International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 482–490
DOI: https://doi.org/10.2991/ijcis.d.201222.001; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Human Body Multiple Parts Parsing for Person Reidentification Based on Xception

Sibo Qiao[1], Shanchen Pang[1,*], Xue Zhai[1], Min Wang[2], Shihang Yu[3], Tong Ding[4], Xiaochun Cheng[5,*]

[1]*College of Computer Science and Technology, China University of Petroleum, Qingdao, Shandong, China*
[2]*College of Control Science and Engineering, China University of Petroleum, Qingdao, Shandong, China*
[3]*College of Mechanical Engineering, Tiangong University, Tianjin, China*
[4]*College of Software, Shandong University, Jinan, Shandong, China*
[5]*School of Science and Technology, Middlesex University, The Burroughs, Hendon, London*

ARTICLE INFO

Article History
Received 09 June 2020
Accepted 21 Dec 2020

Keywords
Person reidentification
Semantic parsing
Global representations
Local representations

ABSTRACT

A mass of information grows explosively in socially networked industries, as extensive data, such as images and texts, is captured by vast sensors. Pedestrians are the main initiators of various activities in socially networked industries, hence, it is very important to quickly obtain relevant information of pedestrians from a large number of images. Person reidentification is an image retrieval technology, which can immediately retrieve target person in abundant images. However, due to the complexity of many important factors especially of changeful poses, occlusion and background clutter, person reidentification still faces extensive challenges. Considering these challenges, robust and distinguishing person representations are hard to be extracted well to identify different people. In this paper, to obtain more discriminative representations, we propose a human body multiple parts parsing (BMPP) architecture, which captures local pixel-level representations from body parts and global representations from whole body simultaneously. Additionally, a straightforward preprocessing method is adopted in this paper to improve the resolution of images in person reidentification benchmarks. To eliminate the negative effects of changeful poses, a simple yet effective representation fusion strategy is used for the original and horizontally flipped images to get final representations. Experimental results indicate that the method proposed in this article attains superior performance to most of state-of-the-art methods on CUHK03 and Market-1501.

1. INTRODUCTION

Nowadays, to ensure the safety of industrial production, surveillance cameras can be seen almost everywhere in industries. With the popularity of surveillance cameras, a huge social industrial network has gradually formed around people. The data such as videos and pictures provided by this network can be used to locate timely the person who causes dangerous events. However, if there is not an effective person retrieval method, a large amount of time is spent in filtering useless information, which leads to a considerable cost in human and material resources during the process. Person reidentification refers to retrieve all the images of the same person from a whole gallery based on a given query person image. The images in benchmarks (e.g., Market-1501 [1] and CUHK03 [2]) are captured by various surveillance cameras with different backgrounds, body poses and angles in socially networked industries. These images always have overlapping fields of view, hence, person reidentification can be viewed as a cross-camera image retrieval task. To complete the image retrieval task, two person benchmarks, namely Market-1501 [1] and CUHK03 [2], which are utilized in this paper. Several person images in the two benchmarks are shown in Figure 1. From Figure 1, the query contains four different person images. Furthermore, the gallery consists of their corresponding retrieved images which are taken from different online cameras.

Currently, online security cameras have been used in public places such as roads, schools and industries, which ensures our society more transparent and safe. The policeman can quickly retrieve the suspected person through person reidentification technology from the large quantity of images when a criminal case occurs. However, many factors render person reidentification an extremely challenging task. Firstly, when a single camera captures a person, the images taken at different times vary greatly due to lighting conditions, occlusion, background clutter, human postures and other factors. Secondly, the effect of aforementioned factors is more remarkable when a person is captured by diverse cameras. Hence, there are two things that can happen, one is that the images of one person captured by two or more surveillance cameras can be identified as different individuals, and the other is that the images captured of two or more different individuals can be identified as one person. Last but not the least, person images taken by surveillance cameras

* *Corresponding author. Email: pangsc@upc.edu.cn and x.cheng@mdx.ac.uk*

Figure 1 Several person images in the two benchmarks. The first column is four query person images. The second to fifth columns correspond to their retrieved images, respectively.

usually have poor qualities, making it considerably difficult to learn effective representations to distinguish identities. Consequently, we argue that an outstanding person reidentification model can obtain excellent representations to identify different individuals.

Several traditional models distinguish one identity from another by learning low-level representations such as color, shape or texture, but the performance of these models is poor [3,4]. Today, due to the rapid development of deep learning, it has been proved that a high-level representation of image learned through deep convolutional neural network (CNN) is more robust. Several methods have made a commendable improvement on the person reidentification and medical and transportation problems based on a designed CNN architecture [5–14]. Recently, the global-level representation of human image is utilized by most of the existing deep learning methods [15–17]. It is considered that the representations learned from deep CNN architecture should capture the most significant cues to the identities of different individuals. However, none of these deep learning methods has a highly satisfactory performance due to the global-level representations extracted including the human body parts and the background regions simultaneously. The background area contains a lot of clutter, which may add several noises to the final global-level representation.

