Improve Person Re-Identification With Part Awareness Learning

Houjing Huang, *Student Member, IEEE,* Wenjie Yang, *Graduate Student Member, IEEE,* Jinbin Lin, Guan Huang, Jiamiao Xu, Guoli Wang, Xiaotang Chen, *Member, IEEE,* and Kaiqi Huang, *Senior Member, IEEE*

***Abstract*— Person re-identification (ReID) aims to predict whether two images from different cameras belong to the same person. Due to low image quality and variance in view point and body pose, it remains a difficult task. To solve the task, a model is supposed to appropriately capture features that describe body regions for identification. With the simple intuition that explicitly incorporating ReID model with part awareness could be beneficial for learning a more discriminative feature space, we propose part segmentation as an assistant body perception task during the training of a ReID model. Specifically, we add a lightweight segmentation head to the backbone of ReID model during training, which is supervised with part labels. Note that our segmentation head is only introduced during training and that it does not change network input or the way of extracting ReID feature. Experiments show that part segmentation considerably improves the performance of ReID. Through quantitative and qualitative analyses, we further reveal that body part perception helps ReID model to capture a set of more diverse features from the body, with decreased similarity between part features and increased focus on different body regions. We experiment with various representative ReID models and achieve consistent improvement on several large-scale datasets including Market1501, CUHK03, DukeMTMC-reID and MSMT17. *E.g.* on MSMT17, our method increases Rank-1 Accuracy of GlobalPool-ResNet-50, PCB and MGN by 2.3%, 2.9% and 3.9%, respectively. Incorporated with MGN, our model achieves state-of-the-art performance, with Rank-1 Accuracy 95.8%, 78.8%, 90.0% and 84.0% on four datasets, respectively.**

***Index Terms*— Person re-identification, part awareness, part segmentation, multi-task learning.**

Manuscript received October 18, 2019; revised May 3, 2020; accepted June 3, 2020. Date of publication June 24, 2020; date of current version July 13, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001005, in part by the National Natural Science Foundation of China under Grant 61673375, Grant 61721004, and Grant 61876181, in part by the Projects of Chinese Academy of Science under Grant QYZDB-SSW-JSC006, and in part by the Youth Innovation Promotion Association CAS. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lucio Marcenaro. *(Corresponding author: Kaiqi Huang.)*

Houjing Huang and Wenjie Yang are with the Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: houjing.huang@nlpr.ia.ac.cn; wenjie.yang@nlpr.ia.ac.cn).

Jinbin Lin and Guoli Wang are with Horizon Robotics, Inc., Beijing 100036, China (e-mail: jinbin.lin@horizon.ai; guoli.wang@horizon.ai).

Guan Huang was with Horizon Robotics, Inc., Beijing 100036, China. He is now with the Algorithm Department, Xforward AI Technology Company Ltd., Beijing 100081, China (e-mail: guan.huang@horizon.ai).

Jiamiao Xu was with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China, and also with Horizon Robotics, Inc., Beijing 100036, China. He is now with the Deep Learning Department, DeepRoute.ai, Shenzhen 518000, China (e-mail: jiamiaoxu_93@163.com).

Xiaotang Chen is with the Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xtchen@nlpr.ia.ac.cn).

Kaiqi Huang is with the Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: kqhuang@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2020.3003442

I. Introduction

PERSON re-identification is a fundamental task in video surveillance and smart retail, providing support for pedestrian retrieval and cross-camera tracking [1], [2], *etc*. It aims to predict whether two images from different cameras belong to the same person. With large-scale datasets, as well as improved feature extraction and metric learning methods, recent years have seen great progress in this task [3]–[10]. However, due to degraded image quality, pose and view point variation, *etc*., it still remains a tough problem.

Generally speaking, it is desirable for a ReID model to capture discriminative features that well represent body regions, in order for accurate identification. From this perspective, we believe that the awareness of body parts should be an underlying capability of the model. However, in most existing methods, the model is merely supervised by identity labels. We argue that these models may be short of part sensitivity. To enhance such ability of a ReID model, we propose to train ReID with an additional task of part perception. Concretely, we connect a lightweight segmentation head to the backbone and supervise it with part labels, during the training of a normal ReID model. The idea is illustrated in Fig. 1. Here we achieve part awareness by ensuring that the model understands which body part the current pixel on the feature map belongs to.

In our framework, the whole backbone is shared between the tasks of ReID and part segmentation. Hence after training is finished, it is appropriately enhanced with part knowledge. The proposed method can be viewed as a regularization for feature learning, without altering network input, the way of extracting ReID feature, or ReID optimization settings. In addition, after training, the segmentation head is simply removed. These merits make our method easy to integrate with existing models.

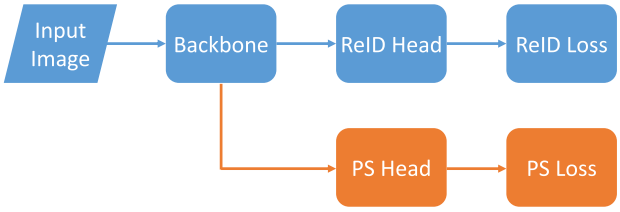


Fig. 1. Overview of our method. Generally speaking, distinguishing between different identities requires a model to capture discriminative features from body regions. We believe that perception of body structure should be an underlying capability of a ReID model. To enhance such capability, we propose to extend the backbone with a Part Segmentation (PS) head during training. In the figure, the blue branch is the normal routine of a ReID model, while the orange one is what we propose in this paper. Note that our method does not change network input, how ReID features are extracted, or optimization settings of ReID. Besides, the PS branch can be simply stripped off after training is finished.

We validate the proposed method with several representative ReID models including GlobalPool [11], PCB [9] and MGN [10], on four large-scale datasets Market1501 [12], CUHK03 [13], DukeMTMC-reID [14] and MSMT17 [15]. The proposed part segmentation constraint consistently improves upon these models on all datasets (Fig. 6). On the largest dataset MSMT17, our method boosts Rank-1 Accuracy by 2.3%, 2.9% and 3.9% for GlobalPool-ResNet-50, PCB and MGN, respectively. When applied to MGN, our model achieves state-of-the-art performance, with Rank-1 Accuracy 95.8%, 78.8%, 90.0%, 84.0% (and mAP 88.7%, 74.4%, 79.9%, 62.4%) on four benchmarks, respectively.

To reveal the changes that part segmentation brings to ReID features in a quantitative way, we base on PCB and calculate part similarity in feature space. Results show that a set of more diverse features are learned (Fig. 9). We reckon that the increased diversity between part features in turn spans a larger and more discriminative space for identification. Through Grad-cam [16] visualization on MGN, we also discover that the proposed method helps ReID model to emphasize on more regions on human body. We believe that it reduces the risk of overfitting to salient body regions and facilitates learning comprehensive ReID features. Extensive ablation experiments are also conducted to analyze key factors of the proposed method, including part granularity in segmentation supervision, structure of the segmentation head, impact on each part, *etc*. To be complete, we also confirm that the improvement in ReID is generalizable across domains.

The contribution of this paper is as follows. 1) We propose to equip ReID model with part awareness by explicit part segmentation supervision. 2) Extensive experiments are conducted to reveal the mechanism of improvement brought by part awareness learning. 3) The proposed method consistently improves over several representative ReID models. Our final model achieves state-of-the-art performance on four large-scale benchmarks.

II. Related Work

A. Person Re-Identification

Person re-identification is an important task in video surveillance, which supports pedestrian retrieval and multi-target-multi-camera tracking (MTMCT) [1], [2]. Its purpose is to predict whether two images from different cameras belong to the same person. Both discriminative feature extractor and effective metric learning are indispensable for the task. For feature extraction, the common baseline [1], [5], [6], [17] is to perform global average (or max) pooling on the feature map of a backbone and obtain one feature vector for an image. The backbone, *e.g*. ResNet-50 [18], is originally designed for common object recognition. To improve upon this paradigm, there is one group of works [9], [10], [19]–[21] paying attention to multiple body regions, which extract multiple feature vectors from different image (or body) regions and concatenate them into a final feature representation. Besides, there is a method [22] that considers body orientation and extracts distinct features for different views. Another line of works [23], [24] devise backbones which are more suitable for ReID, with the benefits of attention mechanism, multi-scale feature, or being lightweight, *etc*. In terms of metric learning, the most representative work is triplet loss [25], which constrains the distance relation among a triplet of samples. A triplet $\langle anchor, positive, negative\rangle$ consists of two persons, where $anchor$ and $positive$ are from the same person, and $negative$ from the other. Triplet loss requires the distance between $\langle anchor, negative\rangle$ to be larger than that between $\langle anchor, positive\rangle$ by a margin. Hermans *et al*. [5] construct online and hard triplets inside each batch composed of P identities with K samples for each. To pay attention to more triplets within a batch, Wang *et al*. [26] loop through all $\langle anchor, positive\rangle$ pairs, while sampling negative instances under a gaussian distribution, whose standard deviation is controlled in the manner of curriculum learning. Chen *et al*. [27] proposes quadruplet loss, which extends a triplet with a sample from the third identity, to ensure larger inter-class variation and smaller intra-class variation.

