applied sciences

MDPI

Article

Parts Semantic Segmentation Aware Representation Learning for Person Re-Identification

Hua Gao [1], **Shengyong Chen** [1,2,*] **and Zhaosheng Zhang** [3,4]

[1] College of Computer Science, Zhejiang University of Technology, Hangzhou 310014, China; ghua@zjut.edu.cn
[2] College of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China
[3] Zhejiang Jieshang Vision Technology Co., Ltd., Hangzhou 311121, China; zszhang@icarevision.cn
[4] Collaborative Innovation Center for Economics Crime Investigation and Prevention Technology, Nanchang 330000, China
* Correspondence: sy@ieee.org

Received: 26 February 2019; Accepted: 22 March 2019; Published: 25 March 2019

Abstract: Person re-identification is a typical computer vision problem which aims at matching pedestrians across disjoint camera views. It is challenging due to the misalignment of body parts caused by pose variations, background clutter, detection errors, camera point of view variation, different accessories and occlusion. In this paper, we propose a person re-identification network which fuses global and local features, to deal with part misalignment problem. The network is a four-branch convolutional neural network (CNN) which learns global person appearance and local features of three human body parts respectively. Local patches, including the head, torso and lower body, are segmented by using a U_Net semantic segmentation CNN architecture. All four feature maps are then concatenated and fused to represent a person image. We propose a DropParts method to solve the parts missing problem, with which the local features are weighed according to the number of parts found by semantic segmentation. Since three body parts are well aligned, the approach significantly improves person re-identification. Experiments on the standard benchmark datasets, such as Market1501, CUHK03 and DukeMTMC-reID datasets, show the effectiveness of our proposed pipeline.

Keywords: person re-identification; representation learning; parts alignment; occlusion handling

1. Introduction

Person re-identification is a typical computer vision problem which aims at matching pedestrians across disjoint camera views. It has attracted a lot of research interest due to its significant application potentials, such as in visual recognition and surveillance [1,2]. One of the most important tasks that person re-identification is shouldering is to learn generic and robust feature representations of people. Recently, the methods based on deep learning learn feature representation directly from tasks and have shown significant improvement compared with hand-crafted feature extractors. State-of-the-art CNN network architectures, such as Inception network [3–5], Resnet network [6,7], are applied to learn feature representation for person re-identification.

Person re-identification is a challenging task due to the misalignment of body parts caused by poses variation, background clutter, detection errors, camera point of view variation, different accessories and occlusion. Figure 1 illustrates some examples in person re-identification tasks. There are images of two persons in Figure 1, with one person in each row. Images in the top row are from the Market1501 dataset [8], and those in the bottom row are from DukeMTMC-reID dataset [9]. In the top row, poses, background, detection, camera viewpoints and accessories are quite different. In the bottom row, poses variation, background clutter, detection errors, camera point of view variation occlusion also occur. Part misalignment occurs frequently and degrades the performance of person re-identification.

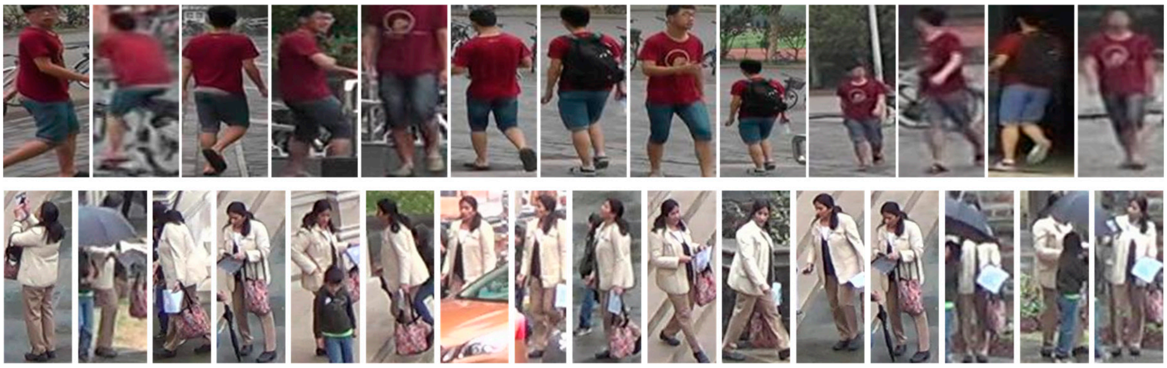


Figure 1. Examples of part misalignment caused by pose variation, background clutter, detection errors, camera point of view variation, different accessories and occlusion. Images in top row are from Market1501 dataset [8], and images in bottom row from DukeMTMC-reID dataset [9].

To solve this problem, many scholars focus on person re-identification based on part alignment recently. Some methods divide the person image into many stripes or grids to reduce the effects of part misalignment [7,10]. The division of grids or strips is predefined and heuristic, which can't locate the parts precisely. Pose-based methods [5,11] employ a pose estimation model to infer corresponding bounding boxes. However, parts missing is ineluctable; it causes the convolutional neural network to not work properly.

This paper focuses on the problem of body part misalignment. It proposes a human parts semantic segmentation aware representation learning method for person re-identification. We employ semantic segmentation network to infer corresponding bounding boxes, and propose a DropParts method to solve the part missing problem. Experiments on the standard benchmark datasets show the effectiveness of our proposed pipeline. The contributions of this paper are as follows:

(1) We design a four-branch convolutional neural network to deal with parts misalignment problem. The four-branch CNN network learns a person's appearance features globally and using the features of three local body parts. The bounding boxes of three body parts are inferred from human parts semantic segmentation results, which are learned with a popular U_Net [12] semantic segmentation network.

(2) We propose a DropParts method to solve the part missing problem, with which the local features are weighed due to the appearance vector and fused with global feature. The DropParts method makes the four-branch convolutional neural network work properly when part missing occurs. On the other hand, it improves the performance of person re-identification.

2. Related Work

In this section, we present a brief review of works in feature exaction and part alignment for person re-identification.

At the beginning of the study, hand-crafted features extractors, such as color histogram [13], Scale-Invariant Feature Transform (SIFT) [14], Local Binary Patterns (LBP) features [15], Bag of Word (BoW) [8] and Local Maximal Occurrence (LOMO) [16] are employed for the person representations. Recently, the methods based on deep learning learn feature representation directly from tasks and have shown significant improvement compared with hand-crafted feature extractors. All kinds of popular CNN network architectures, such as Inception network [3–5], Resnet network [6,7], are applied to learn feature representation for person re-identification. Additionally, different loss functions, such as Softmax loss [17], Siamese loss [4,18], Cluster loss [19], Triplet loss [20] and their combination [21] are used to improve the discriminative feature learning in person re-identification tasks. Softmax loss [17] function is the common loss function used in recognition tasks.

