A Part Power Set Model for Scale-Free Person Retrieval

Yunhang Shen[1] , **Rongrong Ji**[1,2*] , **Xiaopeng Hong**[3,4] , **Feng Zheng**[5]
Xiaowei Guo[6] , **Yongjian Wu**[6] and **Feiyue Huang**[6]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, 361005, China
[2]Peng Cheng Laborotory, China
[3]Xi'an Jiaotong University, China
[4]University of Oulu, Finland
[5]Southern University of Science and Technology
[6]Tencent Youtu Lab, Tencent Technology (Shanghai) Co., Ltd.
shenyunhang01@gmail.com, rrji@xmu.edu.cn, hongxiaopeng@mail.xjtu.edu.cn, zfeng02@gmail.com,
{scorpioguo, littlekenwu, garyhuang}@tencent.com

Abstract

Recently, person re-identification (re-ID) has attracted increasing research attention, which has broad application prospects in video surveillance and beyond. To this end, most existing methods highly relied on well-aligned pedestrian images and hand-engineered part-based model on the coarsest feature map. In this paper, to lighten the restriction of such fixed and coarse input alignment, an end-to-end part power set model with multi-scale features is proposed, which captures the discriminative parts of pedestrians from global to local, and from coarse to fine, enabling part-based scale-free person re-ID. In particular, we first factorize the visual appearance by enumerating k-combinations for all k of n body parts to exploit rich global and partial information to learn discriminative feature maps. Then, a combination ranking module is introduced to guide the model training with all combinations of body parts, which alternates between ranking combinations and estimating an appearance model. To enable scale-free input, we further exploit the pyramid architecture of deep networks to construct multi-scale feature maps with a feasible amount of extra cost in term of memory and time. Extensive experiments on the mainstream evaluation datasets, including Market-1501, DukeMTMC-reID and CUHK03, validate that our method achieves the state-of-the-art performance.

1 Introduction

Person retrieval, *a.k.a.*, person re-identification (re-ID), aims at retrieving pedestrians across non-overlapping camera views distributed at distinct locations. To calculate the similarities between person images, visual features play a central role. In the past decade, deep features have shown their superiority in person re-ID [Li *et al.*, 2014; Chen *et al.*, 2017a; He *et al.*, 2018]. Despite the remarkable progress, person re-ID is still suffering from large variations on persons such as pose, occlusion, clothes, background clutters, and detection failures. To handle such variations, recent advances have advocated the use of part-level features to offer fine-grained description [Sun *et al.*, 2018; Suh *et al.*, 2018].

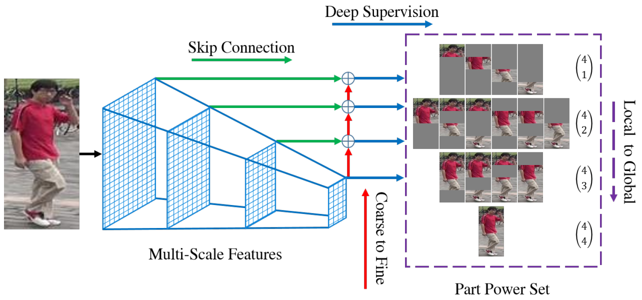


Figure 1: Overview of the proposed PPS model with scale-free feature learning. Given a set of feature maps, we enumerate k-combinations for all k of n horizontal body parts, which captures global to local visual appearances for person retrieval. Furthermore, skip connections (green arrows) are used to construct multi-scale feature maps, which associate bottom, fine spatial, and weak semantic features with top, coarse spatial, and strong semantic features efficiently in a coarse-to-fine fashion (red arrows).

A common practice for part-based models is to partition the intermediate feature maps of an input image into multiple horizontal parts uniformly [Sun *et al.*, 2018]. Subsequently, each partial region is used for identity classification independently. To that effect, various architectures for partitioning person images are exploited recently [Wang *et al.*, 2018a; Fu *et al.*, 2019]. Such a pipeline has three key drawbacks: Firstly, the overall performance seriously depends on how well the pedestrian parts are aligned. Therefore, in challenging scenes, most existing methods suffer from part misalignment due to inaccurate detection, pose variation, occlusion, *etc*. Secondly, most methods attempt to combine features ex-

[*] Corresponding author.

tracted from different manually selected partial regions. Such handcrafted designs require extra human efforts and may lead to suboptimal solutions. Thirdly, most works only exploit the coarsest feature map extracted from the top of the backbone network. Thus, spatial details in finer-resolution feature maps are lost. In fact, finer-resolution feature maps contain meaningful low-level information, such as colour and texture, which are crucial to distinguish similar pedestrians.

