# Stargan-based camera style transfer for person retrieval

Yuanyuan Wang*

College of Computer and Software Engineering,
Huaiyin Institute of Technology,
Huaian, China
Email: zhfwyy@hyit.edu.cn
and
College of Computer and Information,
Hohai University,
Nanjing, China
*Corresponding author

Zhijian Wang

College of Computer and Information,
Hohai University,
Nanjing, China
Email: zhjwang@hhu.edu.cn

Mingxin Jiang

College of Electronic Information Engineering,
Huaiyin Institute of Technology,
Huaian, China
Email: jiangmingxin@126.com

**Abstract:** Person retrieval is also known as person re-identification (ReID) aiming to match person among cross cameras. Although the results of the person ReID have performed well in small datasets, the issues of the large number of identities in real scenarios or with more cameras have not been fully investigated. Being an image retrieval task under cross multi-cameras of intelligent video security, person ReID is influenced by the image style change caused by different camera illumination and view angles. The number of cameras in the latest datasets is increasing and more camera transfer models need to be trained. Traditional methods of generative adversarial network (GAN) can only handle transfer of two domains. To facilitate the research towards solving these problems, we use star generative adversarial networks (StarGAN) to transfer the image from one camera to another camera in the latest large benchmark datasets. We train multiple transfer models simultaneously, minimising the bias among different cameras. Label smooth regularisation (LSR) algorithm is utilised to mitigate the effects of noise in the model. We learn part-based descriptors from pedestrian samples to generate robust feature representation. Our work is competitive compared to the state-of-the-art.

**Keywords:** StarGAN; person retrieval; LSR.

**Reference** to this paper should be made as follows: Wang, Y., Wang, Z. and Jiang, M. (2020) 'Stargan-based camera style transfer for person retrieval', *Int. J. Information and Communication Technology*, Vol. 16, No. 1, pp.1–16.

**Biographical notes:** Yuanyuan Wang is currently pursuing her PhD in Computer Science and Technology from College of Computer and Information, Hohai University of China. She received her MS in Computer Technology from Nanjing University of Science and Technology in 2010. She is currently a Lecturer with the College of Computer and Software Engineering of Huaiyin Institute of Technology of China. Her current research focuses on computer vision and person re-identification.

Zhijian Wang received his MS and PhD in Computer Science from Nanjing University, China. He is currently a Professor at the College of Computer and Information, Hohai University, China. His research interests include machine learning and computer application.

Mingxin Jiang received her PhD in Signal and Information Processing from Dalian University of Technology, China, in 2013. She was a post-doctoral researcher with the Department of Electrical Engineering in Dalian University of Technology from 2013 to 2015. She is currently an Associate Professor in College of Electronic Information Engineering at Huaiyin Institute of Technology. Her research interests include multi-object tracking and vision sensors for robotics.

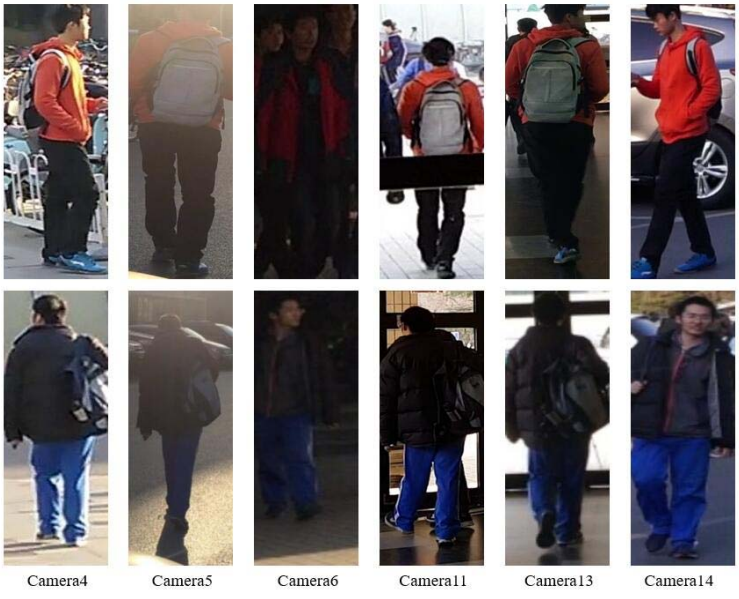
## 1 Introduction

Person re-identification (ReID) is widely considered to be a sub-question for image retrieval (Li et al., 2018; Wang et al., 2018). Recently, person ReID has received a great attention for advantages in the field of intelligent transportation and video surveillance. Retrieving the same pedestrians across the camera is one of the most crucial issues in Person ReID. Person ReID spots the appearance from another camera view for a query person image. In real scenes and large pedestrian datasets, cameras differ from each other regarding resolution, indoor and outdoor environment changes, different periods of light, viewing angles, weather conditions, etc., resulting in severe lighting changes. In the person ReID task, person images often undergo major changes in appearance as shown in Figure 1.

To address the variations of cameras, some literature [e.g., person transfer generative adversarial network (PTGAN) (Wei et al., 2018), CamStyle (Zhong et al., 2018), SPGAN (Deng et al., 2018)] selects a strategy using CycleGAN (Zhu et al., 2017b) with deep representation learning methods. To learn robust features of camera variations, these methods generate style-transferred samples in the style of other cameras using CycleGAN. SPGAN (Deng et al., 2018) improved CycleGAN and converted the source domain image into the target domain style. CycleGAN is used to generate new training samples, and the styles between different cameras are treated as different domains.

