HUMAN PARSING BASED ALIGNMENT WITH MULTI-TASK LEARNING FOR OCCLUDED PERSON RE-IDENTIFICATION

Houjing Huang [1,2], *Xiaotang Chen* [1,2*] *and Kaiqi Huang* [1,2,3]

[1] CRISE, CASIA [2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
{houjing.huang, xtchen, kaiqi.huang}@nlpr.ia.ac.cn

ABSTRACT

Person re-identification (ReID) has obtained great progress in recent years. However, the problem caused by occlusion, which is frequent under surveillance camera, is not sufficiently addressed. When human body is occluded, extracted features are flooded with background noise. Moreover, without knowing location and visibility of parts, directly matching partial images with others will cause misalignment. To tackle the issue, we propose a model named HPNet to extract part-level features and predict visibility of each part, based on human parsing. By extracting features from semantic part regions and perform comparison with consideration of visibility, our method not only reduces background noise but also achieves alignment. Furthermore, ReID and human parsing are learned in a multi-task manner, without the need for an extra part model during testing. In addition to being efficient, the performance of our model surpasses previous methods by a large margin under occlusion scenarios.

Index Terms— Person Re-identification, Partial, Occlusion, Human Parsing, Multi-task

1. INTRODUCTION

Person re-identification (ReID) [1] aims at matching person images across cameras to determine whether they are depicting the same identity. It has attracted increasing attention in recent years, due to its great potential in video surveillance tasks *e.g*. person retrieval, cross-camera tracking and activity recognition, *etc*. Recent progress in ReID is mainly concentrated on full-body images, which are filtered detection or manually cropped. In real application, however, environmental occlusion and degraded localization of detectors are both inevitable, especially under surveillance scenario, as illustrated in Fig. 1. When body occlusion occurs, the extracted features are flooded with noise. Moreover, directly matching two images without taking into account part location and visibility would cause spatial misalignment.

To cope with the problem, researchers resort to part-level features. Zheng *et al*. [2] first manually crop out visible body region. Then a two-stream model with local-to-local and local-to-global matching is developed. The first is based on sparse coding, and the second on sliding window matching. This method requires manually cropping, and sliding window is time consuming. He *et al*. [5] formulate the problem as reconstructing each patch of query image using patches of gallery image. It still requires solving reconstruction for each pair of query and gallery images during testing, whose time consumption is exaggerated when a large number of queries are required. Sun *et al*. [6] utilize self-supervised learning to predict location and visibility of each part. The drawback is that self supervision assumes too much of body alignment in training set. Miao *et al*. [7] make use of body keypoints to pool part features and indicate visibility. Nevertheless, it is necessary to pass each image through an extra pose estimation model during testing.

(a) Partial-REID [2] (b) Partial-iLIDS'19 [3] (c) Occluded-REID [4]

Fig. 1: Body occlusion caused by environment or detection error is frequent in surveillance scenario. First row contains occluded persons, second row holistic ones.

Considering the benefits brought by part assistance, as well as the necessity of being efficient, we devise a multi-task model with co-training of ReID and human parsing. In our method, four body parts are separately learned, with a branch for each part. The segmentation task shares backbone with ReID, predicting one mask for each part. Predicted masks are not only used for pooling ReID features from corresponding regions, but also for deducing part visibility. During training, ReID loss is calculated only for visible parts. In testing, for a pair of images, we first compute their distance for each part. Then the overall distance is calculated as the average of their commonly visible parts. Compared to previous methods,

*Corresponding author

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

ours does not require strict alignment in training set or an independent part model during testing. Moreover, our distance computation of image pairs can be simply accomplished by matrix multiplication, which is scalable with the number of queries. In other words, we achieve a comprehensive model in an efficient way.

Since part annotations are not available in ReID datasets, we train a human parsing model on COCO [8] and use the trained model to predict pseudo labels on ReID images. The preparation is finished before multi-task training begins. In order to improve the localization precision of the final model under various occlusion scenarios, we think about how we can make full use of COCO images which shows large diversity in human pose and comes with ground truth part labels. Owing to the multi-task characteristic of our framework, we further propose to train human parsing on COCO images as a regularization task.