B. Body Knowledge Assisted ReID

Additional body information has been widely adopted in person ReID. Su *et al*. [28] use key points to crop body parts, which are then normalized and combined into a new image for network input. Kalayeh *et al*. [29] train a part segmentation model on human parsing dataset LIP [30] to predict 4 body parts as well as foreground. These body masks are then used to perform local region pooling on ReID feature maps. Xu *et al*. [31] share similar idea, but with regions generated from key points. Besides, part visibility is also integrated for computing the final feature. Sarfraz *et al*. [22] directly concatenate 14 key point confidence maps with the image as network input, letting the ReID model learn alignment in an automatic way. Suh *et al*. [32] propose a two-stream network, a ReID stream and a pose estimation stream, and use bilinear pooling to obtain part-aligned feature. Recently, Zhang *et al*. [33] propose DSA-reID which contains two streams. The first stream utilizes only the original image, while the second stream uses UV coordinates to cut out and resize 24 body parts before feeding them to the ReID network.

With the guidance of the second stream during training, the first stream alone can achieve spatial alignment during testing. **Comparison with Our Method.** 1) The methods mentioned above mainly focus on fine-grained feature and/or part alignment, while ours focuses on enhancing ReID model with part awareness. 2) The former modifies network input or the way of pooling ReID feature, while the latter does not. Consequently, the method proposed in this paper can be easily incorporated into those aforementioned models to achieve the benefits of both worlds. 3) From the perspective of efficiency, all these methods, except DSA-reID, set the need for an additional part segmentation or key point estimation model during both training and testing. Moreover, the additional model is independent of the ReID model in terms of model structure and parameters. Our method, in the manner of multi-task learning, does not require an extra model, for both training and testing. The proposed lightweight head can even be a single 1×1 Conv classifier with equal benefit (Section IV-G), which is simply stripped off after training.

III. Methodology

In person re-identification, a model measures similarity between images in order to determine whether they are from the same identity. Basically, it is desirable for the model to capture discriminative features representing body regions. From this point, we believe that perception of body parts should be an underlying ability of a ReID model. To enhance such a capability, we take into account part awareness learning during ReID training. To be more specific, we expect a model to understand which body part it is processing, for each pixel on the feature map. As a straightforward solution, we propose to extend the backbone with a Part Segmentation (PS) head during the training of a ReID model. Note that our method does not modify network input or the way of extracting ReID features. Once training is finished, body awareness is already embedded in the backbone, and the extra head can be removed. An overview of the method is depicted in Fig. 1. The integration with GlobalPool, PCB and MGN are illustrated in Fig. 4, 5 and 8, respectively.

At the current step, we mainly intend the PS head to work as some constraint on ReID backbone, not to perfectly predict a high resolution label map for downstream tasks. As a result, we do not devise it so sophisticated as is done in standard pixel level predicting tasks like semantic segmentation or super resolution, *etc*. We design four types of head structures with different depth and output resolution as in Fig. 2. These variants are analyzed in Section IV-G, with type (c) in our final models.

Since existing ReID datasets do not come with part annotations, we resort to some public part segmentation dataset. Specifically, we utilize COCO Densepose dataset [34] and re-arrange annotations to have 7 body parts, as shown in Fig. 3a. We then train DANet [35], a model originally designed for semantic segmentation, on this part segmentation dataset. With this trained model, we predict pseudo part labels on ReID datasets. The pseudo labels predicted on ReID images are illustrated in Fig. 3b. Details of training DANet can be found in Section IV-B.

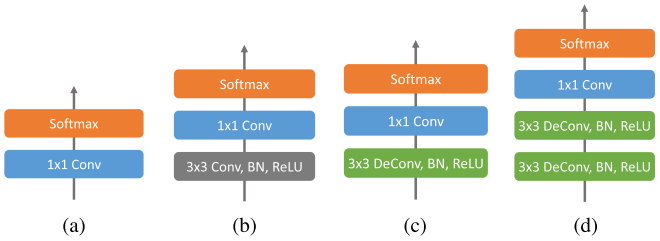


Fig. 2. Four types of part segmentation heads used in our experiments. All 3×3 convolution and deconvolution layers have 256 output channels. Deconvolution is for learnable upsampling, which results in higher output resolution. The 1×1 convolution is a pixel-wise classifier, predicting which of the 7 parts or background each pixel belongs to. (a) Only classifier. (b) With an additional Conv layer. (c) With an additional DeConv layer. (d) With two additional DeConv layers.

Fig. 3. We re-arrange COCO Densepose [34] annotations to have 7 body parts (a), train a part segmentation model, and predict pseudo labels for ReID datasets (b).

The overall loss function of a model with part awareness learning can be denoted by

$$\mathcal{L} = \mathcal{L}^{reid} + \lambda \mathcal{L}^{ps}, \tag{1}$$

where $\mathcal{L}^{reid}$ is ReID loss of the original model, and $\mathcal{L}^{ps}$ is the newly introduced part segmentation loss. λ is the loss weight that balances the importance of ReID loss and segmentation loss. By default, we set $\lambda = 1$, and other values will be discussed in Section IV-H.

We denote a training set as $\{(\mathcal{I}_i, y_i, \mathcal{S}_i) | i = 1, 2, \ldots, N\}$, where N is the number of images. $\mathcal{I}_i$ is the i-th image with identity label $y_i \in \{1, 2, \ldots, C\}$, and segmentation label being a 2-dim map $\mathcal{S}_i \in \{1, \ldots, M\}^{H \times W}$. Here C is the total number of identities, and M is the number of body part classes plus background. By default $M = 8$, while analysis of other cases is conducted in Section IV-F.

A. $\mathcal{L}^{reid}$ for GlobalPool Model

The GlobalPool model predicts a probability distribution p_i for image $\mathcal{I}_i$, where $p_i \in \mathbb{R}^C$. Multi-class cross entropy loss is adopted, which is negative log likelihood of the output node corresponding to ground truth. The loss over a batch is computed as

$$\mathcal{L}^{reid} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\log(p_{i,y_i}), \tag{2}$$

in which N_b is number of images in a batch, and p_{i,y_i} is the y_i-th element of p_i.

B. $\mathcal{L}^{reid}$ for PCB

PCB evenly divides feature map of Conv5 into six horizontal stripes and performs feature learning inside each stripe. It predicts six probability distributions $\{p_i^j | j = 1, 2, \ldots, 6\}$ for image $\mathcal{I}_i$, where $p_i^j \in \mathbb{R}^C$. Multi-class cross entropy loss is adopted for each stripe. The loss over a batch is computed as

$$\mathcal{L}^{reid} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{j=1}^{6}\log(p_{i,y_i}^j). \tag{3}$$

C. $\mathcal{L}^{reid}$ for MGN

There are three branches in MGN, as illustrated in Fig. 8. In each branch, features obtained by global max pooling are supervised with both cross entropy loss and triplet loss. The second branch further splits feature map into two stripes, and third branch into three stripes. Each of these five stripes are supervised by cross entropy loss.

Consider a triplet $(\mathcal{I}_{i1}, \mathcal{I}_{i2}, \mathcal{I}_{i3})$ within a batch, where $(\mathcal{I}_{i1}, \mathcal{I}_{i2})$ have the same identity while $(\mathcal{I}_{i1}, \mathcal{I}_{i3})$ are from different persons. The loss imposed on this triplet, w.r.t. global feature $(f_{i1}^j, f_{i2}^j, f_{i3}^j)$, $j \in \{1, 2, 3\}$, is

$$\mathcal{L}_{tri}^j(\mathcal{I}_{i1}, \mathcal{I}_{i2}, \mathcal{I}_{i3}) = [\delta + d(f_{i1}^j, f_{i2}^j) - d(f_{i1}^j, f_{i3}^j)]_+, \tag{4}$$

in which j indexes three branches, $\delta = 0.1$ is a margin, and $d(\cdot, \cdot)$ is euclidean distance. According to BatchHard [5] sampling strategy, the number of triplets in a batch is the same as batch size N_b. The triplet loss inside a batch is thus calculated as

$$\mathcal{L}_{tri} = \frac{1}{3N_b}\sum_{i1,i2,i3}\sum_{j=1}^{3}\mathcal{L}_{tri}^j(\mathcal{I}_{i1}, \mathcal{I}_{i2}, \mathcal{I}_{i3}). \tag{5}$$

We denote the probability distributions predicted from five stripes plus three global ones for image $\mathcal{I}_i$ as $\{p_i^j, j = 1, 2, \ldots, 8\}$. The cross entropy loss over a batch is computed as

$$\mathcal{L}_{ce} = -\frac{1}{8N_b}\sum_{i=1}^{N_b}\sum_{j=1}^{8}\log(p_{i,y_i}^j). \tag{6}$$