Existing methods [e.g., CycleGAN (Zhu et al., 2017b), DiscoGAN (Kim et al., 2017), DualGAN (Yi et al., 2017)] have shown success in handling image translation under two domains. However, the above methods are inefficient and time consuming when each training in dealing with more domains. To handle more than two domain transformations, existing methods generally build separate cross-domain models for each pair of image domains. A domain conversion is trained separately, and data in other fields cannot be fully utilised to increase generalisation ability. This training method which can only train one-to-one image translation models has limited effect. $N * (n - 1)$ generators should be trained to learn mappings among n domains. Meanwhile, each generator learns from two domains out of n. The dataset cannot be fully exploited. This method can only learn from two domains, which further limit the quality of the generated images. In addition, existing methods cannot jointly train domains from different datasets. The latest large-scale person ReID datasets and real surveillance scene involve larger numbers of cameras and identities. The newest person ReID datasets MSMT17 (Wei et al., 2018) contains 15 cameras and 4,101 identities. The model of StarGAN (Choi et al., 2018) is proposed to train multiple domains within a single network simultaneously. StarGAN learns the mappings among multiple domains using just a generator and a discriminator. To generate multiple transfer models simultaneously, we employ StarGAN to learn the style mapping functions among many different cameras. The training set is a combination of the original training images and the style-transferred images. In the actual scene, there are few pedestrians with the same id under different cameras which occurs over-fitting. The camera style transfer method is beneficial in reducing over-fitting risk.

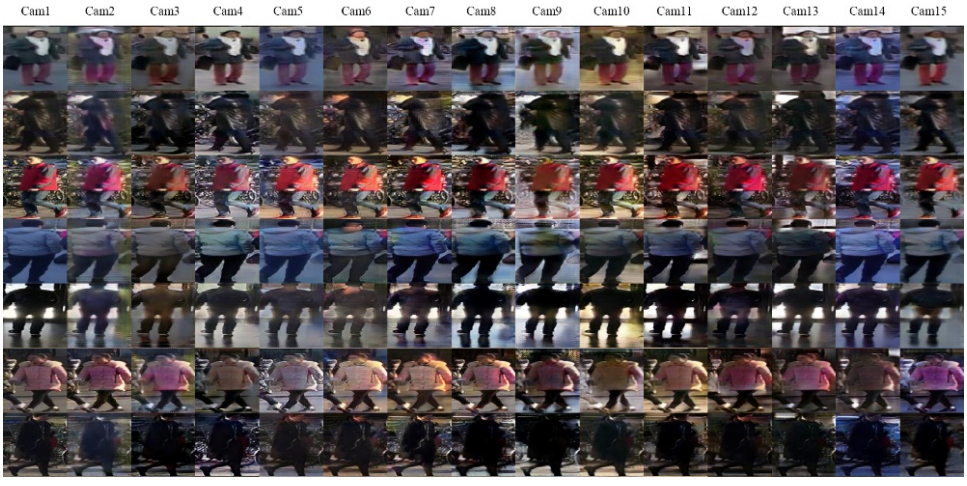
**Figure 1** Example images under different six cameras from MSMT17 (see online version for colours)




Note: Images in the same row represent the same person.

*Source:* Wei et al. (2018)

**Figure 2** Examples of camera style transfer on MSMT17 dataset (see online version for colours)




*Source:* Wei et al. (2018)

Figure 2 illustrates examples of camera style transfer using StarGAN on MSMT17 dataset (Wei et al., 2018). Style transfer and cross domain image generation can also be considered as image-to-image translation (Zhong et al., 2018). We only use one model to apply an image to an image transformation under multiple cameras. We learn invariant features across different cameras and add more generated person samples to the training set. Utilising StarGAN to generate more samples without manual annotating, the cost is lower and more efficient. Therefore, this paper mainly explores the use of StarGAN to resolve the domain gaps among different cameras. To mitigate the noise of the style transfer images, we use the label smoothing regularisation (LSR) (Szegedy et al., 2016) to re-weight the generated style transferred samples. Recently, the LSR method has been re-proposed, which reduces the over-fitting problem of the model by adding noise to the output. In the training phase, employing human part model (Sun et al., 2017b), real images and generated fake images are both fed into base network. Then, part-based features are extracted in proportion. In this work, we employ the LSR to help the chief cross-entropy loss deal with training data.

Image captured in camera 1 is translated to styles in other fourteen cameras using StarGAN (Choi et al., 2018). The images in the first column are the images captured from camera1. From the image of camera 1 to the style transfer of camera 2 – camera 15 are shown in the second column to the fifteenth column. Yet, the noise of the style transfer images needs to be solved.

The main contributions of our work are the following list.

1 Firstly, we propose a framework to learn camera style transfer model using StarGAN. It is regarded as an efficient method of data augmentation in multiple cameras or real-world scenarios.

2 Secondly, during the training phase, local body-part feature fusion descriptors are learned for robust feature representation. We employ the different LSR on the style-transferred samples to alleviate the impact of noise. It can be regarded as a data augmentation scheme.