Many scholars focus on person re-identification based on part alignment [7,10,17,22,23]. Early works divide the person image into many stripes or grids to reduce the effects of part misalignment. Article [10] divides the person image into three horizontal stripes and extracts CNN features of each strip. After that, they concatenate and fuse them with a fully connected layer to represent a person image. Meanwhile, DeepReID method [17] also divides the person image into horizontal stripes and carries out patches matching within each stripe. On the other hand, SpindleNet [22] takes the human body structure information into person re-identification pipeline to help align body part features of images. The features of different semantic levels are merged by a tree-structured fusion network based on human body region which is guided by multi-stage feature decomposition and tree-structured competitive feature fusion, to represent a person image. IDLA method [23] captures local relationships between the two input images on the basis of mid-level features of each input image, and computes a high-level summary of the outputs of this layer by a layer of patch summary features, which are then spatially integrated with subsequent layers. More stripes- and grids-based methods can be found in [7]. Although stripes- and grids-based methods reduce the risk of part misalignments, the division of grids or strips is predefined and heuristic, which can't locate parts precisely.

Pose-based person re-identification methods leverage external cues from human pose estimation. Article [11] incorporates a simple cue of the person's coarse pose (i.e., the captured view with respect to the camera) and the fine body pose (i.e., joint locations) to learn a discriminative representation of person image. PDC method [5] leverages the human part cues to alleviate the pose variations and learn feature representations from both the global image and different local parts. To match the features from global human body and local body parts, a pose driven feature weighting sub-network is further designed to learn adaptive feature fusions. Pose-based methods leverage human pose estimation to infer the location of body parts. However, parts missing is ineluctable, it makes the convolutional neural network not work properly. And it is hard to find the right body part in the crowd because there may be several parts of the same semantic label in an image.

Attention mechanism has a large impact on neural computation, which selects the most pertinent pieces of information and focuses on specific parts of their visual inputs to compute the adequate responses [24–29]. Article [25] decomposes the human body into regions following the learned person re-identification sensitive attention maps. Accordingly, it computes the representations over the regions, and aggregates the similarities computed between the corresponding regions of a pair of probe and gallery images as the overall matching score. The PersonNet method [26] learns attention map from different scales for each module and applies the attention map to different layers of the network. At the end, they learn features by fusing three attention modules with Softmax loss. Moreover, HydraPlus-Net method [27] has several local feature extraction branches which learn a set of complementary attention maps in which hard attention is used for the local branch and soft attention for the global branch, respectively. More methods based on the attention mechanism can be found in [28,29]. Methods based on the attention mechanism highlight the important region information of person images, but they also increase the number of feature maps by several times, and bring risks of over-fitting.

We use a semantic segmentation network to infer human body parts in this paper. Due to the ensemble effects of label of each pixel, bounding boxes inferred from semantic segmentation map are stable. We propose a DropParts method to solve the part missing problem; the method makes the four-branch convolutional neural network work properly when part missing occurs.

3. The Proposed Method

3.1. Overview of the Proposed Method

Given a probe person image, person re-identification targets the most similar persons from gallery sets according to the distance between appearance representations. Our object is to learn the generic and robust feature representations of person.

Figure 2 illustrates the architecture of the proposed parts aware person re-identification network, consisting of four CNN branches which learn person appearance and three body parts feature maps. The four feature maps are fused to an image descriptor. Three local patches, including head patch, torso patch and lower-body patch, are inferred from a semantic segmentation map. Four image patches, including whole person image and three image patches, are resized to the fixed size and then input into the proposed four-branch network. Each branch learns the representation of one part and finally is fused by a concatenation layer and a fully connected layer. A softmax layer is used to classify person ID.

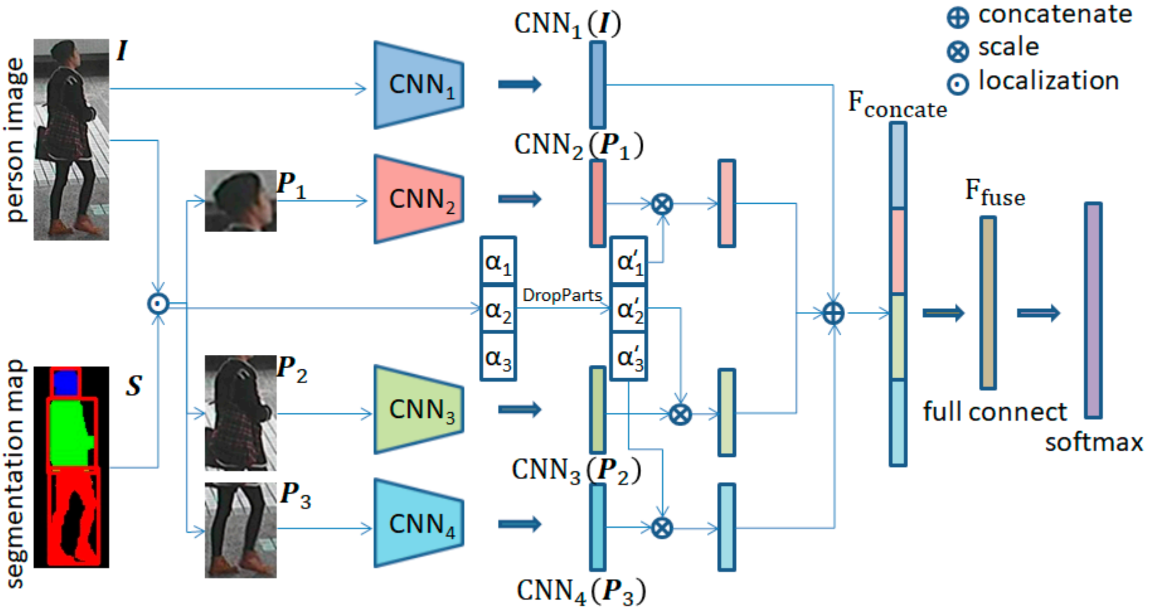


Figure 2. The architecture of proposed parts aware person re-identification network. The network consists of four convolutional neural network branches which learn person appearance and three body parts feature maps respectively, and then fuses four feature maps to an image descriptor. Its input images include a whole person image, head part patch, torso part patch and lower-body part patch. Each branch learns the representation of the whole person image or part patch and finally is fused by a concatenation layer and a fully connected layer.

INPUT: Given an image $\boldsymbol{I} \in R^{M\times N}$ and its semantic segmentation map $\boldsymbol{S} \in \{0,1,2,3\}^{M\times N}$, where semantic labels 0, 1, 2 and 3 represent background, head, torso and lower-body pixels of person image respectively; M and N are the height and width of person image, respectively.

NETWORK: The bounding boxes $\{\boldsymbol{BB}_i\}_{i=1,2,3}$ of the three local parts are fixed by the minimum enclosing rectangles of pixels with the same semantic label. The corresponding image patches are denoted as $\{\boldsymbol{P}_i\}_{i=1,2,3}$ ($\boldsymbol{P}_i$ is a null matrix if its corresponding part is missing). The person image $\boldsymbol{I}$ and three local parts patches $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3\}$ go through four network branches $\{\text{CNN}_i\}_{i=1,2,3,4}$, each image passes through one branch. The feature vectors of four network branches are $\text{CNN}_1(\boldsymbol{I})$, $\text{CNN}_2(\boldsymbol{P}_1)$, $\text{CNN}_3(\boldsymbol{P}_2)$ and $\text{CNN}_4(\boldsymbol{P}_3)$ respectively, $\text{CNN}_i(\cdot) \in \mathrm{R}^{d_i\text{x}1}$.