As a result, the total ReID loss for MGN is

$$\mathcal{L}^{reid} = \mathcal{L}_{ce} + \mathcal{L}_{tri}. \tag{7}$$

D. Segmentation Loss $\mathcal{L}^{ps}$

Suppose the output probability tensor of the segmentation head for image $\mathcal{I}_i$ is $\mathcal{G}_i \in \mathbb{R}^{M\times H\times W}$. Consider a spatial location (h, w) of $\mathcal{G}_i$, the corresponding probability vector is denoted by $g \in \mathbb{R}^M$, and the ground truth label by $s \in \{1, 2, \ldots, M\}$. The segmentation loss at this location is negative log likelihood $\mathcal{L}^{ps}(i, h, w) = -\log(g_s)$. To aggregate segmentation loss over locations and the batch of images, we propose to treat different types of body parts fairly. Concretely, instead of naively averaging $\mathcal{L}^{ps}(i, h, w)$ across spatial and batch dimensions, we first calculate the sum of loss for each part, $\mathcal{L}_{sum,s}^{ps} = \sum_{\mathcal{S}_{i,h,w}=s}\mathcal{L}^{ps}(i, h, w)$, and the corresponding number of locations $\Omega_s = |\mathcal{S}_{i,h,w} = s|$, $s = 1, 2, \ldots, M$. The average loss of part s is then computed as

$$\mathcal{L}_{avg,s}^{ps} = \begin{cases}\frac{1}{\Omega_s}\mathcal{L}_{sum,s}^{ps}, & \text{if } \Omega_s > 0,\\ 0, & \text{otherwise}\end{cases}. \tag{8}$$

Finally, the overall loss of the batch is calculated as

$$\mathcal{L}^{ps} = \frac{1}{\sum_{s=1}^{M}\mathbb{1}\{\Omega_s > 0\}}\sum_{s=1}^{M}\mathcal{L}_{avg,s}^{ps}, \tag{9}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function which gives 1 if the condition holds and 0 otherwise. The above computation averages inside each part before averaging across parts, to avoid large-size classes dominating the loss, *e.g. background* and *torso*. It is important for small-size classes like *foot* and *head*, which also contain much discriminative information for ReID and should be equally attended. Note that for MGN, the segmentation loss is calculated for three branches and then averaged.

During multi-task training of ReID and part segmentation, the optimization settings remain the same as original ReID training, *e.g.* optimizer, batch size, learning rate, training iteration, *etc*. The only difference is the additional PS loss mentioned above.

IV. Experiment

The experiments are organized as follows. Section IV-A introduces ReID datasets used in this paper. Section IV-B details how to obtain pseudo part labels for ReID datasets. Effectiveness of our method can be found in Section IV-C, and comparison with state-of-the-art methods in Section IV-D. In Section IV-E we try to answer why part awareness learning is beneficial for ReID. Some component analyses are also included, *i.e.* part granularity in segmentation supervision (Section IV-F), structure of the segmentation head (Section IV-G), loss weight of segmentation (Section IV-H), ReID improvement for each part (Section IV-I), and choice of domain adaptation method when training DANet on COCO (Section IV-J). We also verify that the improvement in ReID is generalizable across domains, in Section IV-K. Finally, in Section IV-L, we demonstrate the segmentation result as well as some ReID test cases. Note that our experiments in Fig. 6, Fig. 7, TABLE II, TABLE III, TABLE IV, TABLE V, TABLE IX are run for five times with different random seeds, whose scores are then averaged and reported.

TABLE I

STATISTICS OF FOUR REID DATASETS, WITH EACH ENTRY DENOTING #IDENTITIES / #IMAGES

Dataset	Training	Testing	
		Query	Gallery
Market1501	751 / 12,936	750 / 3,368	750 / 15,913
CUHK03	767 / 7,365	700 / 1,400	700 / 5,332
DukeMTMC-reID	702 / 16,522	702 / 2,228	1,110 / 17,661
MSMT17	1,041 / 32,621	3,060 / 11,659	3,060 / 82,161

A. Datasets and Evaluation Metrics

We conduct our experiments on four large-scale person ReID datasets, Market1501 [12], CUHK03 [13], DukeMTMC-reID [14] and MSMT17 [15]. **Market1501** contains 12,936 training images from 751 persons, 29,171 test images from another 750 persons. The query set has 3,368 images and gallery set has 15,913. To increase the difficulty of retrieval, the gallery set contains 2,798 distractor images with just background or part of body. A total of 6 cameras are involved, and each identity appears at most under 6 cameras. The images are detected by Deformable Part Model (DPM) [36]. **CUHK03** contains DPM detected and hand-cropped images, both with 14,096 images from 1,467 identities. Images of each person come from two disjoint cameras. Following [17] to obtain a larger test set, we adopt the new train/test protocol with 767 training identities and 700 testing ones. We experiment on the detected images, since it is closer to real scenario. **DukeMTMC-reID** contains 16,522 training images from 702 persons, 19,889 test images from another 1110 persons. The query set has 2,228 images of 702 persons and gallery set has 17,661 images of 1110 persons. 408 persons in gallery set are distractors, without sharing identity with query set. A total of 8 cameras are involved. The images are cropped from frames by human. **MSMT17** is currently the largest dataset with challenging conditions. A total of 126,441 bounding boxes of 4,101 identities are annotated, which involve 15 cameras, wide light variety, and different weather conditions. In the training set, there are 32,621 bounding boxes of 1,041 identities. In test set, there are 93,820 bounding boxes of 3,060 identities in total, 11,659 images as query set and 82,161 as gallery. Different from other datasets, bounding boxes are detected by Faster RCNN [37] in MSMT17. Statistics of three datasets are listed in TABLE I. **Metrics**. Two common evaluation metrics are used, Cumulative Match Characteristic (CMC) [38] for which we report the Rank-1, -5 and -10 accuracy, and mean Average Precision (mAP) [12].

B. Details of Training DANet on COCO

We use DANet [35] to train a part segmentation model on COCO Densepose [34] data. **Model.** The backbone is ResNet-50. For simplicity, we do not use Channel Attention Module, thus there is only one loss term to optimize. The multi-dilation parameter is set to (2, 4, 8). **Dataset.** The COCO Densepose dataset contains 46,507 person bounding boxes for training and 2,243 for validation. It annotates segmentation labels for

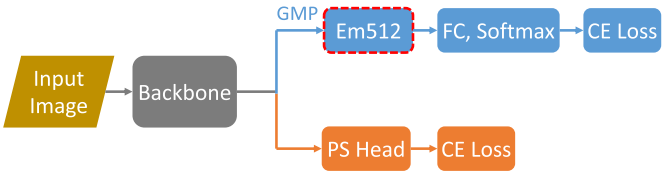


Fig. 4. Adding part segmentation head to GlobalPool [11] baseline. **GMP:** Global Max Pooling, **Em512:** 512-dimension (FC, BN, ReLU) embedding, **FC:** Fully Connected Layer, **CE Loss:** Cross Entropy Loss. The output of the embedding (with dashed red boundary) is used for ReID testing. The orange branch is what we propose in this paper.

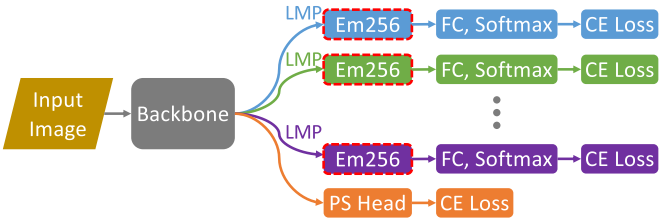


Fig. 5. Adding part segmentation head to PCB [9]. **LMP:** Local Max Pooling, **Em256:** 256-dimension (FC, BN, ReLU) embedding, **FC:** Fully Connected Layer, **CE Loss:** Cross Entropy Loss. There are six ReID heads in total, three of which are omitted. The outputs of the six embedding layers (with dashed red boundaries) are used for ReID testing. The orange branch is what we propose in this paper.