The efficacy of the approach is verified on four commonly used datasets, Partial-REID [2], Partial-iLIDS'18 [5], Partial-iLIDS'19 [3], and Occluded-REID [4]. We are able to reach best performance on all these datasets. Extensive experiments are also conducted to analyze design choice of main components. The contribution of this paper is three-fold.

- To the best of our knowledge, we are the first to adopt human parsing to address the problem of occlusion in ReID and verify its efficacy.
- We propose a multi-task implementation, which is not only efficient but also enables training precise part localization.
- Our model achieves state-of-the-art performance on commonly used benchmarks, surpassing previous methods by a large margin.

2. RELATED WORK

Person Re-identification (ReID). In order to improve the discrimination ability of ReID models, Sun *et al*. [9] propose to evenly partition feature maps into rectangles and pool fine-grained features within each region. Considering that backbones originally designed for ImageNet classification may not be the best choice for ReID, Zhou *et al*. [10] develops their own backbones with multi-scale features and lightweight implementation. During image matching, key points and part segmentation masks are often used to extract fine-grained features and facilitate part alignment [11]. Sampling methods and loss formulations are also critical for optimizing a discriminative feature space [12].

Occluded ReID. Zheng *et al*. [2] first propose the problem of occluded ReID, or partial ReID. They formulate the protocol as searching for a partially occluded query image among a gallery set of full-body images. A model consisting of two streams is developed accordingly, which takes manually cropped query image and full gallery image as input. In the first stream, both images are partitioned into patches to perform local-to-local matching. In the second stream, the query directly works as a template searching upon the gallery image in sliding window manner. He *et al*. [5, 3] formulate the problem as reconstructing each patch of query image using patches of gallery image, in the feature space. Reconstruction error thus indicates image similarity in a negatively correlated way. Foreground mask is further proposed to eliminate the distraction from background patches [3]. Under the assumption that training images in regular ReID datasets are well aligned, Sun *et al*. [6] employ self-supervised learning to predict location and visibility of each part. A set of part features are predicted and visibility is considered during distance computing. With the assistance of a pose estimation model, Miao *et al*. [7] generate gaussian masks centered at keypoints for extracting local features.

Although segmentation masks have been used in SPReID [13] to assist alignment, our framework has clear distinctions. First, our model is able to tackle the problem of occlusion by disentangling parts and predicting visibility. Moreover, the implementation for discriminative feature learning is carefully designed accordingly. Finally, our multi-task framework avoids additional segmentation model during testing.

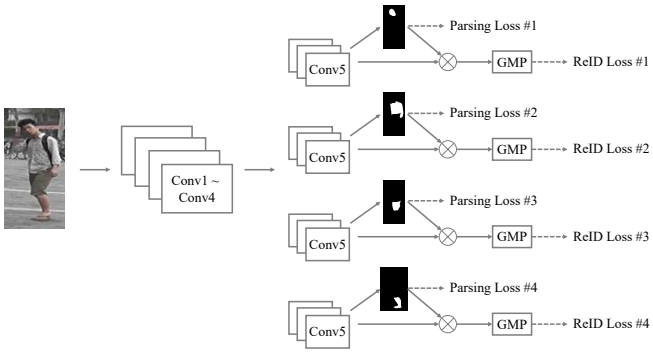

Fig. 2: Overview of our method. We formulate a multi-task framework where ReID and human parsing are simultaneously trained, with part-level feature learning and visibility consideration. For clarity, triplet loss upon concatenated part features is omitted in the figure.