In order to overcome the above limitations, we propose an end-to-end Part Power Set (PPS) Model, which enables robust and scale-free part-based person re-ID, as shown in Fig. 1. Firstly, to deal with misalignments of parts, our method enumerates k-combinations for all k of n body parts coming from a uniform partition, result in $\sum_{k=1}^{n} C_n^k$ combinations in total. As illustrate in Fig. 2, the PPS model is able to align discriminative parts in pedestrian images and exploit global-to-partial information elegantly. Enumerating alignment of body parts with different scales and ratios would have provided an almost infinite candidate set of combinations. However, PPS efficiently filter the candidates in accordance with a power set of $2^n - 1$ combinations. Secondly, we introduce a combination ranking module (CRM) to guide the model training with all combinations of body parts. It considers inter-class relations by constructing two subnets to perform pedestrian identification and combination ranking in parallel. Compared with manually-designed part-based models [Wang *et al.*, 2018a; Fu *et al.*, 2019], this pivotal module relieves the requirement of human efforts in designing specific part-based model, and thus has potential to achieve global optimal. Thirdly, we merge features extracted from different scales of CNNs with a feasible amount of extra cost in terms of memory and time, while gaining stronger abstract semantic as well as higher resolution features. In particular, before feeding features at different scales of the backbone network to PPS, we take advantage of the feature maps from the top layers to enhance the abstract semantics for the bottom layers. It derives from our observation that, feature maps in the top layers have strong abstract semantic but are coarse due to loss of details, while maps in the bottom layers have fine resolutions but suffer from lacking of abstract semantics.

In sum, the main contributions of this paper are three-fold:

- We propose an end-to-end part power set model by enumerating k-combinations for all k of n body parts, which exploits, in an all-inclusive manner, from the global to various levels of the partial information to learn discriminative features.
- Our method introduces a combination ranking module to guide the optimization to converge rapidly and stably, which performs combination identification and ranking in parallel. It replaces handcrafted designs and selections that require extra human efforts and may lead to suboptimal solutions.
- We exploit multi-scale features of the backbone network with single stream and image, to further boost the performance. It is trained end-to-end by combining all scales to enhance feature representation in a deeply supervised fashion.

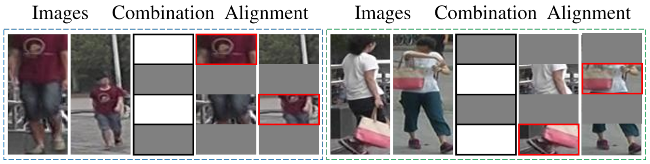


Figure 2: Misalignment of discriminative parts in pedestrian images. For example, the discriminative parts (T-shirt) in the first image pairs are misaligned by occlusion. While the discriminative parts (handbag) in the second image pairs are misaligned by large pose variation. By enumerating different combinations of local parts, PPS aligns the discriminative parts (red boxes).

We perform extensive quantitative evaluation on mainstream datasets, including Market-1501, DukeMTMC-reID and CUHK03, with comparisons to cutting-edge methods. Results show our method achieves the state-of-the-art performance. Code and models will be made publicly available.

2 Related Work

Person re-ID. Deep learning methods have dominated person re-ID in the last decade. Deep neural network is first employed by Li *et al.* [Li *et al.*, 2014] to determine if a pair of input images belong to the same ID. Different loss functions have been designed for network optimization, *e.g.*, the siamese loss [Radenović *et al.*, 2016], triplet loss [Hermans *et al.*, 2017], and quadruplet loss [Chen *et al.*, 2017a]. Many efforts focus on reducing the impact of the misalignment and occlusion [Su *et al.*, 2017; He *et al.*, 2018].

Part-based Model. Several recent works propose to generate deep representation from body parts as fine-grained discriminative features of pedestrians. Such part-based models can be divided into three groups. The first one leverages external cues to partition pedestrian parts, *e.g.*, assistance from the latest progress on human pose estimation [Qian *et al.*, 2018; Suh *et al.*, 2018; Su *et al.*, 2017; Xu *et al.*, 2018]. The second group utilizes attention-based methods to handle the misaligned matching in re-ID [Li *et al.*, 2018; Xu *et al.*, 2018]. The third group crops the intermediate feature maps into pre-defined patches (*i.e.*, patches) [Sun *et al.*, 2018; Wang *et al.*, 2018a; Fu *et al.*, 2019]. However, most methods usually require extra human efforts to examine and select the combinations of parts, and thus may lead to suboptimal solutions. To solve all these problems, we propose to factorize the visual appearance by enumerating k-combinations for all k of n body parts.

Multi-scale Feature Learning. Most existing methods typically consider only one resolution of person appearance by a standard scale normalisation process. It discards the potentially useful information of other different scales. To mine complementary information across different scales, one of the earliest endeavors is [Li *et al.*, 2015], which jointly trains multi-scale images. Recent works [Liu *et al.*, 2016; Chen *et al.*, 2017b] use an image pyramid to build multi-scale features by designing multi-scale streams. Instead, our method works in a single stream with top-down pathway, *i.e.*, taking a single-scale image of an arbitrary size as input, which is more feasible in termed of time and memory than [Liu *et al.*, 2016; Chen *et al.*, 2017b]. Qian *et al.* [Qian *et al.*, 2017] designs a

