Learning to Know Where to See: A Visibility-Aware Approach for Occluded Person Re-identification

Jinrui Yang[1,2,3,7], Jiawei Zhang[1,2,3], Fufu Yu[7], Xinyang Jiang[6], Mengdan Zhang[7], Xing Sun[7*], Yingcong Chen[4,5], and Wei-Shi Zheng[1,2,3*]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China, [2] Pazhou Lab
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4] The Hong Kong University of Science and Technology
[5] The Hong Kong University of Science and Technology (Guangzhou)
[6]Microsoft Research Asia, [7]Youtu Lab, Tencent
{yangjr27,zhangjw67}@mail2.sysu.edu.cn, fufuyu@tencent.com, xinyangjiang@microsoft.com,
{davinazhang,winfredsun}@tencent.com, yingcong.ian.chen@gmail.com, wszheng@ieee.org

Abstract

Person re-identification (ReID) has gained an impressive progress in recent years. However, the occlusion is still a common and challenging problem for recent ReID methods. Several mainstream methods utilize extra cues (e.g., human pose information) to distinguish human parts from obstacles to alleviate the occlusion problem. Although achieving inspiring progress, these methods severely rely on the fine-grained extra cues, and are sensitive to the estimation error in the extra cues. In this paper, we show that existing methods may degrade if the extra information is sparse or noisy. Thus we propose a simple yet effective method that is robust to sparse and noisy pose information. This is achieved by discretizing pose information to the visibility label of body parts, so as to suppress the influence of occluded regions. We show in our experiments that leveraging pose information in this way is more effective and robust. Besides, our method can be embedded into most person ReID models easily. Extensive experiments validate the effectiveness of our model on common occluded person ReID datasets.

1. Introduction

Person re-identification (ReID) [6, 26] aims to match images of a person across disjoint cameras, which is widely used in video surveillance, security and smart city. Despite the great progress in the recent years, there is still some practical problems that remain unsolved. Especially, occlusion [31] is one common problem that deteriorates the person ReID performance in real-word scenarios. In practice, people can be easily occluded by some obstacles (e.g. baggage, counters, cars, trees) or walk out of the camera fields, making the prediction results prone to error.

*Corresponding author: Wei-Shi Zheng, Xing Sun

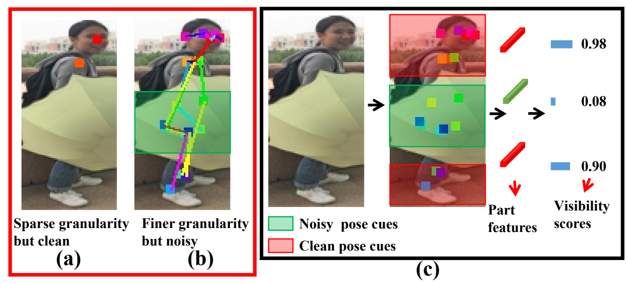

Figure 1. (a), (b) illustrate the relationship between the pose granularity and the estimation error on existing person ReID datasets using the off-the-shelf pose estimator. The existing state-of-the-art methods heavily rely on the fine-grained pose cues but suffer noisy pose detections. Thus, they are hard to directly learn an occlusion-robust feature. (c) illustrates our basic idea. Our model learns a robust mapping from imperfect pose cues to the visibility of body parts to alleviate the effect of occlusion.

In recent years, plenty of efforts [31, 32, 17, 4, 21, 11] have been made to solve this occluded person ReID problem. Usually, they adopt extra cues, such as pose or human parsing, to assist in judging the occlusion scenarios. However, these methods assume the existence of fine-grained and error-free extra cues, which is difficult to achieve in practical scenarios. For example, [17, 4] utilizes 18 pose keypoints. And if we use less keypoints, the performance could drop. However, such fine-grained keypoints may come with high estimation error in practice. As illustrated in Fig. 1 (a) to (b), we can see that as the granularity of pose improves, the estimation error also increases, and the estimation error may degrade the robustness of the existing ReID model. In this paper, we aim to find a solution to handle the occlusion problem without heavily relying on the pose information, i.e., it can achieve comparable/superior performance with coarse pose information.

