

# Person Transfer GAN to Bridge Domain Gap for Person Re-Identification

Longhui Wei<sup>1</sup>, Shiliang Zhang<sup>1</sup>, Wen Gao<sup>1</sup>, Qi Tian<sup>2</sup>

<sup>1</sup>School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China

<sup>2</sup>Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249-1604, USA

Flonghuiwei, slzhang, jdl, wgao@pku.edu.cn, qi.tian@utsa.edu

## Abstract

*Although the performance of person Re-Identification (ReID) has been significantly boosted, many challenging issues in real scenarios have not been fully investigated, e.g., the complex scenes and lighting variations, viewpoint and pose changes, and the large number of identities in a camera network. To facilitate the research towards conquering those issues, this paper contributes a new dataset called MSMT17<sup>1</sup> with many important features, e.g., 1) the raw videos are taken by an 15-camera network deployed in both indoor and outdoor scenes, 2) the videos cover a long period of time and present complex lighting variations, and 3) it contains currently the largest number of annotated identities, i.e., 4,101 identities and 126,441 bounding boxes. We also observe that, domain gap commonly exists between datasets, which essentially causes severe performance drop when training and testing on different datasets. This results in that available training data cannot be effectively leveraged for new testing domains. To relieve the expensive costs of annotating new training samples, we propose a Person Transfer Generative Adversarial Network (PTGAN) to bridge the domain gap. Comprehensive experiments show that the domain gap could be substantially narrowed-down by the PTGAN.*

## 1. Introduction

Person Re-Identification (ReID) targets to match and return images of a probe person from a large-scale gallery set collected by camera networks. Because of its important applications in security and surveillance, person ReID has been drawing lots of attention from both academia and industry. Thanks to the development of deep learning and the availability of many datasets, person ReID performance has been significantly boosted.

Although the performance on current person ReID datasets is pleasing, there still remain several open issues

Figure 1: Illustration of the domain gap between CUHK03 and PRID. It is obvious that, CUHK03 and PRID present different styles, e.g., distinct lightings, resolutions, human race, seasons, backgrounds, etc., resulting in low accuracy when training on CUHK03 and testing on PRID.

hindering the applications of person ReID. First, existing public datasets differ from the data collected in real scenarios. For example, current datasets either contain limited number of identities or are taken under constrained environments. The currently largest DukeMTMC-reID [41] contains less than 2,000 identities and presents simple lighting conditions. Those limitations simplify the person ReID task and help to achieve high accuracy. In real scenarios, person ReID is commonly executed within a camera network deployed in both indoor and outdoor scenes and processes videos taken by a long period of time. Accordingly, real applications have to cope with challenges like a large number of identities and complex lighting and scene variations, which current algorithms might fail to address.

Another challenge we observe is that, there exists domain gap between different person ReID datasets, i.e., training and testing on different person ReID datasets results in severe performance drop. For example, the model trained on CUHK03 [20] only achieves the Rank-1 accuracy of 2.0% when tested on PRID [10]. As shown in Fig. 1, the domain gap could be caused by many reasons like different lighting conditions, resolutions, human race, seasons, backgrounds, etc. This challenge also hinders the applications

<sup>1</sup>The dataset is available at <http://www.pkumc.com>.

of person ReID, because available training samples cannot be effectively leveraged for new testing domains. Since annotating person ID labels is expensive, research efforts are desired to narrow-down or eliminate the domain gap.

Aiming to facilitate the research towards applications in realistic scenarios, we collect a new Multi-Scene Multi-Time person ReID dataset (*MSMT17*). Different from existing datasets, *MSMT17* is collected and annotated to present several new features. 1) The raw videos are taken by an 15-camera network deployed in both the indoor and outdoor scenes. Therefore, it presents complex scene transformations and backgrounds. 2) The videos cover a long period of time, *e.g.*, four days in a month and three hours in the morning, noon, and afternoon, respectively in each day, thus present complex lighting variations. 3) It contains currently the largest number of annotated identities and bounding boxes, *i.e.*, 4,101 identities and 126,441 bounding boxes. To our best knowledge, *MSMT17* is currently the largest and most challenging public dataset for person ReID. More detailed descriptions will be given in Sec. 3.

To address the second challenge, we propose to bridge the domain gap by transferring persons in dataset *A* to another dataset *B*. The transferred persons from *A* are desired to keep their identities, meanwhile present similar styles, *e.g.*, backgrounds, lightings, *etc.*, with persons in *B*. We model this transfer procedure with a Person Transfer Generative Adversarial Network (PTGAN), which is inspired by the Cycle-GAN [42]. Different from Cycle-GAN [42], PTGAN considers extra constraints on the person foregrounds to ensure the stability of their identities during transfer. Compared with Cycle-GAN, PTGAN generates high quality person images, where person identities are kept and the styles are effectively transformed. Extensive experimental results on several datasets show PTGAN effectively reduces the domain gap among datasets.

