Eliminating Background-bias for Robust Person Re-identification

Maoqing Tian[1], Shuai Yi[1], Hongsheng Li[2], Shihua Li[3],
Xuesen Zhang[1], Jianping Shi[1], Junjie Yan[1], Xiaogang Wang[2]
[1] SenseTime Research, [2] Chinese University of Hong Kong, [3] Shenzhen Municipal Public Security Bureau
tianmaoqing@sensetime.com, yishuai@sensetime.com, jasonlee@szga.gov.cn

Abstract

Person re-identification is an important topic in intelligent surveillance and computer vision. It aims to accurately measure visual similarities between person images for determining whether two images correspond to the same person. State-of-the-art methods mainly utilize deep learning based approaches for learning visual features for describing person appearances. However, we observe that existing deep learning models are biased to capture too much relevance between background appearances of person images. We design a series of experiments with newly created datasets to validate the influence of background information. To solve the background bias problem, we propose a person-region guided pooling deep neural network based on human parsing maps to learn more discriminative person-part features, and propose to augment training data with person images with random background. Extensive experiments demonstrate the robustness and effectiveness of our proposed method. [1] [2]

1. Introduction

Person re-identification aims to identify a person of interest from a large gallery image database collected from different cameras, when given a probe image of the person. Person re-identification is conducted by estimating the visual similarities between person images. According to the similarities between a probe image and gallery images, the gallery images can be ranked in the descending order of the similarities as re-identification results. Such a task has extensive applications in intelligent surveillance. For instance, it can be used to search criminal suspects or missing persons from a large surveillance camera network efficiently and effectively. In recent years, the performance of person re-identification has been improved significantly because of the emergence of deep learning techniques and more advanced computational hardware.

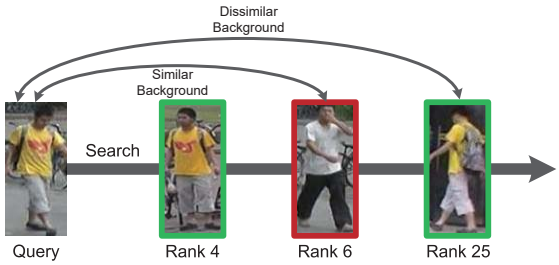

Figure 1. One example to show the background bias in person re-identification. We rank all the gallery pictures by the distance between probe picture ascending and list them from left to right. (a). The probe pictures. (b). The same person from the same camera. (c).The different person from the same camera having similar background with probe pictures. (d). The same person from different camera which has bigger distance than (c) because of the background bias.

Deep learning based approaches were proposed to learn feature representations for describing persons' visual differences. The similarities between person images can then be efficiently calculated as the distances between their visual features. However, there remain challenges that existing methods fail to investigate (see Figure 1).

Firstly, the influence of background (context) regions of person images is mostly ignored by existing methods. Person images for re-identification algorithms are usually obtained by cropping at person detection bounding boxes. Existing methods generally treat the whole person images as individual data samples and therefore all pixels in each image have equal influence. However, in our experiments, we observed that, for a large person image database consisting of person images with different backgrounds, treating all pixels in images equally would bias the learning algorithms to generate high similarities between images with similar

[1] S. Yi and H. Li are the corresponding authors.

[2] This work is supported in part by SenseTime Group Limited, in part by the Shenzhen Municipal Public Security Bureau, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14213616, CUHK14206114, CUHK14205615, CUHK419412, CUHK14203015, CUHK14239816, CUHK14207814, CUHK14208417, CUHK14202217 in part by the Hong Kong Innovation and Technology Support Programme Grant ITS/121/15FX, and in part by the China Postdoctoral Science Foundation under Grant 2014M552339.

backgrounds. We argue that such a phenomenon is a major drawback for existing methods and was not thoroughly investigated before.

Secondly, the background-bias phenomenon can hardly be observed in individual academic datasets. This is because that existing datasets contain person images with similar background captured by a small number of cameras. For instance, CUHK03 [10] dataset consists of images from 2 cameras, CUHK01 [9] is collected from 2 cameras, and Market-1501 [29] is collected from 6 cameras. Deep learning models trained on one dataset usually perform poorly on the other datasets with different background. The background-bias phenomenon cannot be effectively tested on such individual datasets.

Thirdly, most deep learning-based person re-identification methods focus on learning visual features for overall person appearance and lack a mechanism to focus on specific person regions. For instance, two person images might have similar overall appearances but can be easily distinguished based on differences of their hair styles. Although there are attempts to adopt the attention mechanism to automatically identify regions of interest, such implicitly learned regions are not always reliable and sometimes might be associated to unrelated regions.