3 Last but not the least, the proposed model is compared with the latest methods on benchmark datasets and has achieved competitive outcomes.

## 2 Related work

Deep learning-based descriptors has become a dominant method in the field of person ReID (Zheng et al., 2016, 2017a; Cheng et al., 2016; Varior et al., 2016; Xiao et al., 2016). Deep metric learning is also a widely method for person ReID which computes descriptors distance among pedestrian images. Specifically, similarity of different images of the same pedestrian needs to be higher than that of different pedestrians. Various metric learning loss methods [e.g., triplet loss (Liu et al., 2017), quadruplet loss (Chen et al., 2017a), margin sample mining loss (Xiao et al., 2017), triplet hard loss (Hermans et al., 2017) and angular loss (Wang et al., 2017)] are proposed to measure the similarity.

The variant of the above metric learning ignores the local information of body. Human parts methods have achieved much better performance than traditional algorithms (Sun et al., 2017b; Cheng et al., 2016; Wei et al., 2017; Li et al., 2017a; Zhang et al., 2017; Zhao et al., 2017b). Part-based convolutional baseline (PCB) (Sun et al., 2017b) are proposed to learn part-based samples descriptor. The method of PCB has basically achieved the performance of human-level in common pedestrian datasets. In this work, we employ PCB (Sun et al., 2017b) to train the ReID CNN. Human pose and posture detection have achieved improvement (Zhao et al., 2017a; Su et al., 2017; Qian et al., 2017). Yet, little literature focuses on the impact of different cross cameras on the person ReID results.

In the training phase, sufficient training data and annotations cannot be ignored for well performance. However, the number of the same training person images in existing person ReID datasets is limited. Therefore, more training data is helpful for improving accuracy of person ReID. In addition, manually annotations for multiple camera views in dataset is costly. Since generative adversarial networks (GAN) (Goodfellow et al., 2014) have already been adopted and have been approved in many researches recently. They have achieved impressive success in image generation.

Targeting to solve the above issue, some variants of GAN (Wei et al., 2018; Zhong et al., 2018; Zheng et al., 2017b; Zhu et al., 2017a; Ding et al., 2018; Huang et al., 2018) have achieved good performance in data augmentation and regularisation for person ReID. Zheng et al. (2017b) generated new unlabeled GAN images. However, the quality of the image generated by this method is poor. This is the earliest work of using GAN to generate training data in person ReID. They combined the generated unlabelled fake samples with the actual person images for data augmentation and regularise the network. Zhu et al. (2017a), Ding et al. (2018) and Huang et al. (2018) proposed pseudo labelling for semi-supervised to enrich the insufficient training data. Wei et al. (2018) proposed person transfer generative adversarial network (PTGAN)-based CycleGAN to bridge the domain gap among different datasets. Zhong et al. (2018) proposed camera style (CamStyle) adaptation employing CycleGAN to solve image style variations caused by different cameras.

However, image styles differ under cross cameras. The number of cameras in current new large pedestrian datasets and real scenarios is increasing. These issues make it difficult to apply multiple style transfer with a mapping function. To address these problems, considering the style differences among crossing cameras simultaneously to get robust mapping functions. In the current literature for solving the problem of pedestrian ReID, CycleGAN is generally used to independently construct multi models for each pair of images field under the camera. Existing models are ineffective in multi-domain style transfer tasks. More recently, Choi et al. (2018) proposed StarGAN which allows multiple different domains simultaneously be trained using a single model. Inspired by the existing excellent job, we apply StarGAN to learn mappings simultaneously among multiple domains using one network. To the best of our knowledge, this is the first work that StarGAN has been applied to person ReID.

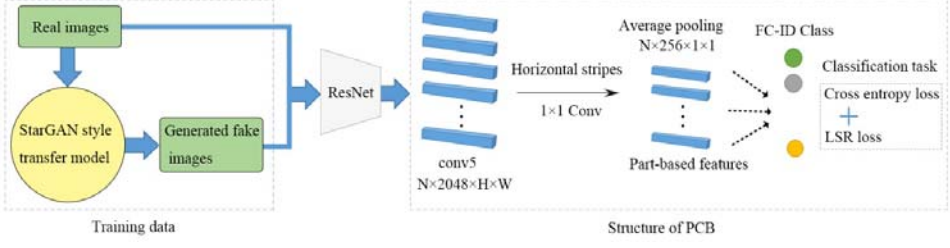
## 3 Methodology

This framework as shown in Figure 3 which addresses both data augmentation and robust discriminative feature generation. The two phases of network relate to, i.e.,

1 Efficient generation of camera-aware data using the StarGAN model. We train model using StarGAN which enables image-image translation between each camera pair.

2 Local body-part feature representation with the cross-entropy loss is used in the training phase.

**Figure 3** The framework of our proposed network (see online version for colours)




The StarGAN (Choi et al., 2018) is first employed to learn mappings among multiple different camera domains. Then, we apply the generated model to generate fake images that fit the style of other cameras simultaneously. Next, real images and generated fake images are fed into ResNet-50 to extract feature maps. We employ a successful PCB (Sun et al., 2017b) model to train the augmented dataset. Each of the combined images is equally partitioned into several horizontal parts. After a 1 × 1 kernel-sized convolutional layer and a fully connected (FC) layer, a classifier is trained with cross-entropy loss and LSR loss.