To address the problems mentioned above, local-level representations extracted from human body parts are leveraged by some recent methods to discuss the person reidentification problem [5,8,18–20]. These methods can better locate the human body parts, which can capture discriminative features and reduce the negative effects of clutter to some extent. From the performance of these methods, we can observe that local-level representations have a stronger robustness. By studying these methods, we find that almost all of the works adopt bounding boxes to locate human body parts automatically. However, the bounding boxes are coarse due to including background clutter or incomplete person, which cannot capture exquisite representations of human body parts. In addition, the person reidentification systems proposed in these works are very complex CNN architectures. As described in literature [21], these systems have a good performance due to sub-models that contain lots of complicated training stages.

Through a set of experiments, we observe that the poor resolution of person images is a key factor that affects the accuracy of the model. In this work, we first do a simple preprocessing of person images to improve the resolution of the images. Then, we introduce human body multiple parts parsing (BMPP) architecture that merges human classification model and human parsing model together. The human classification and the human parsing model is used to exploit global-level representations and local-level cues for human body, respectively. The human parsing instead of bounding boxes is used to extract local body information, mainly because human parsing is a pixel-level method that can accurately locate body parts in variable environments. We separate person body into 5 parts such as head, upper-body, lower-body, shoes and foreground, which is inspired by literature [21].

To relieve the complexity of the proposed model, we use a popular deep convolutional model of Xception [22] as the backbone model with minor modifications to research person reidentification. We also demonstrate that the popular model has an excellent performance with no bells and whistles, when it operates on full images in the benchmarks such as Market-1501 [1] and CUHK03 [2].

The main contributions of this paper are as follows:

1. To improve the resolution of images in person benchmarks, we adopt a straightforward super-resolution method to preprocess these person images. The preprocessing strategy is demonstrated that it can make the performance of the model considerably better than the most existing methods through a group of experiments.
2. We propose a human BMPP model, where explores local-level representations from human multiple body parts and global-level representations from whole human body, simultaneously. Human semantic parsing provides complementary representations for the global features.
3. To improve the discrimination of person representations, a representation fusion strategy is proposed in this paper. An original person image and the same image flipped horizontally go through the BMPP architecture to obtain the final representation by the strategy. Our experiments demonstrate that the performance of BMPP with image-flipped technique proposed in this paper outperforms that of BMPP without image-flipped technique.

The remainder of this paper is organized as follows. Several related works on person reidentification are introduced in Section 2. In Section 3, we present the BMPP architecture for person reidentification. In Section 4, we give experimental results to show the performance of the BMPP architecture. In Section 5, we conclude the paper and have a prospect of future work.

2. RELATED WORK

Nowadays, deep learning has penetrated into various computer areas, in particular person reidentification, and plays a significant role in their development. The problem of person reidentification has achieved magnificent progress due to the rapid evolution of deep CNNs. Next, we provide an overview of literature on the person reidentification.

Recently, a growing number of new works have been focused on capturing human body parts to obtain robust representations. In work [18–20,23], to extract human body parts, they adopt predefined horizontal stripes to slice the feature maps of input human image. Specifically, Li *et al.* [18] exploit a spatial transform network [24] to learn human body parts including in three parts such as head-shoulder, upper-body and lower-body. The spatial transform network [24] can adaptively transform and align part-body data in space, improving the accuracy of the model. Such method can only learn human body representations roughly. Taking some improvements, several works [5,16,25,26] explore human body-part cues to research person reidentification. Multiple patches of a human image are extracted in [5,25] to capture local cues. Yao *et al.* [26] use region proposal network to detect body-part regions as well as generate local cues. Xu and Srivastava [27] design an automatic recognition algorithm for signs images, which can accurately extract images regions of interest and automatic recognition images. Generally speaking, these models are very complex and have a multistage training process.

To solve the problem of bounding boxes not accurate enough, several works attempt to use attention mechanism to study person reidentification [28–32]. Li *et al.* [28] propose an integrated attention model, which combines both soft and hard attention mechanism. Liu *et al.* [29] propose a multidirectional attention model, which captures attentive representations through masking various levels of representations with attention map. To preferably extract global-level and local-level features simultaneously, a multi-scale body-part mask guided attention architecture is proposed in [32], in which body parts masks are utilized to direct the training of corresponding attention.

To handle the problem of background clutter caused by bounding boxes, semantic segmentation is first introduced by several works, which has excellent performance on person reidentification task [21,33]. Nowadays, only a few works pay close attention to semantic segmentation in person reidentification. Semantic segmentation is naturally more suitable for locating human body parts than bounding boxes, which is attributed to its pixel-level accuracy. It has stronger robustness for the changes of human body posture. Human semantic parsing model is proposed in [21] to extract local feature masks from human body, which yields state-of-the-art performance.