3. METHODOLOGY

We implement a multi-task framework, as shown in Fig. 2, where ReID and human parsing share the same backbone and are simultaneously trained. The four body parts are 1) *head*, 2) *torso and arms*, 3) *upper legs*, and 4) *lower legs and feet*, as illustrated in Fig. 3. It is intuitive that four parts contain distinct feature sets. For example, the first part mainly involves features like hair length, color and shape, as well as facial characteristics, *etc*., while the fourth part is related to shoes types, color, patterns as well as texture of lower part of trousers, *etc*. From this perspective, we propose to model each part with a separate branch, to encourage learning dis-

entangled part features. As a result, we initiate four parallel Conv5 modules without sharing parameters. In each branch, upon Conv5, a lightweight head performs segmentation, whose result indicates the region of the corresponding part. The segmentation mask with binary value is then multiplied with Conv5 feature maps, with a following global max pooling layer to obtain the part feature vector. The visibility of one part is computed as the maximum value of the mask, either 0 or 1. During training, ReID loss is only imposed on visible parts; In testing, distance of a pair of images is calculated using their commonly visible parts. Our model possesses the merits of high efficiency and being effective in learning discriminative features. Details are given below.

3.1. Segmentation Based Part Features

Formally, we denote a training set as $\{(\mathcal{I}_i, y_i, \mathcal{S}_i)|i = 1, 2, ..., N\}$, where N is the number of images, and $\mathcal{I}_i$ is the i-th image with its identity label being y_i. $\mathcal{S}_i = \{\mathcal{S}_i^1, \mathcal{S}_i^2, \mathcal{S}_i^3, \mathcal{S}_i^4\}$ is the binary part labels, where $\mathcal{S}_i^j \in \mathbb{R}^{H\times W}$, H and W being output resolution of Conv5 layer. Note that (pseudo) part labels are predicted by a COCO-trained human parsing model, as in Fig. 3. There are C identities in total and $y_i \in \{1, 2, ..., C\}$. For image $\mathcal{I}_i$, the Conv1$\sim$Conv4 layers first transform it into feature maps which are further processed by four parallel Conv5 layers into $\mathcal{F}_i^j \in \mathbb{R}^{D\times H\times W}, j \in \{1, 2, 3, 4\}$, where D, H, W are number of channels, height and width respectively, with j indexing parts. A part parsing branch consists of {3×3 Conv, BN, ReLU, 1×1 Conv, Sigmoid} layers, taking $\mathcal{F}_i^j$ as input and predicting a probability map $\mathcal{G}_i^j \in \mathbb{R}^{H\times W}$. To indicate region of the part, we discretize $\mathcal{G}_i^j$ into a binary mask $\widehat{\mathcal{G}}_i^j = \mathbb{1}\{\mathcal{G}_i^j > T\}$, where $T = 0.5$ is a threshold and $\mathbb{1}\{\cdot\}$ is an indicator function which gives 1 if the condition is satisfied, and 0 otherwise. In order to extract features of the j-th part, we first multiply the binary mask with each Conv5 feature map, *i.e.* $\mathcal{H}_{i,k}^j = \mathcal{F}_{i,k}^j \bigotimes \widehat{\mathcal{G}}_i^j$ for $k \in \{1, 2, ..., D\}$, in which $\bigotimes$ represents element wise multiplication. The resulting tensor $\mathcal{H}_i^j$ has the same size as $\mathcal{F}_i^j$. Then we perform Global Max Pooling (GMP) and Batch Normalization (BN) upon the result, to obtain a feature vector $\text{BN}(\text{GMP}(\mathcal{H}_i^j)) \rightarrow f_i^j$, where $f_i^j \in \mathbb{R}^D$ is a D-dim vector for the j-th part of the i-th image. The predicted visibility of that part is determined by $v_i^j = max(\widehat{\mathcal{G}}_i^j)$, where $v_i^j \in \{0, 1\}$ has binary value. During testing, distance of a pair of images $(\mathcal{I}_{i1}, \mathcal{I}_{i2})$ is the average over their commonly visible parts, computed as

$$d(\mathcal{I}_{i1}, \mathcal{I}_{i2}) = \begin{cases} \frac{\sum_{j=1}^4 \left\| \overline{f}_{i1}^j - \overline{f}_{i2}^j \right\| v_{i1}^j v_{i2}^j}{\sum_{j=1}^4 v_{i1}^j v_{i2}^j}, & \text{if } \sum_{j=1}^4 v_{i1}^j v_{i2}^j > 0 \\ +\infty, & \text{otherwise} \end{cases}, \tag{1}$$

where $\overline{f}_i^j = \frac{f_i^j}{\|f_i^j\|}$ is a L2-normalized feature vector. An illustration is given in Fig. 4. Note that, when the two images share no visible parts, we set their distance to $+\infty$, since no hint is available to predict them as the same person.