Our contributions can be summarized into three aspects. 1) A new challenging large-scale *MSMT17* dataset is collected and will be released. Compared with existing datasets, *MSMT17* defines more realistic and challenging person ReID tasks. 2) We propose person transfer to take advantages of existing labeled data from different datasets. It has potential to relieve the expensive data annotations on new datasets and make it easy to train person ReID systems in real scenarios. An effective PTGAN model is presented for person transfer. 3) This paper analyzes several issues hindering the applications of person ReID. The proposed *MSMT17* and algorithms have potential to facilitate the future research on person ReID.

## 2. Related Work

This work is closely related with descriptor learning in person ReID and image-to-image translation by GAN. We briefly summarize those two categories of works in this sec-

tion.

### 2.1. Descriptor Learning in Person ReID

Deep learning based descriptors have shown substantial advantages over hand-crafted features on most of person ReID datasets. Some works [33, 41, 28] learn deep descriptors from the whole images with classification models, where each person ID is treated as a category. Some other works [40, 6] combine verification models with classification models to learn descriptors. Hermans *et al.* [9] show that triplet loss effectively improves the performance of person ReID. Similarly, Chen *et al.* [1] propose the quadruplet network to learn representations.

The above works learn global descriptors and ignore the detailed cues which might be important for distinguishing persons. To explicitly utilize local cues, Cheng *et al.* [2] propose a multi-channel part-based network to learn a discriminative descriptor. Wu *et al.* [32] discover hand-crafted features could be complementary with deep features. They divide the global image into five fixed-length regions. For each region, a histogram descriptor is extracted and concatenated with the global deep descriptor. Though the above works achieve good performance, they ignore the misalignment issue caused by fixed body part division. Targeting to solve this issue, Wei *et al.* [31] utilize Deeppercut [11] to detect three coarse body regions and then learn an global-local-alignment descriptor. In [38], more fine-grained part regions are localized and then fed into the proposed Spindle Net for descriptor learning. Similarly, Li *et al.* [18] adopt Spatial Transform Networks (STN) [13] to detect latent part regions and then learn descriptors on those regions.

### 2.2. Image-to-Image Translation by GAN

Since GAN proposed by Goodfellow *et al.* [7], many variants of GAN [24, 25, 30, 36, 17, 34, 16, 5, 21, 35, 14, 42, 23] have been proposed to tackle different tasks, *e.g.*, natural style transfer, super-resolution, sketch-to-image generation, image-to-image translation, *etc.* Among them, image-to-image translation has attracted lots of attention. In [12], Isola *et al.* propose conditional adversarial networks to learn the mapping function from input to output images. However, this method requires paired training data, which is hard to acquire in many tasks [42]. Targeting to solve the unpaired image-to-image translation task, Zhu *et al.* [42] propose cycle consistency loss to train unpaired data. Also, the works [35, 14] propose a similar framework to solve the task. Our proposed PTGAN is similar to Cycle-GAN [42] in that, it also performs image-to-image translation. Differently, extra constraints on person identity are applied to ensure the transferred images can be used for model training. Zheng *et al.* [41] adopt GAN to generate new samples for data augmentation in person ReID. Their work differs from ours in both motivation and methodology. As far as

we know, this is an early work on person transfer by GAN for person ReID.

### 3. MSMT17 Dataset

#### 3.1. Overview of Previous Datasets

Current person ReID datasets have significantly pushed forward the research on person ReID. As shown in Table 1, *DukeMTMC-reID* [41], *CUHK03* [20], and *Market1501* [39] involve larger numbers of cameras and identities than *VIPeR* [8] and *PRID* [10]. The amount of training data makes it possible to develop deep models and show their discriminative power in person ReID. Although current algorithms have achieved high accuracy on those datasets, person ReID is far from being solved and widely applied in real scenarios. Therefore, it is necessary to analyze the limitations of existing datasets.

Compared with the data collected in real scenarios, current datasets present limitations in four aspects: 1) The number of identities and cameras are not large enough, especially when compared with the real surveillance video data. In Table 1, the largest dataset contains only 8 cameras and less than 2,000 identities. 2) Most of existing datasets cover only single scene, *i.e.*, either indoor or outdoor scene. 3) Most of existing datasets are constructed from short-time surveillance videos without significant lighting changes. 4) Their bounding boxes are generated either by expensive hand drawing or out-dated detectors like Deformable Part Model (DPM) [4]. Those limitations make it necessary to collect a larger and more realistic dataset for person ReID.