### *3.1 Review of the StarGAN model*

To solve image-to-image transformations of multiple domains in one model, StarGAN trains one generator which learns the mappings among different domains. StarGAN is a star network structure. The generation network *Generator* is implemented as a star structure. StarGAN only needs one *Generator* to learn the transition between all pairs of fields. In order to have the ability to learn multiple domain transformations, the following three folded changes are required for the generation network *Generator* and the discrimination network *Discriminator*. These changes mainly contain the following three aspects.

1 Add target domain information to the input of *Generator*. That is, the generated model needs to obtain information about which domain the image needs to be translated.

2 In addition to determining whether the image is true, *Discriminator* also has to determine which class the image belongs to. This will ensure the same input image in *Generator*, and generate different effects depending on the target field.

3 Finally, in the image translation process, it is necessary to save the real image and only the part of the domain with differences is changed. Image reconstruction is translated from domain A to domain B using image reconstruction, and then translated back without any change.

Given multiple domain datasets, the StarGAN train Generator to convert an image into the target domain image conditional on another domain label. On the other hand, *Discriminator* produces a probability distribution on the source and domain labels. The totally loss function to optimise G and D is expressed respectively as equation (1) and equation (2).

$$L_D = -L_{adv} + \lambda_{cls} L_{cls}^r \tag{1}$$

$$L_{\mathrm{G}} = L_{adv} + \lambda_{cls} L_{cls}^f + \lambda_{rec} L_{rec}, \tag{2}$$

where $L_{adv}$ is an adversarial loss that makes the generated fake images virtually indistinguishable from the actual images of the original GAN (Goodfellow et al., 2014). The general function of GAN $L_{adv}$ determines whether the output image is true or not. $L_{cls}^r$ and $L_{cls}^f$ both are the domain classification loss. For the real images, $L_{cls}^r$ is defined as a domain classification loss. The *Discriminator* is trained in the original field using real images. On the other side, for the fake images, $L_{cls}^r$ is defined as the loss function for the domain classification. The *Generator* is trained in the target domain using fake images. $L_{rec}$ is the reconstruction loss to minimise the adversarial and classification losses using the concept of cycle loss. The reconstruction function is similar to the forward function in CycleGAN (Zhu et al., 2017b). The weight of domain classification and reconstruction losses is set by $\lambda_{cls}$ and $\lambda_{rec}$. More derivation and proof of StarGAN can be accessed in Choi et al. (2018).

### 3.2 Method of camera transfer

The superiority of human part detection in person ReID has been demonstrated in some of the newest work (Sun et al., 2017b; Cheng et al., 2016; Wei et al., 2017; Zhang et al., 2017; Bai et al., 2017b; Dai et al., 2016). To preserve the spatial contextual information, combined person real image and fake images are fed into base network to extract features in this framework. As described in Figure 3, ResNet-50 is used for learning the low-level features.

In our method, we apply StarGAN (Choi et al., 2018) as an augmentation method to the training samples. The difference in style among different cameras is considered a different domain in the model. We reorganised the dataset of the pedestrian ReID so that the new dataset is organised according to different camera views. The goal of our method is to utilise StarGAN to simultaneously learn the image conversion model among multiple cameras. The model takes image and domain information as input instead of learning a fixed transformation and is trained to convert the training pedestrian image to the proper domain. Information in the domain is indicated by labels. We randomly generate an objective domain label in the training stage. The model is trained to transform the input pedestrian samples to the target camera domain. Thus, domain labels can be used to control the conversion of testing samples into any targeted domain.

Following the training methodology in Choi et al. (2018), we resize training person images to 256 × 256. As illustrated in Figure 3, real training pedestrian images under different cameras should be fed into the StarGAN model firstly. We use the same model as StarGAN for camera transformation network. Then, fake images are generated with the learned StarGAN model. As illustrated in Figure 2, images captured from camera1 are translated to styles in other 14 cameras, which are presented from the second column to the fifteenth column. In this way, training data of our framework is composed of fake images generated by StarGAN and the real images. The generated style-transferred fake images have the same id as the real images, so no manual labelling is required.

Similar to the experimental observation in Zhong et al. (2018), the generated camera style fake pedestrian images also have noise samples due to occlusion and detection errors. Considering the noise introduced by the fake samples, we further apply LSR (Szegedy et al., 2016) on samples of style transitions so that their labels are gently distributed during training. The method of LSR is not necessary for actual images since their labels are correct. Zhong et al. (2018) applied LSR to all generated fake images. However, we have experimentally proved that not all images will produce noise during the transfer process. We find that the different images in the dataset produced different noise during the camera style transfer process. We perform different LSR operations on different images in the experiment and the performance is further improved. The distribution of LSR is established as equation (3).

$$q_{LSR}(p) = \begin{cases} \dfrac{\varepsilon}{P}, & p \neq y \\ 1 - \varepsilon + \dfrac{\varepsilon}{P}, & p = y \end{cases}, \tag{3}$$

where $p \in \{1, 2, 3 \ldots, P\}$ represents the identity id of the training sample images, and $P$ is the total number of classes, $y$ is the ground truth of labels. $\varepsilon \in [0, 1]$ is a hyper-parameter which is set according to the confidence of the fake images. In the

training phase, the value of $\varepsilon$ is adjusted as 0 for the actual images since the label and the real image are correctly match. In our work, different from (Zhong et al., 2018), the value of $\varepsilon$ is set to different values according to the quality of generated images in the experiment. Fake images are generated according to camera styles. The usage of StarGAN ensures that the generated images remain the main characteristics of the pedestrian images. Our method has achieved better results in the baseline dataset such as MSMT17 with many cameras and large changes in style among cameras.