This paper uses a 3-dimensional vector to represent the absence of all 3 parts:

$$PA = [\alpha_1, \alpha_2, \alpha_3] \tag{1}$$

where $\alpha_i = \begin{cases} 0, \ \mathrm{size}(\mathrm{P_i}) = 0 \\ 1, \ \mathrm{size}(\mathrm{P_i}) > 0 \end{cases}$.

The proposed DropParts method (detailed in Section 3.2) maps parts absence vector *PA* to another 3-dimensional vector $\widetilde{\mathrm{PA}}$:

$$\widetilde{\mathrm{PA}} = \left[\alpha_1', \alpha_2', \ \alpha_3'\right] \tag{2}$$

Scale the part feature vectors and concatenate them with the whole image feature vector, get a fusion vector:

$$\widetilde{\mathrm{F}}_{\mathrm{concate}}(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3) = \begin{bmatrix} Normalize(\mathrm{CNN}_1(\boldsymbol{I})) \\ \alpha_1' \cdot Normalize(\mathrm{CNN}_2(\boldsymbol{P}_1)) \\ \alpha_2' \cdot Normalize(\mathrm{CNN}_3(\boldsymbol{P}_2)) \\ \alpha_3' \cdot Normalize(\mathrm{CNN}_4(\boldsymbol{P}_3)) \end{bmatrix} \tag{3}$$

where $Normalize(\cdot)$ is a normalized operator. This paper uses batch normalization method [27] to normalize features of each part branch.

And then a fully connected layer which functions as metric learning [10,30], is used to fuse the features of the whole person image and three body part patches:

$$\widetilde{\mathrm{F}}_{\mathrm{fuse}}\left(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3 \middle| \tilde{\mathbf{W}}, \tilde{\mathbf{b}}\right) = \tilde{\mathbf{W}}\widetilde{\mathrm{F}}_{\mathrm{concate}}(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3) + \tilde{\mathbf{b}} \tag{4}$$

where $\widetilde{\mathrm{W}} \in \mathrm{R}^{\mathrm{d_f} \times (\mathrm{d_1}+\mathrm{d_2}+\mathrm{d_3}+\mathrm{d_4})}$, $\widetilde{\mathrm{b}} \in \mathrm{R}^{\mathrm{d_f} \times 1}$.

The object of this paper is to learn stable and discriminative person representation $\widetilde{\mathrm{F}}_{\mathrm{fuse}}(\mathrm{I}, \mathrm{P_1}, \mathrm{P_2}, \mathrm{P_3} | \mathrm{W}, \mathrm{b})$.

OUTPUT: At last, a softmax classifier [17] is used to discriminate different person IDs according to their fused CNN features.

3.2. Person Parts Localization and Parts Alignment

Semantic segmentation associates each pixel of an image with a class label. Due to the ensemble effects of label of each pixel, bounding boxes inferred from semantic segmentation map are more stable and accurate than detection methods. This paper uses semantic segmentation map to find the bounding boxes of human body parts.

U_Net [12] is a popular semantic segmentation method which is good at biomedical image segmentation. U_Net architecture consists of a contracting path to capture the context and a symmetric expanding path that enables precise localization. We make three modifications to adopt it for the person parts segmentation. At first, we reduce the number of pooling operators due to the small size of person image. Next, we add two residual structures to compensate for the depth reduction. Third, we do not reduce the size of feature maps by 2 when passing through convolutional layers; as a result, the output segmentation maps have the same size as input images. Figure 3 illustrates the U_Net structure we used. We use its segmentation maps to find the bounding boxes of human body parts. Person images are resized to 192 × 88 and pass through the U_Net network. The size of the output semantic segmentation maps is also 192 × 88, and then the segmentation maps are resized to the same size as the original person images. Figure 4 illustrates some examples of part segmentation by super-pixels.

Bounding boxes of person parts are fixed by the parts semantic segmentation map. For the stable feature extractor, there are two points need to be considered: (1) Large scale differences make extracted feature instable; (2) Large aspect ratio changes lead to part misalignment. This paper gives up two kinds of part regions: (1) the part region whose area bellows 5‰ of its corresponding person image; (2) the part region whose aspect ratio beyond reasonable scope. We set reasonable scopes [0.75, 1.33] for head region, [1, 3] for torso region and [1, 3] for lower-body region. We crop the person images with minimum circumscribed rectangle of its corresponding parts if they are complete.

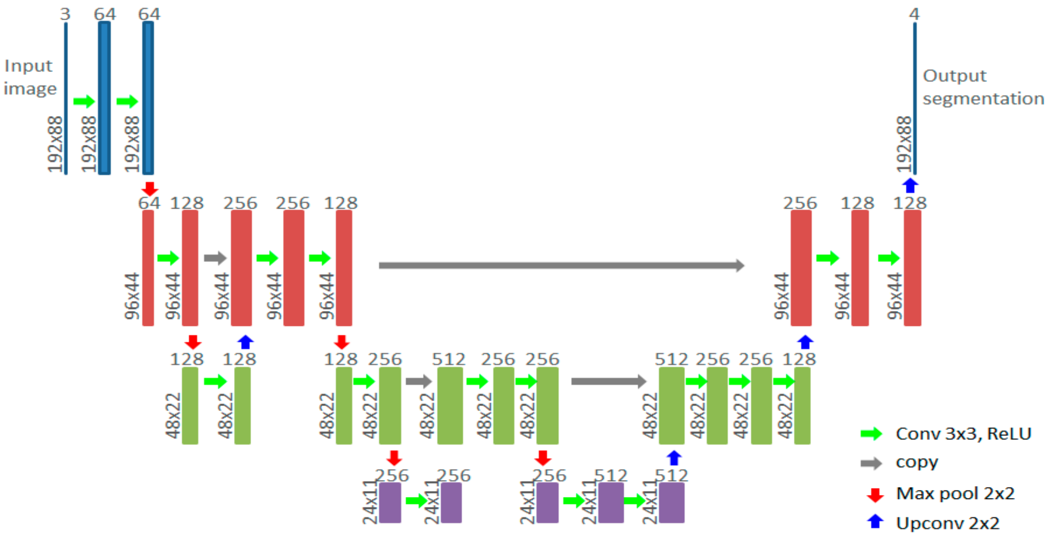


Figure 3. U-net architecture used for part segmentation. Each box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The height and width are provided at the lower left edge of the box. The arrows denote the different operations.

