analysis. In the testing stage, 3368 images from the remaining 750 identities are chosen as query images to retrieve the matching persons in a gallery set of 15,913 images.

### 5.3. Impact of Hyper-Parameters

The sensitivities of hyper-parameters in loss functions are analyzed in this section: i.e., the weight of camera classification loss $\lambda_{\text{cam}}$ and gradient penalty $\lambda_{\text{gp}}$, which control the relative importance of different losses when generating camera style transferred features (Section 3.2), and the weight of generated feature loss $\lambda_{\text{gen}}$, which controls the contribution of generated features when jointly training (Section 4.3). The experiments are carried out on the validation set, and the results are shown in Figure 5–7.
Figure 5. Hyper-parameter analysis of the weight of the camera classification loss $\lambda_{\text{cam}}$ on Market-1501. Best performance achieved when $\lambda_{\text{cam}} = 0.3$.

When $\lambda_{\text{cam}} = 0.3$, the proposed model reaches its best results. It is worth mentioning that the camera style information is not identity-related, so the weight of camera classification loss $\lambda_{\text{cam}}$ should not be too large. Otherwise, the generator in the generative model tends to learn overcharged camera information, which does not exist in the input identity feature and corrupts the training process.

$\lambda_{\text{gp}}$ is introduced into the network to maintain 1-Lipschitz continuity. The proposed model performs best when $\lambda_{\text{gp}} = 1.0$.

The best performance can be obtained when $\lambda_{\text{gen}} = 0.5$. Assigning an excessive value for $\lambda_{\text{gen}}$ reduces the result. This suggests that the proposed generative model can act as an effective data augmentation scheme in the appropriate range.

Figure 6. Hyper-parameter analysis of gradient penalty $\lambda_{\text{gp}}$ on Market-1501. Best performance achieved when $\lambda_{\text{gp}} = 1.0$. 

Experimental results are shown in Table 1. Despite the high baseline adopted in this experiment, the proposed methods still achieve remarkable improvement over the previous works. Compared to the IDE baseline with Cross Entropy (CE) loss (IDC-CE), by introducing CST, the mAP and Rank-1 improve by +3.9% and +2.2%. When compared to IDE with triplet loss (IDE-Trip) and IDE with both losses (IDE-CE-Trip), a similar increase can be observed. The authors further illustrate the performance of the proposed method adapted to the PCB baseline; the performance of the proposed CST surpasses the baseline by +1.6% in mAP and +0.9% in Rank-1. For the MGN baseline, the CST approach also outperforms the baselines in both mAP and Rank-1.

Some state-of-the-art methods are also listed in Table 2 for comparison on the Market-1501 dataset. The first group includes methods that did not use generative models as data augmentation, and the second group includes GAN-based methods that use generated or style transferred images. As we can see, the proposed CST applied to any baseline method outperforms most of the other SOTA methods. These quantitative results show that the proposed CST approach is suitable for both global and partial features; thus, it can be applied in multiple person re-ID baselines, and it outperforms state-of-the-art methods.

**Table 1.** The performance comparison of baseline methods and the proposed proposed CST on the Market1501 dataset. Results of the proposed methods are marked as +CST, and better results are marked in bold. * The results of MGN are reported by our reproducing.

| Market-1501 | mAP  | Rank-1 | Rank-5 | Rank-10 |
|-------------|------|--------|--------|---------|
| IDE-CE [1]  | 75.0 | 89.1   | 96.4   | 97.9    |
| IDE-CE + CST| 78.9 | 91.3   | 97.4   | 98.5    |
| IDE-Trip    | 74.9 | 88.4   | 95.4   | 96.8    |
| IDE-Trip + CST| 76.7 | 89.0   | 95.6   | 97.0    |
| IDE-CE-Trip | 79.8 | 91.6   | 97.0   | 98.1    |
| IDE-CE-Trip+CST | 82.3 | 92.8   | 97.6   | 98.7    |
| PCB [7]     | 77.4 | 92.3   | 97.2   | 98.2    |
| PCB + CST   | 79.0 | 93.2   | 97.8   | 98.8    |
| MGN * [31]  | 86.3 | 93.9   | 97.3   | 98.5    |
| MGN + CST   | 87.9 | 95.0   | 98.1   | 99.0    |
Table 2. The performance comparison between the proposed methods and other state-of-the-art methods on the Market1501 dataset. Group 1: the methods not using generated data. Group 2: GAN-based methods. Results of the proposed methods are in the last group with +CST mark, and better results are marked in bold.