## 4 Results and discussion

### *4.1 Datasets*

Our experiment is performed on three benchmark person ReID datasets including MSMT17 (Wei et al., 2018) Market1501 (Zheng et al., 2015) and DukeMTMC-reID (Zheng et al., 2017b; Ristani et al., 2016). A brief description of them is included as follows.

The MSMT17 dataset (Wei et al., 2018) is the newly-released the largest person ReID as far as we know. It comes closer to the environment under realistic monitoring conditions. There are more cameras, the same identity is located in both indoor and outdoor, and there is a time span of morning, noon, and afternoon. The latest pedestrian detection technology Faster RCNN (Ren et al., 2017) is employed to obtain pedestrian bounding box in the MSMT17. The identity image generated by this method is more accurate and efficient. MSMT17 dataset includes 126,441 annotated bounding boxes of 4,101 identities obtained from twelve outdoor cameras and three indoor cameras. Compared with the previous pedestrian dataset, the background of person images in the MSMT17 dataset are more complicated, and the variation among the cameras is larger. Therefore, the MSMT17 dataset is more challenging.

The Market1501 dataset (Zheng et al., 2015) contains 32,668 annotated bounding boxes of 1,501 identities which collected from six cameras. The approach of deformable part model (DPM) is used to detect hand-drawn boxes automatically. There are two evaluation settings for a single query and multiple queries (Zheng et al., 2015).

The DukeMTMC-reID dataset is a subset of the large DukeMTMC dataset (Ristani et al., 2016) for multi cross-camera retrieval. The ReID version (Zheng et al., 2017b) is adopted in this work. The DukeMTMC-reID dataset contains 1,404 identities, 16,522 training images, 2,228 queries, and 17,661 gallery images obtained by eight high-resolution cameras.

All training and testing images are normalised to 256 × 256 pixels which are fed to StarGAN model. In the feature learning phase, the combination of real images and fake images is normalised to 256 × 128 pixels. We adopt the re-ranking method shown in Zhong et al. (2017). The re-ranking method, ever since it was proposed in Zhong et al. (2017), has been widely used and become a classical step for person ReID. Zhong et al. (2017) proposed a k-reciprocal encoding method to re-rank the results of person ReID. After obtaining the initial top-k using the normal person ReID method, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbours into a single vector. The re-ranking method with k-reciprocal encoding combines the original distance and Jaccard distance. The advantage of re-ranking method is that no labelled data is required

and no human interaction. The re-ranking method effectively improves the person ReID performance on several large-scale benchmark datasets for person ReID.

### 4.2 Influence of camera style transfer model on person ReID

We further evaluate several important parameters to illustrate the validity. MSMT17 is the latest and most challenging pedestrian dataset, the evaluation is based on MSMT17 dataset.

Table 1 contrasts the mAP and rank-k using the CycleGAN model ('Ours with CycleGAN model'), where the camera style transfer fake images are generated by CycleGAN. Using the StarGAN ('Ours with StarGAN model'), camera style transfer fake images are generated by StarGAN. Comparing results using StarGAN model are superior than CycleGAN model.

**Table 1** Effectiveness of using the StarGAN to generate camera style fake images without re-ranking method on MSMT17 dataset

| *Methods* | *mAP* | *Rank-1* | *Rank-5* | *Rank-10* |
|---|---|---|---|---|
| Ours with CycleGAN model | 37.6 | 68.8 | 73.6 | 86.1 |
| *Ours with StarGAN model* | *39.3* | *71.1* | *76.7* | *87.2* |

### 4.3 Influence of different body part-based features on person ReID

To investigate the contributions of body part-based descriptors, we conduct experiments based on the MSMT17 dataset. We train five different network models, corresponding to the five different parts of the body from top to bottom. Table 2 compares the mAP, and rank-$k$ ($k$ = 1, 5, 10) accuracies obtained with different body parts trained without using re-ranking method. The final layer of the feature map is divided into several horizontal stripes. Several local features are obtained by performing global average pooling on horizontal stripes. As detailed in this table, according to the human body structure, the experimental results are best when the body is divided into five parts. Table 3 shows the comparison results of body part features with and without overlapping ('overlap' in the table) between body parts. Apparently, different body parts contribute differently to the person ReID task. Features of different body parts can be fused in a more effective way and are helpful in improving person ReID performance. For pedestrian images of a size of 256 × 128 pixels, the length of the overlapping is set to be 10. Compared to the models without additional strategies in the structure, it can boost the person ReID accuracy.