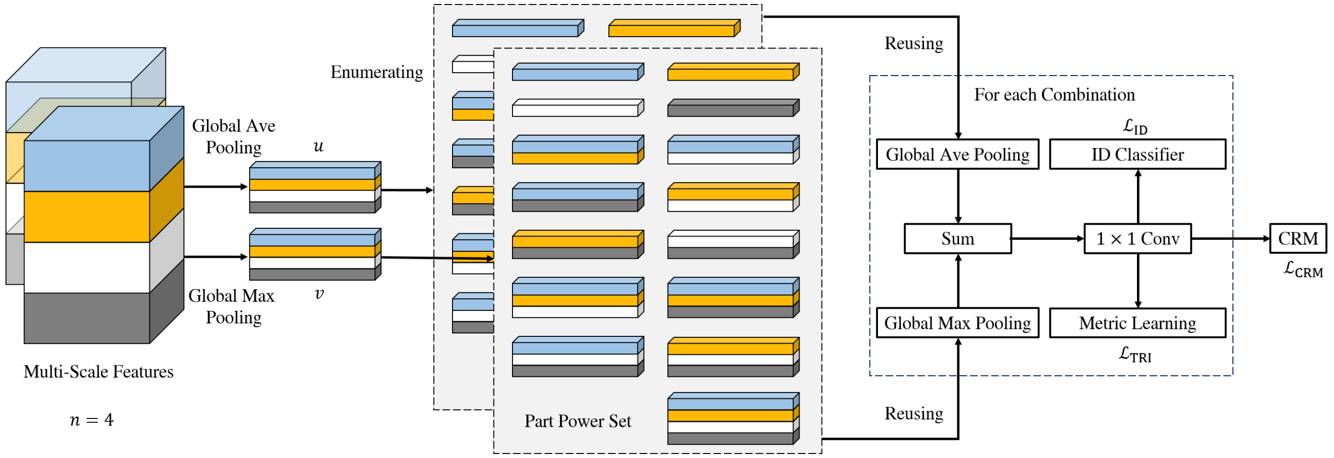


Figure 3: Overview of the proposed model: Given a set of feature maps, our method first crops it into n strips horizontally. In this example, we set n to 4. Then global max and average pooling layers are applied to each strip. After that, we enumerate k-combinations for all k of n strips. For each combination, another group of global max and average pooling layers are applied, followed by the element-wise sum. Then a 1×1 convolutional layer is added to get the final feature for each combination.

deep architecture, which has multiple streams with different sizes of receptive fields. Our method is distinct from [Qian *et al.*, 2017] from the following two aspects. Firstly, our method is independent of the backbone architecture. Secondly, the output feature maps of our method are with proportional sizes and strong semantics at multiple levels, in a fully convolutional fashion.

3 The Method

3.1 Overview

Our method is designed from the following three aspects: First, part power set model relaxes the requirement of well-aligned body parts, and smoothly incorporates global-to-local information by enumerating k-combinations for all k of n body parts. Secondly, a combination ranking module is introduced to guide the training process to converge rapidly and stably. Thirdly, we exploit multi-scale features with a feasible amount of extra cost in terms of memory and time.

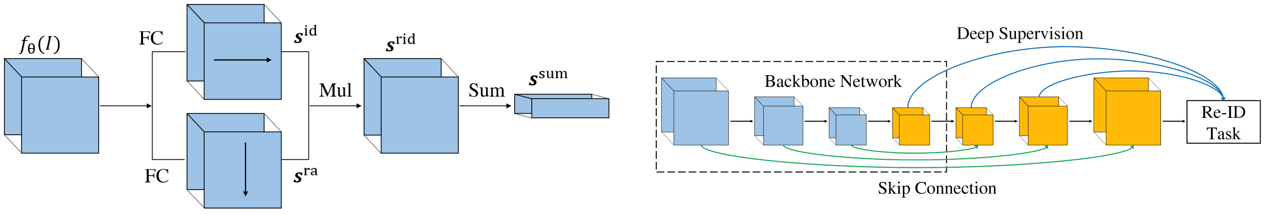
3.2 Part Power Set Model

Our model is built on a feature map M extracted from the backbone network. M is a 3-dimensional tensor of the size $C \times H \times W$, where C is the number of the channels, W and H are the spatial width and height, respectively. The feature map M is divided into n horizontal parts uniformly with a fixed size of $C \times (H/n) \times W$. Then we enumerate k-combinations for all k of n body parts, where $1 \leq k \leq n$ and n is a hyper-parameter, as illustrate in the middle of Fig. 3. It can be viewed as a specific global-to-local architecture, in which each component captures the discriminative information at different spatial scales. For example, n is set to 4 and thus 15 combinations, *i.e.*, $\sum_{k=1}^{4} \binom{4}{k}$, are enumerated in total, which is also illustrated in Fig. 1. For each combination, the feature map(s) of corresponding body part(s) is cropped and concatenated. Then, global max and average pooling layers are applied to the feature maps of each combination, followed by element-wise sum. A convolutional layer followed by a batch normalization and a ReLU activation layer further reduces the dimension and produces the final feature vector for the re-ID task.