To investigate and handle the above-mentioned challenges, we study the influence of background (context) information in person images to the re-identification performance, which was ignored in existing work. A deep human parsing network is trained to obtain background and foreground (human) regions of the person images. Based on such human parsing maps, we create background-influence datasets to study the adverse influence of background regions to existing deep learning based person re-identification methods, which leads us to the conclusion of existing methods being biased to learning too much relevance between background appearances. To mitigate the background-bias problem, we propose a novel deep neural network for person re-identification based on the human parsing, which generates human parsing maps depicting background and different foreground regions for person images. We integrate the human parsing maps into our deep neural network to guide feature pooling from intermediate feature maps. This mechanism help the network focus on informative regions of input images for learning more discriminative features for describing person appearance and robust re-identification. Extensive experiments demonstrate the effectiveness of our proposed method on mitigating the background-bias problem and show state-of-the-art performance by our proposed method on multiple datasets.

The contribution of this paper is twofold.

- We investigate the influence of background regions to the person re-identification performance by deep learning-based methods. Background-influence datasets are created based on human parsing maps to identify the background-bias problem.
- A novel deep neural network with a human parsing module and random-background data augmentation are introduced for solving the person re-identification task, which is robust to background variations and is able to focus on informative foreground regions to achieve state-of-the-art performance.

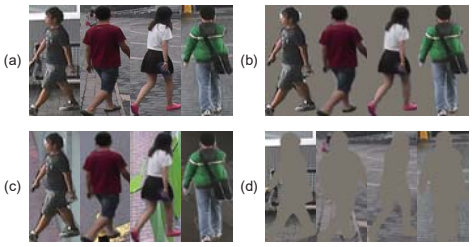

Figure 2. Four examples of our background-influence datasets. For each example, we show (a) the original image; (b) the mean-background image; (c) the random-background image; and (d) the background-only image.

2. Investigations on Background-bias

Deep learning based person re-identification methods focus on learning visual features to distinguish different persons. However, existing methods generally treat the whole input person image as individual data sample for learning. Background and foreground pixels in each image have the same influence to the learning algorithms. The influence of background (context) information to the re-identification performance was not studied in existing work. In order to study such influence and guide the design of more robust re-identification algorithms, new datasets are created based on foreground-background human parsing masks and a series of experiments are conducted on the created datasets to discover the background bias problem.

2.1. Datasets and experimental setup

Four types of datasets are created from existing re-identification datasets (CUHK03 [10] and Market-1501 [29]) based on the human parsing masks produced by [4]. Following the strategy introduced in [24], the created person images from CUHK03 and Market-1501 datasets are merged to the joint training datasets. Joint training datasets should be more suitable to study the influence of background information, because each of the existing dataset is captured from several nearby cameras and the background appearances are quite similar.

Deep human parsing network is utilized to generate foreground and background parsing masks for each image. The following four joint datasets are created based on existing datasets and the parsing masks for the background-