**Table 2** The comparison of mAP and rank-1, rank-5, rank-10 accuracies of person ReID obtained on MSMT17 dataset using models trained with the features extracted from body parts

| *Methods* | *mAP* | *Rank-1* | *Rank-5* | *Rank-10* |
|---|---|---|---|---|
| Ours with global full-body features | 35.1 | 61.9 | 77.7 | 82.4 |
| Ours with two part of body features | 36.6 | 63.8 | 79.6 | 86.2 |
| Ours with three part of body features | 37.2 | 68.5 | 81.3 | 86.9 |
| *Ours with five part of body features* | *39.3* | *71.1* | *83.9* | *87.2* |

**Table 3** Effectiveness of using the complementary advantages of different features on MSMT17 dataset

| *Methods* | *mAP* | *Rank-1* | *Rank-5* | *Rank-10* |
|---|---|---|---|---|
| Ours without overlap | 36.2 | 68.6 | 74.4 | 84.1 |
| Ours + 10% overlap | 37.5 | 69.8 | 75.3 | 86.1 |
| Ours + 20% overlap | 39.3 | 71.1 | 83.9 | 87.2 |
| Ours + 25% overlap | 36.9 | 69.2 | 74.5 | 85.4 |
| *Ours + 20% overlap re-rank* | *62.6* | *79.1* | *87.3* | *90.7* |

## *4.4 Influence of LSR selection on person ReID*

In addition, we found that all images are more or less will produce noise during the transfer process. Zhong et al. (2018) performed the same LSR loss operation on all generated fake images. However, we discovered that the images under different cameras produced different degrees of noise during the transfer process in the experiment. Different from Zhong et al. (2018), we performed different LSR operations on the images under different cameras, and the performance is further improved on the MSMT17 dataset. We perform a statistical analysis of the image clarity of the MSMT17 dataset. Then we adjust the value of $\varepsilon$ according to the quality of the real and fake images. In equation (3), $\varepsilon \in [0, 1]$. $\varepsilon$ is set to 0 for real images. With the method of LSRO in Zheng et al. (2017b), the generated fake image has no label, $\varepsilon$ is set to 1 in equation (3). If the image obtained from the camera is in poor quality, the generated fake image is more prone to noise. The parameter $\varepsilon$ setting with high image quality tends to 0. Setting different parameter values has certain benefits for achieving higher accuracy.

## *4.5 Comparison with the state-of-the-art*

We evaluate the proposed method against the recently published works on three datasets. We compare our result with the recently published work on the MSMT17 dataset, Market-1501 and DukeMTMC-reID dataset. As depicted in Table 4 and Table 5, our method obtains competitive performance on three datasets. Specifically, our method attains 39.3% in terms of mAP and 71.1% matching rate at rank-1 under the single query setting on MSMT17, and outperforms all other person ReID methods. Combined with the re-ranking approach, our performance is further improved, reaching 79.1% matching rate at rank-1 with single query mode. Our approach also outperforms other methods in terms of both rank-5 and rank-10 matching rates. The comparison with several existing models on the Market-1501 and DukeMTMC-reID dataset are presented in Table 5. As shown in this table, our method outperforms all other methods with a large margin, achieving 72.5% and 55.7% in mAP after using re-ranking method on the Market-1501 and DukeMTMC-reID dataset, respectively. These results demonstrate that our proposed model has consistent superiority and robustness over the existing methods. As depicted in Figure 4, we further visualise some retrieval results on MSMT17 (Wei et al., 2018), Market1501 (Zheng et al., 2915) and DukeMTMC-reID (Zheng et al. 2017b; Ristani et al., 2016).

**Table 4** Comparison of results of single query on MSMT17 dataset

| *Methods* | *mAP* | *Rank-1* | *Rank-5* | *Rank-10* |
|---|---|---|---|---|
| PAN (GoogLeNet) (Szegedy et al., 2015) | 23.0 | 47.6 | 65.0 | 71.8 |
| PDC* (CaffeNet) (Su et al, 2017) | 29.7 | 58.0 | 73.6 | 79.4 |
| GLAD* (GoogLeNet) (Wei et al., 2017) | 34.0 | 61.4 | 76.8 | 81.6 |
| *Ours (ResNet-50)* | *39.3* | *71.1* | *83.9* | *87.2* |
| *Ours+re-rank (ResNet-50)* | *62.6* | *79.1* | *87.3* | *90.7* |

Notes: *Denotes the use of deep learning methods for body-part features. Base networks are annotated, e.g., ResNet-50.