However, due to the time and memory constraints, it is infeasible to directly crop and concatenate feature maps of corresponding body parts from feature maps of the backbone network. The reason is clear: as the number of k-combinations ($1 \leq k \leq n$) is the coefficient of the x^k term in the polynomial expansion of the binomial power $(1+x)^n$, the total number of available combinations is $\sum_{k=1}^{n} \binom{n}{k} = 2^n - 1$, where $\binom{n}{0}$ is abandoned. As a result, the size of intermediate feature maps for all combinations before two pooling operations is $C(H/n)W \sum_{k=1}^{n} k\binom{n}{k}$, which is also the number of elements that need to be stored in memory and be accessed by each pooling operation.

To address the above limitations, we propose a novel paradigm by reusing the pre-computed pooling values of each part, as illustrate in Fig. 3. More specially, a group of global max and average pooling layers are applied to the feature maps of each body part to get two sets of vectors: $u = \{\mathbf{u}_i | 1 \leq i \leq n, \mathbf{u}_i \in \Re^C\}$ and $v = \{\mathbf{v}_i | 1 \leq i \leq n, \mathbf{v}_i \in \Re^C\}$, respectively. Then for each combination, the pooled vectors of the corresponding body parts in u and v are picked and reduced by another group of global max and average pooling layers, respectively. Finally, an element-wise sum is applied to the output vectors. As the size of intermediate feature maps after first two pooling operations, is $2nC$. This paradigm significantly reduces the size of intermediate feature maps for all combinations from $C(H/n)W \sum_{k=1}^{n} k\binom{n}{k}$ to $2C \sum_{k=1}^{n} k\binom{n}{k}$. And it minimizes the number of elements that the four pooling operations need to access to $2HWC + 2C \sum_{k=1}^{n} k\binom{n}{k}$. Thus, this paradigm is a tensor decomposition to decouple the hyper-parameter n and spatial size of feature maps HW.

To empower our model to be sufficiently discriminative, a softmax cross-entropy identification loss is introduced in the fully-connected layer, which has the feature vector of combi-



(a) (b)

Figure 4: (a): Overview of the combination rank module. FC, Mul and Sum denote the fully-connected, element-wise product and sum pooling layer, respectively. (b): Overview of the multi-scale feature learning scheme. Given a set of feature maps from different scales, we construct a deeper, finer and better semantic feature representation. This figure is best viewed in color.

nations as the input:

$$\mathcal{L}_{\mathrm{ID}} = \frac{1}{N_{im}} \sum_{i=1}^{N_{im}} \sum_{c=1}^{2^n-1} -\ln\left(\frac{\exp\left((W^c_{y_i})^T f^c_\theta(I_i)\right)}{\sum_{j=1}^{N_{id}} \exp\left((W^c_j)^T f^c_\theta(I_i)\right)}\right), \tag{1}$$

where N_{im} and N_{id} are the number of images and person identities, respectively. And y_i, $W^c_{y_i}$, and $f^c_\theta(I_i)$ denote the identity of the i-th image, the weight matrix of the fully-connected layer for the y_i-th identity, and the feature vector of the i-th image in the c-th combination, respectively. On the other hand, a triplet loss based metric learning is imposed to the feature maps:

$$\begin{aligned}\mathcal{L}_{\mathrm{TRI}} = \frac{1}{N_{im}} \sum_{i=1}^{N_{im}} \sum_{c=1}^{2^n-1} \Big[& d\big(f^c_\theta(I_i), f^c_\theta(I^p_i)\big) - \\ & d\big(f^c_\theta(I_i), f^c_\theta(I^n_i)\big) + \delta \Big]_+ \end{aligned}, \tag{2}$$

where I_i and I^p_i have the same identity, I_i and I^n_i are the images of different identities, $d(\cdot)$ is the normalized Euclidean distance, $[\cdot]_+$ is the hinge loss, and δ denotes the margin hyper-parameter to control the distance differences.

The extra computations are mainly on fully connected layers, which classify each combination. The extra FLOPs [Molchanov *et al.*, 2017] are computed as: $(2^n-1)(2I-1)O$, where $I = 128$ is the input feature size and $O = 752$ is the number of identity in Market1501. The extra FLOPs are 6×10^6 and 0.2×10^9 for $n = 5$ and $n = 10$, respectively, while the original ResNet50 has 3.8×10^9 FLOPs. Therefore, the increased computational complexity is marginal.