#### 3.2. Description to MSMT17

Targeting to address above mentioned limitations, we collect a new Multi-Scene Multi-Time person ReID dataset (*MSMT17*) by simulating the real scenarios as much as possible. We utilize an 15-camera network deployed in campus. This camera network contains 12 outdoor cameras and 3 indoor cameras. We select 4 days with different weather conditions in a month for video collection. For each day, 3 hours of videos taken in the morning, noon, and afternoon, respectively, are selected for pedestrian detection and annotation. Our final raw video set contains 180 hours of videos, 12 outdoor cameras, 3 indoor cameras, and 12 time slots. Faster RCNN [26] is utilized for pedestrian bounding box detection. Three labelers go through the detected bounding boxes and annotate ID label for 2 months. Finally, 126,441 bounding boxes of 4,101 identities are annotated. Some statistics on *MSMT17* are shown in Fig. 3. Sample images from *MSMT17* are shown and compared in Fig. 2. Compared with existing datasets, we summarize the new features in *MSMT17* into the following aspects:

1) *Larger number of identities, bounding boxes, and cameras.* To our best knowledge, *MSMT17* is currently

Figure 2: Comparison of person images in *CUHK03*, *Market1501*, *DukeMTMC-reID*, and *MSMT17*. Each column shows two sample images of the same identity. It is obvious that, *MSMT17* presents a more challenging and realistic person ReID task.

the largest person ReID dataset. As shown by the comparison in Table 1, *MSMT17* contains 126,441 bounding boxes, 4,101 identities, which are significantly larger than the ones in previous datasets.

2) *Complex scenes and backgrounds.* *MSMT17* contains the largest number of cameras, *i.e.*, 15 cameras placed in different locations. It is also constructed with both indoor and outdoor videos, which has not been considered in previous datasets. Those considerations result in complex backgrounds and scene variations, also make *MSMT17* more appealing and challenging.

3) *Multiple time slots result in severe lighting changes.* *MSMT17* is collected with 12 time slots, *i.e.*, morning, noon, and afternoon in four days. It better simulates the real scenarios than previous datasets, but brings severe lighting changes.

4) *More reliable bounding box detector.* Compared with hand drawing and DPM detector, Faster RCNN [26] is a better choice for bounding box detection in real applications, *e.g.*, easier to implement and more accurate.

#### 3.3. Evaluation Protocol

We randomly divide our dataset into training set and testing set, respectively. Different from dividing the two parts equally in previous datasets, we set the training and testing ratio as 1:3. We use this setting because of the expensive data annotation in real scenarios, and thus want to en-

Table 1: Comparison between *MSMT17* and other person ReID datasets.

Dataset	MSMT17	Duke [41]	Market [39]	CUHK03 [20]	CUHK01 [19]	VIPeR [8]	PRID [10]	CAVIAR [3]
BBoxes	<b>126,441</b>	36,411	32,668	28,192	3,884	1,264	1,134	610
Identities	<b>4,101</b>	1,812	1,501	1,467	971	632	934	72
Cameras	<b>15</b>	8	6	2	10	2	2	2
Detector	<b>Faster RCNN</b>	hand	DPM	DPM, hand	hand	hand	hand	hand
Scene	<b>outdoor, indoor</b>	outdoor	outdoor	indoor	indoor	outdoor	outdoor	indoor

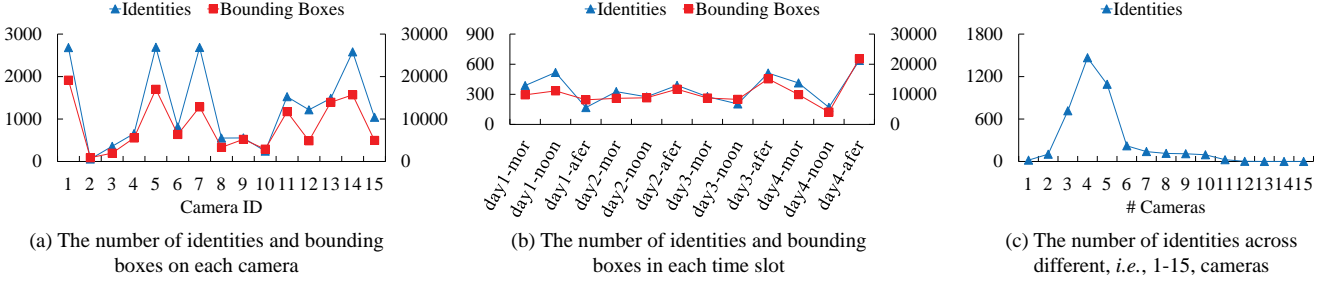


Figure 3: Statistics of *MSMT17*.

courage more efficient training strategies. Finally, the training set contains 32,621 bounding boxes of 1,041 identities, and the testing set contains 93,820 bounding boxes of 3,060 identities. From the testing set, 11,659 bounding boxes are randomly selected as query images and the other 82,161 bounding boxes are used as gallery images.