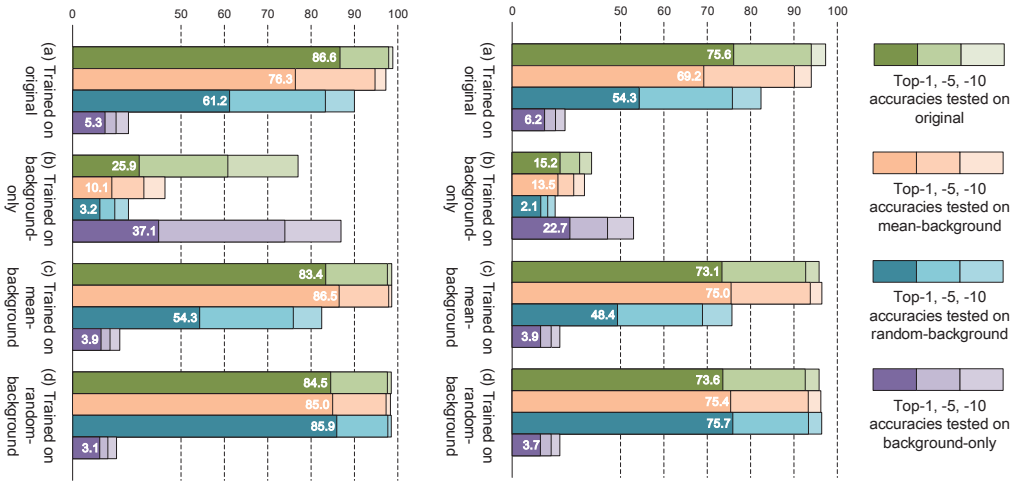

Figure 3. Results of the background-influence investigations. Four models are trained from different background-influence joint datasets and are evaluated on the four modified CUHK03 datasets **(left column)** and the four modified Market-1501 datasets **(right column)**. The models trained with original datasets, background-only datasets, mean-background datasets, and random-background datasets are shown in (a-d), respectively. The results evaluated on original data, mean-background data, random-background data, and background-only data are marked by different colors. The results demonstrate that existing models **fails** due to background-bias and existing datasets **fails** to evaluate the background-bias problem.

influence study. 1) The first joint dataset keeps all person images unchanged (denoted as *original*); 2) the second joint dataset fills the foreground (person) regions with mean pixel value and keep the background (context) regions unchanged (denoted as *background-only*); 3) the third one keeps the foreground regions unchanged and fills the background regions with mean pixel value (denoted as *mean-background*); 4) the fourth one keeps the foreground and replaces the background with one of 100 randomly collected images from the Internet (denoted as *random-background*). Example images from each joint dataset are shown in Figure 2. By training deep neural networks for re-identification on each of the joint datasets, one could better study the impact of background information.

Following the network structure in [24], we train four deep models with cross-entropy classification loss on each of the four joint training datasets. Then the four trained models are tested on the four modified testing datasets from CUHK03 and the four modified datasets from Market-1501 (e.g., original dataset, background-only dataset, mean-background dataset, and random-background dataset). The features from the 2nd topmost layer of the trained deep networks are used as persons' visual feature vector, and visual similarities between person images are calculated as cosine distances between the feature vectors. The testing performance are shown in Figure 3.

2.2. Existing model fails due to background bias

In Figure 3(a), we show the re-identification accuracies by training the deep neural network on the original joint dataset and testing on the four modified datasets from CUHK03 (left column) and the four modified datasets from Market-1501 (right column). When testing the trained deep model on the mean-background and random-background datasets, although the two testing datasets have the same foreground as the original testing dataset, the top-k accuracies drop significantly by a large margin. For the modified CUHK03 testing datasets, the top-1 accuracies decreases by 10.3% on the mean-background dataset and by 28.4% on the random-background dataset. We also observe that testing on the background-only datasets also results in accuracies much higher than random guess, e.g., the top-1 accuracy is 5.3%. We argue that such performance gap from original to mean-background and random-background datasets and the higher-than-random-guess accuracy on the background-only dataset is caused by overfitting the trained model to background appearance. In other words, when training the deep neural networks with the original person images, the trained networks are biased towards capturing too much relevance between background appearances of different images. It means that even though existing methods trained on standard datasets can achieve quite good performance, the trained model may easily **fail** in real application where the background appearance can be quite different for the same person. Moreover, such problem can not be easily observed by existing standard datasets. We call such a problem the *background-bias* problem.

2.3. Background can help distinguish persons?

To further validate our claim on the background-bias problem. We train deep neural networks on the background-only joint dataset and test the trained models on the modi-

fied testing datasets. The results are shown in Figure 3(b). The top-1 accuracies on the testing datasets are significantly higher than those of random guess. For example, the trained model achieves top-1 accuracies of 25.9% and 37.1% on CUHK03 original and background-only datasets. The experimental results validate our claim. It also demonstrates that in existing datasets, the background appearances of the same persons are usually similar and the deep neural networks are easily **biased** by such similar backgrounds.

2.4. A way to eliminating background-bias

A naive solution would be isolating the influence of similar background appearances from training the deep models. We therefore train deep models on the mean-background and the random-background datasets, which fill the background regions with the mean pixel value or replace them with random background images. In this way, we force the deep models to focus only on the foreground regions. The testing results of the trained models are shown in Figures 3(c) and 3(d). The deep model trained on the mean-background dataset performs better on mean-background and random-background test sets than the models trained with the original joint dataset. However, there is still large performance drop when it is applied to the random-background test set. On the other hand, the deep model trained with the random-background joint dataset shows consistent performance on all original, mean-background, and random-background test sets, which demonstrates that training with the random-background dataset could isolate the influence of background similarities.

There are 2%-3% performance drop on the original test sets compared with the deep model trained with original images. The performance drop is also reasonable, because the misleading background information is removed from the model training process. The trained model can only rely on foreground information for the identification task. We believe the trained model from random-background data should be much more robust to new scenes.