3.3 Combination Ranking Module

The proposed PPS model is imposed to classify each combination in the power set independently. However, the discriminative powers of different partial combinations are not the same. We except that the model highlights discriminative combinations during learning. To this end, we alternate between ranking all combinations and estimating an appearance model using the weighted combinations. More concretely, we construct two subnets to perform such simultaneous identity classification and combination ranking, as illustrate in Fig. 4(a). The first subnet outputs an identity score for each combination individually. In particular, we get the j-th identity prediction probability for c-th combinations by applying a fully-connected layer to the feature vector, followed by a softmax operator, defined as: $\mathbf{s}^{id}_{cj} = \frac{\exp\left((W^{id}_j)^T f^c_\theta(I_i)\right)}{\sum_{l=1}^{N_{id}} \exp\left((W^{id}_l)^T f^c_\theta(I_i)\right)}$, where W^{id} is the weight matrix of the fully-connected layer. The second subnet has a similar structure with the first one. However, in order to rank combinations, we use a different axis for normalization in the softmax operator: $\mathbf{s}^{ra}_{cj} = \frac{\exp\left((W^{ra}_j)^T f^c_\theta(I_i)\right)}{\sum_{l=1}^{2^n-1} \exp\left((W^{ra}_j)^T f^l_\theta(I_i)\right)}$, where W^{ra} is the weight matrix of the fully-connected layer in the second subnet. The ranking matrix $\mathbf{s}^{ra}$ is then used to weight the identity matrix $\mathbf{s}^{id}$ by taking the element-wise (Hadamard) product: $\mathbf{s}^{rid} = \mathbf{s}^{ra} \odot \mathbf{s}^{id}$. It's noted that in CRM, all combinations share the same parameters, *i.e.*, W^{id} and W^{ra}. Finally, a cross-entropy loss function can be applied to $\mathbf{s}^{rid}$. However, the computed loss may fail to converge and oscillates model parameters continuously, since each combination is competed to each other. Therefore, a sum pooling layer is further applied: $\mathbf{s}^{sum}_j = \sum_{c=1}^{2^n-1} \mathbf{s}^{rid}_{cj}$, which unifies the identity prediction vector to a single scalar for each identity. Then we define the loss function L_{CRM} as:

$$\mathcal{L}_{\mathrm{CRM}} = \frac{1}{N_{im}} \sum_{i=1}^{N_{im}} \log \mathbf{s}^{sum}_{y_i}, \tag{3}$$

where y_i is the identity of the i-th image.

3.4 Multi-Scale Feature Learning

To overcome the major limitation of multi-scale discriminative feature learn as we analyzed in Section 1, inspired by [Honari *et al.*, 2016], we introduce a multi-scale deep feature architecture to capture the determinative feature of pedestrians from coarse to fine, as illustrated in Fig. 4(b). Specifically, for ResNets, we use the feature maps output by the last residual block of the last 4 stages to construct multi-scale feature maps by the following steps: Firstly, starting from the coarsest feature map, we use a 1×1 convolutional layer to reduce the channel dimensions, and upsample the spatial resolution by a factor of 2. The reduced channel dimensions and upsampled spatial resolution are determined by the corresponding bottom feature maps. Secondly, the upsampled maps are merged with the corresponding bottom feature maps by element-wise addition with skip connection. We iterate the above two steps until all feature maps from the backbone network are traversed. Finally, the output feature maps are fed to PPS model for re-ID. To enforce direct and early supervision for both the output layer and the new layers, instead of using only the last feature map, we impose re-ID to all the new feature maps including the coarsest map from the backbone network with shared parameters.

4 Experimental Evaluation

4.1 Dataset and Evaluation Protocol

Market-1501. Market-1501 [Zheng *et al.*, 2015] contains bounding boxes from a person detector, which have been selected based on their intersection-over-union overlap with annotated bounding boxes. It has $1,501$ persons and is split into training/test sets of $12,936$/$19,732$ images.

DukeMTMC-reID. DukeMTMC-reID [Ristani *et al.*, 2016; Zheng *et al.*, 2017] is a subset of Duke-MTMC for person re-ID. It contains $36,411$ annotated images of $1,812$ different identities captured by eight high-resolution cameras. A total of $1,404$ identities are observed by at least two cameras, and the remaining 408 identities are distractors. The training set contains $16,522$ images of 702 identities and the test set contains the other 702 identities.

CUHK03. CUHK03 [Li *et al.*, 2014] consists of $14,097$ cropped images from $1,467$ identities. For each identity, images are captured from two cameras, and there are about 5 images for each view. Two ways are used to produce the cropped images, *i.e.*, human annotation and detection DPM. We follow the new training/test protocol, which has 767 identities for training and 700 identities for testing. Datasets named as *labelled* and *detected* are both used for training and testing.

Evaluation metrics. We adopt the Cumulative Matching Characteristics (CMC) at *Rank-1*, *Rank-5*, and *Rank-10*, and mean Average Precision (*m*AP). It is worth noting that all our results are obtained in a single-query setting. For the proposed method, no re-ranking is used in all the experiments.

4.2 Implementation Details

Our experiments are implemented based on the Caffe2 framework. For the backbone network, we use Resnet50 initialized with the weights pretrained on ImageNet. We remove the last fully-connected layer and the average pooling layer and set the stride of last resent *conv5_1* from 2 to 1. The feature maps are reduced to a 128-dimensional vector using a 1×1 convolutoinal layer, followed by Batch Normalization and ReLU layers. We also apply dropout with a ratio of 0.2 on the output feature vector before the fully-connected layer. All images are resized into a resolution of 384×128 by following [Sun *et al.*, 2018]. The training images are augmented with horizontal flipping. We use a step strategy with mini-batch Stochastic Gradient Descent (SGD) to train the neural networks on a Tesla V100 GPU. Parameters of the maximum number of epochs, batch size, momentum, weight decay factor and base learning rate are set as 120, 64, 0.9, 0.0005 and 0.01, respectively. The base learning rate is dropped by a half every 10 epochs from epoch 60 to epoch 90. The learning rate for all new layer is set to $10\times$ the base learning rate. The margin in the triplet loss is 1.4 in all our experiments. Multi-loss dynamic training [Zheng *et al.*, 2019] is also used. We use the normalized feature for retrieval evaluation.