Similar with most of previous datasets, we utilize the Cumulated Matching Characteristics (CMC) curve to evaluate the ReID performance. For each query bounding box, multiple true positives could be returned. Therefore, we also regard person ReID as a retrieval task. mean Average Precision (mAP) is thus also used as the evaluation metric.

#### 4. Person Transfer GAN

To better leverage the training set of dataset *A* in person ReID tasks on dataset *B*, we propose to bridge the domain gap by transferring persons in *A* to *B*. As illustrated in Fig. 1, different datasets present distinct styles due to multiple reasons such as backgrounds, lighting conditions, resolutions, *etc.* Imagine that, if persons in *A* were captured by the cameras of *B*, the style of those person images would be consistent with the style of *B*. Our person transfer tries to simulate this procedure, *i.e.*, learning a transfer function to 1) ensure the transferred person images show similar styles with the target dataset, and 2) keep the appearance and identity cues of the person during transfer.

This transfer task seems easy, *e.g.*, can be finished by cropping the person foregrounds from *A* and paste them on the backgrounds on *B*. However, it is difficult to deal with multiple reasons of domain gap in a rule-based algorithm. Moreover, there could be complicated style variations on *B*,

*e.g.*, different backgrounds and lighting conditions between two cameras of *PRID* in Fig. 1. Our algorithm is inspired by the popularity of GAN models, which have been proven effective in generating the desired image samples. We hence design a Person Transfer GAN (PTGAN) to perform person transfer from *A* to *B*.

Based on the above discussions, PTGAN is constructed to satisfy two constraints, *i.e.*, the style transfer and person identity keeping. The goal of style transfer is to learn the style mapping functions between different person datasets. The goal of person identity keeping is to ensure the identity of one person remains unchanged after transfer. Because different transferred samples of one person are regarded as having the same person ID, the constraint on person identity is important for person ReID training. We thus formulate the loss function of PTGAN as, *i.e.*,

$$\mathcal{L}_{PTGAN} = \mathcal{L}_{Style} + \lambda_1 \mathcal{L}_{ID}, \quad (1)$$

where  $\mathcal{L}_{Style}$  denotes the style loss and  $\mathcal{L}_{ID}$  denotes the identity loss, and  $\lambda_1$  is the parameter for the trade-off between two losses.

ReID datasets do not contain paired person images, *i.e.*, images of the same person from different datasets. Therefore, the style transfer can be regarded as an unpaired image-to-image translation task. Because of the good performance of Cycle-GAN in unpaired image-to-image translation task, we employ Cycle-GAN to learn the style mapping functions between dataset *A* and *B*. Suppose  $\mathcal{G}$  represents the style mapping function from *A* to *B* and  $\bar{\mathcal{G}}$  represents the style mapping function from *B* to *A*.  $\mathcal{D}_A$  and  $\mathcal{D}_B$  are the style discriminators for *A* and *B*, respectively. The



objective function of style transfer learning can be formulated as follows:

$$\begin{aligned} L_{\text{Style}} = & L_{\text{GAN}}(G, D_B, A, B) \\ & + L_{\text{GAN}}(\bar{G}, D_A, B, A) \\ & + \lambda L_{\text{cyc}}(G, \bar{G}), \end{aligned} \quad (2)$$

Where  $L_{\text{GAN}}$  represents the standard adversarial loss [7], and  $L_{\text{cyc}}$  represents the cycle consistency loss [42]. For more details, please refer to the Cycle-GAN [42].