Fig. 3: Demonstration of part labels of COCO dataset [8] (1st row), and predicted labels on Market1501 [14] (2nd row).

3.2. Loss Functions

Identification Loss formulates each identity as one class, utilizes multi-class classifier, and adopts cross-entropy loss to optimize the network. As demonstrated in PCB [9], employing an independent ID classifier for each part benefits learning discriminative part features. The explanation is that if features in each region alone are able to distinguish between identities, then they would be expressive enough. Following this paradigm, we construct four classifiers $g^j : \mathbb{R}^D \rightarrow \mathbb{R}^C, j \in \{1, 2, 3, 4\}$, each containing an FC and Softmax layer. The j-th part feature vector is taken as classifier input to obtain a probability distribution $p^j = g^j(f_i^j)$, where $p^j \in \mathbb{R}^C$. The final identification loss is computed as

$$\mathcal{L}_{ide} = -\frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{j=1}^{4} \log(p_{y_i}^j) v_i^j, \tag{2}$$

in which N_b is number of images in a batch. Note that, only those visible parts contribute to the loss.

Triplet Loss. Identification loss relies on classifier as a bridge to learn identity distinguishable features, while triplet loss [12] directly imposes constraint on sample feature distance to pull close instances of the same identity and push away those from different persons. According to previous practice [15], it is a recommended choice to apply triplet loss to features before BN, which is $\text{GMP}(\mathcal{H}_i^j)$ in our model. We denote it by $h_i^j \in \mathbb{R}^D$. The loss of a triplet $(\mathcal{I}_{i1}, \mathcal{I}_{i2}, \mathcal{I}_{i3})$, where $(\mathcal{I}_{i1}, \mathcal{I}_{i2})$ is a positive pair with $y_{i1} = y_{i2}$ and $(\mathcal{I}_{i1}, \mathcal{I}_{i3})$ a negative pair with $y_{i1} \neq y_{i3}$, is calculated as

$$\mathcal{L}_{tri}^{(\mathcal{I}_{i1}, \mathcal{I}_{i2}, \mathcal{I}_{i3})} = [M + \hat{d}(\mathcal{I}_{i1}, \mathcal{I}_{i2}) - \hat{d}(\mathcal{I}_{i1}, \mathcal{I}_{i3})]_+, \tag{3}$$

in which $M = 0.3$ is a margin, and distance of an image pair is computed as

$$\hat{d}(\mathcal{I}_{l1}, \mathcal{I}_{l2}) = \begin{cases} \frac{\sum_{j=1}^4 \left\| h_{l1}^j - h_{l2}^j \right\| v_{l1}^j v_{l2}^j}{\sum_{j=1}^4 v_{l1}^j v_{l2}^j}, & \text{if } \sum_{j=1}^4 v_{l1}^j v_{l2}^j > 0 \\ -\infty, & \text{else if } y_{l1} = y_{l2} \\ +\infty, & \text{else} \end{cases}. \tag{4}$$

When a pair of images do not share visible parts, we have to exclude them from participating in triplet loss. Since we use Batch-Hard [12] to form triplets, we can simply set the distance to $-\infty$ if it is a positive pair, and $+\infty$ if negative, so that the pair will not be chosen when selecting hardest positive

Vanilla $$\frac{\|\overline{f}_1^1-\overline{f}_2^1\|+\|\overline{f}_1^2-\overline{f}_2^2\|+\|\overline{f}_1^3-\overline{f}_2^3\|+\|\overline{f}_1^4-\overline{f}_2^4\|}{4}$$

Visibility Aware $$\frac{\|\overline{f}_1^1-\overline{f}_2^1\|+\|\overline{f}_1^2-\overline{f}_2^2\|}{2}$$

Fig. 4: Illustration of vanilla and visibility aware image distance (Equation 1). For the former, visibility is neglected; for the latter, only commonly visible parts are considered.

and negative. The triplet loss inside a batch is thus computed as

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

According to Batch-Hard sampling strategy, there are exactly N_b triplets inside a batch of N_b images. Note that, different from independent identification loss for each part, our triplet loss considers four parts of an image at the same time.