**Table 5** Comparison of results of single query on Market1501 dataset and DukeMTMC-reID dataset

| *Methods* | *Market1501* | | *DukeMTMC-reID* | |
|---|---|---|---|---|
| | *mAP* | *Rank-1* | *mAP* | *Rank-1* |
| BoW + KISSME (Zheng et al., 2015) | 20.7 | 44.4 | 12.1 | 25.1 |
| Gated S-CNN (Varior et al., 2016) | 39.5 | 65.8 | - | - |
| P2S (ResNet-50) (Zhou et al., 2017) | 44.2 | 70.7 | - | - |
| CADL (CaffeNet) (Lin et al., 2017) | 47.1 | 73.8 | - | - |
| Spindle Net* (Zhao et al., 2017a) | - | 76.9 | - | - |
| GAN (ResNet-50) (Zheng et al., 2017b) | 56.2 | 78.0 | 47.1 | 67.6 |
| TOMM (ResNet-50) (Zheng et al., 2017a) | 59.8 | 79.5 | - | - |
| Quad (ResNet-50) (Chen et al., 2017a) | 61.1 | 80.0 | - | - |
| MSCAN* (CaffeNet) (Li et al. 2017a) | 57.5 | 80.3 | - | - |
| PAR* (ResNet-50) (Zhao et al., 2017b) | 63.4 | 81.0 | - | - |
| SSM (ResNet-50) (Bai et al., 2017a) | 68.8 | 82.2 | - | - |
| SVDNet (ResNet-50) (Sun et al. 2017a) | 62.1 | 82.3 | 56.8 | 76.7 |
| PAN (ResNet-50) (Szegedy et al., 2015) | 63.3 | 82.8 | 51.5 | 71.5 |
| PDC* (CaffeNet) (Su et al. 2017) | 63.4 | 84.4 | - | - |
| TriNet (ResNet-50) (Hermans et al., 2017) | 69.1 | 84.9 | - | - |
| JLML (ResNet-39) (Li et al., 2017b) | 65.5 | 85.1 | - | - |
| Angular (GoogLeNet) (Wang et al., 2017) | 69.7 | 85.5 | - | - |
| MultiScale* (ResNet-50) (Chen et al., 2017b) | 73.1 | 88.9 | - | - |
| GLAD* (GoogLeNet) (Wei et al., 2017) | 73.9 | 89.9 | - | - |
| CamStyle (ResNet-50) (Zhong et al., 2018) | 68.7 | 88.1 | 53.4 | 75.2 |
| *Ours (ResNet-50)* | *70.2* | *89.4* | *54.7* | *78.6* |
| *Ours + re-rank (ResNet-50)* | *72.5* | *90.9* | *55.7* | *79.9* |

Notes: *Denotes the use of deep learning methods for body-part features. Base networks are annotated, e.g., ResNet-50.

**Figure 4** Examples of pedestrian retrieval result on three datasets using the proposed method in single query mode (see online version for colours)


Notes: The images in the first column are the query images. The top-10 retrieved images are sorted and shown in the second column to the eleventh column according to the similarity scores from high to low. The correct and false matches are shown in green solid bounding boxes and red dash bounding boxes, respectively.

## 5 Conclusions

This paper has presented a framework which combines the part-based discriminative descriptors and camera style transfer method to solve the problem of variations among

different cameras in person ReID. StarGAN model is adopted to efficiently generate fake camera style images. Our approach is also a complement to the data augmentation approach. It is shown that overlapping part-based body feature fusion descriptor representations capture the discriminant details of pedestrian images and achieve better performance in experiments. Different LSR losses are employed to reduce the noise caused by StarGAN model. Comparison on three benchmark datasets have shown the competitiveness of the proposed framework over the state-of-the-art methods.

## Acknowledgements

The authors acknowledge the Science and Technology Projects of Huaian Jiangsu (Grant: HAG201602), Major Program of Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (Grant: 18KJA520002), the Six Talent Peaks Project in Jiangsu Province (Grant: 2016XYDXXJS-012), the Natural Science Foundation of Jiangsu Province (Grant: BK20171267), and the 533 Talents Engineering Project in Huaian (Grant: HAA201738).

## References

Bai, S., Bai, X. and Tian, Q. (2017a) 'Scalable person re-identification on supervised smoothed manifold', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Vol. 6, pp.2530–2539.

Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R. and Xu, Y. (2017b) 'Deep-person: learning discriminative deep features for person re-identification', *arXiv preprint arXiv:1711.10658*.

Chen, W., Chen, X., Zhang, J. and Huang, K. (2017a) 'Beyond triplet loss: a deep quadruplet network for person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.403–412.

Chen, Y., Zhu, X. and Gong, S. (2017b) 'Person re-identification by deep learning multi-scale representations', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.2590–2600.

Cheng, D., Gong, Y., Zhou, S., Wang, J. and Zheng, N. (2016) 'Person reidentification by multi-channel parts-based cnn with improved triplet loss function', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.1335–1344.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S. and Choo, J. (2018) 'Stargan: unified generative adversarial networks for multi-domain image-to-image translation', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.8789–8797.

Dai, J., Li, Y., He, K. and Sun, J. (2016) 'R-fcn: object detection via region-based fully convolutional networks', in *Advances in Neural Information Processing Systems (NIPS)*, pp.379–387.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y. and Jiao, J. (2018) 'Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.994–1003.

Ding, G., Zhang, S., Khan, S., Tang, Z., Zhang, J. and Porikli, F. (2018) 'Feature affinity based pseudo labeling for semi-supervised person reidentification', *arXiv preprint arXiv:1805.06118*.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative adversarial nets', in *Neural Information Processing Systems (NIPS)*, pp.2672–2680.

Hermans, A., Beyer, L. and Leibe, B. (2017) 'In defense of the triplet loss for person re-identification', *arXiv preprint arXiv:1703.07737*.

Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z. and Zhang, J. (2018) 'Multi-pseudo regularized label for generated samples in person reidentification', *arXiv preprint arXiv:1801.06742*.

Kim, T., Cha, M., Kim, H., Lee, J.K. and Kim, J. (2017) 'Learning to discover cross-domain relations with generative adversarial networks', in *Int. Conf. on Machine Learning (ICML)*, Vol. 4, No. 4, pp.2941–2949.

Li, D., Chen, X., Zhang, Z. and Huang, K. (2017a) 'Learning deep context-aware features over body and latent parts for person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.384–393.