In this article, we are inspired by the literature [21] to study the person reidentification problem adopting semantic segmentation method. We present an integrated architecture named BMPP, including human parsing model and human classification model. The human parsing model and human classification model are used to extract local-level representations from human multiple body parts and global-level representations from whole human body, respectively. Additionally, to improve the resolution of images in benchmarks, we adopt a simple super-resolution method to preprocess images. At the same time, to eliminate the negative effects of changeful poses, a simple yet effective representation fusion strategy is used for the original and horizontally flipped images to get the final representations. Then, we provide details for our methods on person reidentification.

3. OUR METHODS

In this work, Xception [22] is adopted as a backbone architecture, which is used for both human parsing and human classification models. Hence, we first make a simple introduction to Xception architecture. Subsequently, we describe our human classification and human parsing model in detail. Finally, just like reference [21], we give some details about how to combine the human parsing model with human classification model, which generates our proposed final person reidentification architecture.

3.1. Human Classification Model

In this paper, the backbone architecture of human classification model is Xception [22]. Hence, we give a detail of Xception [22] architecture below. Xception [22] is a 36-layers deep CNN architecture, which is based entirely on depthwise separable convolution layers. The depthwise separable convolution has a similar performance with regular convolution operation yet the former has a smaller parameter count than the latter. Therefore, Xception [22] architecture is considerably less computational cost than current popular neural network framework, just as Inception-V3 [34], ResNet-50 [35] and ResNet-152 [35]. These convolutional architectures are deeper than Xception [22], which Inception-V3 [34] has 48 layers, ResNet-50 [35] has 50 layers and ResNet-152 [35] is a deep convolutional architecture with 152 layers. While being shallower than these popular network architectures, our experiments demonstrate that it gives a better result than these architectures. The quantitative comparison is performed by a set of tests with different choices of backbone architectures in this work.

The Xception [22] architecture is inspired by Inception-V3 [34], where Inception modules are replaced with extremely separable convolutions. In these extremely separable convolutions, 1×1 convolution is first used to map cross-channel dependencies, and then the spatial correlations of each output channel are mapped separately. An extreme convolution is almost consistent with a depthwise separable convolution, an operation that has been used in [22]. The Xception [22] architecture has three modules, which are entry flow, middle flow and exit flow. Given some images as input, they first go through the entry flow, then the outputs of entry flow pass through the middle flow which is repeated 8 times. Finally, the intermediate outputs flow through the exit flow to get extracted features. The three modules consist of 14 submodules, all of which have linear residual connections expect for the first and last submodules. This is only a brief introduction to the concept of Xception [22] architecture. For more information on the framework, readers may refer to [22].

In human classification model, to obtain a pretrained model, we make some modifications to the baseline Xception [22] architecture for person reidentification. The last classification layer is removed by us in the original Xception [22] model. Then, we add a fully-connected layer with 512 neural units followed by a ReLU

nonlinearity after global average pooling layer. To mitigate overfitting of the baseline model, a dropout layer with 0.5 decay rate is adopted after the activation layer. Finally, a fully-connected layer is also added with 2048 neural units after the dropout layer. Hence, we can learn about that the human classification model generates a 2048-D global representation in the end. The 2048-D global representation is passed to a multi-class classification layer with softmax nonlinearity function when we train the human classification network. Simultaneously, we should be aware that the final 2048-D representation is employed to retrieve correct matches of a target person from the benchmarks when testing the performance of the designed model. The performance of designed model is showed in Section 4 when we use Xception [22], Inception-V3 [34], ResNet-50 [35] and ResNet-152 [35] as the backbone architecture one by one.

3.2. Human Parsing Model

To explore local body parts cues for person reidentification, just as literature [21], human parsing is utilized in this work. Human parsing is a pixel-level classification of a person image. It can capture a more comprehensive representation from person image. We insist that pixel-level operation has not only a considerable improvement in the performance than bounding box, but also significant robustness to the variation of person poses and background clutter.

Xception [22] architecture is adopted as the backbone of the human parsing model. And we use the architecture inspired by Deeplabv3+ [36], a prestigious Encoder-Decoder model for semantic segmentation. To apply to the human parsing task, several modifications are made to Xception [22] architecture in this work. The performance of human parsing task depends largely on whether the image with effective resolution is obtained. Hence, in order to reduce the representations loss caused by pooling operations, we replace all the max pooling operations with depthwise separable convolutions. The strides of these convolutions are seted to 2. Simultaneously, extra batch normalization [37] and ReLU method are added after each 3×3 depthwise separable convolution, increasing the validity of forward propagation through the architecture and cut down on overfitting. In this work, to integrate into human classification model, we do not need a parsing output that is the same size as the original input person image. In addition, a parsing output in this literature has 5 object classes that are background, head, upper-body, lower-body and shoes. Therefore, we remove the last upsampling operation and add a convolution layer, including 5 filters with dimensions of 1×1 and stride of 1.