After parts localization, the person image and three local patches are propagated forward through the proposed four-branch network, which completes parts alignment. An example illustrated in Figure 4. Figure 4a is an example of parts misalignment. It is the head region in red rectangle region of left image while it is background in the same location in right image. We locate three body parts and combine them with the whole image as the input of proposed four-branch CNN network. Figure 4b,c illustrate two input of the proposed network, which corresponds to two images of Figure 4a. As seen from Figure 4b,c, the input patches are well aligned.

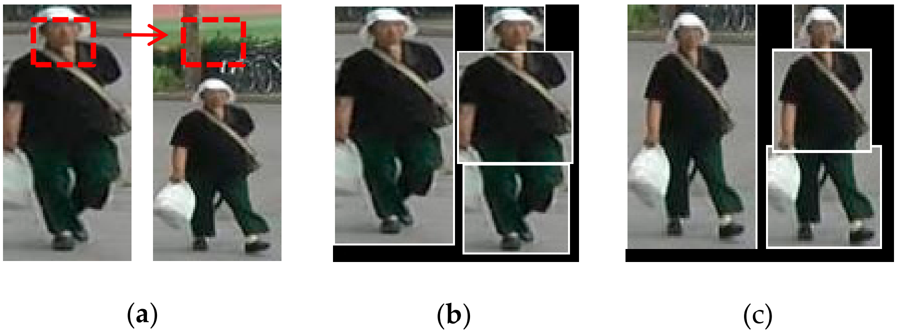


Figure 4. Example of body part alignment based on part segmentation. (**a**) An example of part misalignment; (**b**) Aligned image and part patches of left image in (**a**); (**c**) Aligned image and part patches of right image in (**a**).

3.3. Part Missing Representation and DropParts Method

Part missing is another problem of person re-identification in a complex environment, which happens when meeting with occlusion or parts region is small enough. It degrades the performance of person re-identification. This paper proposes the DropParts Method to solve part missing problem.

A normal feature fusion and metric learning are formulated as follows:

$$\hat{\mathrm{F}}_{\mathrm{concate}}(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3) = \begin{bmatrix} \mathrm{CNN}_1(\boldsymbol{I}) \\ \mathrm{CNN}_2(\boldsymbol{P}_1) \\ \mathrm{CNN}_3(\boldsymbol{P}_2) \\ \mathrm{CNN}_4(\boldsymbol{P}_3) \end{bmatrix} \tag{5}$$

$$\hat{\mathrm{F}}_{\mathrm{fuse}}(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3 | \mathbf{W}, \mathbf{b}) = \overset{\wedge}{\mathbf{W}} \hat{\mathrm{F}}_{\mathrm{concate}}(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3) + \overset{\wedge}{\mathbf{b}} \tag{6}$$

In Equation (5), both normalization and non-normalization of whole person image and part patches vectors $\mathrm{CNN}_1(\boldsymbol{P}_\mathrm{i})$ are feasible, because subsequent metric learning Equation (6) layer will reweigh them.

When meeting with parts missing, the usual method set its corresponding patch or feature a zero matrix or a zero vector. However, it takes the risk of unstable training when all the numbers in a big block are zero. Norms of feature fusion vector $\hat{\mathrm{F}}_{\mathrm{concate}}(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3)$ with zero blocks and without zero blocks are quite different, as a result, parameters $\overset{\wedge}{\mathbf{W}}$ and $\overset{\wedge}{\mathbf{b}}$ cannot meet the demands of parts missing and part non-missing and a compromising solution degrades the performance.

The key is to make the norms of feature fusion vector $\hat{\mathrm{F}}_{\mathrm{concate}}(\boldsymbol{I}, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3)$ stable when part missing happens. In this paper, inspired by Dropout [31], we propose a DropParts method to deal with the parts missing problem.

Dropout [31] is a technique to deal with the over-fitting problem of deep neural networks with a large number of parameters. For example, the l+1 th original hidden layer is formulated as:

$$\mathrm{z}_i^{(l+1)} = \mathrm{W}_i^{(l+1)} \mathrm{y}^{(l)} + \mathrm{b}_i^{(l+1)} \tag{7}$$

$$\mathrm{y}_i^{(l+1)} = \mathrm{f}\left(\mathrm{z}_i^{(l+1)}\right) \tag{8}$$

where $\mathrm{z}_i^{(l+1)}$ denotes ith node of layer l + 1, $\mathrm{y}_i^{(l+1)}$ denotes the ith active value of layer l + 1, $\mathrm{W}_i^{(l+1)}$ and $\mathrm{b}_i^{(l+1)}$ are the weights and biases of layer l + 1 respectively. $f(\cdot)$ is the active function.

The key idea of Dropout is to randomly drop units (along with their connection) with probability p from the neural network during training. During training, dropout samples from an exponential number of different thinned networks. With Dropout, the l + 1 th hidden layer is illustrated as:

$$\mathrm{r}_j^{(l)} \sim \mathrm{Bernoulli}(p) \tag{9}$$

$$\widetilde{\mathrm{y}}^{(l)} = \mathrm{r}_j^{(l)} * \mathrm{y}^{(l)} \tag{10}$$

$$\mathrm{z}_i^{(l+1)} = \mathrm{W}_i^{(l+1)} \widetilde{\mathrm{y}}^{(l)} + \mathrm{b}_i^{(l+1)} \tag{11}$$

$$\mathrm{y}_i^{(l+1)} = \mathrm{f}\left(\mathrm{z}_i^{(l+1)}\right) \tag{12}$$

At test time, approximate the effect of averaging the predictions of all these thinned networks by simply using a single un-thinned network that has smaller weights.

$$\mathrm{w}_{\mathrm{test}}^{(l)} = p \mathrm{W}^{(l)} \tag{13}$$

Dropout significantly reduces risks of over-fitting and gives major improvements over other regularization methods.

In the proposed DropParts method, we formulate the feature fusion of the whole feature and local part features as Equation (14).

$$\mathrm{F}_{\mathrm{concate}}(\boldsymbol{I},\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3)=\begin{bmatrix} Normalize(\mathrm{CNN}_1(\boldsymbol{I})) \\ \frac{\alpha_1}{|PA|}\cdot Normalize(\mathrm{CNN}_2(\boldsymbol{P}_1)) \\ \frac{\alpha_2}{|PA|}\cdot Normalize(\mathrm{CNN}_3(\boldsymbol{P}_2)) \\ \frac{\alpha_3}{|PA|}\cdot Normalize(\mathrm{CNN}_4(\boldsymbol{P}_3)) \end{bmatrix} \tag{14}$$

where $|\cdot|$ is the L1 norm operator. $Normalize(\cdot)$ is a normalization operator, and this paper uses the batch normalization method [32] to normalize the features. Here, normalization $Normalize(\cdot)$ is important, because it maintains the stability of L2-norm of feature vectors. The character of $PA = [\alpha_1, \alpha_2, \alpha_3]$ is normalized too by been divided by its L1 norm. After this, norms of feature fusion vector $\mathrm{F}_{\mathrm{concate}}(\boldsymbol{I},\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3)$ is stable.