Li, S., Bak, S., Carr, P. and Wang, X. (2018) 'Diversity regularized spatiotemporal attention for video-based person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.369–378.

Li, W., Zhu, X. and Gong, S. (2017b) 'Person re-identification by deep joint learning of multi-loss classification', in *Proc. Int. Joint Conf. on Artif. Intell. (IJCAI)*, pp.2194–2200.

Lin, J., Ren, L., Lu, J., Feng, J. and Zhou, J. (2017) 'Consistent-aware deep learning for person re-identification in a camera network', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Vol. 6, pp.5771–5780.

Liu, H., Feng, J., Qi, M., Jiang, J. and Yan, S. (2017) 'End-to-end comparative attention networks for person re-identification', *IEEE Transactions on Image Processing (TIP)*, Vol. 26, No. 7, pp.3492–3506.

Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G. and Xue, X. (2017) 'Pose-normalized image generation for person re-identification', *arXiv preprint arXiv:1712.02225*.

Ren, S., Girshick, R., Girshick, R. and Sun, J. (2017) 'Faster r-cnn: towards realtime object detection with region proposal networks', *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 39, No. 6, pp.1137–1149.

Ristani, E., Solera, F., Zou, R., Cucchiara, R. and Tomasi, C. (2016) 'Performance measures and a data set for multi-target, multi-camera tracking', in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp.17–35.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W. and Tian, Q. (2017) 'Pose-driven deep convolutional model for person re-identification', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.3980–3989.

Sun, Y., Zheng, L., Deng, W. and Wang, S. (2017a) 'Svdnet for pedestrian retrieval', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.3800–3808.

Sun, Y., Zheng, L., Yang, Y., Tian, Q. and Wang, S. (2017b) 'Beyond part models: person retrieval with refined part pooling', *arXiv preprint arXiv:1711.09349*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) 'Going deeper with convolutions', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) 'Rethinking the inception architecture for computer vision', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.2818–2826.

Varior, R.R., Haloi, M. and Wang, G. (2016) 'Gated siamese convolutional neural network architecture for human re-identification', in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp.791–808.

Wang, J., Zhou, F., Wen, S., Liu, X. and Lin, Y. (2017) 'Deep metric learning with angular loss', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.2593–2601.

Wang, Y., Wang, Z., Jia, W., He, X. and Jiang, M. (2018) 'Joint learning of body and part representation for person re-identification', *IEEE Access*, Vol. 6, pp.44199–44210.

Wei, L., Zhang, S., Gao, W. and Tian, Q. (2018) 'Person transfer gan to bridge domain gap for person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.79–88.

Wei, L., Zhang, S., Yao, H., Gao, W. and Tian, Q. (2017) 'Glad: global-local-alignment descriptor for pedestrian retrieval', in *Proceedings of the 2017 ACM on Multimedia Conference*, pp.420–428.

Xiao, Q., Luo, H. and Zhang, C. (2017) 'Margin sample mining loss: a deep learning based method for person re-identification', *arXiv preprint arXiv:1710.00478*.

Xiao, T., Li, H., Ouyang, W. and Wang, X. (2016) 'Learning deep feature representations with domain guided dropout for person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.1249–1258.

Yi, Z., Zhang, H., Tan, P. and Gong, M. (2017) 'Dualgan: unsupervised dual learning for image-to-image translation', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.2868–2876.

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C. and Sun, J. (2017) 'Alignedreid: surpassing human-level performance in person re-identification', *arXiv preprint arXiv:1711.08184*.

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X. and Tang, X. (2017a) 'Spindle net: person re-identification with human body region guided feature decomposition and fusion', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1077–1085.

Zhao, L., Li, X., Wang, J. and Zhuang, Y. (2017b) 'Deeply-learned part-aligned representations for person re-identification', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.3239–3248.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q. (2015) 'Scalable person re-identification: a benchmark', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.1116–1124.

Zheng, L., Yang, Y. and Hauptmann, A.G. (2016) 'Person re-identification: past, present and future', *arXiv preprint arXiv:1610.02984*.

Zheng, Z., Zheng, L. and Yang, Y. (2017a) 'A discriminatively learned CNN embedding for person reidentification', *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 14, No. 8, pp.1–10.

Zheng, Z., Zheng, L. and Yang, Y. (2017b) 'Unlabeled samples generated by gan improve the person re-identification baseline in vitro', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.3774–3782.

Zhong, Z., Zheng, L., Cao, D. and Li, S. (2017) 'Re-ranking person reidentification with k-reciprocal encoding', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.3652–3661.

Zhong, Z., Zheng, L., Zheng, Z., Li, S. and Yang, Y. (2018) 'Camera style adaptation for person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pp.5157–5166.

Zhou, S., Wang, J., Wang, J., Gong, Y. and Zheng, N. (2017) 'Point to set similarity based deep feature learning for person re-identification', in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Vol. 6, pp.3741–3750.

Zhu, F., Kong, X., Fu, H. and Tian, Q. (2017a) 'Pseudo-positive regularization for deep person re-identification', *Multimedia Systems*, Vol. 24, No. 4, pp.477–489.

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A. (2017b) 'Unpaired image-to-image translation using cycle-consistent adversarial networks', in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp.2242–2251.