| Market-1501          | mAP | Rank-1 | Rank-5 | Rank-10 |
|----------------------|-----|--------|--------|---------|
| LGMANet [33]         | 82.7| 94.0   | -      | -       |
| MTNet [34]           | 81.5| 93.9   | -      | -       |
| RIN [35]             | 67.6| 86.1   | -      | -       |
| RNLSTM [36]          | 76.9| 90.6   | -      | -       |
| Gconv [37]           | 72.3| 88.1   | 95.1   | 96.8    |
| Top-DB-Net [38]      | 85.8| 94.9   | -      | -       |
| DTr [39]             | 76.7| 91.2   | -      | -       |
| MSFANet [40]         | 82.3| 92.9   | -      | -       |
| DeformGAN [41]       | 61.3| 80.6   | -      | -       |
| LRSO [42]            | 66.1| 84.0   | -      | -       |
| CamStyle [14]        | 71.6| 89.5   | -      | -       |
| FD-GAN [43]          | 77.7| 90.5   | -      | -       |
| DRANet [44]          | 79.7| 92.9   | 97.2   | 98.3    |
| MANet [45]           | 78.8| 93.8   | -      | -       |
| IDE-CE-Trip + CST    | 82.3| 92.8   | 97.6   | 98.7    |
| PCB + CST            | 79.0| 93.2   | 97.8   | 98.8    |
| MGN + CST            | 87.9| 95.0   | 98.1   | 99.0    |

Furthermore, some qualitative comparisons are presented to vividly show the improvement brought by the proposed CST approach. Figure 8 shows the retrieved results of two example query images in the Market-1501 dataset. As we can see in the figure, the baseline model tends to be affected by camera style-related factors, e.g., posture, background, illumination, etc. The proposed CST can generate a large amount of camera style transferred features to alleviate this problem so that the model can overcome the effect of camera style variance and perform more accurate matching.

Figure 8. Qualitative comparison on Market-1501 dataset. The left-most column are query images, the rest are a matching list ranked by similarity. Images enclosed by green boxes represent the correct matches and those in red indicate the wrong matches. Best viewed in color.
5.5. Computational Costs Analysis

The proposed feature-level CST is compared with some image-level generation model on the Market-1501 dataset. The time and memory costs as well as re-ID results are listed in Table 3. The results show that feature-level CST costs much less time and less memory compared with image-level generation models while achieving better re-ID performance.

| Method               | mAP | Rank-1 | Time Cost (h) | Memory Cost (MB) |
|----------------------|-----|--------|---------------|------------------|
| CamStyle [14]        | 71.6| 89.5   | >24           | >11,000          |
| FD-GAN [43]          | 77.7| 90.5   | >24           | >10,000          |
| IDE-CE + CST         | 78.9| 91.3   | 1.5           | 6600             |

5.6. Ablation Study

5.6.1. Necessity of Independent Camera Feature Extractor $E_C$

When trying to extract camera style features, the authors first tried to add a new branch at the identity feature extractor. However, the re-ID accuracy drops dramatically. That is because person identity classification and camera style classification are two different tasks, and they cannot be fulfilled in the same network. The gradient brought by the camera style classifier affects the identity classification. Then, the authors tried to stop the back propagation of the camera style classifier to eliminate its influence to identity classification, but the accuracy of camera style classification decreased. Because the original feature extractor is trained to extract person identity features, camera style information might be lost during training. Since the camera style feature extractor is the input of the generative model, the quality of the camera style feature is crucial for the CST model. The authors decide to add a light-weighted independent feature extractor $E_C$ to the framework. The comparison with the former setting shows that by introducing an independent feature extractor $E_C$, the camera classification accuracy can reach almost 100%, and the quality of the identity feature is not influenced.