Then, the metric learning is:

$$\mathrm{F}_{\mathrm{fuse}}(\boldsymbol{I},\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3|\mathbf{W},\mathbf{b})=\mathbf{W}\mathrm{F}_{\mathrm{concate}}(\boldsymbol{I},\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3)+\mathbf{b} \tag{15}$$

Parts missing samples are not frequent, which leads to imbalanced sample problem. To solve this problem, during training, we drop bins of the absence vector *PA*, and normalize it:

$$\mathrm{r_j} \sim \mathrm{Bernoulli}(p) \tag{16}$$

$$PD=\frac{\boldsymbol{r}*\boldsymbol{PA}}{|\boldsymbol{r}*PA|} \tag{17}$$

$$\widetilde{\mathrm{F}}\left(\mathrm{I},\mathrm{P}_1,\mathrm{P}_2,\mathrm{P}_3\middle|\widetilde{\mathrm{W}},\widetilde{\mathrm{b}}\right)=\widetilde{\mathrm{W}}\begin{bmatrix} Normalize(\mathrm{CNN}_1(\boldsymbol{I})) \\ \alpha_1'\cdot Normalize(\mathrm{CNN}_2(P_1)) \\ \alpha_2'\cdot Normalize(\mathrm{CNN}_3(P_2)) \\ \alpha_3'\cdot Normalize(\mathrm{CNN}_4(P_3)) \end{bmatrix}+\widetilde{\mathrm{b}} \tag{18}$$

Part missing can be regarded as an example of DropParts during training. So, at test time, the fusion feature extractor uses the same parameters $\widetilde{\mathrm{W}}$ and $+\widetilde{\mathrm{b}}$:

$$\widetilde{\mathrm{F}}(\mathrm{I},\mathrm{P}_1,\mathrm{P}_2,\mathrm{P}_3)=\widetilde{\mathrm{W}}\mathrm{F}_{\mathrm{concate}}(\boldsymbol{I},\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3)+\widetilde{\mathrm{b}} \tag{19}$$

4. Experiment

4.1. Network Structure and Experiment Settings

Any network can be used as the baseline of our proposed network. Take 34-layer ResNet [33] as an example, the architecture of our four-branch network and its feature map sizes (on Market-1501 dataset [8]) of input, hidden and output layers are illustrated in Table 1.

The person image size of the input layer (Branch01) is fixed by the average aspect ratio of all images of the dataset, and then the sizes of input layer Branch02, Branch03, Branch04 are fixed by width/2 × width/2, height/2 × width/2, and height/2 × width/2 respectively. The person image and part patches are resized to the input sizes of the corresponding CNN branch. In consideration of the small feature size of res4, we remove the res5 module in Branch02, Branch03, and Branch04. Pool5 layer is the results of global pooling of their previous feature map. We apply our DropParts method to pool5 feature maps of Branch02, Branch03 and Branch04 to get their scaled_pool5 feature maps, then concatenate them with pool5 of Branch01 to get F_concate feature map. An inner product operator is used to map the 1280-dimensional F_concate layer to 512-dimensional F_fuse layer. At last, we use Softmax loss function to train the model. When testing, we use F_concate feature map, normalized by

L2-norm, as the features of a person for person re-identification experiments. Euclidean distance is employed to measure the differences between person features.

Table 1. The architecture of our four-branch network and its feature map size (Market-1501).

NetModule	Branch01	Branch02	Branch03	Branch04
input	3 × 224 × 112	3 × 56 × 56	3 × 112 × 56	3 × 112 × 56
conv1	64 × 112 × 56	64 × 28 × 28	64 × 56 × 28	64 × 56 × 28
res2	64 × 56 × 28	64 × 14 × 14	64 × 28 × 14	64 × 28 × 14
res3	128 × 28 × 14	128 × 7 × 7	128 × 14 × 7	128 × 14 × 7
res4	256 × 14 × 7	256 × 4 × 4	256 × 7 × 4	256 × 7 × 4
res5	512 × 7 × 4	-	-	-
pool5	512 × 1 × 1	256 × 1 × 1	256 × 1 × 1	256 × 1 × 1
scaled_pool5	-	256 × 1 × 1	256 × 1 × 1	256 × 1 × 1
F_concate	1280 × 1 × 1			
F_fuse	512 × 1 × 1			
softmax	751 × 1 × 1			

Our CNN networks are trained on Caffe framework [34] with a TITAN X GPU. We perform stochastic gradient descent (SGD) [35] to perform weight updates. Start with a base learning rate of $\eta_0 = 0.01$ and gradually decrease it as the training progresses using a step policy: $\eta_i = \eta_0 pow(\gamma, floor(i/step\_size))$, where $\gamma = 0.0001$, $step\_size$ = 10,000, i is the current mini-batch iteration. We use a momentum of $\mu = 0.1$ and weight decay $\lambda = 0.0005$.

Training data augmenting often leads to better generalization. We carry out several primary kinds of data augmentation in experiments when training our networks: rotation, shifting, blurring, color jittering and flipping. For rotation, we rotate the image by random degrees between $-30°$ and $30°$. For shifting, we shift the image to the left, right, top and bottom at most 5% of its width or height. For blurring, we blur the image with a 3 × 3, 5 × 5 or 7 × 7 sized Gaussian kernel. For color jittering, we change the brightness, saturation, and contrast by at most 5% of its original value. For flipping, we flip the images horizontally with probability 0.5.

4.2. Modified U_Net Performance

At first, we perform experiments on public LIP dataset [36]. There are 20 semantic labels in LIP dataset: background, hat, hair, glove, sunglasses, upper clothes, dress, coat, socks, pants, jumpsuits, scarf, skirt, face, left-arm, right-arm, left-leg, right-leg, left-shoe, and right-shoe. We change the output num of the last layer in U-net architecture (Figure 3) from 4 to 20, to adopt it for the semantic segmentation tasks on LIP dataset. Our proposed method is compared with current state-of-the-art methods, including SegNet [37], FCN-8s [38], DeepLabV2 [39], Attention [40], DeepLabV2 + SSL [36], Attention + SSL [36] and standard U_Net [12]. From Table 2 it can be observed that standard U_Net network [12] outperforms the state-of-the-art networks on human semantic segmentation dataset LIP [36], our modified U_Net network outperforms the standard U_Net network by 0.23% at overall accuracy, 0.26% at mean accuracy and 0.35% at mean IoU index.

We group the 19 semantic labels of LIP dataset into 3 labels: head (hat, hair, sunglasses, scarf, face), torso (glove, upper clothes, dress, coat, left-arm, right-arm) and lower-body (socks, pants, jumpsuits, skirt, left-leg, right-leg, left-shoe, right-shoe), and train the modified U_Net network on LIP dataset with grouped labels at first. We randomly chose 300 images of people from the trainset of Market-1501 [8], CUHK03 [17], and DukeMTMC-reID [9], and then labelled them with 4 semantic labels. Finally, we fine-tuned the modified U_Net network model on LIP dataset with labeled data. We use the fine-tuned model for part segmentation in the proposed person re-identification method.