Solely considering style transfer may result in ambiguous person ID labels in transferred person images. We thus compute the identity loss to ensure the accuracy of person ID labels in the transferred data. The person identity loss is computed by first acquiring the foreground mask of a person, then evaluating the variations on the person foreground before and after person transfer. Given the data distribution of  $A$  as  $a \sim p_{\text{data}}(a)$  and the data distribution of  $B$  as  $b \sim p_{\text{data}}(b)$ . The objective function of identity loss can be formulated as follows:

$$\begin{aligned} L_{\text{ID}} = & \int_a \int_b p_{\text{data}}(a) p_{\text{data}}(b) ||(G(a) - G(b)) - M(a) \odot M(b)||_2 \\ & + \int_a \int_b p_{\text{data}}(a) p_{\text{data}}(b) ||(\bar{G}(b) - \bar{G}(a)) - M(b) \odot M(a)||_2, \end{aligned} \quad (3)$$

where  $G(a)$  represents the transferred person image from image  $a$ , and  $M(a)$  represents the foreground mask of person image  $a$ .

Because of its good performance on segmentation task, we use PSPNet [37] to extract the mask on person images. On video surveillance data with moving foregrounds and fixed backgrounds, more accurate and efficient foreground extraction algorithms can be applied. It can be observed that, PTGAN does not require person identity labels on the target dataset  $B$ . The style discriminator  $D_B$  can be trained with unlabeled person images on  $B$ . Therefore, PTGAN is well-suited to real scenarios, where the new testing domains have limited or no labeled training data.

We show some sample results generated by PTGAN in Fig. 4. Compared with Cycle-GAN, PTGAN generates images with substantially higher quality. For example, the appearance of person is maintained and the style is effectively transferred toward the one on *PRID* camera1. The shadows, road marks, and backgrounds are automatically generated and are similar with the ones on *PRID* camera1. It is also interesting to observe that, PTGAN still works well with the noisy segmentation results generated by PSPNet. This implies that, PTGAN is also robust to the segmentation errors. More detailed evaluation of PTGAN will be given in Sec. 5.4.

## 5. Experiments

### 5.1. Datasets

In addition to the *MSMT17*, four widely used person ReID datasets are employed in our experiments.

Figure 4: Comparison of the transferred images by PTGAN and Cycle-GAN from *CUHK03* to *PRID-cam1*. The second row shows the segmentation results by PSPNet. The pink regions are segmented as person body regions.

*DukeMTMC-reID* [41] is composed of 1,812 identities and 36,411 bounding boxes. 16,522 bounding boxes of 702 identities are used for training. The rest identities are included in the testing set. *DukeMTMC-reID* is also denoted as *Duke* for short.

*Market-1501* [39] contains 1,501 identities and 32,668 bounding boxes. The training set contains 12,936 bounding boxes of 751 identities. The rest 750 identities are included in the testing set. *Market-1501* is also denoted as *Market* for short.

*CUHK03* [20] consists of 1,467 identities and 28,192 bounding boxes generated by both DPM and hand. Following the work [33], 26,264 bounding boxes of 1,367 identities are used for training, and 1,928 bounding boxes of 100 identities are used for testing.

*PRID* [10] is composed of 934 identities from two cameras. Our experiments use the bounding boxes of 200 persons shared by both cameras as testing set.

### 5.2. Implementation Details

PTGAN uses similar network architecture with the one in Cycle-GAN [42]. For the generator network, two stride-2 convolutions, 9 residual blocks, and two stride- $\frac{1}{2}$  fractionally-strided convolutions are designed. Two parts are included in the discriminator network. PatchGAN [12] is adopted as one part. The PatchGAN classifies whether a  $70 \times 70$  patch in an image is real or fake. For the other part,  $L_2$  distance between the transferred image and input image is computed on the foreground person.

Adam solver [15] is adopted in PTGAN. For the generator network, the learning rate is set as 0.0002. The learning rate is set as 0.0001 for the discriminator network. We set

Table 2: The performance of the state-of-the-art methods on *MSMT17*. R-1 represents the Rank-1 accuracy.

Methods	mAP	R-1	R-5	R-10	R-20
GoogLeNet [29]	23.0	47.6	65.0	71.8	78.2
PDC [27]	29.7	58.0	73.6	79.4	84.5
GLAD [31]	<b>34.0</b>	<b>61.4</b>	<b>76.8</b>	<b>81.6</b>	<b>85.9</b>

$\gamma_1 = 10$ , and  $\gamma_2 = 10$ . The size of input image is  $256 \times 256$ . Finally, we train PTGAN for 40 epochs.

### 5.3. Performance on MSMT17

As described in Sec. 3, *MSMT17* is challenging but close to the reality. This section verifies this claim by testing existing algorithms on *MSMT17*.

We go through the state-of-the-art works published in 2017 and 2016. Among those works, the GLAD proposed by Wei *et al.* [31] achieves the best performance on *Market*, and the PDC proposed by Su *et al.* [27] achieves the best performance on *CUHK03*.<sup>2</sup> We thus evaluate those two methods on *MSMT17*. In most of person ReID works, GoogLeNet [29] is commonly used as the baseline model. We thus also use GoogLeNet [29] as our baseline.