4.3 Comparison with State-of-the-Art Methods

Market-1501. Tab. 1 shows the performance comparison between our method and the state of the arts on Market-1501. It can be observed that, being facilitated by the proposed method, our performance surpasses all the state-of-the-art approaches in all the evaluation metrics. We have improved the baseline method namely the PCB model [Sun *et al.*, 2018] by 7.92% and 2.04% in terms of *m*AP and *Rank-1*, respectively. It shows that both our PPS model play a crucial role beyond the backbone architecture. Fig. 5 shows the top-10 ranking results for some exemplar queries. It demonstrates that the proposed model is more robust to pose variation, blur, and occlusion than the baseline.

Method		*mAP*	*Rank 1*	*Rank 5*	*Rank 10*
DaRe	[Wang *et al.*, 2018b]	74.2	88.5	-	-
DuATM	[Si *et al.*, 2018]	76.6	91.4	97.1	-
HA-CNN	[Li *et al.*, 2018]	75.7	91.2	-	-
KPM	[Shen *et al.*, 2018c]	75.3	90.1	96.7	97.9
AACN	[Xu *et al.*, 2018]	66.9	85.9	-	-
GSRW	[Shen *et al.*, 2018a]	82.5	92.7	96.9	98.1
DNN_CRF	[Chen *et al.*, 2018]	81.6	93.5	-	-
CamStyle	[Zhong *et al.*, 2018]	71.6	89.5	-	-
MLFN	[Chang *et al.*, 2018]	74.3	90.0	-	-
AOS	[Huang *et al.*, 2018]	70.4	86.5	-	-
BraidNet	[Wang *et al.*, 2018c]	69.5	83.7	-	-
HAP2S	[Yu *et al.*, 2018]	74.5	89.7	-	-
PN-GAN	[Qian *et al.*, 2018]	72.6	89.4	-	-
PCB	[Sun *et al.*, 2018]	77.4	92.3	97.2	98.2
PCB RPP	[Sun *et al.*, 2018]	81.6	93.8	97.5	98.5
SGGNN	[Shen *et al.*, 2018b]	82.8	92.3	96.1	97.4
Local CNN	[Yang *et al.*, 2018]	77.7	91.5	-	-
HPM	[Fu *et al.*, 2019]	82.7	94.2	97.5	98.5
PPS		**85.32**	**94.34**	**97.68**	**98.72**

Table 1: Comparison results (%) on Market-1501.

Method		*mAP*	*Rank 1*	*Rank 5*	*Rank 10*
DaRe	[Wang *et al.*, 2018b]	63.0	79.1	-	-
DuATM	[Si *et al.*, 2018]	64.6	81.8	90.2	-
HA-CNN	[Li *et al.*, 2018]	63.8	80.5	-	-
KPM	[Shen *et al.*, 2018c]	63.2	80.3	89.5	91.9
AACN	[Xu *et al.*, 2018]	59.2	76.9	-	-
GSRW	[Shen *et al.*, 2018a]	66.4	80.7	-	-
DNN_CRF	[Chen *et al.*, 2018]	69.5	84.9	-	-
CamStyle	[Zhong *et al.*, 2018]	57.6	78.3	-	-
MLFN	[Chang *et al.*, 2018]	74.3	90.0	-	-
AOS	[Huang *et al.*, 2018]	62.1	79.2	-	-
BraidNet	[Wang *et al.*, 2018c]	59.5	76.4	-	-
HAP2S	[Wang *et al.*, 2018c]	62.6	80.3	-	-
PN-GAN	[Qian *et al.*, 2018]	53.2	73.6	88.8	-
PCB	[Sun *et al.*, 2018]	66.1	81.8	-	-
PCB RPP	[Sun *et al.*, 2018]	69.2	83.3	-	-
SGGNN	[Shen *et al.*, 2018b]	68.2	81.1	88.4	91.2
Local CNN	[Yang *et al.*, 2018]	62.8	81.0	-	-
MGN	[Wang *et al.*, 2018a]	**78.4**	**88.7**	-	-
HPM	[Fu *et al.*, 2019]	74.3	86.6	-	-
PPS		75.94	88.20	**95.39**	**95.83**

Table 2: Comparison results (%) on DukeMTMC-reID.

DukeMTMC-reID. Tab. 2 shows the performance comparison between our approach and the state of the arts on DukeMTMC-reID. Our method consistently achieves competitive performance on all metrics. In particularly, we achieve absolute improvement of 9.84% and 6.40% for *m*AP and *Rank-1*, respectively, compared with the baseline PCB model. It demonstrates the effectiveness of the proposed PPS model on person re-ID.