14 parts, *i.e.* {*torso, right hand, left hand, left foot, right foot, right upper leg, left upper leg, right lower leg, left lower leg, left upper arm, right upper arm, left lower arm, right lower arm, head*}. To make the segmentation model easier to train, we fuse left/right parts into one class and fuse *hand* into *lower arm*, getting 7 parts eventually. **Style Augmentation.** In our experiments, we find that model trained on COCO images has pleasing performance on COCO validation set, but fails in some cases of ReID data, sometimes having noisy prediction. We hypothesize that low resolution of ReID images is a key factor. We try to blur COCO images, but the results do not improve obviously. To train a model most suitable for ReID datasets, we transform COCO images to the style of ReID datasets, using SPGAN [39]. We then train a segmentation model with the combination of original and transferred COCO images, with 186,028 training images in total. We find this method obviously improves prediction on ReID datasets. **Common Augmentation.** The original DANet model targets scene segmentation which tends to require high-resolution images, while we tackle person part segmentation with a bounding box input each time. So we can use much smaller images. We denote a variable *base size* by $S_{base} = 192$. For each image in the current batch, we randomly select a value in interval $[0.75 \times S_{base}, 1.25 \times S_{base}]$ as the shortest size and resize the image, without changing the $height/width$ ratio. Afterwards, the image is rotated by a random degree in range $[-10, 10]$. Denoting another variable *crop size* by $S_{crop} = 256$, if any image side is smaller than S_{crop}, we have to pad the image with zeros. After padding, we randomly crop out a $S_{crop} \times S_{crop}$ square region, which is normalized by ImageNet image mean and standard deviation before being fed to the network. Random horizontal flipping is also used

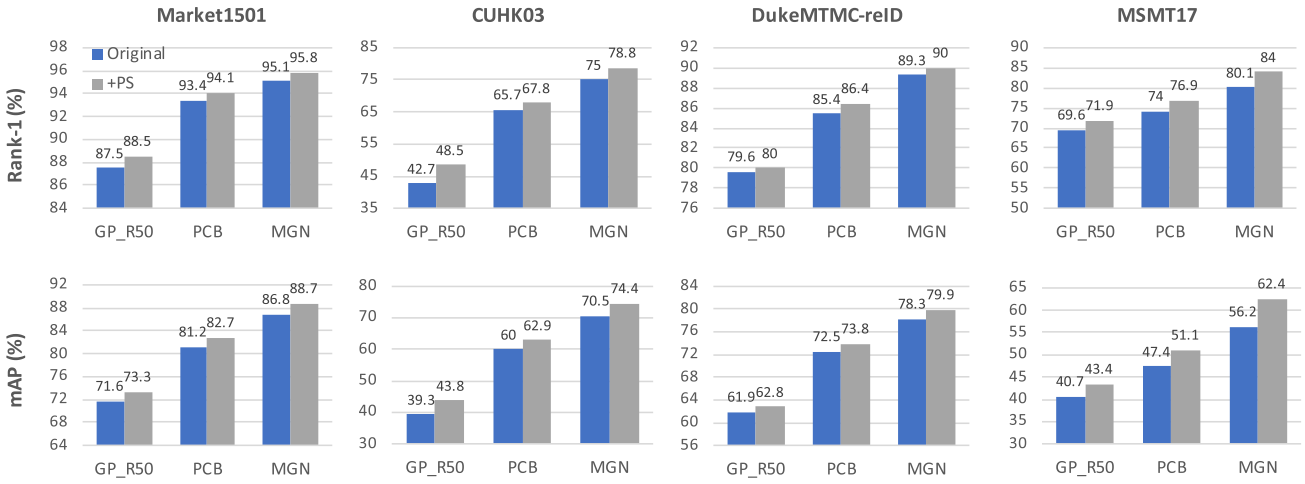


Fig. 6. Effectiveness of part awareness learning for ReID, on three representative models and four large-scale benchmarks. **GP_R50:** GlobalPool model based on ResNet-50.

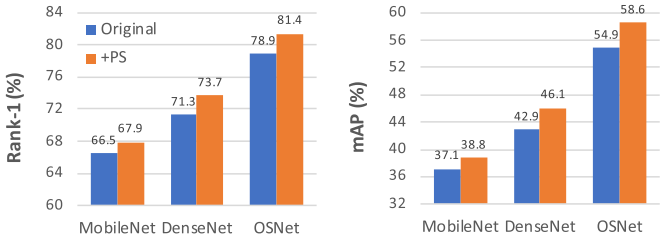


Fig. 7. Effectiveness of part awareness learning for ReID across different backbones. The results are obtained with GlobalPool on MSMT17.

for augmentation. **Optimization.** We use SGD optimizer, with learning rate 0.003, which is multiplied by 0.6 after every epoch. The training takes 5 epochs. The batch size is set to 16, and two GPUs are used for training. **Testing.** During testing, we simply resize each image to have shortest size as S_{base}, *i.e.* 192, while keeping the aspect ratio. No cropping or any other augmentation is applied. The final pixel accuracy on COCO val set (original COCO images, without changing style) is 90.3%, and mIoU is 66.8%.

C. Effectiveness of Part Awareness

To verify the effectiveness of incorporating ReID models with part awareness, we experiment on three representative ReID models, *i.e.* GlobalPool [11], PCB [9] and MGN [10]. The models during training, with part segmentation heads, are depicted in Fig. 4, Fig. 5 and Fig. 8, respectively. To verify that our method is suitable for combining with different types of backbones, in GlobalPool we experiment with ResNet-50 [18], MobileNetV2 [40], DenseNet-121 [41] and OSNet [24]. The first three come from ImageNet recognition, with ResNet-50 most commonly used in ReID models, while OSNet is specially designed with multi-scale features for ReID. In this section, the part segmentation head is type (c) from Fig. 2, and loss weight λ in Equation 1 is set to 1.

1) GlobalPool: The integration of part segmentation with GlobalPool [11] model is shown in Fig. 4. As common practice to increase feature resolution in deep layers, when using ResNet-50, MobileNetV2 or DenseNet-121 as the backbone, we remove the final feature downsampling, *i.e.* changing the stride of corresponding convolution layer from 2 to 1 for ResNet-50 and MobileNetV2, and omitting the corresponding pooling operation for DenseNet-121. Global max pooling (or average pooling for OSNet) is performed after the last convolution layer, output of which is then sent to an embedding layer and a classifier in turn. **Optimization for ResNet-50, MobileNetV2 and DenseNet-121.** We use SGD optimizer with a momentum of 0.9 and weight decay of 5e-4. Newly added layers have initial learning rate of 0.02, while layers to fine-tune use 0.01, all of which are multiplied by 0.1 after every 25 epochs. The training is terminated after 60 epochs. Batch size is set to 32. Input images are resized to $w \times h = 128 \times 256$. Only random flipping is used as data augmentation during training. **Optimization for OSNet.** According to the official code of OSNet, AMSGrad [42] is used as optimizer. Cosine annealing is adopted as the learning rate scheduler, with an initial learning rate of 0.0015, step size of 20 epochs, and 250 training epochs in total. In the first 10 epochs, only the classifier is optimized. Label smoothing is also applied. Batch size is set to 64. Input images are resized to $w \times h = 128 \times 256$. Both random flipping and random erasing [43] are used as data augmentation. **Result.** The scores of GlobalPool based on ResNet-50 (GP_R50), with and without part segmentation (PS), are recorded in Fig. 6. It can be seen that our PS head brings obvious improvement on four datasets, with Rank-1 Accuracy boost 1.0%, 5.8%, 0.4%, 2.3% respectively, and mAP improvement 1.7%, 4.5%, 0.9%, 2.7% respectively. The experiments of GlobalPool based on MobileNetV2, DenseNet-121 and OSNet are conducted on MSMT17, as reported in Fig. 7. The improvement brought by PS is consistent across backbone types, increasing Rank-1 Accuracy by 1.4%, 2.4%, 2.5%, and mAP by 1.7%, 3.2%, 3.7% for MobileNetV2, DenseNet-121 and OSNet respectively.

2) PCB: The PCB [9] model, as well as our part segmentation head denoted in orange, are illustrated in Fig. 5. PCB uses ResNet-50 [18] as the backbone, with stride set to 1 in Conv5. However, instead of obtaining a spatially global

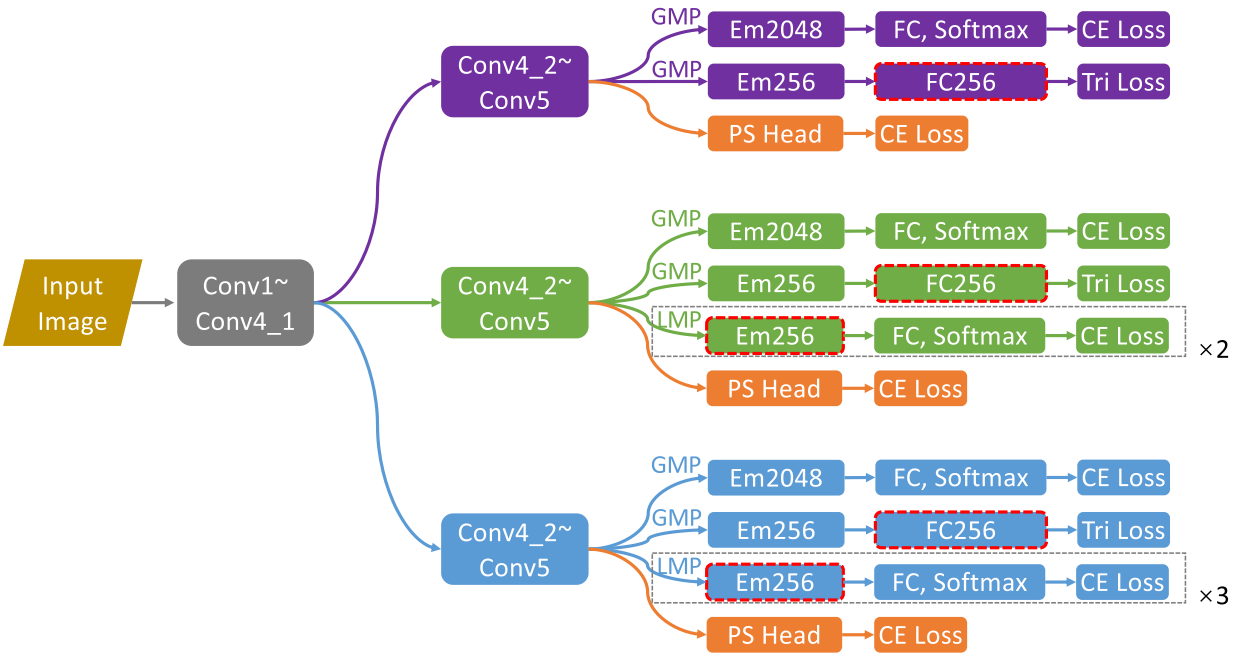


Fig. 8. Adding part segmentation head to MGN [10]. **GMP:** Global Max Pooling, **LMP:** Local Max Pooling, **Em#:** #-dimension (FC, BN, ReLU) embedding, **FC:** Fully Connected Layer, **FC256:** 256-dimension FC, **CE Loss:** Cross Entropy Loss, **Tri Loss:** Triplet Loss. ×**M:** M parallel heads operating on M different regions of feature map. The outputs of the three FC and five embedding layers with dashed red boundaries are used for ReID testing. The orange branches are what we propose in this paper. Note that three PS heads share parameters.