3. The Proposed Framework

To solve the discovered background-bias problem, we propose a novel deep neural network with a person-region guided pooling mechanism, which is based on person region parsing maps generated by a human parsing network. We also augment training images with random background to achieve state-of-the-art performance, and more importantly, to achieve robustness against background variations.

3.1. Person-region guided pooling network

The overall structure of our proposed deep network is illustrated in Figure 4. It can be generally decomposed into three parts, the whole-person main network, the person-region parsing network, and the person-region guided pooling sub-network.

The whole-person main network is designed to capture the overall appearance of persons (upper part of Figure 4). It takes person images of size $96 \times 96 \times 3$ as input. The images are first processed by three convolution layers with 5×5 kernels followed by a 2×2 max-pooling layer. The resulting $64 \times 48 \times 48$ feature maps are then input to three inception modules. Each of the inception modules reduces the spatial resolution by half and consists of two blocks, where there are four convolution layers in the first block and three in the second one with the same number of input and output channels. The last inception module is followed by a 6×6 average pooling layer and a fully connected layer to output the final 256-dimensional feature vector. Within the main network, ReLU is utilized as the activation function with Batch Normalization following each nonlinear operation.

The person-region parsing network generates the part regions parsing map for each input person image. Each person is partitioned into three regions, i.e., head, upper-torso, and lower-torso regions. Each region's parsing map is converted into a binary parsing mask to guide the pooling of person appearance feature maps. Each binary parsing mask is first down-sampled and then used to gate the output feature maps from the Inception-1 block of the main network by element-wise multiplication. The resulting three feature maps after gating therefore correspond to visual features from head, upper-torso and lower-torso regions, respectively (left part of Figure 4). The person-region parsing network is pre-trained separately and its parameters are fixed during training the overall network. Example parsing masks by the network are shown in Figure 5.

The person-region guided pooling sub-network obtains visual features for each person region from the feature maps of the main network (see lower part of Figure 4). The sub-network has three branches corresponding to the three person regions, which take the gated output of Inception-1 module in the main network as inputs and has the same structures of Inpcetion-2, Inception-3, global pooling and fully connected layers as the main network. However, the three branches have independent parameters for different person regions, each of which generates a 256-dimensional visual feature for the corresponding region. By utilizing the proposed guided pooling sub-network, each branch is forced to focus on specific regions to learn their corresponding visual features for more accurate re-identification.

The 256-d features from the main network and the three 256-d features from the sub-network are then concatenated and transformed to a final 256-d feature by a fully connected layer. In this way, the visual features capture both whole person as well as person-region appearances. A final linear layer is added and the whole network is trained to classify the person's identifications with the cross-entropy loss. Af-

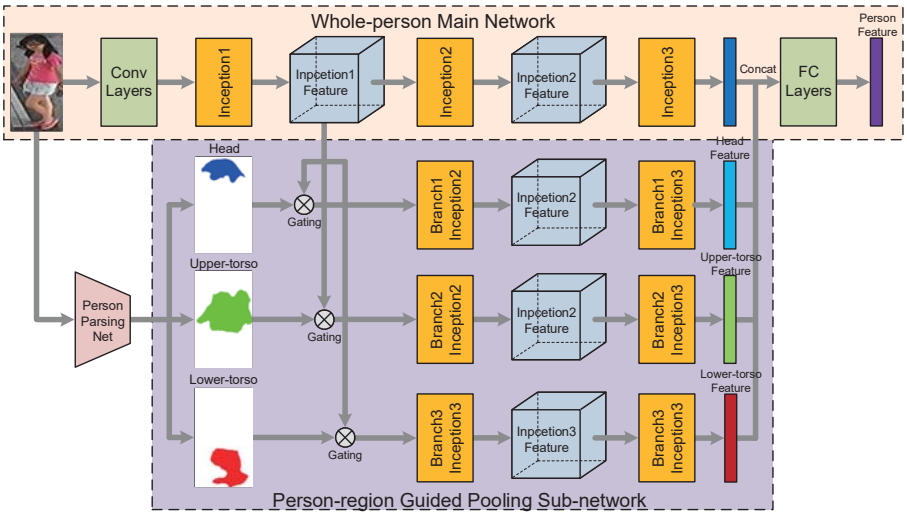

Figure 4. The structure of our proposed person-region guided pooling network, which consists of three main part: the whole-person main network, the person parsing network, and the person-region guided pooling sub-network.

ter training convergence, the 2nd topmost 256-d visual features can be used to represent each person's appearance and the visual similarities are calculated as the cosine distances between such 256-d features.