CUHK03. Performance comparison on CUHK03 is shown in Tab. 3. We report the results with two types of boxes: man-

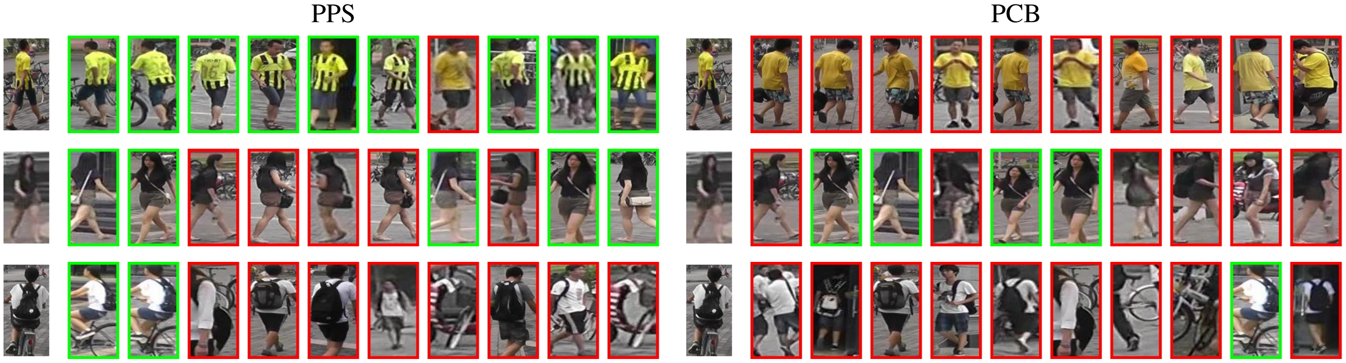


Figure 5: Top-10 ranking list for example query images on Market-1501 by our method and the baseline PCB model. The green/red rectangles indicate that images have the same/different identities as the query.

Method	Labelled		Detected	
	mAP	*Rank 1*	*mAP*	*Rank 1*
DaRe [Wang *et al.*, 2018b]	60.2	64.5	58.1	61.6
HA-CNN [Li *et al.*, 2018]	41.0	44.4	38.6	41.7
MGCAM [Song *et al.*, 2018]	50.2	50.1	46.9	46.7
MLFN [Chang *et al.*, 2018]	49.2	54.7	47.8	52.8
AOS [Huang *et al.*, 2018]	-	-	43.3	47.1
PCB [Sun *et al.*, 2018]	-	-	54.2	61.3
PCB RPP [Sun *et al.*, 2018]	-	-	57.5	63.7
Local CNN [Yang *et al.*, 2018]	53.8	58.7	51.6	56.8
MGN [Wang *et al.*, 2018a]	67.4	68.0	66.0	66.8
HPM [Fu *et al.*, 2019]	57.5	63.1	-	-
PPS	**72.66**	**75.64**	**70.56**	**73.75**

Table 3: Comparison results (%) on CUHK03 using new protocol .

ually labeled and detected. We achieve an *m*AP of 72.66% for manually labeled boxes, outperforming the previous best *m*AP reported in [Wang *et al.*, 2018a; Yang *et al.*, 2018] by 5.26% and 18.86%, respectively. It is noted that CUHK03 is the most challenging benchmark among the three datasets. And the result indicates that our proposed PPS model is more robust than all other methods.

4.4 Ablation Study

Influence of the body part number. As showed in Tab. 4, with fewer body partitions $n \leq 3$, the performance drops dramatically, due to that local feature is not fully exploited. With an increasing n ($n \leq 6$), the performance consistently boosts and converges to about 83% in terms of *m*AP. It is excepted that the more local feature is exploited, the more PPS model is robust to the misalignment of body parts. When n increases to 10, which enumerates $1,023$ ($2^{10} - 1$) combination and extracts $1,023$ different local feature, it is observed that the performance drops about 5% to 11% in terms of *m*AP. We argue that imposing too many combinations to identify person may cause the optimization being hard to converge.

Impact of combination ranking module. As shown in Tab. 4, CRM is able to improve the performance especially with a large n. Results also demonstrate that CRM refines the model to focus on discriminative combinations. Our method achieves the best performance when n is set to 5.

Multi-scale feature versus single-scale feature. Finally, we explore the multi-scale features to help re-ID. Tab. 4 shows that the proposed multi-scale architecture improves the results by 0.29% to 0.81% in terms of *m*AP.