We summarize the experimental results in Table 2. As shown in the table, the baseline only achieves mAP of 23% on *MSMT17*, which is significantly lower than its mAP of 51.7% on *Market* [6]. It is also obvious that, PDC [27] and GLAD [31] substantially outperform the baseline performance by considering extra part and regional features. However, the best performance achieved by GLAD, *e.g.*, mAP of 34%, is still substantially lower than its reported performance on other datasets, *e.g.*, 73.9% on *Market*. The above experiments clearly show the challenges of *MSMT17*.

We also show some sample retrieval results in Fig. 5. From the samples, we can conclude that although challenging, the ReID task defined by *MSMT17* is realistic. Note that, in real scenarios distinct persons may present similar clothing cues, and images of same person may present different lightings, backgrounds, and poses. As shown in Fig. 5, the false positive samples do show similar appearances with the one of query person. Some true positives present distinct lightings, poses, and backgrounds from the query. Therefore, we believe *MSMT17* is a valuable dataset to facilitate the future research on person ReID.

### 5.4. Performance of Person Transfer

Person transfer is performed from dataset *A* to *B*. The transferred data is hence used for training on *B*. To ensure there is enough transferred data for training on *B*, we test person transfer in two cases, *i.e.*, 1) transferring from a large

<sup>2</sup>The work [22] reports better performance, but it is trained on an augmented data including training sets from three datasets.

Figure 5: Sample person ReID results generated by the method of GLAD [31] on *MSMT17*.

Figure 6: Sample transferred person images from *CUHK03* to *PRID-cam2*. Each sample shows an image from *CUHK03* in the first column, and the transferred image in the second column.

A to a small B, and 2) transferring from a large A to a large B. In the following experiments, we use the training set provided by *A* for person transfer.

#### 5.4.1 Transfer from Large Dataset to Small Dataset

This part tests the performance of transferred person data from *CUHK03* and *Market* to a small dataset *PRID*. As shown in Fig. 1, person images captured by two cameras on *PRID* show different styles. Therefore, we perform person transfer to those two cameras, *i.e.*, *PRID-cam1* and *PRID-cam2*, respectively.

We first perform person transfer from *CUHK03* to *PRID-cam1* and *PRID-cam2*. Samples of the transferred person images to *PRID-cam1* are shown in Fig. 4. We additionally show samples of transferred person images from *CUHK03*

Table 3: Performance of GoogLeNet tested on *PRID* but trained with different training sets.  $\cdot$  denotes the transferred dataset. For instance, the subscript *cam1* represents the transferred target dataset *PRID-cam1*. “cam1/cam2” means using images in *PRID-cam1* as query set and images from *PRID-cam2* as gallery set.

Training Set	cam1/cam2		cam2/cam1	
	R-1	R-10	R-1	R-10
<i>CUHK03</i>	2.0	11.5	1.5	11.5
<i>CUHK03</i> <sub>cam1</sub>	18.0	43.5	6.5	24.0
<i>CUHK03</i> <sub>cam2</sub>	17.5	53.0	22.5	54.0
<i>CUHK03</i> <sub>cam1</sub> + <i>CUHK03</i> <sub>cam2</sub>	<b>37.5</b>	<b>72.5</b>	<b>37.5</b>	<b>69.5</b>
<i>Market</i>	5.0	26.0	11.0	40.0
<i>Market</i> <sub>cam1</sub>	17.5	50.5	8.5	28.5
<i>Market</i> <sub>cam2</sub>	10.0	31.5	10.5	37.5
<i>Market</i> <sub>cam1</sub> + <i>Market</i> <sub>cam2</sub>	<b>33.5</b>	<b>71.5</b>	<b>31.0</b>	<b>70.5</b>

to *PRID-cam2* in Fig. 6. It is clear that, the transferred person images to those two cameras show different styles, which are consistent with the ones on *PRID*. We also transfer *Market* to *PRID-cam1* and *PRID-cam2*, respectively. Samples of the transferred person images from *Market* are shown in Fig. 7, where similar results can be observed as the ones in Fig. 4 and Fig. 6, respectively.