3.2. Person parsing network

In our proposed framework, the person parsing network is used for two purposes. The first is to generate foreground-background binary parsing maps. The background parsing maps can be used to replace background regions of person images with randomly chosen images to create the random-background dataset as mentioned in the previous section. The second purpose is to generate parsing maps for the three person regions, i.e., the head, upper-torso, and lower-torso regions, which can be used in the overall network for pooling features from specific person regions.

Our parsing network have similar structure as our whole-person main network, with small modifications to output parsing maps with the same spatial size as the inputs and to achieve better parsing performance. 1) We double the number of channels in each convolution layers in the network because of the complexity of the parsing problem. 2) The average pooling layer in the feature learning network is replaced by a pyramid pooling network with 6×6, 3×3, 2×2 kernels to capture context information from different receptive fields. The three pooled feature maps are then upsampled to the same 96×96 size and concatenated along the channel dimension. The 4-class parsing map (background + 3 foreground regions) are obtained by a final 1×1 convolutional layer and a cross-entropy classification loss.

Figure 5. Some examples of the person parsing network results. (Upper row) original images. (Lower row) person parsing maps by our person parsing network.

3.3. Random-background based data augmentation

In our investigations, we observe that the baseline deep model trained on the original images shows poor performance on random-background images. On the other hand, the model trained with random-background datasets shows most robust performance on both original and random-background images. Therefore, we propose to augment training data with random background.

An online random-background data augmentation strategy is adopted by setting a hyper parameters p to represent the probability of replacing one images's background with random-background. Before the training process, for one input image $\mathcal{I}_{in}$, its background-foreground binary mask $\mathcal{M}$ is first generated. During the training process, if $\mathcal{I}_{in}$ is determined to have its background replaced, one background image $\mathcal{B}$ will be randomly selected from 100 background images collected from real surveillance scenes, and one region $\mathcal{R}$ will be randomly cropped from $\mathcal{B}$ with the same height and width as $\mathcal{I}_{in}$. Then the output image $\mathcal{I}_{out}$

can be generated by $\mathcal{I}_{in}$, $\mathcal{M}$, and $\mathcal{R}$ as

$$\mathcal{I}_{out} = \mathcal{I}_{in} \odot \mathcal{M} + \mathcal{R} \odot (1 - \mathcal{M}), \quad (1)$$

where $\odot$ is element-wise multiplication. If $\mathcal{I}_{in}$ is not determined to have its background replaced, $\mathcal{I}_{in}$ is directly used for training. The probability p is set as 0.5 to achieve the optimal performance, and more details will be further discussed.

3.4. Training scheme

The overall framework is trained in four stages using mini-batch Stochastic Gradient Descent. The weight decay is set to 5×10^{-4} for all stages. In Stage I, the person parsing network is pre-trained with LIP [6], HumanParsing [12], and MS-COCO [14] human parsing datasets. The initial learning rate is set to 0.1 and is halved for five times every 20,000 iterations. After convergence, the person parsing network's parameters are fixed for the following stages. In Stage-II, the whole-person main network is train independently with an initial learning rate of 0.1, which is halved for five times at every 20,000 iterations. For Stage-III, the main network's parameters are fixed and only the person-region guided pooling sub-network is trained. The trained parameters of the main network are copied to the three branches in the guided-pooling sub-network as initial point. The initial learning rate is set to 0.01 and decreased to 1/10 of its previous value at every 20,000 iterations for four times. In Stage-IV, the main network and the guided-pooling sub-network are trained in an end-to-end manner with the same learning rate policy as that in Stage-III.

4. Experiments

4.1. Traditional datasets and evaluation protocol

The proposed person re-identification method, together with several comparisons, are evaluated on five public datasets, including CUHK03 [10], CUHK01 [9], VIPER [7], 3dpes [2], and Market-1501 [29]. For Market-1501, we use the official train/test split protocol, and for all the other datasets, we follow the same train/validation/test split protocol as in [24]. We choose the commonly used CMC method for evaluation. Top-1, top-5, top-10, and top-20 accuracies are reported for each dataset.

The HumanParsing [12], LIP [6], and MS-COCO [14] datasets are used to train the foreground-background binary parsing masks for data augmentation. The HumanParsing [12] and LIP [6] datasets are used to train the four-class (background, head, upper-torso, lower-torso) parsing masks for the guided-pooling process.