3.3. Overall Architecture for Person Reidentification

The overall architecture for person reidentification proposed in this article consists of two different branches, which are human classification model for global-level representations and human parsing model for local-level body parts representations. The overall architecture is named as BMPP, which is human BMPP for person reidentification. The BMPP architecture is illustrated in Figure 2. As shown in Figure 2, human classification model, based on Xception model, extract global activation values from input person image. Simultaneously, human parsing model, based on Xception* model, obtain masks of different human body parts. The Xception* model is modified from Xception model to parse human body parts, as described in Section 3.2. The final person representation is generated from raw image and its horizontally flipped image through the feature fusion strategy, as described in Section 3.3.

To explore the person body multiple local parts visual cues, we employ the mask maps obtained from human parsing model. These mask maps are associated with 5 different regions that are foreground, head, upper-body, lower-body and shoes. In this work, we also argue that segmenting the human body into 5 different parts, such as head, upper-body, lower-body, shoes and foreground, can bring a better performance to the BMPP model without large computational cost. In BMPP, we do the same operations to semantic regions as the literature [21]. The output activations of human classification model are pooled with one of five mask maps multiple times. This contrasts with global average pooling, which is unconscious of the spatial domain activation. In this article, we also see the specific pooling activation within 5 different parsing regions as a weighted sum operation. It is easy to understand that the mask maps in here are utilized as weights to get more effective and robust person representations. In other words, the weighted sum operation is consistent with a matrix multiplication between the output of human classification and human parsing architecture. We can find that their corresponding spatial domain is flattened through the matrix multiplication operation, and five 2048-D feature vectors are generated that represents one human body. The five 2048-D feature vectors respectively represent foreground, head, upper-body, lower-body and shoes. Next, the element-wise max operation is applied to the feature vectors such as head, upper-body, lower-body and shoes, generating one 2048-D feature vector named wise-element. In the end, we obtain two 2048-D feature vectors that are foreground and wise-element.

In this work, a final person image representation is identified by an original person image and the same person image flipped horizontally, shown in Figure 2. A detail of description of such a procedure is given as follows. Firstly, the two representations of foreground-original and wise-element-original are generated simultaneously by the BMPP architecture. Similarly, the two representations of foreground-flipped and wise-element-flipped are generated simultaneously by a horizontally flipped person image through the BMPP architecture. Then, the two representations of foreground-original and foreground-flipped are computed by the feature fusion strategy getting the final representation of foreground. The final representation of wise-element is computed by two representations of wise-element-original and wise-element-flipped in the same way. Finally, we concatenate the representation of foreground with the outcome of wise-element getting the final one 2048-D person image representation. Our experiments demonstrate that the performance of BMPP with image-flipped technique proposed in this paper is superior to that of BMPP without image-flipped technique. The strategy of feature fusion is described as follows:

$$X_a = \frac{X_1}{\sqrt{\sum_{i=1}^{n} x_i}} + \frac{X_2}{\sqrt{\sum_{j=1}^{n} x_j}}, x_i \in X_1, x_j \in X_2 \tag{1}$$