To further evaluate whether the domain gap is reduced through PTGAN. We conduct comparisons between GoogLeNet trained with the training sets on *CUHK03* and *Market*, and GoogLeNet trained on their transferred training sets, respectively. The experimental results are summarized in Table 3. As shown in the table, GoogLeNet trained on the *CUHK03*, only achieves the Rank-1 accuracy of 2.0% on *PRID*, which implies substantial domain gap between *CUHK03* and *PRID*. With training data transferred by PTGAN, GoogLeNet achieves a significant performance boost, *e.g.*, the Rank-1 accuracy is improved from 2.0% to 37.5%, the Rank-10 accuracy is improved from 11.5% to 72.5%. Similar improvements can be observed from the results on *Market*, *e.g.*, the Rank-1 accuracy is significantly improved from 5.0% to 33.5% after person transfer. The substantial performance improvements clearly indicate the shrunken domain gap. Moreover, this experiment shows that even without using labeled data on *PRID*, we can achieve reasonable performance on it using training data from other datasets.

From Table 3, we also observe an interesting phenomenon, *i.e.*, combining the transferred datasets on two cameras results in better performance. This might be due to two reasons: 1) the combined dataset has more training samples, thus helps to train a better deep network, and 2) it enables the learning of style differences between two cameras. In the combined dataset, each person image has two transferred samples on camera1 and camera2, respectively with different styles. Because those two samples have the

Figure 7: Sample transferred person images from *Market* to *PRID-cam1* and *PRID-cam2*. Images in the first column are from *Market*. Transferred images to *PRID-cam1* and *PRID-cam2* are shown in the second and third columns, respectively.

same person ID label, this training data enforces the network learning to gain robustness to the style variations between camera1 and camera2.

#### 5.4.2 Transfer from Large Dataset to Large Dataset

This part simulates a more challenging scenario commonly existing in real applications, *i.e.*, the available training data on a large testing set is not provided. We thus test the performance of PTGAN by conducting person transfer among three large datasets.

The large person ReID dataset commonly contains a large number of cameras, making it expensive to perform person transfer to each individual camera. Therefore, different from the experimental settings in Sec. 5.4.1, we do not distinguish different cameras and directly transfer person images to the target dataset with one PTGAN. Obviously, this is not an optimal solution for person transfer. Our experimental results are summarized in Fig. 8. It is obvious that GoogLeNet trained on transferred datasets works better than the one trained on the original training sets. Sample transferred images are presented in Fig. 9. It is obvious that, although we use a simple transfer strategy, PTGAN still generates high quality images. Possible better solutions for person transfer to large datasets will be discussed as our future work in Sec. 6.

#### 5.5 Performance of Person Transfer on MSMT17

We further test PTGAN on *MSMT17*. We use the same strategy in Sec. 5.4.2 to conduct person transfer. As shown in Table 4, the domain gaps between *MSMT17* and the other three datasets are effectively narrowed-down by PTGAN. For instance, the Rank-1 accuracy is improved by 4.7%, 6.8%, and 3.7% after performing person transfer

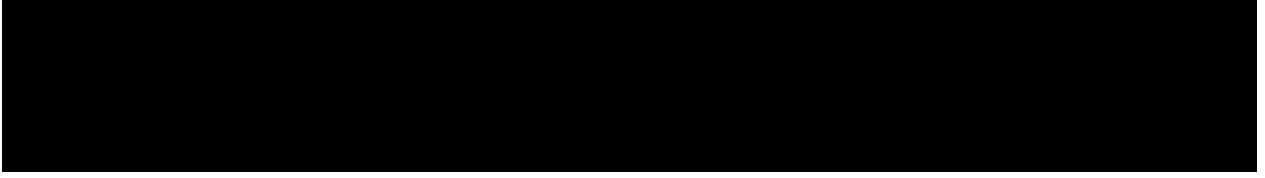


Figure 8: Rank-1 and Rank-10 accuracies of GoogLeNet on *CUHK03*, *Market*, and *Duke*. The subscripts C, Ma, and D denote the transferred target dataset is *CUHK03*, *Market*, and *Duke*, respectively.

Table 4: The performance of GoogLeNet tested on *MSMT17*. The subscript MS denotes the transferred target dataset *MSMT17*.

	<i>Duke</i>	<i>Duke</i> <sub>MS</sub>	<i>Market</i>	<i>Market</i> <sub>MS</sub>	<i>CUHK03</i>	<i>CUHK03</i> <sub>MS</sub>
R-1	7.1	<b>11.8</b>	3.4	<b>10.2</b>	2.8	<b>6.5</b>
R-10	17.4	<b>27.4</b>	10.0	<b>24.4</b>	8.6	<b>17.2</b>
mAP	1.9	<b>3.3</b>	1.0	<b>2.9</b>	0.7	<b>1.7</b>

Figure 9: Illustration of the transferred person images to *Duke*. The images in first row are from *Duke*. The images in second and third rows are transferred images from *Market* to *Duke*. Obviously, those images have the similar styles, *e.g.*, similar backgrounds and lightings, *etc.*

Table 5: The performance of GoogLeNet for weakly supervised learning on *MSMT17*.