Table 2. Performance of human semantic segmentation on the validation split of LIP.

Method	Overall Accuracy	Mean Accuracy	Mean IoU
SegNet [37]	0.6904	0.2400	0.1817
FCN-8s [38]	0.7606	0.3675	0.2829
DeepLabV2 [39]	0.8266	0.5167	0.4164
Attention [40]	0.8343	0.5255	0.4244
DeepLabV2 + SSL [36]	0.8316	0.5255	0.4244
Attention + SSL [36]	0.8436	0.5494	0.4473
U_Net [12]	0.8499	0.5625	0.4677
U_Net (ours)	**0.8522**	**0.5651**	**0.4712**

Figure 5 illustrates some examples of parts semantic segmentation map and corresponding bounding boxes of human parts. The top row images are person images, and the bottom images illustrate their part segmentation by super-pixels and bounding boxes of person parts are demonstrated with red rectangles. It illustrates the results of parts localization in different situations, including normal situation (1st column), leg occlusion (2nd and 3rd columns), head occlusion (4th and 5th columns), detection mistakes (6th to 9th columns) and crowds (10th and 11th columns). As seen from the localization results in different situations, semantic segmentation-based part localization is stable and accurate. There are also some mistakes. As seen from the 6th column and the 10th column, there are some segmentation mistakes in the torso part, which result in the width of the bounding box of torso part reduced by 7.14% in the 6th column, and height of bounding box of lower-body part increased by 8.26% in the 10th column. We then randomly chose another images of people from the trainset of Market-1501 [8], CUHK03 [17], and DukeMTMC-reID [9], and labelled their part bounding boxes to evaluate the performance of part location with modified U_Net. The mean IoU between labeled bounding boxes and inferred ones are 69.15% for head, 82.57% for torso and 76.78% for low-body, respectively. This is acceptable for part location and can be treated with data augmentation.

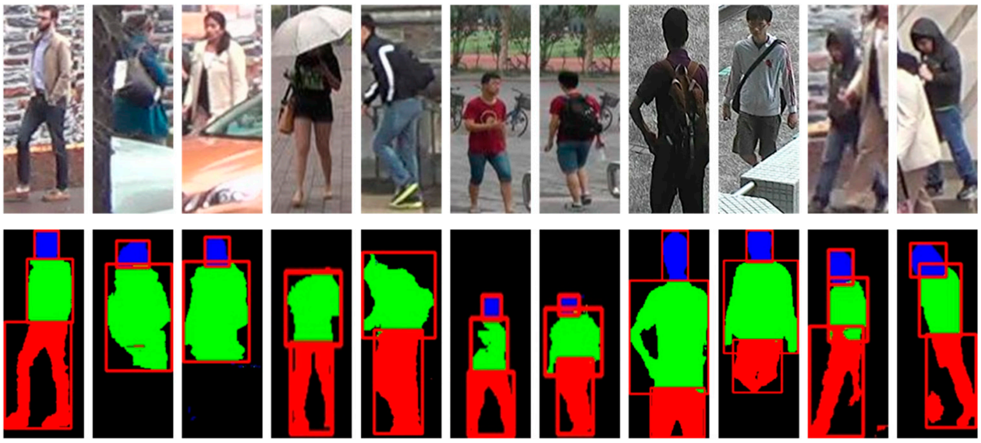


Figure 5. Examples of part semantic segmentation maps and corresponding bounding boxes of parts. The top row images are RGB images of person, and the bottom images illustrate their part segmentation by super-pixels and bounding boxes of human parts are demonstrated with red rectangles.

4.3. Person Re-Identification Performance

We performed experiments on three public person re-identification datasets: Market-1501 [8], CUHK03 [17], and DukeMTMC-reID [9].

Market-1501 dataset [8] consists of images of 1,501 persons 32,668 images which cropped with bounding-boxes predicted by DPM detector [41]. These images are captured from 6 different cameras, including 5 high-resolution cameras, and one low-resolution camera. Overlap exists among different cameras. The whole dataset is divided into training set with 12,936 images of 751 persons and testing set with 3368 query images and 19,732 gallery images of 750 persons.

The **CHUK03** dataset [17] includes 13,164 images of 1,360 people captured by six cameras. Each identity appears in two disjoint camera views (i.e., 4.8 images in each view on average). Our dataset is partitioned into training set (1160 persons), validation set (100 persons), and test set (100 persons).

The **DukeMTMC-reID** dataset [9] consists of 1,404 identities appearing in more than two cameras and 408 identities (distractor ID) who appear in only one camera. There are 16,522 training images of 702 identities, 2,228 query images of the other 702 identities and 17,661 gallery images (702 ID + 408 distractor ID).

Our proposed method is compared with current state-of-the-art methods, including IDLA [23], Part-Aligned [25], PersonNet [26], HydraPlus-Net [27], BoW+kissme [8], Basel. + LSRO [9], DCSL [42], PDC [5], PSE [11], SVDNet [43], PAN [44] and ATWL [45] to show our considerable performance advantage over all the existing competitors. In our experiments, we report the cumulative matching characteristics (CMC) rank-1, rank-5, rank-10 and mean average precision (mAP) to evaluate the performances of person re-identification methods. And we use state-of-the-art network architectures, such as VGG [46], ResNet [33], DenseNet [47] and Inception v3 network [48], as our baseline network to test the performance of different networks in our approach. The results are summarized in Tables 3–5, where we denote our method as PADP (Parts Alignments with DropParts).

Table 3. Rank-1, rank-5, rank-10 and mAP of various methods on the Market-1501 dataset.

Method	Rank-1	Rank-5	Rank-10	mAP
PersonNet [26]	0.3721	-	-	0.1857
BoW+kissme [8]	0.4725	-	-	0.2188
PAN [44]	0.7159	0.8389	-	0.5151
HydraPlus-Net [27]	0.7690	0.9130	0.9450	-
Part-Aligned [25]	0.8100	0.9200	0.9470	0.6340
SVDNet [9]	0.8230	0.9230	0.9520	0.6210
PDC [5]	0.8414	0.9273	0.9492	0.6341
AACN [29]	0.8590	-	-	0.6687
PSE [11]	0.8770	0.9450	0.9680	0.6900
PADP (VGG19+BN)	0.8821	0.9530	0.9697	0.7086
PADP (Resnet18)	0.8771	0.9510	0.9658	0.6902
PADP (Resnet34)	**0.8895**	0.9541	**0.9742**	0.7093
PADP (Resnet50)	0.8824	**0.9561**	0.9718	0.7028
PADP (Densenet121)	0.8881	0.9513	0.9685	**0.7133**
PADP (Inception_v3)	0.8884	0.9543	0.9700	0.7056

Table 4. Rank-1, rank-5 and rank-10 of various methods on the CUHK03 dataset.