4.2. Evaluation results on traditional datasets

We compare our proposed method with state-or-art methods, including LOMO [13], Bow-best [29], SCSP [3], TMA [16], WARCA [8], NFST[27], TCP [5], DGD [24], PersonNet [23], GOG [17], S-CNN [22], SSM [1], SpindleNet [28], on all the above-mentioned datasets. The results are listed in Table 1. Following the experimental setup in [24] and [28], our model is trained using the joint dataset (denoted as "Ours") and then tested on individual datasets. The result by including our proposed data augmentation mechanism is denoted as "Ours w/ data aug.".

We can see that our proposed method outperform most

CUHK03	**Top-1**	**Top-5**	**Top-10**	**Top-20**
WARCA-χ^2 [8]	78.4	94.6	-	-
PersonNet [23]	64.8	89.4	94.9	98.2
S-CNN [22]	61.8	80.9	88.3	-
DGD [24]	75.3	-	-	-
SpindleNet [28]	88.5	97.8	98.6	99.2
SSM [1]	76.6	94.6	98.0	-
Ours	**91.7**	98.2	98.7	99.0
Ours w/ data aug.	**92.5**	98.4	98.9	99.5
CUHK01	**Top-1**	**Top-5**	**Top-10**	**Top-20**
NFST [27]	69.1	86.9	91.8	95.4
PersonNet [23]	71.1	90.1	95.0	98.1
TCP [5]	53.7	84.3	91.0	96.3
DGD [24]	66.6	-	-	-
SpindleNet [28]	79.9	94.4	97.1	98.6
Ours	**80.7**	95.0	97.5	98.9
Ours w/ data aug.	**82.5**	96.1	98.2	99.0
VIPeR	**Top-1**	**Top-5**	**Top-10**	**Top-20**
TMA [16]	48.2	-	87.7	95.5
NFST [27]	51.2	82.1	90.5	96.0
LOMO+XQDA [13]	40.0	-	80.5	91.1
GOG+XQDA [17]	49.7	79.7	88.7	94.5
TCP [5]	47.8	74.7	84.8	91.1
SpindleNet [28]	**53.8**	74.1	83.2	92.1
Ours	50.6	70.3	79.1	88.0
Ours w/ data aug.	**51.9**	74.4	84.8	90.2
3DPeS	**Top-1**	**Top-5**	**Top-10**	**Top-20**
WARCA-χ^2 [8]	51.9	75.6	-	-
WARCA-L [8]	43.6	68.3	-	-
SCSP [3]	57.3	79.0	-	91.5
DGD [24]	56.0	-	-	-
SpindleNet [28]	62.1	83.4	90.5	95.7
Ours	**64.1**	87.4	90.4	93.7
Ours w/ data aug.	**65.6**	88.1	91.5	95.0
Market-1501	**Top-1**	**Top-5**	**Top-10**	**Top-20**
WARCA-L [8]	45.2	68.2	-	-
NFST [27]	61.0	-	-	-
S-CNN [22]	65.9	-	-	-
Bow-best [29]	42.6	-	-	-
SpindleNet [28]	76.9	91.5	94.6	96.7
SSM [1]	**82.2**	-	-	-
Ours	80.5	94.3	96.8	98.2
Ours w/ data aug.	**81.2**	94.6	97.0	98.3

Table 1. Experimental results of the proposed method and compared methods on five public datasets. The CMC Top-1, -5, -10, -20 accuracies are reported.

Method	Test on CUHK03	Top-1	Top-5	Top-10
DGD [24]	original	**87.2**	97.4	98.3
DGD [24]	mean-background	77.2	93.7	97.0
DGD [24]	random-background	**64.0**	85.5	91.7
Ours	original	**92.5**	98.4	98.9
Ours	mean-background	91.7	97.7	98.6
Ours	random-background	**91.1**	97.6	98.5
Method	**Test on Market-1501**	**Top-1**	**Top-5**	**Top-10**
DGD [24]	original	**79.1**	94.1	96.6
DGD [24]	mean-background	71.1	90.3	94.2
DGD [24]	random-background	**58.4**	80.1	86.4
Ours	original	**81.2**	94.6	97.0
Ours	mean-background	80.5	93.9	98.2
Ours	random-background	**79.8**	93.5	98.0

Table 2. Experimental results by the DGD [24] model and our proposed method on the proposed CUHK03 [10] and Market-1501 [29] background-influence datasets. The CMC Top-1, -5, -10, -20 accuracies are reported.

comparisons on the testing datasets. Compared with SpindleNet [28], which uses person landmarks to pool features from person regions, our proposed method is able to achieve higher re-identification performance. This is because the human parsing maps are able to provide more accurate part layout than the rough human joints. More importantly, SpindleNet and existing methods do not consider how to handle the background-bias problem. They are all influenced by the same background-bias problem as the baseline deep model in the background-influence section.