This study is carried on the IDE baseline; the camera style classification accuracy and re-ID results are listed in Table 4, where ‘branch’ means adding a branch to the identity classifier, ‘stop bp’ means stopping the back-propagation of the camera classifier, and ‘independent’ means using an extra independent classifier for the camera style.

| Method   | Acc | mAP | Rank-1 |
|----------|-----|-----|--------|
| branch   | 99.5| 58.2| 66.4   |
| stop bp  | 95.8| 70.2| 83.4   |
| independent | 99.8| 76.4| 90.8   |

5.6.2. Effectiveness of ABN Layer

To verify the effectiveness of the proposed ABN layer, mAP and Rank-k results are compared between different camera style transfer strategies, and the results are shown in Table 5, where ‘none’ means not using camera style information, ‘one-hot’ represents a one-hot camera label, ‘concat’ means the concatenation of identity features and camera features, and ABN is the proposed Adaptive Batch Normalization layer.

The study is carried on the IDE baseline. We can see that the proposed ABN layer performs best among these camera style transfer strategies, showing that the ABN layer can successfully inject camera style information into the identity feature.
5.6.3. CST Applied on Different Branches of MGN

To further verify the generalization capability of the proposed CST approach, CST is applied for each network branch of the original MGN baseline; the results are shown in Table 6.

From the results, we can see that the CST approach improves all branches of the MGN re-ID baseline, proving that the CST approach is suitable for both global features and partial features.

Table 5. The performance comparison between different camera style transfer strategies. Better results are marked in bold.

| Dataset   | Method  | mAP  | Rank-1 | Rank-5 | Rank-10 |
|-----------|---------|------|--------|--------|---------|
| Market-1501 | none    | 75.0 | 89.1   | 96.4   | 97.9    |
|           | one-hot | 76.1 | 90.6   | 96.5   | 98.0    |
|           | concat  | 76.0 | 90.6   | 96.4   | 97.7    |
|           | ABN     | 78.9 | 91.3   | 97.4   | 98.5    |

Table 6. The performance comparison when we applied the proposed CST on different branches of MGN. Results of the proposed methods are marked as +CST, and better results are marked in bold. * The results of MGN are reported by our reproducing.

| Dataset   | Method         | mAP  | Rank-1 |
|-----------|----------------|------|--------|
| Market-1501 | MGN * [31]   | 86.3 | 93.9   |
|           | MGN + CST     | 87.9 | 95.0   |
|           | MGN-global    | 83.5 | 92.7   |
|           | MGN-global + CST | 84.3 | 93.2   |
|           | MGN-part2     | 75.2 | 90.3   |
|           | MGN-part2 + CST | 76.3 | 91.0   |
|           | MGN-part3     | 77.2 | 91.2   |
|           | MGN-part3 + CST | 77.3 | 91.8   |

5.6.4. Comparisons on RAP Dataset

The RAP dataset [46] is a newly published large-scale dataset which contains 84,928 images of 2589 identities. It is designed for both pedestrian attribute recognition and person re-ID tasks. The proposed CST approach is further evaluated on the RAP dataset. Following the dataset split protocol suggested by the authors of the RAP dataset [46], and the training settings suggested in Yaghoubi et al. [47], the experiments are carried out using the MGN baseline [31]. The comparison results are listed in Table 7. It is shown that the CST approach improves the performance of the MGN baseline and outperforms the SOTA methods. The significant improvement in mAP demonstrates that the proposed CST approach can alleviate camera style variance in general.