Human Parsing Loss. As mentioned above, the predicted parsing probability for j-th part of image $\mathcal{I}_i$ is $\mathcal{G}_i^j \in \mathbb{R}^{H\times W}$, with the corresponding ground truth being $\mathcal{S}_i^j$. For clarity, we denote value of $\mathcal{G}_i^j$ and $\mathcal{S}_i^j$ at location (m,n) by r and t respectively, where $m \in \{1,\ldots,H\}$ and $n \in \{1,\ldots,W\}$. The binary cross entropy (BCE) between r and t is described as

$$\mathcal{L}_{seg}^{i,j,m,n}=-\left[t\log r+(1-t)\log(1-r)\right]. \tag{6}$$

Since foreground region only occupies small portion of each ground-truth mask, the remaining belonging to background. It is a typical case of imbalance between positive and negative samples, for which Focal Loss [16] has proven to be an effective solution. A balancing variable η based on predicted probability is utilized, as follows.

$$\psi=(1-r)t+r(1-t), \tag{7}$$

$$\eta=\left[\alpha t+(1-\alpha)(1-t)\right]\psi^{\gamma}. \tag{8}$$

In the equation, $\alpha=0.25$ and $\gamma=2$ are two hyper parameters. The overall parsing loss of a batch is given below.

$$\mathcal{L}_{seg}=\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{j=1}^{4}\frac{1}{\sum\mathcal{S}_i^j}\sum_{m=1}^{H}\sum_{n=1}^{W}\mathcal{L}_{seg}^{i,j,m,n}, \tag{9}$$

where $\sum\mathcal{S}_i^j$ means the number of positive pixels on $\mathcal{S}_i^j$.

Precise human parsing is critical for localizing body parts, the following feature extraction, and final matching. To this end, we propose to additionally train human parsing on COCO images which not only come with human annotated part labels but also encompass large pose diversity and various occlusion scenarios. We denote the parsing loss on COCO images by $\mathcal{L}_{seg}^{coco}$.

Dataset	Training	Testing	
		Query	Gallery
Market1501 [14]	751 / 12,936	750 / 3,368	750 / 15,913
Partial-REID [2]	-	60 / 300	60 / 300
Partial-iLIDS'18 [5]	-	119 / 119	119 / 119
Partial-iLIDS'19 [3]	-	107 / 238	109 / 238
Occluded-REID [4]	-	200 / 1,000	200 / 1,000

Table 1: Statistics (#Identities / #Images) of ReID datasets.

Final Loss. The overall loss of the multi-task framework is given by

$$\mathcal{L}=\lambda_{ide}\mathcal{L}_{ide}+\lambda_{tri}\mathcal{L}_{tri}+\lambda_{seg}\mathcal{L}_{seg}+\lambda_{seg}^{coco}\mathcal{L}_{seg}^{coco}, \tag{10}$$

where λ_{ide}, λ_{tri}, λ_{seg} and λ_{seg}^{coco} are constants to balance the importance of different tasks, which are set to 1, 0.1, 1, and 1 respectively by default in our experiments. Forward and backward computation is first done for a ReID batch, and then the following COCO batch. The gradients of two batches are summed up before updating the parameters.

4. EXPERIMENT

4.1. Implementation Details

We use ResNet-50 [17] as the backbone, changing the stride of Conv5 from 2 to 1 and discarding the original classifier. SGD optimizer with a momentum of 0.9 and weight decay of 5e-4 is adopted. We initialize learning rates of the backbone and newly added layers to 0.01 and 0.02 respectively, which would be multiplied by 0.1 after 240 epochs. The model is trained for 300 epochs in total. A batch of ReID images are composed of 16 identities each with 4 images randomly sampled. A batch of 32 COCO images are also fed to the network to compute parsing loss $\mathcal{L}_{seg}^{coco}$. We resize images to $width \times height = 128 \times 384$. Random flipping is used as data augmentation during training.