$$X = \frac{X_a}{\sqrt{\sum_{k=1}^{n} x_k}}, x_k \in X_a \tag{2}$$

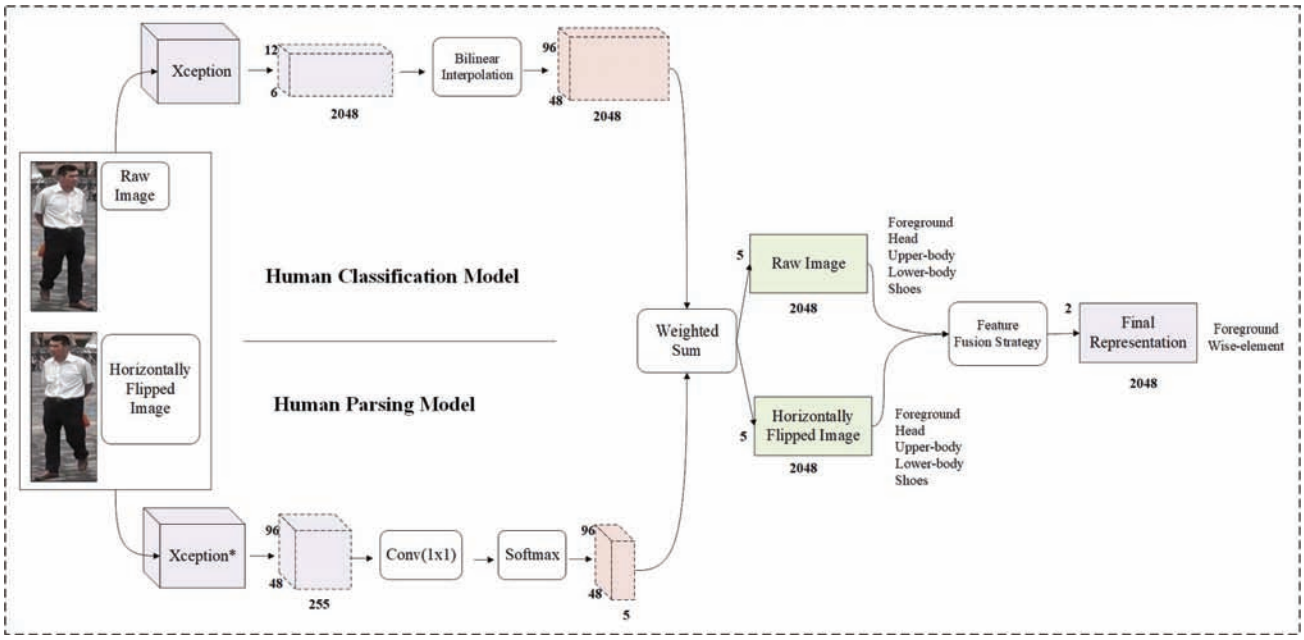

Figure 2 | BMPP architecture proposed in the paper.

Where both X_1 and X_2 are feature vectors of an original person image and the same person image flipped horizontally, respectively. n represents the dimension of a feature vector, x is a value of a feature vector.

4. EXPERIMENTS

4.1. Data Sets and Evaluation Strategy

To evaluate BMPP architecture proposed in this article, experiments are performed on two public person reidentification benchmarks, Market-1501 [1] and CUHK03 [2]. In Market-1501 [1] benchmark, 32,668 person images of 1,501 identities are included in the benchmark, which are captured by 5 high-resolution cameras and 1 low-resolution camera. Additionally, Deformable Part Model (DPM) [38] is used in this dataset to acquire bounding boxes of person, which generates several misaligned bounding boxes. In this work, we adopt 751 identities that have 12,936 person images to train our model. When we test the model, we divide person images of 750 identities into gallery data sets and query data sets. 19,734 and 3,368 person images consist in the gallery data sets and the query data sets respectively. These images are not used during training our model. The CUHK03 [2] dataset consists of 1,467 identities captured by 6 various cameras with a total of 13,164 person images. In this dataset, each person is recorded from 2 different views. On the average, each identity has 4.48 images from each view.

To estimate the performance of our method, two popular methods are adopted in this paper, which are Cumulative Matching Characteristic (CMC) and mean average precision (mAP). In the evaluation report, we give rank-1, rank-5, rank-10 and mAP results, which all are under single-query setting.

4.2. Training the Network

Keras, which is a deep learning framework, is adopted in this paper to deploy our BMPP architecture for these experiments. To improve the resolution of images in benchmark datasets, we adopt the residual dense network (RDN) [39] to preprocess these images. The RDN is a super-resolution CNN, which can generate a high-resolution image from its low-resolution image. Therefore, we obtain relatively higher-resolution person images with a scale magnified 4 times via RDN, which the model proposed in this paper can be trained with these higher-resolution images to achieve better performance. In short, all experiments in this article are conducted on the preprocessed person images.

To train the BMPP architecture proposed in this work, we first train one of the branches of BMPP, namely human classification model, on full person images. We train the human classification model using input images of size 384×192. The training data sets are divided into many mini-batches, and a mini-batch size is set to 32. The Adam [40] algorithm is adopted to update weights, which are initialized with pretrained parameters on ImageNet [41]. In addition, to further fight overfitting, the data augmentation technique is used in this paper to increase human data sets.

Then, we train another branch of BMPP named human parsing model on the large Active Template Regression (ATR) dataset [42] for human parsing. Totally, 17,709 images with 17 semantic labels are included in the ATR dataset [42]. In this paper, we use 12,000 images for training, 5,000 images for validation and 709 images for testing. When we train the model, a mini-batch size is set to 12 in this work. To parse 5 human body parts for person reidentification, we make several modifications to 17 semantic labels. The original head parts, such as hat, hair, sunglasses, face and scarf, are all grouped into the head class. The original upper-body parts, such as upper-clothes, left-arm and right-arm are labeled upper-body class.

Figure 3 Examples of parsing mask generated by our human parsing model. The original person images are shown in the first row, and its parsing masks are shown in the second row. The masks of the third row are generated by original person images after supper-resolution.

The original lower-body regions, such as skirt, pants, dress, belt, left-leg and right-leg, are merged into lower-body class. The original left-shoe and right-shoe classes are grouped into shoe class. To reduce the impact of bags, we put the bags in the same class as background in this work. We train the human parsing model using input images of size 512 × 512. The Adadelta [43] algorithm is adopted to update weights initialized with pretrained parameters on ImageNet [41]. After that, we use the trained model to obtain 5 body parts for person reidentification. Experiments show that the human parsing model is able to extract various body parts well. As shown in Figure 3, the human parsing model segments body parts on several person images in Market-1501 benchmark. We observe that the effect of segmentation on the preprocessed person images is superior to that of original person images.