Method	Rank-1	Rank-5	Rank-10
IDLA [23]	0.5474	0.8650	0.9388
PersonNet [26]	0.6480	0.8940	0.9492
DCSL [42]	0.8020	0.9773	0.9917
SVDNet [43]	0.8180	0.9520	0.9720
Basel. + LSRO [9]	0.8460	0.9760	0.9890
Part-Aligned [25]	0.8540	0.9760	0.9940
PDC [5]	0.8870	0.9861	0.9924
PADP (VGG19+BN)	0.9029	0.9879	0.9943
PADP (Resnet18)	0.8923	0.9820	0.9923
PADP (Resnet34)	**0.9083**	**0.9896**	**0.9953**
PADP (Resnet50)	0.9001	**0.9869**	0.9951
PADP (Desnet121)	0.9021	0.9842	0.9941
PADP (Inception_v3)	0.9053	0.9849	0.9944

Table 5. Rank-1, rank-5, rank-10 and mAP of various methods on the DukeMTMC-ReID dataset.

Methods	Rank-1	Rank-5	Rank-10	mAP
Basel. + LSRO [9]	0.6768	-	-	0.4713
PAN [44]	0.7159	0.8389	-	0.5151
SVDNet [43]	0.7670	0.8640	0.8990	0.5680
AACN [29]	0.7684	-	-	0.5925
PSE [11]	0.7980	**0.8970**	0.9220	0.6200
ATWL(2-stram) [45]	0.7980	-	-	0.6340
HA-CNN [28]	0.8050	-	-	0.638
PADP (VGG19+BN)	0.8134	0.8720	0.9264	0.6444
PADP (Resnet18)	0.8076	0.8726	0.9221	0.6407
PADP (Resnet34)	0.8156	0.8742	0.9286	**0.6455**
PADP (Resnet50)	0.8150	**0.8755**	0.9223	0.6452
PADP (Desnet121)	**0.8191**	0.8752	**0.9287**	**0.6455**
PADP (Inception_v3)	0.8128	0.8748	0.9285	0.6424

From Tables 3–5, it can be seen that the proposed algorithm with different network architectures outperforms the current state-of-the-art person re-identification methods on average. Due to its simpler structure, 18-layer Resnet method performs worse than other networks, by 0.9% on Market-1501 dataset, 1.14% on the CUHK03 dataset, 0.76% on the DukeMTMC-ReID dataset at rank-1. However, it also outperforms the current state-of-the-art person re-identification methods on three datasets. On the Market-1501 dataset, our method with 34-layer Resnet outperforms the second best method by 1.25% at rank-1, and the Densenet121 based method outperforms the second best method by 2.33% at mAP. On the CUHK03 dataset, our method with 34-layer Resnet outperforms the second best method by 2.13% at rank-1. On the DukeMTMC-ReID dataset, our method with Densenet121 outperforms the second best method by 1.41% at rank-1, and 0.75% at mAP.

5. Discussion

To better understand the proposed method, we analyzed it in two aspects: the effect of part alignment, and the effect of DropParts.

5.1. Effect of Part Alignment

The analysis is performed on DukeMTMC-ReID dataset [9]. The proposed network is compared with two networks, 34-layer Resnet network and four-branch parts fusion network without DropParts. The architectures are as follows:

(1) BASE network architecture: Branch01 in Table 1, including input, conv1, res2, res3, res4, res5, pool5, F_fuse and Softmax layers.

(2) ALIGN network architecture: same as architecture in Table 1 but without scaled_pool5 layer; when meeting part occlusion, a zero patch is used to replace it.

(3) ALIGN +DROP network architecture: same as architecture in Table 1.

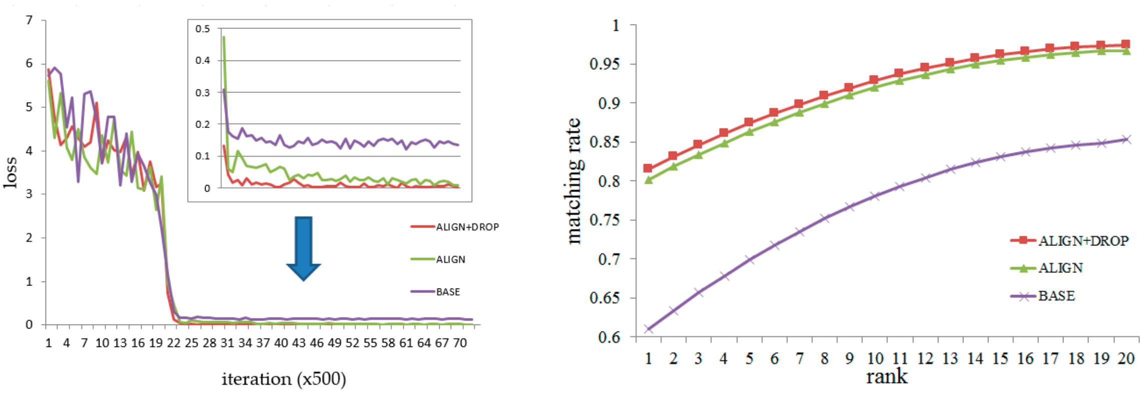
In this experiment, we report the loss curve during training and CMC curve to evaluate the performances of three networks above.

From Figure 6a, we can see that after 13000 iterations of training, the losses of ALIGN and ALIGN +DROP network reach the very low level (<0.02) while the loss of BASE network is above 0.12, which signifies under-fitting of BASE network. As a result, seen from Figure 6b, rank-1 accuracy of BASE network is lower than of ALIGN and ALIGN +DROP by almost 19%, and CMC curves of ALIGN and ALIGN +DROP networks are above the CMC curve of BASE network all the time. As seen from the loss curves and CMC curves, the addition of part alignment and feature fusion result in significant improvement in the person re-identification performance.

5.2. Effect of DropParts

We analyzed the role of DropParts by comparing ALIGN with the ALIGN +DROP network. In Figure 6a, the loss of the ALIGN network can be very low, but the loss of ALIGN +DROP network is even lower, i.e., below 0.01. Another point is that the losses of ALIGN +DROP network during training are more stable than ones of ALIGN network which oscillate even at the end of training. These two points signify well-fitting and easy training of ALIGN +DROP network. This validates the first function of the DropParts method that it makes the four-branch convolutional neural network work properly when part missing occurs.

In Figure 6b, CMC curve of ALIGN +DROP network is always above the CMC curve of BASE network. The addition of DropParts results in improvement in the training and recognition performance of person re-identification, by 1.37% at rank-1.



(**a**) (**b**)

Figure 6. Loss curves and CMC curves. (**a**) Loss curves of BASE, ALIGN and ALIGN +DROP network during training; (**b**) CMC curves of BASE, ALIGN and ALIGN +DROP network during testing.

In order to further analyze the role of DropParts, we investigate the performances on samples with missing part. Table 6 demonstrates statistics of part missing on Market-1501 [8], CUHK03 [17] and DukeMTMC-reID [9] datasets. As seen from Table 6, part missing does not always occur.