4.3. Evaluations on background influence datasets

We also evaluate our proposed method and the compared DGD model [24] on the random-background and mean-background datasets to test its performance against the background-bias phenomenon.

As shown by the results in Table 2, our proposed method can achieve similar performance on the three types of datasets, which demonstrate the robustness of our proposed method and it is not inclined to capture too much background relevance. However, the DGD model [24] suffer great performance drop if adopted to the mean-background and random-background datasets. It demonstrates that our model can be more suitable for real world ReID applications while the testing scenarios have quite different background.

4.4. Structure component analysis

We also conduct a series of experiments on the Market-1501 to study the individual components of the proposed method. The first experiment is to study whether the three branches in the guided-pooling sub-network that try to capture region visual features can indeed provide additional information and improve the re-identification performance of the main network. The accuracies by the main network and by our proposed network are reported in Table 3.

Different part	Top-1	Top-5	Top-10	Top-20
Main-net only	75.6	91.9	95.6	97.0
1-branch foreground	77.5	92.7	95.7	97.6
2-branch upper+lower	80.0	94.2	96.8	98.2
Ours (3-branch)	80.5	94.3	96.8	98.2

Table 3. Experimental results by different guided-pooling sub-network structures on Market-1501 [29] dataset. The CMC Top-1, -5, -10, -20 accuracies are reported.

Aug. strategies	Top-1	Top-5	Top-10	Top-20
Main-net only	75.6	91.9	95.6	97.0
online (0.25)	79.2	93.7	96.3	97.9
online (0.5)	79.5	93.9	96.5	98.1
online (0.75)	78.8	93.6	96.5	98.1
offline (1:1)	76.1	92.4	95.6	97.6
offline (1:2)	77.3	93.1	96.0	97.7

Table 4. Segmentation performance by different random background augmentation strategies on Market-1501 [29] dataset. The CMC Top-1, -5, -10, -20 accuracies are reported.

The second experiment is designed to study whether the three branches in the guided-pooling sub-network that learn visual features specifically for the head, upper-torso, and lower-torso regions are necessary. We therefore design two additional networks whose guided-pooling sub-networks pool from only two foreground regions (head+upper-torso and lower-torso) or only one foreground region (the whole-person foreground region) instead of the three regions as in our proposed network. The two network are trained in the same way as our proposed network and its accuracies on Market-1501 dataset is reported in Table 3. The results show that our proposed three-region pooling network results in the highest accuracies and the number of person regions should not be decreased. This is because each of the three regions contains discriminative visual features to determine whether two images correspond to the same person or not. Since the features in each region are averaged before visual similarity calculation, decreasing the number of foreground regions would result in averaging unrelated features from multiple person parts.

4.5. Background augmentation strategy analysis

A series of experiments are conducted on Market-1501 [29] to compare different random-background data augmentation strategies, including the proposed online generating strategy and offline generating strategies. For the online generating strategies, different replacing probability p are evaluated including 0.25, 0.5 and 0.75. For the offline generating strategies, different ratio between original data and random-background data are tested, such as 1:1 and 1:2. In order to better analyze the effects of different strategies, all the compared experiments are conducted with the main branch network, and the results are listed in Table 4.

Foreground-background	Acc.	Pre.	Rec.	F1
HumanParsing [12]	0.976	0.965	0.969	0.967
LIP [6]	0.892	0.891	0.893	0.891
MS-COCO [14]	0.908	0.908	0.908	0.908
Four-class	**Acc.**	**Pre.**	**Rec.**	**F1**
HumanParsing [12]	0.965	0.919	0.891	0.904
LIP [6]	0.846	0.809	0.775	0.790

Table 5. Experimental results by foreground-background binary model and four-class parsing model. The mean accuracy (Acc.), precison (Pre.), recall (Rec.) and F1 score (F1) are reported.

From the results, we can see that both the offline and online strategies have better performance than the Main-net model without random background augmentation. Moreover, the online strategies have better performance than offline strategies. Different replacing probability p of online strategies have little impact (1%) on the final results, which demonstrates that the final results are robust to the probability p.