In both training stages, the learning rate is set to 0.001 at first. We reduce learning rate by a factor of 0.2 when the validation loss has stopped declining. The Early Stopping technique is used with patience 10 when the validation accuracy has stopped improving. After training human classification model and human parsing model, the BMPP architecture aggregated of these two models is fine-tuned on Market-1501 and CUHK03 datasets separately.

4.3. The Performance of BMPP

We have a further study of the performance of various baseline models. In the Table 1, we show the results achieved by different baseline architectures based on the original-resolution person benchmarks. Table 2 shows the effect of various baseline architectures based on the super-resolution person benchmarks. From

Table 1 The performance of original-resolution person reidentification benchmarks on various baseline models.

Market-1501			
Model	**mAP (%)**	**Rank-1**	**Rank-10**
Xception	65.33	85.34	98.35
Inception-V3	62.32	84.30	96.20
ResNet-50	58.35	73.34	95.98
ResNet-152	**65.54**	**85.89**	**98.43**
CUHK03			
Xception	–	86.63	98.12
Inception-V3	–	85.09	97.35
ResNet-50	–	84.45	94.04
ResNet-152	–	**86.85**	**98.98**

mAP, mean average precision.

Table 2 The performance of super-resolution person reidentification benchmarks on various baseline models.

Market-1501			
Model	**mAP (%)**	**Rank-1**	**Rank-10**
Xception	**74.85**	**90.05**	**98.95**
Inception-V3	73.06	88.87	97.00
ResNet-50	66.32	85.10	95.95
ResNet-152	72.95	88.33	98.88
CUHK03			
Xception	–	90.87	98.80
Inception-V3	–	89.91	98.41
ResNet-50	–	85.88	99.19
ResNet-152	–	**91.01**	**99.27**

mAP, mean average precision.

Tables 1 and 2, we observe that the baseline models, owing to the representation fusion strategy, outperform most of the current state-of-the-art. The results show that the performance of the model is greatly improved by using super-resolution preprocessing. Furthermore, we can draw a conclusion that Xception [22] has an extremely competitive effectiveness than ResNet-152 [35], in despite of its shallower framework. It also considerably outperforms Inception-V3 [34] and ResNet-50 [35], which has almost the same depth as those two backbone models. Additionally, we observe that the size of Xception [22], Inception-V3 [34], ResNet-50 [35] and ResNet-152 [35] is 88MB, 92MB, 98MB and 232MB respectively. We can conclude that the Xception [22] is the lightest among these models. Therefore, from the computation and performance point of view, Xception architecture is employed in our backbone architecture.

In the Table 3, we show the performance comparison between the BMPP architecture and the Xception baseline model. From the Table 3, we observe that the performance of BMPP architecture is superior to that of the Xception model, regardless of whether the image-flipped technique is adopted. The BMPP architecture with and without image-flipped technique is denoted as BMPP $^{w/flipped}$ and BMPP $^{wo/flipped}$ respectively. From the point of view, the human parsing adopted in BMPP has a significantly positive effect on the baseline model for person reidentification. In addition, the performance of BMPP $^{w/flipped}$ outperforms BMPP $^{wo/flipped}$, which demonstrates that the representation fusion strategy aggregating original and horizontally flipped image representation is more effective for person reidentification.

In the Figure 4, we show the top-5 retrieve results achieved by the BMPP architecture. From Figure 4, we list four person images from query set on Market-1501 and CUHK03. The images from gallery set are retrieved given a query image. As shown in Figure 4, the red rectangle represents the correct retrieved image. On the contrary, the green rectangle represents the incorrect matching image. The top-5 retrieve results for the two person images from Market-1501 are all correct. However, there is a matching error for one person images from CUHK03. From the picture content, they may have great similarity in appearance, which makes the BMPP to mismatch images of different pedestrians.

4.4. Comparison with the State-of-the-Art

Our proposed human BMPP method is compared with current state-of-the-art on CUHK03 and Market-1501 benchmarks. Table 4 shows a performance comparison between the BMPP architecture and the current state-of-the-art. We can observe that the BMPP architecture proposed in this paper outperforms most of state-of-the-art with large margin, adopting our super-resolution preprocessing and representation fusion strategy. In addition, the performance of our model is better than that of SPReID [21] on Market-1501, but not as good as SPReID [21] on CUHK03. We

Figure 4 | The top-5 ranking list for the query images on Market-1501 and CUHK03 by our body multiple parts parsing (BMPP). The first column is query images. The gallery images are in the second to fifth column

Table 3 The performance of BMPP architecture on the person reidentification benchmarks.

Market-1501			
Model	**mAP (%)**	**Rank-1**	**Rank-10**
Xception	74.85	90.05	98.95
BMPP$^{wo/flipped}$	80.02	91.32	97.75
BMPP$^{w/flipped}$	**81.50**	**93.04**	**98.01**
CUHK03			
Xception	–	90.87	98.80
BMPP$^{wo/flipped}$	–	85.88	98.79
BMPP$^{w/flipped}$	–	**92.67**	**98.95**

BMPP, body multiple parts parsing; mAP, mean average precision.