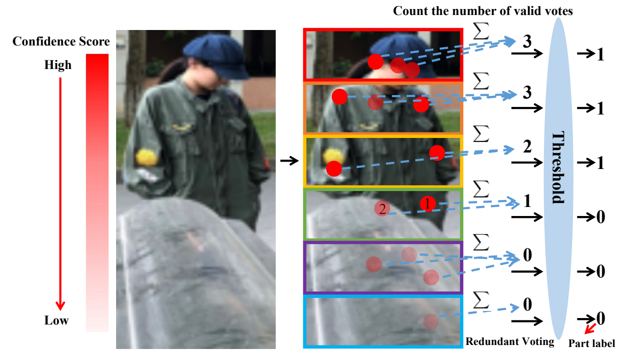

Figure 2. An illustration of the mechanism of part label generation. The red dots represent key points. The shade of the color represents confidence score. When confidence score of a key-point is smaller than the λ, its voting weight is set to 0. Otherwise, it is 1. Thus we can see fourth part contains two key-points(1 and 2). But the confidence score of 2 is very low. Thus the total voting weight of the fourth part is 1. After obtaining the total voting weight of each part, we set a voting weight threshold to produce the part label.

We design a simple but effective framework to accurately perceive the occluded region of person with the help of a region visibility discriminator. In particular, we firstly learn a part label generator, as illustrated in Fig. 2, to generate part labels of different body parts. We then develop a region visibility discriminator based on the generated part labels, which learns a robust mapping from pose to visibility score of horizontal stripes. More concretely, we incorporate the off-the-shelf pose estimator to obtain pose keypoints of person. For each keypoint, the pose estimator predicts its coordinates and confidence score. We use redundant voting to utilize these information to determine whether a part is visible or occluded. We then can obtain the coarse part labels of different body parts. Then we utilize these coarse part labels to optimize the region visibility discriminator in the training stage. Finally, in the test stage, we use the learned region visibility discriminator to predict the visibility scores of person parts without using any keypoint detectors. We conduct comprehensive evaluations on four benchmark datasets, which proves the effectiveness of our part-label based ReID algorithm.

The main contributions of this work are summarized as follows: 1) To the best of our knowledge, we are the first to quantitatively explore how the quality of pose information can influence the performance in occluded person ReID. 2) We propose a novel pose discretization based approach that is robust to the quality of pose information. Even if the keypoints are sparse and inaccurate, our model still achieves promising performance. 3) Our model achieves superior performance on popular occluded person ReID datasets. Besides, our model runs 404.0 FPS during inference, which is about 10 times faster than HOReID [21] (35.3 FPS)[1].

[1]The GPU device of the evaluation is NVIDIA GeForce RTX 2080 Ti.

2. Related work

Part-based Re-Identification. Part-based person ReID methods focus on utilizing local descriptors from different regions to improve the discriminative ability and robustness of the algorithm. Sun *et al.*[20] propose a part-based convolutional baseline (PCB), which partitions feature maps into several horizontal stripes and learns each part-level feature by multiple classifiers. Based on PCB, Fu *et al.*[3] introduce a multi-scale and more fine-grained partition method to enhance the discriminative capabilities of various person parts. Yao *et al.* [22] design a part loss network to minimize the empirical classification risk on the training set and gain the discriminative power on unseen persons. Zheng *et al.* [24] propose a coarse-to-fine pyramid model to incorporate local information, global information and the gradual cues between them and to alleviate the effect of inaccurate bounding boxes. Although these methods can achieve good results in non-occluded person images, they don't explicitly deal with contaminated person regions, leading to obvious mismatch in occluded scenarios.