n	$\mathcal{L}_{\text{CRM}}$	MS	$\mathcal{L}_{\text{TRI}}$	*mAP*	*rank 1*	*rank 5*	*rank 10*	time(ms)
1				74.13	89.04	95.46	97.27	8
2				80.09	92.58	97.48	98.31	9
3				82.77	93.65	97.86	98.66	10
4				82.89	93.38	97.36	98.34	12
5				83.48	93.82	97.65	98.57	15
6				82.79	93.44	97.39	98.43	24
7				78.11	90.91	96.67	97.68	45
8				76.13	90.47	96.17	97.48	86
9				74.30	89.76	96.05	97.39	164
10				72.32	88.66	95.81	97.30	329
2	✓			81.02	92.89	97.51	98.31	9
3	✓			83.02	93.72	97.80	98.67	10
4	✓			83.23	93.56	97.89	98.57	12
5	✓			83.59	93.59	97.51	98.49	17
6	✓			83.07	93.05	97.15	98.34	26
7	✓			80.80	92.49	97.18	98.22	47
8	✓			77.10	91.72	96.59	97.60	86
9	✓			75.22	89.79	96.23	97.54	169
10	✓			74.32	89.66	95.91	97.55	333
5		✓		83.91	93.65	97.66	98.58	34
5	✓	✓		84.50	94.39	97.66	98.61	34
5	✓		✓	85.03	95.27	97.67	98.71	17
5	✓	✓	✓	85.32	94.34	97.68	98.72	34

Table 4: Ablation study of our method on Market-1501.

5 Conclusion

In this paper, we propose an end-to-end part power set model for part-based person retrieval, which is robust to misalignment of body parts. In particular, a combination ranking module is introduced to perform identify classification and combination ranking in parallel. We further exploit how to extract scale-free features with deep supervision, which has received little attention so far. Extensive experiments demonstrate that our method substantially outperforms the state of the arts.

Acknowledgements

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

[Chang *et al.*, 2018] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-Level Factorisation Net for Person Re-Identification. In *CVPR*, 2018.

[Chen *et al.*, 2017a] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017.

[Chen *et al.*, 2017b] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2017.

[Chen *et al.*, 2018] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group Consistent Similarity Learning via Deep CRF for Person Re-Identification. In *CVPR*, 2018.

[Fu *et al.*, 2019] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal Pyramid Matching for Person Re-identification. In *AAAI*, 2019.

[He *et al.*, 2018] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-Free Approach. In *CVPR*, 2018.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv*, 2017.

[Honari *et al.*, 2016] Sina Honari, Jason Yosinski, Pascal Vincent, and Christopher Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *CVPR*, 2016.

[Huang *et al.*, 2018] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially Occluded Samples for Person Re-identification. In *CVPR*, 2018.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[Li *et al.*, 2015] Xiang Li, Wei-shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale Learning for Low-resolution Person Re-identification. In *ICCV*, 2015.

[Li *et al.*, 2018] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-Identification. In *CVPR*, 2018.

[Liu *et al.*, 2016] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-Scale Triplet CNN for Person Re-Identification. In *ACMMM*, 2016.

[Molchanov *et al.*, 2017] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning Convolutional Neural Networks for Resource Efficient Inference. In *ICLR*, 2017.

[Qian *et al.*, 2017] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale Deep Learning Architectures for Person Re-identification. In *ICCV*, 2017.

[Qian *et al.*, 2018] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-Normalized Image Generation for Person Re-identification. In *ECCV*, 2018.

[Radenović *et al.*, 2016] Filip Radenović, Giorgos Tolias, and Ond\vrej Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016.

[Ristani *et al.*, 2016] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *ECCV*, 2016.

[Shen *et al.*, 2018a] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep Group-shuffling Random Walk for Person Re-identification. In *CVPR*, 2018.

[Shen *et al.*, 2018b] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person Re-identification with Deep Similarity-Guided Graph Neural Network. In *ECCV*, 2018.

[Shen *et al.*, 2018c] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-End Deep Kronecker-Product Matching for Person Re-identification. In *CVPR*, 2018.

[Si *et al.*, 2018] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification. In *CVPR*, 2018.

[Song *et al.*, 2018] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided Contrastive Attention Model for Person Re-Identification. In *CVPR*, 2018.

[Su *et al.*, 2017] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven Deep Convolutional Model for Person Re-identification. In *ICCV*, 2017.

[Suh *et al.*, 2018] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-Aligned Bilinear Representations for Person Re-identification. In *ECCV*, 2018.

[Sun *et al.*, 2018] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). In *ECCV*, 2018.

[Wang *et al.*, 2018a] Guanshuo Wang, Yufeng Yuan, Xiong Chen, and Jiwei et al. Li. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *ACMMM*, 2018.

[Wang *et al.*, 2018b] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. Resource Aware Person Re-identification across Multiple Resolutions. In *CVPR*, 2018.

[Wang *et al.*, 2018c] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person Re-identification with Cascaded Pairwise Convolutions. In *CVPR*, 2018.

[Xu *et al.*, 2018] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-Aware Compositional Network for Person Re-identification. In *CVPR*, 2018.

[Yang *et al.*, 2018] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, and Jianqiang et al. Huang. Local Convolutional Neural Networks for Person Re-Identification. In *ACMMM*, 2018.

[Yu *et al.*, 2018] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-Aware Point-to-Set Deep Metric for Person Re-identification. In *ECCV*, 2018.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[Zheng *et al.*, 2017] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[Zheng *et al.*, 2019] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training. In *CVPR*, 2019.

[Zhong *et al.*, 2018] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera Style Adaptation for Person Re-identification. In *CVPR*, 2018.