feature for each image, PCB evenly divides feature map of Conv5 into six horizontal stripes and performs pooling inside each stripe. The purpose of this kind of local pooling is to explicitly capture features from multiple body regions. What's more, feature pooled from each region is supervised by its own identity loss. In other words, the model urges each region to extract features that are discriminative enough on their own for identification. **Optimization.** The optimization is the same as in ResNet-50 based GlobalPool, except that input image has resolution $w \times h = 128 \times 384$, and that embedding size is now 256. **Result.** The scores of PCB and PCB+PS are recorded in Fig. 6. We observe that PCB has huge improvement over GlobalPool. Upon this strong model, our proposal still brings non-trivial boost, with 0.7%, 2.1%, 1.0%, 2.9% increase in Rank-1 Accuracy for four datasets respectively, and 1.5%, 2.9%, 1.3%, 3.7% in mAP.

3) MGN: Another local feature based model MGN [10] also draws much attention from the literature. For clarity, we show the model in Fig. 8. The improvement of MGN upon PCB is multi-fold. First, MGN proposes multi-granularity feature representation by pooling from multiple levels of stripe sizes. It has a branch that splits feature map into two stripes, and another branch into three stripes. Second, it emphasizes not only local features, but also global ones, with the resurgence of global max pooling. Finally, MGN integrates the benefits of both cross entropy loss and triplet loss, which is supposed to learn a more discriminative feature space. **Optimization.** For training, PK sampling [5] is adopted, with $P = 16$ persons and $K = 8$ images per person in a batch. Input images have size $w \times h = 128 \times 384$. Both random flipping and random erasing [43] are used during training. Test-time flipping is also performed. We use SGD optimizer with a momentum of 0.9 and weight decay of 5e-4. For ReID, cross entropy loss and triplet loss are iteratively trained, base learning rates being 0.1 and 0.01 respectively. We adopt warmup in the first 20 epochs, decay learning rates (×0.1) at 140, 180 epochs, and terminate training at 200 epochs. MGN also bases on ResNet-50, whose stride in Conv5 is set to 1 in all three branches in our experiments. **Result.** Refer to Fig. 6, we see that scores of MGN are significantly superior to PCB. Nonetheless, our part awareness learning again brings consistent improvement. Compared to MGN, MGN+PS increases Rank-1 Accuracy by 0.7%, 3.8%, 0.7%, 3.9%, and mAP by 1.9%, 3.9%, 1.6%, 6.2%, on four datasets respectively. On the largest dataset MSMT17, the benefit of our method is even much more prominent than on other datasets.

D. Comparison With State-of-the-Art Methods

We compare our final model MGN+PS with state-of-the-art methods (SOTA) on Market1501, CUHK03, DukeMTMC-reID and MSMT17 in TABLE II, III, IV, V repectively. The methods are separated into groups, including those using one global feature for one image (G), those learn multiple regions without assistance of body annotation (LMR), pose-guided methods (PG), mask guided methods (MG), methods that rigidly define multiple regions on the feature map to extract features (RDMR), and attention based methods (A). Our method achieves SOTA scores on four datasets, achieving first place on three out of four datasets, and second on CUHK03. On Market1501, we are slightly superior to previous SOTA Pyramid [21], with 0.1% (95.8% *vs*. 95.7%) increase in Rank-1 Accuracy, and 0.5% (88.7% *vs*. 88.2%) in mAP. On CUHK03, we are slightly worse than Pyramid, with a gap of 0.1%

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON MARKET1501. **G**: GLOBAL FEATURE, **LMR**: LEARNED MULTIPLE REGIONS, **PG**: POSE GUIDED, **MG**: MASK GUIDED, **RDMR**: RIGIDLY DEFINED MULTIPLE REGIONS, **A**: ATTENTION. IN EACH COLUMN, THE 1st AND 2nd HIGHEST SCORES ARE IN BOLD AND WITH GRAY BACKGROUND, RESPECTIVELY

Methods		Publication	Rank-1	mAP
G	DaRe [44]	CVPR18	86.4	69.3
	AOS [45]	CVPR18	86.5	70.4
	Cam-GAN [6]	CVPR18	89.5	71.6
	OSNet [24]	ICCV19	94.8	84.9
	DG-Net [46]	CVPR19	94.8	86.0
LMR	MSCAN [19]	CVPR17	80.3	57.5
	AANet [47]	CVPR19	93.9	82.5
	CAMA [48]	CVPR19	94.7	84.5
PG	PDC [28]	ICCV17	84.4	63.4
	AACN [31]	CVPR18	85.9	66.9
	PSE [22]	CVPR18	87.7	69.0
	PN-GAN [49]	ECCV18	89.4	72.6
	PABR [32]	ECCV18	91.7	79.6
MG	MGCAM [50]	CVPR18	83.8	74.3
	SPReID [29]	CVPR18	92.5	81.3
	DSA-reID [33]	CVPR19	95.7	87.6
RDMR	PCB [9]	ECCV18	92.3	77.4
	PCB+RPP [9]	ECCV18	93.8	81.6
	HPM [51]	AAAI19	94.2	82.7
	MGN [10]	MM18	95.7	86.9
	Pyramid [21]	CVPR19	95.7	88.2
A	HA-CNN [23]	CVPR18	91.2	75.7
	Mancs [26]	ECCV18	93.1	82.3
	CASN [52]	CVPR19	94.4	82.8
	IANet [53]	CVPR19	94.4	83.1
	MGN (Our Imp.)	-	95.1	86.8
	MGN+PS (Ours)	-	**95.8**	**88.7**

(78.8% *vs*. 78.9%) in Rank-1 Accuracy, and 0.4% (74.4% *vs*. 74.8%) in mAP. On DukeMTMC-reID, our method surpasses Pyramid by 1.0% (90.0% *vs*. 89.0%) in Rank-1 Accuracy, and 0.9% (79.9% *vs*. 79.0%) in mAP. In the case of MSMT17, the proposed model surpasses all existing methods by a large margin. The superiority over OSNet [24] reaches 5.3% (84.0% *vs*. 78.7%) in Rank-1 Accuracy and 9.5% (62.4% *vs*. 52.9%) in mAP. Note that MSMT17 is currently the largest dataset with enormous divergence in terms of cameras, dates, weather conditions, scenes, *etc*. From this perspective, we believe our method will have considerable advantage in practice.

E. Influence of Part Awareness Learning on ReID Feature

In this section, we analyze in which way part awareness learning affects ReID feature. We pay attention to PCB and MGN in particular. **Analysis on PCB.** PCB has six horizontal parts and we analyze similarity between part features. Since Conv5 is the deepest layer shared by all six ReID heads, for each image, we calculate cosine similarity between its part features pooled from Conv5, obtaining a 6×6 matrix. To analyze the statistical property, we average similarity matrices of images over the whole test set. The overall similarity matrices of PCB and PCB+PS are illustrated in Fig. 9. We have following observations. First, we notice that the results show large absolute values, with a minimum of 0.78. That is partially because the last step of ResNet Conv5 is a ReLU function

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS ON CUHK03 *detected* SUBSET UNDER THE 767/700 PROTOCOL

Methods		Publication	Rank-1	mAP
G	AOS [45]	CVPR18	47.1	43.3
	MLFN [54]	CVPR18	52.8	47.8
	DaRe [44]	CVPR18	55.1	51.3
	DG-Net [46]	CVPR19	65.6	61.1
	OSNet [24]	ICCV19	72.3	67.8
LMR	CAMA [48]	CVPR19	66.6	64.2
RDMR	PCB [9]	ECCV18	61.3	54.2
	PCB+RPP [9]	ECCV18	63.7	57.5
	HPM [51]	AAAI19	63.9	57.5
	MGN [10]	MM18	66.8	66.0
	Pyramid [21]	CVPR19	**78.9**	**74.8**
MG	DSA-reID [33]	CVPR19	78.2	73.1
A	HA-CNN [23]	CVPR18	41.7	38.6
	Mancs [26]	ECCV18	65.5	60.5
	CASN [52]	CVPR19	71.5	64.4
	MGN (Our Imp.)	-	75.0	70.5
	MGN+PS (Ours)	-	78.8	74.4