Compared with previous background augmentation strategy [18], we do not use segmentation annotation information on person ReID datasets. State-of-art deep learning segmentation models are used to generate foreground-background masks automatically.

4.6. Human parsing performance

Our human parsing models are evaluated on several standard human parsing datasets. The foreground-background binary parsing model is tested on HumanParsing [12], LIP [6], and MS-COCO [14] datasets, while the four-class (background, head, upper-torso, lower-torso) parsing model is tested on the HumanParsing [12] and LIP [6] datasets. 10% images are randomly selected for testing, and the mean accuracy, mean precision, mean recall, and mean F1 score are listed in Table 5. We can observe that both of the two models can achieve satisfactory parsing performance.

5. Related Work

5.1. Person re-identification

Person re-identification has drawn increasing attention in recent years. Existing person re-identification methods can be generally classified into two categories. The first category of methods focus on learning better feature representations by deep learning. The Deep-ID method [10] was one of the first methods that trains deep neural networks for learning feature representations for person re-identification. [26] trained siamese deep networks to determine if two images are from the same person. [23] proposed an end-to-end deep neural network to simultaneously learn high-level features and a similarity metric for person re-identification. [24] aimed to solve the problem of domain gap between multiple person re-id datasets with a domain guided dropout mechanism. [28] integrated the person pose estimation results into a person re-identification network for pooling part-related features for robust re-identification. [21] presented a pose-driven deep convolutional model, which use weighted global human body and local body parts as feature representations. However, compared with our proposed method, the human pose can only provide general layout of human regions and experimental results show that our person-region guided pooling network could achieve better retrieval accuracy with accurate parsing maps.

The second category of methods aim to learn better distance metric between person images. [20] proposed to learn Mahalanobis distance to measure person similarities. [25] evaluated and used several kernel-based distance learning approaches with a ranking ensemble voting scheme for person re-identification. [13] proposed the Local Maximal Occurrence (LOMO) features and a cross-view quadratic discriminant analysis method for learning subspace and metric. [19] directly optimized the commonly used Cumulative Matching Characteristic (CMC) curve for re-identification with a structural learning based approach.

5.2. Human parsing

The human parsing algorithms aim to label every pixel in an input person image into background or one of the pre-defined person-part label (e.g., head, leg, upper-torso, shoe, etc.). In recent years, the human parsing problem is mainly solved by deep learning based approaches. [15] adopted a parametric and non-parametric approach, which first retreive most similar images from an human parsing database and used deep learning based matching to transfer labels from database images to the input person images. [11] utilized Long Short-Term Memory network to incorporate short-distance and long-distance spatial dependencies for human parsing. [12] proposed a Contextualized Convolutional Neural Network architecture, which integrates the cross-layer context, global image-level context, within-super-pixel context and cross-super-pixel neighborhood context into a unified network for solving the human parsing problem. [6] proposed a new person parsing benchmark and a self-supervised deep learning based approach for this task.

6. Conclusion

In this paper, we for the first time identify and study the background-bias problem in person re-identification when training deep learning models on existing academic datasets. Different background-influence datasets are created to study the phenomenon. Motivated by our study, we proposed the person-region guided pooling network and a random-background data augmentation scheme for robust person re-identification. Extensive experiments and component analysis show the state-of-the-art performance by our proposed method and the necessity of our network design.

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. *arXiv preprint arXiv:1703.08359*, 2017. 6

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011. 6

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1268–1277, 2016. 6

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 2

[5] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 6

[6] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *arXiv preprint arXiv:1703.05446*, 2017. 6, 8

[7] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007. 6

[8] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *European Conference on Computer Vision*, pages 875–890. Springer, 2016. 6

[9] W. Li and X. Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013. 2, 6

[10] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2, 6, 7, 8

[11] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3185–3193, 2016. 8

[12] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):115–127, 2017. 6, 8

[13] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. 6, 8

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 8

[15] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets knn: Quasi-parametric human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1419–1427, 2015. 8

[16] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*, pages 858–877. Springer, 2016. 6

[17] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016. 6

[18] N. McLaughlin, J. M. Del Rincon, and P. Miller. Data-augmentation for reducing dataset bias in person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015. 8

[19] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015. 8

[20] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. Springer, 2014. 8

[21] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3980–3989. IEEE, 2017. 8

[22] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016. 6

[23] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016. 6, 8

[24] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016. 2, 3, 6, 7, 8

[25] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014. 8

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE, 2014. 8

[27] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016. 6

[28] 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. CVPR, 2017. 6, 7, 8

[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 2, 6, 7