Table 4 The performance comparison of our model with the state-of-the art.

Market-1501				
Model	**mAP (%)**	**Rank-1**	**Rank-5**	**Rank-10**
BoW+Kissme [1]	20.80	44.40	63.90	72.20
Li *et al.* [18]	57.50	80.30	–	–
MGCAM [33]	74.30	83.80	–	–
HA-CNN [28]	75.70	91.20	–	–
AlignedReID [23]	79.40	91.00	96.30	–
SPReID [21]	81.34	92.54	97.15	**98.10**
BMPP$^{w/flipped}$	**81.50**	**93.04**	**97.35**	98.01
CUHK03				
HA-CNN [28]	–	44.40	–	–
MGCAM [33]	–	50.14	–	–
FT-JSTL+DGD [17]	–	75.30	–	–
AlignedReID [23]	–	88.30	97.10	98.50
HydraPlus [29]	–	91.80	98.40	99.10
SPReID [21]	–	**93.89**	**98.76**	**99.51**
BMPP$^{w/flipped}$	–	92.67	97.65	98.95

BMPP, body multiple parts parsing; mAP, mean average precision; MGCAM, mask-guided contrastive attention model; HA-CNN, harmonious attention convolutional neural network.

think that the SPReID [21] model, thanks to train SPReID [21] model on 10 various benchmarks and fine-tune on the Market-1501 and CUHK03 benchmarks, slightly outperforms our model.

5. CONCLUSION

In this paper, we propose a human BMPP architecture, which includes human classification model and human parsing model, simultaneously. Human classification model and human parsing model are used to extract global-level and local-level representations of person images, respectively. In addition, a supper-resolution preprocessing method is adopted to improve the resolution of person benchmarks. Experimental results show that the human parsing model can accurately locate various body parts based on supper-resolution person images. Furthermore, we propose a representation fusion strategy to make the representations more distinguishable. Through an array of experiments, we demonstrate that our architecture has a competitive performance against most of state-of-the-art.

At the same time, our model still has several problems when it explores the local cues of person images. Although several person images with low-resolution have been preprocessed using the supper-resolution method, the BMPP still cannot locate accurately the 5 body parts. In the future, we will adopt attention mechanism (e.g., spatial attention mechanism and channel attention mechanism) to study person reidentification task. The attention mechanism is embedded in the backbone to focus on body local regions of interest, which can efficiently capture more robust and discriminative representations.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Sibo Qiao designs framework of the work; Shanchen Pang makes the manuscript writing; Xue Zhai makes the data experiments; Min Wang and Shihang Yu complete the data analysis and interpretation; Tong Ding constitutes data collection; Xiaochun Cheng is responsible for literature search and research design. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported in part by the Major Science and Technology Innovation Project of Shandong Province (2019TSLH0214), the Tai Shan Industry Leading Talent Project (tscy20180416) and the National Natural Science Foundation of China under Grant 61873281.

REFERENCES

[1] L. Zheng, L. Shen, L. Tian, *et al.*, Scalable person re-identification: a benchmark, in IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 2015, pp. 1116–1124.

[2] W. Li, R. Zhao, T. Xiao, *et al.*, Deepreid: deep filter pairing neural network for person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 152–159.

[3] M. Farenzena, L. Bazzani, A. Perina, *et al.*, Person re-identification by symmetry-driven accumulation of local features, in IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 2360–2367.

[4] O. Hamdoun, F. Moutarde, B. Stanciulescu, *et al.*, Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, in IEEE International Conference on Distributed Smart Cameras, Stanford, CA, USA, 2018, pp. 1–6.

[5] D. Cheng, Y. Gong, S. Zhou, *et al.*, Person re-identification by multichannel parts-based cnn with improved triplet loss function, in IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1335–1344.

[6] Y. Lin, L. Zheng, Z. Zheng, *et al.*, Improving person re-identification by attribute and identity learning, Pattern Recognition, 95 (2019), 151–161.

[7] T. Matsukawa, E. Suzuki, Person re-identification using CNN features learned from combination of attributes, in International Conference on Pattern Recognition, Cancun, Mexico, 2016, pp. 2428–2433.

[8] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in European Conference on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 791–808.