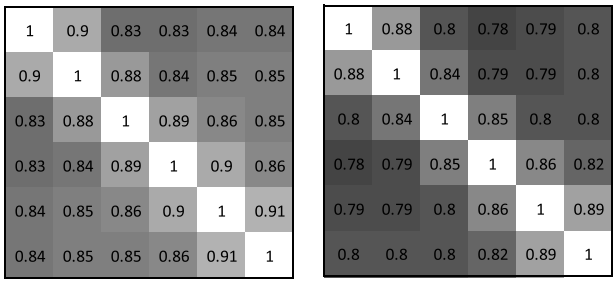
TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS ON DukeMTMC-reID

Methods		Publication	Rank-1	mAP
G	Cam-GAN [6]	CVPR18	78.3	57.6
	AOS [45]	CVPR18	79.2	62.1
	DaRe [44]	CVPR18	80.2	64.5
	DG-Net [46]	CVPR19	86.6	74.8
	OSNet [24]	ICCV19	88.6	73.5
LMR	CAMA [48]	CVPR19	85.8	72.9
	AANet [47]	CVPR19	87.7	74.3
PG	PN-GAN [49]	ECCV18	73.6	53.2
	AACN [31]	CVPR18	76.8	59.3
	PSE [22]	CVPR18	79.8	62.0
	PABR [32]	ECCV18	84.4	69.3
MG	SPReID [29]	CVPR18	84.4	71.0
	DSA-reID [33]	CVPR19	86.2	74.3
RDMR	PCB [9]	ECCV18	81.8	66.1
	PCB+RPP [9]	ECCV18	83.3	69.2
	HPM [51]	AAAI19	86.6	74.3
	MGN [10]	MM18	88.7	78.4
	Pyramid [21]	CVPR19	89.0	79.0
A	HA-CNN [23]	CVPR18	80.5	63.8
	Mancs [26]	ECCV18	84.9	71.8
	IANet [53]	CVPR19	87.1	73.4
	CASN [52]	CVPR19	87.7	73.7
	MGN (Our Imp.)	-	89.3	78.3
	MGN+PS (Ours)	-	**90.0**	**79.9**

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSMT17

Methods		Publication	Rank-1	mAP
G	GoogleNet [55]	CVPR15	47.6	23.0
PG	PDC [28]	ICCV17	58.0	29.7
G	Verif-Identif [56]	TOMM17	60.5	31.6
PG	GLAD [57]	MM17	61.4	34.0
RDMR	PCB [9]	ECCV18	68.2	40.4
A	IANet [53]	CVPR19	75.5	46.8
G	DG-Net [46]	CVPR19	77.2	52.3
G	OSNet [24]	ICCV19	78.7	52.9
	MGN (Our Imp.)	-	80.1	56.2
	MGN+PS (Ours)	-	**84.0**	**62.4**



(a) (b)

Fig. 9. ReID feature similarity between parts, averaged over whole test set of MSMT17. Lighter cell has higher value. Both matrices are symmetric. (a) PCB. (b) PCB+PS.

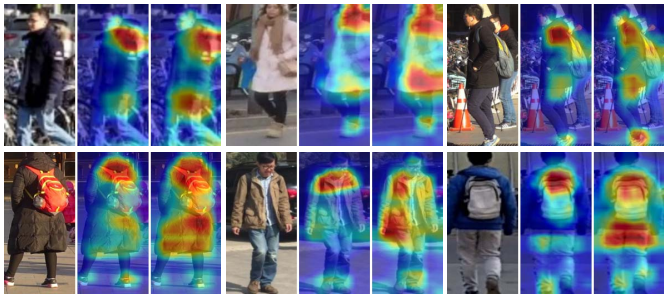


Fig. 10. Visualization of Grad-cam [16] upon MGN and MGN+PS trained on MSMT17. In each of the six cases, the original image, Grad-cam of MGN, and Grad-cam of MGN+PS are shown respectively. Warmer color means the model pays more attention to those regions.

Fig. 11. Different part granularities. We experiment on four types of label granularities, *i.e. Foreground*, *Two Parts*, *Four Parts*, and *Seven Parts*.

which outputs non-negative values, leading to a large offset in similarity value. Second, each part is more similar to adjacent parts than those disjoint ones. That is undoubted because of the spatial continuity of 2D convolution. Finally, it is obvious that the cells on the right matrix are darker than on the left. It means that with the assistance of part segmentation training, part feature similarity is considerably reduced. We reckon that our approach of forcing a segmentation head to predict part semantic at each location would make ReID features from different parts distinct from each other. Quantitatively, the similarity matrices may indicate reduced feature redundancy between parts, which in turn spans a larger and more discriminative space for person re-identification.

Analysis on MGN. In MGN, there is multi-way classification upon global feature during training. To reveal whether part awareness learning makes difference to where the model focuses on the body, we resort to Grad-cam [16], a method that demonstrates which regions of the input image are specially emphasized by the model. We take the activation of Conv5 in the second branch, as well as the gradient backwarded from the corresponding global classifier. The Grad-cam result is illustrated in Fig. 10. There is an obvious phenomenon that the proposed part awareness learning helps the model attend to more regions on human body. In this way, the extracted features could be more comprehensive than those from the original model. As a result, it avoids the model overfitting to only salient regions and improves generalization ability accordingly. Though the above analyses are conducted upon PCB and MGN, we think the implications share with other models as well.

F. Analysis of Part Granularity

In order to analyze the effect of part label granularity on ReID performance, we try three extra granularity types, *i.e. Foreground*, *Two Parts* and *Four Parts*, in addition to the *Seven Parts* used in other sections of the paper. As illustrated in Fig. 11, *Foreground* means we only classify foreground and background; *Two Parts* means to separate out upper body, lower body and background; *Four Parts* denotes that part segmentation distinguishes between head, union of upper torso and arms, upper legs, union of lower legs and feet, and background. We conduct experiments on MSMT17 with PCB and MGN, scores being recorded in TABLE VI. First, we notice that simply distinguishing between foreground and background (PCB + Foreground *vs*. PCB, and MGN + Foreground *vs*. MGN) benefits ReID, which may be due to reducing distraction from background clutter. Second, further separating upper body from lower body shows boost for MGN (MGN + Two Parts *vs*. MGN + Foreground) but only brings marginal difference for PCB (PCB + Two Parts *vs*. PCB + Foreground). Third, dividing the body into four parts has superiority over two parts (Four Parts *vs*. Two Parts). And finally, segmenting the body into seven parts achieves the most improvement. We hold the conjecture that finer granularity

TABLE VI

COMPARISON OF DIFFERENT PART GRANULARITIES FOR TRAINING PS HEAD. RESULTS ARE OBTAINED ON MSMT17

Granularity	Rank-1	Rank-5	Rank-10	mAP
PCB	74.0	84.9	88.2	47.4
PCB + Foreground	75.7	86.0	89.2	49.3
PCB + Two Parts	75.6	86.2	89.3	49.4
PCB + Four Parts	76.1	86.4	89.3	50.0
PCB + Seven Parts	**76.9**	**86.9**	**89.8**	**51.1**
MGN	80.1	89.0	91.5	56.2
MGN + Foreground	82.6	90.7	92.9	60.0
MGN + Two Parts	83.0	90.9	93.1	60.4
MGN + Four Parts	83.6	91.2	93.4	61.7
MGN + Seven Parts	**84.0**	**91.5**	**93.5**	**62.4**

TABLE VII

COMPARISON OF DIFFERENT PART SEGMENTATION HEAD STRUCTURES IN FIG. 2. RESULTS ARE OBTAINED ON MSMT17

Structure	Rank-1	Rank-5	Rank-10	mAP
PCB	74.0	84.9	88.2	47.4
Type (a)	**77.4**	**87.3**	**90.0**	**51.5**
Type (b)	77.0	86.9	89.8	51.0
Type (c)	76.9	86.9	89.8	51.1
Type (d)	76.7	87.1	89.8	51.0
MGN	80.1	89.0	91.5	56.2
Type (a)	**84.4**	91.5	93.6	62.3
Type (b)	84.2	**91.7**	**93.7**	**62.7**
Type (c)	84.0	91.5	93.5	62.4
Type (d)	83.8	91.4	93.4	61.8

in segmentation supervision impels ReID model to extract distinct features from different regions.

G. Analysis of PS Head Structure

As in Fig. 2, we take four structures of PS head into account. The difference between them resides in the number of Conv and DeConv layers. Generally speaking, a deeper PS head implies more independence from the ReID head(s); And more deconvolution layers lead to higher output resolution, which determines the resolution of segmentation supervision. We run the analysis on PCB and MGN, with experiments carried out on MSMT17. The results of four structures can be found in TABLE VII. Surprisingly, directly connecting Conv5 with a raw 1×1 Conv layer as part classifier (type a in Fig. 2) achieves the best results for PCB, and best Rank-1 Accuracy for MGN. The phenomenon indicates strong connection between the two tasks and gets rid of the need for complicated structure design for segmentation. The deepest head (type d), having two DeConv layers and a Conv classifier, still has large improvement over original PCB and MGN, which makes it feasible to generate high-resolution segmentation masks if required by other use cases. As for the choice in our final model, we adopt type c, with a DeConv layer and a Conv classifier, in order to achieve a balance between resolution of mask prediction and ReID performance.