Table 7. The performance comparison between the proposed CST with the MGN baseline and state-of-the-art approaches on the RAP dataset. Results of the proposed method are marked as +CST, and better results are marked in bold.

| RAP                      | mAP  | Rank-1 | Rank-5 | Rank-10 |
|--------------------------|------|--------|--------|---------|
| Yaghoubi et al. (Upper Body) [47] | 42.4 | 64.8   | 81.4   | 85.9    |
| Yaghoubi et al. (Full Body) [47] | 46.3 | 69.0   | 83.6   | 88.1    |
| MGN [31]                 | 48.1 | 71.2   | 85.5   | 90.5    |
| MGN + CST                | 56.2 | 75.3   | 88.1   | 92.2    |
6. Conclusions

In this paper, a feature-level Camera Style Transfer (CST) approach is proposed to alleviate the camera style variance in the person re-ID problem. In addition to the traditional re-ID feature extractor, another feature extractor is used to extract the camera style feature independently. Then, CST takes a source identity feature and a target camera style feature as input, and it generates an identity-preserving feature by a GAN to achieve the camera style transfer. An Adaptive Batch Normalization (ABN) layer is designed to inject camera style information in the feature level. The authors further propose an end-to-end framework to jointly train the generative model with the re-ID model for better performance. Experimental results show that the proposed CST approach can be embedded with multiple re-ID baselines and outperform state-of-the-art methods, proving the effectiveness and generalization capability of the proposed approach.

Author Contributions: Conceptualization, Y.L. and H.S.; methodology, Y.L. and S.W.; validation, Y.L., S.W. and Y.W.; data curation, Y.L. and Y.W.; writing—original draft preparation, Y.L.; writing—review and editing, H.S., Y.W. and S.W.; supervision, Z.X.; project administration, H.S. and Z.X.; funding acquisition, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study is partially supported by the National Key R & D Program of China (No. 2018YFB2100800), the National Natural Science Foundation of China (No. 61872025), and the Science and Technology Development Fund, Macau SAR (File No. 0001/2018/AFJ) and the Open Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2021ZX-03).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thank you for the support from the HAWKEYE Group.

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

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. arXiv 2016, arXiv:1610.02984.
2. Wu, D.; Zheng, S.J.; Zhang, X.P.; Yuan, C.A.; Cheng, F.; Zhao, Y.; Lin, Y.J.; Zhao, Z.Q.; Jiang, Y.L.; Huang, D.S. Deep learning-based methods for person re-identification: A comprehensive review. Neurocomputing 2019, 337, 354–371. [CrossRef]
3. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive margin softmax for face verification. IEEE Signal Process. Lett. 2018, 25, 926–930. [CrossRef]
4. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
5. 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.
6. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. arXiv 2017, arXiv:1703.07737.
7. 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.
8. Chen, B.; Deng, W.; Hu, J. Mixed high-order attention network for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 371–381.
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 2014, 27, 2672–2680.
10. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
11. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial network. arXiv 2016, arXiv:1609.03126.
12. Mescheder, L.; Geiger, A.; Nowozin, S. Which training methods for GANs do actually converge? arXiv 2018, arXiv:1801.04406.
13. Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; Kautz, J. Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2138–2147.
14. 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.

15. 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.

16. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. ACM Trans. Multimed. Comput. Commun. Appl. 2017, 14, 1–20. [CrossRef]

17. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv 2015, arXiv:1511.06434.

18. 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.

19. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. arXiv 2017, arXiv:1710.10196.

20. Abdulrahman, A.A.; Rasheed, M.; Shihab, S. The Analytic of image processing smoothing spaces using wavelet. In Proceedings of the Journal of Physics: Conference Series, Coimbatore, India, 25–26 March 2021; Volume 1879, p. 022118.

21. Rasheed, M.; Ali, A.H.; Alabdali, O.; Shihab, S.; Rashid, A.; Rashid, T.; Hamad, S.H.A. The Effectiveness of the Finite Differences Method on Physical and Medical Images Based on a Heat Diffusion Equation. In Proceedings of the Journal of Physics: Conference Series, Coimbatore, India, 25–26 March 2021; Volume 1999, p. 012080.