4.2. Datasets and Evaluation Metrics

Following previous works, we train the model on Market1501 [14], which mainly consists of holistic persons, and test it on occluded datasets Partial-REID [2], Partial-iLIDS'18 [5], Partial-iLIDS'19 [3] and Occluded-REID [4]. We train segmentation model DANet [18] on COCO Densepose [8] and then predict parsing masks on Market1501, used as ground truth $\mathcal{S}$ in Section 3, before training our multi-task model. Statistics of ReID datasets are listed in Table 1. Two common evaluation metrics are used, Cumulative Match Characteristic (CMC) [19] for which we report the Rank-1, -3, -5 and -10 accuracy, and mean Average Precision (mAP) [14]. Single-gallery-shot and multi-gallery-shot settings [5] are involved, *i.e.* there is (are) one or multiple images of the query person in gallery set, respectively. Multi-gallery-shot setting is utilized by default unless otherwise declared.

	Publication	Partial-REID		Partial-iLIDS'18	
		R1	R3	R1	R3
AMC+SWM [2]	ICCV15	37.3	46.0	21.0	32.8
DSR [5]	CVPR18	56.9	78.5	63.9	74.8
VPM [6]	CVPR19	67.7	81.9	65.5	74.8
HPNet (Ours)		80.6	91.7	68.9	80.7

Table 2: Comparison with SOTA under singlle-gallery-shot setting. 1st and 2nd highest scores are marked by red and blue, respectively.

	Publication	Occluded-REID	
		R1	mAP
AMC+SWM [2]	ICCV15	31.1	27.3
PCB [9]	ECCV18	41.3	38.9
DSR [5]	CVPR18	72.8	62.8
FPR [3]	ICCV19	78.3	68.0
HPNet (Ours)		87.3	77.4

	Publication	Partial-REID		Partial-iLIDS'19	
		R1	mAP	R1	mAP
AMC+SWM [2]	ICCV15	34.3	31.3	38.7	31.3
PCB [9]	ECCV18	56.3	54.7	46.8	40.2
PGFA [7]	ICCV19	68.0	61.8	-	-
DSR [5]	CVPR18	73.7	68.1	64.3	58.1
FPR [3]	ICCV19	81.0	76.6	68.1	61.8
HPNet (Ours)		85.7	81.8	72.0	58.9

Table 3: Comparison with SOTA under multi-gallery-shot setting.

4.3. Comparison with State-of-the-art Methods

The comparison with state-of-the-art (SOTA) algorithms is reported in Table 2 and 3. The methods we compare include AMC+SWM [2], PCB [9], DSR [5], VPM [6], PGFA [7] and FPR [3]. AMC+SWM first crops out visible body region in query image. Then a two-stream matching process is carried out, including local-to-local matching using dictionary learning, and global-to-global matching in sliding window manner. PCB uniformly partitions feature maps into stripes to extract part-level features. Both DSR and FPR treat the query image as reconstruction of patches from a gallery image, and utilize the reconstruction error as indication of similarity. VPM training part localizer with self-supervised constraint, assuming full body and alignment in training images. PGFA adopts an additional pose model to predict body keypoints in order to obtain local features. From Table 2 and 3, it is obvious that our method achieves SOTA performance, surpassing previous algorithms by a large margin. Under **single-gallery-shot setting**, we surpass VPM by 12.9% (80.6 *vs*. 67.7) and 3.4% (68.9 *vs*. 65.5) in Rank-1 on Partial-REID and Partial-iLIDS'18, respectively. Under **multi-gallery-shot setting**, we surpass previous best method FPR by 9.0% (87.3 *vs*. 78.3), 4.7% (85.7 *vs*. 81.0) and 3.9% (72.0 *vs*. 68.1) in Rank-1 on Occluded-REID, Partial-REID and Partial-iLIDS'19, respectively. Our mAP on Partial-iLIDS'19 is 2.9% (58.9 *vs*. 61.8) lower than FPR, while those on Partial-REID and Occluded-REID are 5.2% (81.8 *vs*. 76.6) and 9.4% (77.4 *vs*. 68.0) higher. Note that Occluded-REID is four times the size of Partial-iLIDS'19. In addition to being effective, our model is efficient since we do not require an extra part localization model in testing (Ours *vs*. PGFA), and not requiring time-consuming pairwise reconstruction (Ours *vs*. FPR).