[9] S. Jacob, V.G. Menon, F. Al-Turjman, *et al.*, Artificial muscle intelligence system with deep learning for post-stroke assistance, IEEE Access. 7 (2019), 133463–133473.
[10] Z. Ullah, F. Al-Turjman, L. Mostarda, Applications of artificial intelligence and machine learning in smart cities, Comput. Commun. 11 (2020), 313–323.
[11] M. Arumugam, A. Kumar, Arrhythmia identification and classification using wavelet centered methodology in ecg signals, Concurr. Comput. Pract. Exp. 32 (2019), e5553.
[12] A.K. Sangaiah, M. Arumugam, G.-B. Bian, An intelligent learning approach for improving ecg signal classification and arrhythmia analysis, Artif. Intell. Med. 103 (2020), 101788.
[13] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access. 7 (2019), 81542–81554.
[14] S. Pang, F. Meng, X. Wang, *et al.*, Vgg16-t: a novel deep convolutional neural network with boosting to identify pathological type of lung cancer in early stage by CTimages, Int. J. Comput. Intell. Syst. 13 (2020), 771–780.
[15] S. Pang, S. Qiao, T. Song, *et al.*, An improved convolutional network architecture based on residual modeling for person re-identification in edge computing, IEEE Access. 7 (2019), 106748–106759.
[16] H. Zhao, M. Tian, S. Sun, *et al.*, Spindle net: person re-identification with human body region guided feature decomposition and fusion, in IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 907–915.
[17] H. Zhao, M. Tian, S. Sun, *et al.*, Learning deep feature representations with domain guided dropout for person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1249–1258.
[18] D. Li, X. Chen, Z. Zhang, *et al.*, Learning deep context-aware features over body and latent parts for person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 7398–7404.
[19] Y. Sun, L. Zheng, Y. Yang, *et al.*, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), in European Conference on Computer Vision, Munich, Germany, 2018, pp. 480–496.
[20] G. Wang, Y. Yuan, X. Chen, *et al.*, Learning discriminative features with multiple granularities for person re-identification, in ACM International Conference on Multimedia, Seoul, South Korea, 2018, pp. 274–282.
[21] M.M. Kalayeh, E. Basaran, M. Gokmen, *et al.*, Human semantic parsing for person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1062–1071.
[22] F. Chollet, Xception: deep learning with depthwise separable convolutions, in IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 1800–1807.
[23] X. Zhang, H. Luo, X. Fan, *et al.*, Alignedreid: surpassing human-level performance in person re-identification, arXiv preprint arXiv:1711.08184, 2018. https://arxiv.org/abs/1711.08184
[24] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, Spatial transformer networks, Neural Information Processing Systems, Montreal, Canada. 2015, pp. 2017–2025. https://arxiv.org/abs/1506.02025
[25] F. Zhu, X. Kong, L. Zheng, *et al.*, Part-based deep hashing for large-scale person re-identification, IEEE Trans. Image Process. 26 (2017), 4806–4817.
[26] H. Yao, S. Zhang, Y. Zhang, *et al.*, Deep representation learning with part loss for person re-identification, IEEE Trans. Image Process. 28 (2019), 2860–2871.
[27] H. Xu, G. Srivastava, Automatic recognition algorithm of traffic signs based on convolution neural network, Multimedia Tools Appl. 79 (2020), 11551–11565.
[28] W. Li, X. Zhu, S. Gong, *et al.*, Harmonious attention network for person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2285–2294.
[29] X. Liu, H. Zhao, M. Tian, *et al.*, Hydraplus-net: attentive deep features for pedestrian analysis, in IEEE International Conference on Computer Vision, Venice, Italy, 2017, pp. 350–359.
[30] S. Li, S. Bak, P. Carr, *et al.*, Diversity regularized spatiotemporal attention for video-based person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 369–378.
[31] F. Yang, K. Yan, S. Lu, *et al.*, Attention driven person re-identification, Pattern Recognit. 86 (2019), 143–155.
[32] H. Cai, Z. Wang, J. Cheng, Multi-scale body-part mask guided attention for person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 1555–1564.
[33] C. Song, Y. Huang, W. Ouyang, *et al.*, Mask-guided contrastive attention model for person re-identification, in IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1179–1188.
[34] C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, Rethinking the inception architecture for computer vision, in IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 2818–2826.
[35] K. He, X. Zhang, S. Ren, *et al.*, Deep residual learning for image recognition, in IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778.
[36] L. Chen, Y. Zhu, G. Papandreou, *et al.*, Encoder-decoder with atrous separable convolution for semantic image segmentation, in European Conference on Computer Vision, Munich, Germany, 2018, pp. 833–851.
[37] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, Int. Conf. Mach. Learn. 37 (2018), 448–456. https://arxiv.org/abs/1502.03167
[38] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, *et al.*, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010), 1627–1645.
[39] Y. Zhang, Y. Tian, Y. Kong, *et al.*, Residual dense network for image super-resolution, in IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2472–2481.
[40] D. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2015. https://arxiv.org/abs/1412.6980
[41] J. Deng, W. Dong, R. Socher, *et al.*, Imagenet: a large-scale hierarchical image database, in IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248–255.
[42] X. Liang, S. Liu, X. Shen, *et al.*, Deep human parsing with active template regression, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015), 2402–2414.
[43] M.D. Zeiler, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701, 2012. https://arxiv.org/abs/1212.5701