Table 6. Statistics of part missing on Market-1501, CUHK03 and DukeMTMC-reID datasets.

Dataset	Head Missing	Torso Missing	Lower-Body Missing
Market-1501	12.78%	0.50%	0.65%
CUHK03	0.77%	0.01%	0.59%
DukeMTMC-reID	2.59%	0.03%	0.53%

We evaluate the performance of proposed method with DropParts and without DropParts on samples with missing parts. The results are summarized in Table 7.

Table 7. Performance of ALIGN and ALIGN +DROP methods on part missing samples.

	Market-1501			CUHK03			DukeMTMC-reID		
Methods	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**	**mAP**
ALIGN	0.7334	0.8599	0.6087	0.8589	0.9752	0.8356	0.7379	0.8564	0.5670
ALIGN +DROP	0.8510	0.9286	0.6603	0.8998	0.9803	0.8777	0.7991	0.8962	0.6212

From Table 7, it can be seen that the proposed algorithm outperforms the method without DropParts by 7.32% at rank-1, 3.79% at rank-5 and 4.93% at mAP on average. On the Market-1501 dataset, our method outperforms our method without DropParts by 11.86% at rank-1. It validates that the DropParts method can improve the performance of person re-identification.

6. Conclusions

In this paper, we present a new deep architecture deal with parts misalignment, and propose a DropParts method firstly to solve the parts missing problem. Experiments on standard pedestrian datasets show the effectiveness of our proposed method.

For the future work, we will continue to improve the models of part localization and matching, by:

(1) Dividing person images into more parts, and improving the performance of parts localization.

(2) Designing an end-to-end model that includes both parts segmentation and re-identification tasks.

Author Contributions: Conceptualization, methodology and investigation: H.G. and S.C.; software, validation, formal analysis and data curation: H.G. and Z.Z.; writing, review and editing: H.G., S.C. and Z.Z.

Funding: This research received no external funding.

Acknowledgments: This work was supported by National Key R&D Program of China (2018YFB1305200), Public Welfare Technology Application Research Plan Project of Zhejiang Province (No. LGG18F030002), National Natural Science Foundation of China (No. U1509207), the Opening Project of Collaborative Innovation Center for Economics Crime Investigation and Prevention Technology of Jiangxi Province (No. JXJZXTCX-016).

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

References

1. Saghafi, M.A.; Zaman, H.B.; Saad, M.H.M.; Hussain, A. Review of person re-identification techniques. *IET Comput. Vis.* **2014**, *8*, 455–474. [CrossRef]
2. Bedagkar-Gala, A.; Shah, S.K. A survey of approaches and trends in person re-identification. *Image Vis. Comput.* **2014**, *32*, 270–286. [CrossRef]
3. Tong, X.; Li, H.; Ouyang, W.; Wang, X. Learning deep feature representations with domain guided dropout for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1249–1258. [CrossRef]
4. Li, S.; Ma, H. A Siamese inception architecture network for person re-identification. *Mach. Vis. Appl.* **2017**, *28*, 725–736. [CrossRef]
5. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3980–3989. [CrossRef]
6. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 274–282. [CrossRef]
7. Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; Sun, J. AlignedReID: Surpassing human-level performance in person re-identification. *arXiv*, **2017**, arXiv:1711.08184v2.
8. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1116–1124. [CrossRef]
9. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3774–3782. [CrossRef]
10. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 34–39. [CrossRef]

11. Sarfraz, M.S.; Schumann, A.; Eberle, A.; Stiefelhagen, R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 420–429. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
13. Miri, K.; Jaehoon, J.; Hyuncheol, K.; Joonki, P. Person re-identification using color name descriptor-based sparse representation. In Proceedings of the IEEE 7th Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA, 9–11 January 2017; pp. 1–4. [CrossRef]
14. Zhao, R.; Ouyang, W.; Wang, X. Unsupervised salience learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3586–3593. [CrossRef]
15. Li, W.; Wang, X. Locally aligned feature transforms across views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3594–3601. [CrossRef]
16. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 2197–2206. [CrossRef]
17. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep filter pairing neural network for person re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 152–159. [CrossRef]
18. Wu, L.; Wang, Y.; Gao, J.; Li, X. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Trans. Multimed.* **2018**. [CrossRef]
19. Alex, D.; Sami, Z.; Banerjee, S.; Panda, S. Cluster loss for person re-identification. *arXiv*, **2018**, arXiv:1812.10325.
20. Zhao, C.; Chen, K.; Wei, Z.; Chen, Y.; Miao, D.; Wang, W. Multilevel triplet deep learning model for person re-identification. *Pattern Recogn. Lett.* **2019**, *117*, 161–168. [CrossRef]
21. Wu, D.; Zheng, S.-J.; Bao, W.-Z.; Zhang, X.-P.; Yuan, C.-A.; Huang, D.-S. A novel deep model with multi-loss and efficient training for person re-identification. *Neurocomputing* **2019**, *324*, 69–75. [CrossRef]
22. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; Tang, X. Spindle Net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 907–915. [CrossRef]
23. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 3908–3916. [CrossRef]
24. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and tell: Neural image caption generation with visual attention. *arXiv*, **2015**, arXiv:1502.03044.
25. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3239–3248. [CrossRef]
26. Wu, L.; Shen, C.; van den Hengel, A. PersonNet: Person re-identification with deep convolutional neural networks. *arXiv*, **2016**, arXiv:1601.07255.
27. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. HydraPlus-Net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359. [CrossRef]
28. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 2285–2294. [CrossRef]
29. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018.
30. Brian, K. Metric learning: A survey. *Found. Trends Mach. Learn.* **2013**, *5*, 287–364. [CrossRef]

31. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
32. Sergey, I.; Christian, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*, **2015**, arXiv:1502.03167.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
34. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Science, T.D.J.C. Caffe: Convolutional architecture for fast feature embedding. *arXiv*, **2014**, arXiv:1408.5093.
35. Montavon, G.; Orr, G.B.; Müller, K.-R. Neural networks: Tricks of the trade. *Lect. Notes Comput. Sci.* **2012**, *7700*, 581–598. [CrossRef]
36. Gong, K.; Liang, X.; Zhang, D.; Shen, X.; Lin, L. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6757–6765. [CrossRef]
37. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
38. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
39. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
40. Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649. [CrossRef]
41. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008; pp. 1–8. [CrossRef]
42. Zhang, Y.; Li, X.; Zhao, L.; Zhang, Z. Semantics-aware deep correspondence structure learning for robust person re-identification. In Proceedings of the International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 3545–3551.
43. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. SVDNet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3820–3828. [CrossRef]
44. Zheng, Z.; Zheng, L.; Yang, Y. Pedestrian alignment network for large-scale person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2018**. [CrossRef]
45. Ristani, E.; Tomasi, C. Features for multi-target multi-camera tracking and re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018.
46. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv*, **2014**, arXiv:1409.1556.
47. Huang, G.; Liu, Z.; Maaten, L.v.d.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
48. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]