	mAP	R1	R5	R10
Baseline	54.0	62.1	79.3	85.2
HPNet, test w/o vis	54.8	46.4	82.6	94.4
HPNet, share Conv5	76.3	86.5	93.6	96.3
HPNet, $\lambda_{tri} = 0$	74.3	84.4	93.2	96.0
HPNet, $\lambda_{seg}^{coco} = 0$	70.9	82.4	90.2	93.4
HPNet	77.4	87.3	93.9	96.3

Table 4: Ablation study of our model on Occluded REID.

4.4. Ablation Study

In this section, we conduct ablation study on Occluded REID to analyze components of our model.

HPNet *vs*. Baseline. In case of occlusion, features tend to be contaminated by noisy information. Moreover, without knowing which elements are missing, matching images with these features would cause severe misalignment. To verify the assumption, we implement a global feature based baseline. Concretely, it only consists of a branch with global max pooling to capture features over the whole image. During training, there is no human parsing loss, while other details are the same as HPNet. We denote it by Baseline in Table 4. We observe a huge performance drop, *i.e*. 23.4%, 25.2%, 14.6% and 11.1% in mAP, Rank-1, Rank-5 and Rank-10, respectively. It shows the importance of using part-level features and taking into account part visibility. The retrieving results of two methods are also illustrated in Fig. 5. The baseline is distracted by the plant which occludes the query person, returning images with grass and trees in background. HPNet instead focuses on the visible parts of the person and makes correct decision.

Testing w/ or w/o Part Visibility. During testing, the distance of a pair of images is calculated as Equation 1, *i.e*. averaging over those parts visible in both images. We compare with the vanilla version where distance of two images is just the average of four part distances, neglecting part visibility. The two versions are illustrated in Fig. 4. The testing score of the vanilla version is reported in Table 4, with notation "HPNet, test w/o vis". The scores drop drastically when discarding the visibility deduced from the segmentation masks, with Rank-1 even much lower than Baseline.

Separate *vs*. Shared Conv5 Modules. Different semantic parts contain distinct characteristics. For example, region of head encompasses hair length, color and facial attributes, while torso region is closely related to clothes logos and patterns, *etc*. In our model, we initiate an independent branch for each part to disentangle the feature learning of parts. For comparison, we experiment with four branches sharing the same parameters, represented by "HPNet, share Conv5" in Table 4. We see that sharing parameters leads to a drop of 1.1% in mAP and 0.8% in Rank-1. Since the scale of decreasing is not too serious, and that sharing Conv5 has the advantage of reducing storage and inference time, it could be a good choice in practice to balance between performance and efficiency.

Fig. 5: Retrieving example of Baseline and HPNet. Green (red) bounding box: same (different) identity as query.

Training w/ or w/o Triplet Loss. Identification loss treats each person as a class and adopts a Fully Connected Layer to perform classification. However, during testing the classifier is stripped off and only the features are used for calculating similarity between images. Consequently, it raises a discrepancy between training and testing objectives. By contrast, triplet loss directly optimizes similarity relationship between images in feature space, which is consistent with image ranking in testing phase. Intuitively, combining triplet loss with identification would enjoy the diversity of both and achieve a more discriminative feature space. We carry out an experiment where only identification loss is involved, denoted by "HPNet, $\lambda_{tri} = 0$" in Table 4. We observe that triplet loss brings prominent improvement, *i.e.* 3.1% (77.4 *vs*. 74.3) in mAP and 2.9% (87.3 *vs*. 84.4) in Rank-1.