22. Sohn, K.; Liu, S.; Zhong, G.; Yu, X.; Yang, M.H.; Chandraker, M. Unsupervised domain adaptation for face recognition in unlabeled videos. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3210–3218.

23. Yin, X.; Yu, X.; Sohn, K.; Liu, X.; Chandraker, M. Feature transfer learning for deep face recognition with under-represented data. arXiv 2018, arXiv:1803.09014.

24. Gao, H.; Shou, Z.; Zareian, A.; Zhang, H.; Chang, S.F. Low-shot learning via covariance-preserving adversarial augmentation networks. Adv. Neural Inf. Process. Syst. 2018, 31, 975–985.

25. Chen, Y.; Zhu, X.; Gong, S. Instance-guided context rendering for cross-domain person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October 2019–2 November 2019; pp. 232–242.

26. 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.

27. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7202–7211.

28. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein gans. arXiv 2017, arXiv:1704.00028.

29. Huang, X.; Belongie, S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

31. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 274–282.

32. 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.

33. Sun, L.; Liu, J.; Zhu, Y.; Jiang, Z. Local to Global with Multi-Scale Attention Network for Person Re-Identification. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2254–2258. [CrossRef]

34. Ling, H.; Wang, Z.; Li, P.; Shi, Y.; Chen, J.; Zou, F. Improving person re-identification by multi-task learning. Neurocomputing 2019, 347, 109–118. [CrossRef]

35. Xu, D.; Chen, J.; Liang, C.; Wang, Z.; Hu, R. Cross-view Identical Part Area Alignment for Person Re-identification. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2462–2466. [CrossRef]

36. Yang, W.; Yan, Y.; Chen, S. Adaptive deep metric embeddings for person re-identification under occlusions. Neurocomputing 2019, 340, 125–132. [CrossRef]

37. Yuan, Y.; Zhang, J.; Wang, Q. Deep Gabor convolution network for person re-identification. Neurocomputing 2020, 378, 387–398. [CrossRef]

38. Quispe, R.; Pedrini, H. Top-DB-Net: Top DropBlock for Activation Enhancement in Person Re-Identification. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2980–2987. [CrossRef]

39. Tahir, M.; Anwar, S. Transformers in Pedestrian Image Retrieval and Person Re-Identification in a Multi-Camera Surveillance System. Appl. Sci. 2021, 11, 9197. [CrossRef]
40. Huang, W.; Li, Y.; Zhang, K.; Hou, X.; Xu, J.; Su, R.; Xu, H. An Efficient Multi-Scale Focusing Attention Network for Person Re-Identification. *Appl. Sci.* **2021**, *11*, 2010. [CrossRef]

41. Siarohin, A.; Sangineto, E.; Lathuilière, S.; Sebe, N. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3408–3416.

42. 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, 23–29 October 2017; pp. 3754–3762.

43. Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; Li, H. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Adv. Neural Inf. Process. Syst.* **2019**, *31*, 1222–1233.

44. Liu, S.; Qi, L.; Zhang, Y.; Shi, W. Dual Reverse Attention Networks for Person Re-Identification. In *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 22–25 September 2019; pp. 1232–1236. [CrossRef]

45. Xiong, M.; Gao, Z.; Hu, R.; Chen, J.; He, R.; Cai, H.; Peng, T. A Lightweight Efficient Person Re-Identification Method Based on Multi-Attribute Feature Generation. *Appl. Sci.* **2022**, *12*, 4921. [CrossRef]

46. Li, D.; Zhang, Z.; Chen, X.; Huang, K. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Trans. Image Process.* **2018**, *28*, 1575–1590. [CrossRef] [PubMed]

47. Yaghoubi, E.; Borza, D.; Kumar, S.A.; Proença, H. Person re-identification: Implicitly defining the receptive fields of deep learning classification frameworks. *Pattern Recognit. Lett.* **2021**, *145*, 23–29. [CrossRef]