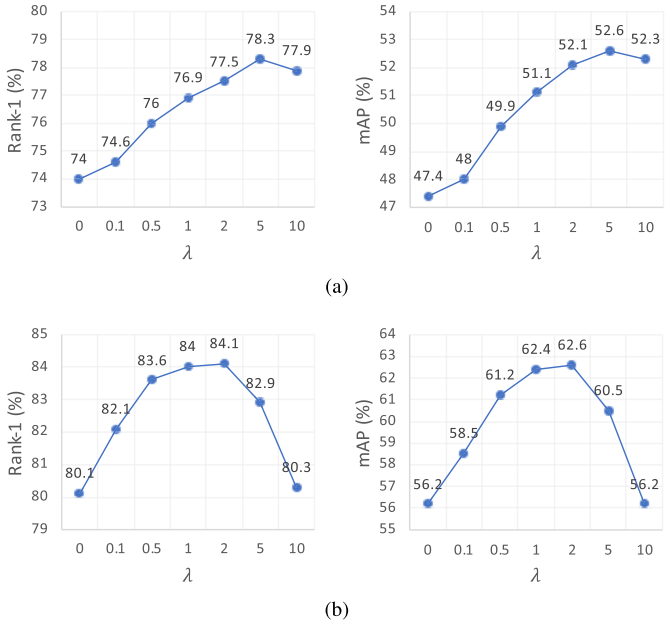


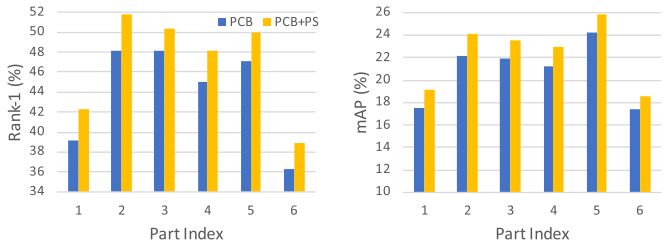
Fig. 12. Effect of PS loss weight λ on ReID performance, reported on MSMT17. (a) PCB+PS. (b) MGN+PS.

H. Influence of PS Loss Weight

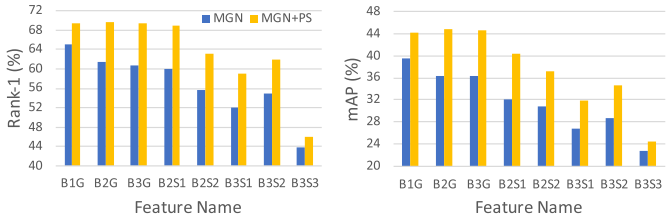
As stated before, the loss weight λ in Equation 1 determines the importance of part segmentation during training. We experiment on PCB+PS and MGN+PS, while varying λ with values 0, 0.1, 0.5, 1, 2, 5, and 10. The ReID scores on MSMT17 are plotted in Fig. 12. The plots show that, at the beginning, ReID performance benefits from increasing segmentation loss weight, reaching a summit where $\lambda = 5$ for PCB and $\lambda = 2$ for MGN. Further increasing loss weight drags the curves downward. One significant discrepancy between PCB+PS and MGN+PS is that, when $\lambda = 10$, PCB+PS still has large improvement (3.9% in Rank-1 Accuracy and 4.9% in mAP) over PCB, while MGN+PS only shows marginal improvement (0.2% in Rank-1 Accuracy and 0.0% in mAP) over MGN. Finally, we also conclude that in a wide range of λ, the proposed segmentation loss satisfactorily enhances the ReID models, which avoids tedious tuning in some way. Since $\lambda = 1$ and $\lambda = 2$ are comparable to each other for MGN+PS, we intuitively prefer $\lambda = 1$ in the final model for simplicity.

I. Does Any Part Especially Benefit From Part Awareness Learning?

Although our part awareness learning takes into account all body parts without bias, we are curious about whether it is especially beneficial for some part compared to others. We work on PCB and MGN. PCB explicitly extracts feature from six parts (stripes), while MGN splits two and three stripes in the second and third branch respectively. During testing, we only adopt feature from one stripe. The performance of each stripe is demonstrated in Fig. 13. Note that for completeness, we also display scores of the three global features (B1G, B2G and B3G) for MGN. We can see that, the increase brought by part segmentation is somewhat uniform across different



(a)



(b)

Fig. 13. ReID testing with feature from only one part, reported on MSMT17. (a) PCB+PS. (b) MGN+PS. Feature name *B1G* means global feature from the first branch, *B2S1* meaning feature of the first stripe of the second branch, and so on.

TABLE VIII

COMPARISON OF DIFFERENT DOMAIN ADAPTATION METHODS FOR TRAINING ON COCO. FINAL REID PERFORMANCE OF MGN+PS ON CUHK03 IS REPORTED

	Rank-1	Rank-5	Rank-10	mAP
No Adaptation	**78.9**	90.6	93.9	73.5
CSPL [58]	77.8	**91.4**	94.4	74.1
RoL [59]	78.7	90.9	94.4	**74.5**
SPGAN [39]	78.8	90.9	**94.6**	74.4

stripes for both PCB and MGN, except that the third stripe of the third branch of MGN (B3S3) fails to achieve as much improvement as other stripes. Besides, we find that scores of the first and last stripes in PCB, approximately corresponding to head and shoes, are much worse than other four stripes. Feature B3S3 of MGN shows similar phenomenon as well. Though the performance gap is interesting, we leave it in future work.

J. Domain Adaptation When Training on COCO

As described in Section IV-B, the segmentation model DANet trained on original COCO images fails in some cases when inferring on ReID images, with noisy prediction. The domain gap between COCO and ReID datasets could be a key factor, *e.g.* image quality of the latter under surveillance scenario is much worse than that of the former. To mitigate the domain gap, we utilize three domain adaptation methods, *i.e.* SPGAN [39], Curriculum Self-Paced Learning (CSPL) [58] and Robust Learning (RoL) [59], in the process of training DANet. For SPGAN, we first transfer COCO images into the style of ReID datasets and then train DANet with original and these styled COCO images. For CSPL, the following steps are performed, 1) training DANet on original COCO, 2) finetuning DANet with styled COCO images, 3) using DANet to predict pseudo labels on ReID datasets, 4) finetuning DANet on ReID images with pseudo labels. For RoL, the procedure is 1) training DANet on original COCO, 2) using DANet to predict soft pseudo labels on ReID datasets, 3) training DANet on both COCO and ReID images according to Equation 4 of paper [59]. We train MGN+PS on CUHK03 with part segmentation labels obtained from different models trained on COCO, and report the final ReID performance in TABLE VIII. We observe that different adaptation methods lead to comparable ReID outcome. Our final choice is SPGAN, which is the simplest implementation among three methods and has 0.9% advantage in mAP over the baseline.

TABLE IX

TRAINING ON MSMT17 AND TESTING ON MARKET1501 AND CUHK03

Method	MSMT17 → Market1501		MSMT17 → CUHK03	
	Rank-1	mAP	Rank-1	mAP
PCB	58.7	30.5	14.3	13.1
PCB+PS	62.2	33.4	15.5	14.2
MGN	60.7	31.8	20.1	17.4
MGN+PS	67.4	39.1	23.5	21.0

Fig. 14. Part segmentation on MSMT17 test images, predicted by our multi-task method PCB+PS.

K. Generalizable Improvement

With the great progress in single-domain ReID, recently more and more researchers are paying attention to cross-domain setting [15], [39], [60]–[63]. In practical scenario, it is desirable if a model trained in one scene can be easily adapted to a new one. Otherwise, the expensive data acquisition for re-training would be inevitable. We train PCB (or PCB+PS) and MGN (or MGN+PS) on MSMT17 and directly test the models on Market1501 and CUHK03. The scores are shown in TABLE IX. Compared to PCB, PCB+PS increases Rank-1 Accuracy by 3.5% and 1.2% for MSMT17 → Market1501 and MSMT17 → CUHK03 respectively, and 2.9% and 1.1% in mAP. Compared to MGN, MGN+PS increases Rank-1 Accuracy by 6.7% and 3.4% for MSMT17 → Market1501 and MSMT17 → CUHK03 respectively, and 7.3% and 3.6% in mAP. We can see that the proposed method is indeed generalizable across domains and is with practical meaning.

L. Visualize Segmentation and ReID Testing

Although the part segmentation head is not used during ReID testing, we here demonstrate the segmentation quality of PCB+PS on test images of MSMT17, as in Fig. 14.