Training w/ or w/o COCO. Precise part localization is crucial for extracting features from corresponding regions, especially under occlusion scenarios. In order to enhance the human parsing capability of the model, we pass COCO images through the model to train with human parsing loss, *i.e.* $\mathcal{L}_{seg}^{coco}$ in Equation 10. The benefit is two-fold. First, part labels of COCO images are annotated by human, which would stabilize model optimization. Second, it encompasses a variety of poses, with ubiquitous body occlusion, which largely diversifies the training data for human parsing. The part segmentation results of the model trained with or without COCO images are illustrated in Fig. 6, where we can observe obvious improvement in segmentation quality brought by COCO. ReID performance of the model trained without COCO images is recorded in Table 4, *i.e.* "HPNet, $\lambda_{seg}^{coco} = 0$". The finer part localization brings significant improvement in ReID scores, increasing mAP by 6.5% (77.4 *vs*. 70.9) and Rank-1 by 4.9% (87.3 *vs*. 82.4).

5. CONCLUSION

In this paper, we address the problem of occlusion in ReID. To get rid of noise and misalignment caused by occlusion, we adopt human parsing to extract part-level features and indicate part visibility for matching. A multi-task framework is implemented to avoid the need for an extra part model during testing. Extensive experiments confirm the efficacy of our method, which surpasses previous algorithms by a large margin on occlusion benchmarks.

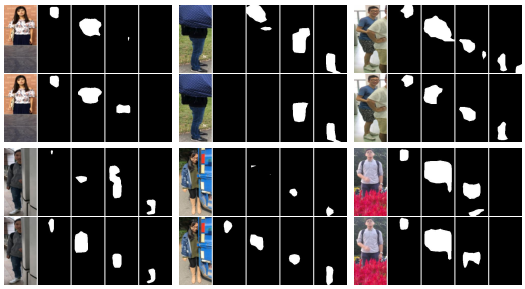

Fig. 6: Human parsing results predicted by the model trained without (1st row) or with (2nd row) COCO images.

6. ACKNOWLEDGEMENT

This work is supported in part by the National Key Research and Development Program of China (Grant No. 2016YFB1001005), the National Natural Science Foundation of China (Grant No. 61673375 and 61721004), and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006).

7. REFERENCES

[1] Zheng et al., "Person re-identification: Past, present and future," *arXiv*, 2016.
[2] Zheng et al., "Partial person re-identification," in *ICCV*, 2015.
[3] He et al., "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *ICCV*, 2019.
[4] Zhuo et al., "Occluded person re-identification," in *ICME*, 2018.
[5] He et al., "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *CVPR*, 2018.
[6] Sun et al., "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *CVPR*, 2019.
[7] Miao et al., "Pose-guided feature alignment for occluded person re-identification," in *ICCV*, 2019.
[8] Güler et al., "Densepose: Dense human pose estimation in the wild," in *CVPR*, 2018.
[9] Sun et al., "Beyond part models: Person retrieval with refined part pooling," in *ECCV*, 2018.
[10] Zhou et al., "Omni-scale feature learning for person re-identification," *CVPR*, 2019.
[11] Zhang et al., "Densely semantically aligned person re-identification," in *CVPR*, 2019.
[12] Hermans et al., "In defense of the triplet loss for person re-identification," *arXiv*, 2017.
[13] Kalayeh et al., "Human semantic parsing for person re-identification," in *CVPR*, 2018.
[14] Zheng et al., "Scalable person re-identification: A benchmark," in *ICCV*, 2015.
[15] Luo et al., "Bag of tricks and a strong baseline for deep person re-identification," in *CVPR Workshops*, 2019.
[16] Lin et al., "Focal loss for dense object detection," in *ICCV*, 2017.
[17] He et al., "Deep residual learning for image recognition," in *CVPR*, 2016.
[18] Fu et al., "Dual attention network for scene segmentation," *arXiv*, 2018.
[19] Gray et al., "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS Workshop*, 2007.