Training Set	R-1	R-10	mAP
<i>MSMT</i> (1%)	0.9	3.6	0.2
<i>MSMT</i> (2.5%)	2.0	7.4	0.5
<i>MSMT</i> (5%)	6.3	18.1	1.9
<i>MSMT</i> (10%)	11.5	26.9	3.7
<i>Duke</i> + <i>MSMT17</i> (10%)	16.1	33.1	5.5
<i>Duke</i> <sub>MS</sub> + <i>MSMT17</i> (10%)	<b>18.0</b>	<b>36.4</b>	<b>6.2</b>
<i>Market</i> + <i>MSMT17</i> (10%)	12.6	28.5	4.4
<i>Market</i> <sub>MS</sub> + <i>MSMT17</i> (10%)	<b>17.7</b>	<b>35.9</b>	<b>6.0</b>
<i>CUHK03</i> + <i>MSMT17</i> (10%)	11.9	28.3	4.1
<i>CUHK03</i> <sub>MS</sub> + <i>MSMT17</i> (10%)	<b>14.3</b>	<b>31.7</b>	<b>4.6</b>

from *Duke*, *Market*, and *CUHK03*, respectively.

In real scenarios, the testing set is commonly large and has limited number of labeled training data. We hence test the validity of person transfer in such case. We first show the person ReID performance using different portions of training data on *MSMT17* in Table 5. From the comparison between Table 4 and Table 5, it can be observed that 10% of *MSMT17* training set gets similar performance with the transferred training set from *Duke*, *e.g.*, both achieve the Rank-1 accuracy of about 11.5% 11.8%. Therefore, 16,522 transferred images from *Duke* achieves similar performance with 2,602 annotated images on *MSMT17*. We can roughly estimate that 6.3 transferred images are equivalent to 1 annotated image. This thus effectively relieves the cost of data annotation on new datasets. The transferred data is then combined with the training set on *MSMT17*. As shown in Table 5, the Rank-1 accuracy is constantly improved by 1.9%, 5.1%, and 2.4%, respectively by combining the transferred data from *Duke*, *Market*, and *CUHK03*, respectively.

## 6. Conclusions and Discussions

This paper contributes a large-scale *MSMT17* dataset. *MSMT17* presents substantially variants on lightings, scenes, backgrounds, human poses, *etc.*, and is currently the largest person ReID dataset. Compared with existing

datasets, *MSMT17* defines a more realistic and challenging person ReID task.

PTGAN is proposed as an original work on person transfer to bridge the domain gap among datasets. Extensive experiments show PTGAN effectively reduces the domain gap. Different cameras may present different styles, making it difficult to perform multiple style transfer with one mapping function. Therefore, the person transfer strategy in Sec. 5.4.2 and Sec. 5.5 is not yet optimal. This also explains why PTGAN learned on each individual target camera performs better in Sec. 5.4.1. A better strategy is to consider the style differences among cameras to get more stable mapping functions.

## ACKNOWLEDGMENTS

This work is supported by National Science Foundation of China under Grant No. 61572050, 91538111, 61620106009, 61429201, and the National 1000 Youth Talents Plan, in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar. We gratefully acknowledge the support from NVIDIA NVAIL program.



## References

- [1] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.
- [2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.
- [3] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011.
- [4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.
- [5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.
- [6] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.
- [7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.
- [8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.
- [9] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [10] M. Hirzer, C. Beleznaï, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.
- [11] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcrut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.
- [12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.
- [13] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.
- [14] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.
- [15] D. Kinga and J. B. Adam. A method for stochastic optimization. In *ICLR*, 2015.
- [16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.
- [17] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016.
- [18] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.
- [19] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.
- [20] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.
- [21] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.
- [22] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017.
- [23] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. *arXiv preprint arXiv:1712.02621*, 2017.
- [24] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.
- [26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [27] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.
- [28] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016.
- [29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [30] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.
- [31] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017.
- [32] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016.
- [33] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.
- [34] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016.
- [35] Z. Yi, H. Zhang, P. T. Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017.
- [36] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixel-level domain transfer. In *ECCV*, 2016.
- [37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.
- [38] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.
- [39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

- [40] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *arXiv preprint arXiv:1611.05666*, 2016.
- [41] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.
- [42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.