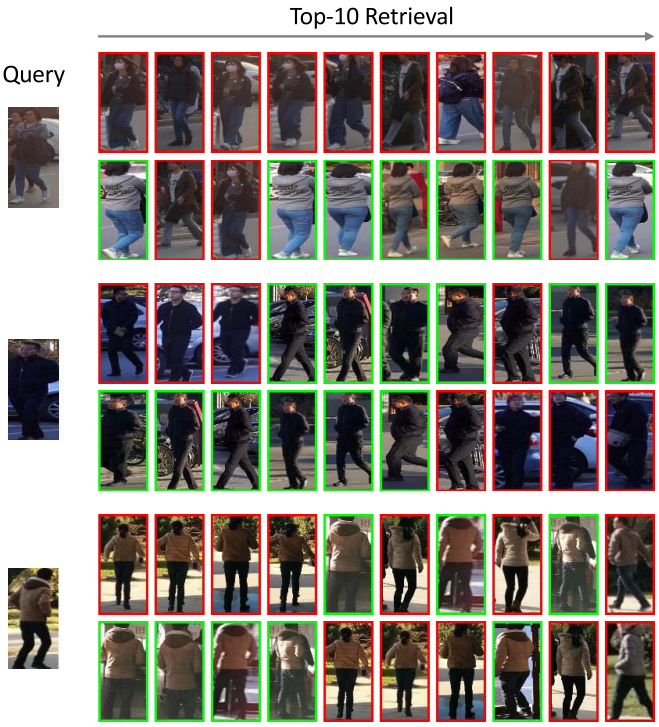


Fig. 15. Three testing cases on MSMT17. In each case, the first and second rows are generated by PCB and PCB+PS respectively. The first column are query images, while the top-10 ranked gallery images are shown to the right. In each row, green (or red) bounding boxes denote same (or different) identity compared to the query image.

From the result we see that the segmentation task is decently trained as well, instead of being sacrificed for the sake of ReID. Since the two tasks can be compatibly trained, we have the chance to utilize segmentation results to perform local pooling or assist other body perception tasks, without the need of an extra backbone, which we believe to be insightful for motivating lightweight implementation in the literature.

We show some ReID test cases upon which PCB+PS makes improvement over PCB, as in Fig.15. The test set is MSMT17, where we can observe various lighting and weather conditions. The failure cases of PCB sometimes seem understandable. For example, in case (b), both the query and top-ranked images are male, wearing similar black clothes, and with short hair. In case (c), PCB may be distracted by analogous background, clothes material and color. The proposed part awareness constraint during training successfully corrects the mistakes with a set of more discriminative features.

V. Conclusion

In this paper, we address person ReID and propose to enhance models with part awareness. Specifically, we embed part knowledge into ReID models by training the additional task of part segmentation. Although being straightforward, our method achieves consistent improvement over three representative ReID models, evaluated on four large-scale benchmarks. When incorporated with MGN, our model obtains state-of-the-art performance. Through quantitative analysis upon PCB, we find our proposal helps to learn a set of more diverse features for identification. Qualitative visualization on MGN also reveals that our method encourages ReID model to attend to more regions on human body, which could reduce the potential of overfitting to salient body regions. To shed light on the mechanism behind the improvement, extensive experiments are carried out to analyze structure of our segmentation head, part granularity in supervision, and loss weight of segmentation task, *etc*. In addition to improving ReID performance, we further demonstrate that the segmentation task is optimized in a decent way as well. In this way, the part labels predicted by our multi-task model may have the chance to provide assistance for more sophisticated model design which requires both additional body information and lightweight implementation. We would embark on the hypothesis in our future work.

References

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: http://arxiv.org/abs/1610.02984

[2] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.

[3] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1249–1258.

[4] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-End comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[5] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[6] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.

[7] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.

[8] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated data in person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1391–1403, Mar. 2019.

[9] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 480–496.

[10] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, Oct. 2018, pp. 274–282.

[11] *Person Re-Identification Baseline in Pytorch*. Accessed: Oct. 10, 2018. [Online]. Available: https://github.com/layumi/Person_reID_baseline_pytorch

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[13] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[14] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[15] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[17] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-Reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[19] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 384–393.

[20] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.

[21] F. Zheng *et al.*, "Pyramidal person re-IDentification via multi-loss dynamic training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8514–8522.

[22] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.

[23] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[24] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[26] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 365–381.

[27] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 403–412.

[28] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3960–3969.

[29] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[30] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 932–940.

[31] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.

[32] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 402–419.

[33] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 667–676.

[34] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[35] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2015, pp. 91–99.

[38] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, Oct. 2007, pp. 1–7.

[39] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[42] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," 2019, *arXiv:1904.09237*. [Online]. Available: http://arxiv.org/abs/1904.09237

[43] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896

[44] Y. Wang *et al.*, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8042–8051.

[45] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5098–5107.

[46] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.

[47] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7134–7143.

[48] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.

[49] X. Qian *et al.*, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 650–667.

[50] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.

[51] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 8295–8302.

[52] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5735–5744.

[53] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-And-Aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.

[54] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.

[55] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[56] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–20, 2017.

[57] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 420–428.

[58] P. Soviany, R. Tudor Ionescu, P. Rota, and N. Sebe, "Curriculum self-paced learning for cross-domain object detection," 2019, *arXiv:1911.06849*. [Online]. Available: http://arxiv.org/abs/1911.06849

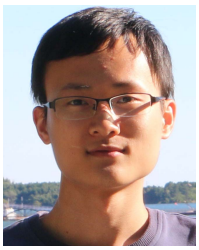
[59] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 480–490.

[60] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.

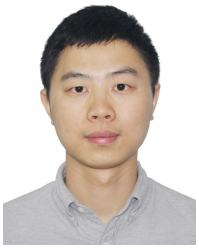
[61] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.

[62] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 719–728.

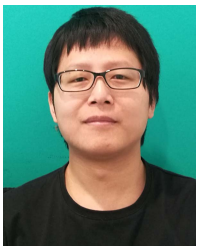
[63] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7202–7211.



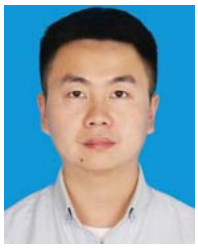
Houjing Huang (Student Member, IEEE) received the B.E. degree from the South China University of Technology (SCUT), Guangzhou, China, in 2015. He is currently pursuing the Ph.D. degree with Center for Research on Intelligent System and Engineering (CRISE), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS). His research interests include deep learning and person re-identification.



Wenjie Yang (Graduate Student Member, IEEE) received the B.E. degree from Beihang University (BUAA) in 2016. He is currently pursuing the Ph.D. degree with Center for Research of Intelligent System and Engineering (CRISE), Institute of Automation, Chinese Academics of Sciences (CASIA), Beijing, China. He is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS). His research interests include computer vision, deep learning, person re-identification, and person search.



Jinbin Lin received the B.S. degree from Sun Yat-sen University (SYSU) in 2011 and the M.S. degree from the Beijing University of Posts and Telecommunications (BUPT) in 2014. He is currently working at Horizon Robotics. His research interests include computer vision and deep learning.



Guan Huang received the B.E. degree from the Huazhong University of Science and Technology in 2013 and the master's degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2016. He is currently the Director of the Algorithm Department, Xforward AI Technology Company, Ltd. He has published research papers in the areas of computer vision and pattern recognition at international journals and conferences such as CVPR, ICCV, AAAI, and the IEEE SIGNAL PROCESSING LETTERS. His current research interests include computer vision, deep learning, object recognition, and face recognition.

Jiamiao Xu received the B.S. and M.S. degrees from the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, in 2016 and 2019, respectively. His research interests include computer vision, pattern recognition, and machine learning.

Guoli Wang received the B.S. degree in electronic information science and technology from the China University of Mining and Technology, Xuzhou, China, in 2011, and the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. His current research interests include computer vision, pattern recognition, and machine learning.

Xiaotang Chen (Member, IEEE) received the B.E. degree from Xidian University in 2008 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2013. In 2013, she joined CASIA as an Assistant Professor. She is currently an Associate Professor with Center for Research on Intelligent System and Engineering. Her current researches focus on computer vision and pattern recognition, including object tracking, person re-identification, and attribute recognition. She served as the technical program committee member of several conferences.

Kaiqi Huang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Nanjing University of Science Technology, China, and the Ph.D. degree from Southeast University. He is currently a Full Professor with Center for Research on Intelligent System and Engineering (CRISE), Institute of Automation, Chinese Academy of Sciences (CASIA). He is also with the University of Chinese Academy of Sciences (UCAS), and the CAS Center for Excellence in Brain Science and Intelligence Technology. He has published over 210 papers in the important international journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, THE IEEE TRANSACTIONS ON IMAGE PROCESSING, THE IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, THE IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, CVIU, ICCV, ECCV, CVPR, ICIP, and ICPR. His current researches focus on computer vision, pattern recognition, and game theory, including object recognition, video analysis, and visual surveillance. He serves as a Co-Chair and a program committee member over 40 international conferences, such as ICCV, CVPR, ECCV, and the IEEE workshops on visual surveillance. He is an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS and *Pattern Recognition*.