Occluded Person Re-identification. Occluded Person Re-identification is first studied by [31]. Early works [31, 32] try to learn occlusion-free and discriminative features to match image pairs in response to diverse occlusions. However, it is difficult to learn such features in a self-learned manner without the help of extra cue. Thus extra cue based methods [17, 4, 21, 11] have been the mainstream for occluded person ReID. Wang *et al.*[21] utilize graph convolutional network to model the high-order relations among keypoint based local semantics for the robust alignment. Gao *et al.*[4] propose to use pose-guided attention mechanism to learn clean and discriminative local features. He *et al.*[11] design a method to reconstruct the feature map of non-occluded regions guided by the extra semantic cues. Miao *et al.*[17] utilize the attention maps of pose joints to disentangle the useful information from the occlusion. However, these methods heavily rely on finer-grained extra cues and are sensitive to the granularity of extra cues. Meanwhile, [4, 21, 17] rely on the off-the-shelf pose estimator in the test stage, and thus noisy pose detections degrade their performance. Specially, compared with [17], our method has three main differences as follows: (1) [17] judges whether or not a part should be considered visible or occluded by directly utilizing these pose estimation detectors in an off-the-shelf manner. However, the underlying gap between datasets for pose estimation and person ReID remains a obvious obstacle. We use the region visibility discriminator, which is optimized on the same source datasets in an end-to-end scheme, to produce accurate visibility information of body parts.(2) Our method is more efficient because the keypoint detector is not needed in the test stage. (3) We evaluate occlusion of each person part smoothly in a continuous way.

In summary, different from the aforementioned paper, we further investigate the influence of granularity of pose

Methods CVPR20	key num	Partial-REID Rank1	Partial-REID Rank3	Occluded-REID Rank1	Occluded-REID mAP	Occluded-Duke Rank1	Occluded-Duke mAP
HR v1	6	69.0	80.0	65.9	61.4	49.2	39.5
HR v2	8	70.3	81.7	70.8	65.4	51.2	41.1
HR v3	10	72.0	82.3	72.1	65.4	53.5	43.7
HR(orig)	14	**77.0**	**85.7**	**76.3**	**68.5**	**55.1**	**43.8**

Table 1. The reproduced results of HOReID [21]. We conduct a series of experiments to observe the effect of granularity of keypoints on HOReID. The number of keypoints changes from 14,10,8 util 6, they are denoted as "HR (orig)","HR v3", "HR v2","HR v1", respectively. The 14 keypoints is the original setup of paper [21].

Methods ICCV19	Partial-REID Rank1	Partial-REID Rank3	Occluded-REID Rank1	Occluded-REID mAP	Occluded-Duke Rank1	Occluded-Duke mAP
PForig	68.0	79.0	63.7	53.2	51.9	37.2
PForig w/ GT	70.3	81.7	68.5	56.9	52.8	37.4

Table 2. The reproduced results of PGFA [17].

information and provide a simple yet effective method to robustly make use of pose information on occluded person ReID. We hope our exploration will inspire more works from the ReID research community.

3. Method

In this section, we first carry out a concrete analysis on the pose information discretization in Sec. 3.1. Then, our part-label based ReID algorithm is introduced in Sec. 3.2, particularly emphasizing on the part-label related modules in Sec. 3.3, Sec. 3.4 and Sec. 3.5. Finally, we elaborate on the overall training processes in Sub. 3.6.

3.1. Discretization analysis

We first carry out a quantitative analysis to explore the effect of the granularity of pose keypoints and the estimation error for the existing state-of-the-art occluded ReID models. We take HOReID [21] and PGFA [17] as examples, which have been introduced in the related work section.

To explore the impact of the granularity of pose keypoints, we change the granularity of pose cues from fine to sparse. Specifically, we divide pedestrians into three regions: top, middle and bottom, as shown in part (c) of Fig. 4. For each region, we randomly drop human joints to obtain different granularity of pose cues. The experimental results are presented in Table 1. Table 1 shows that as the number of keypoints decreases from 14 to 6, there is a significant performance degradation for HOReID.

To explore the impact of noise of pose cues, we reproduce the results of PGFA. The results are shown in Table 2. "PFori" is the results of using the original pose cues, which keeps the original setup of paper [17]. To further quantitatively observe the impact of noise, we manually annotate the three datasets with red rectangle to represent visible regions in pedestrian images. As shown in Fig. 4 (b), the red rectangle box can been as precise extra cues without any noise, which can be seen an approximate ground-truth pose annotation. With this kind of refinement, the results using manual annotation extra cues are denoted by the "PForig w/ G" in Table 2. We can see PGFA obtains consistent performance gains on three different datasets.

In conclusion, the set of control experiments prove the following insights: 1) Although sparse pose granularity contains less noise, it causes an obvious degradation on performance of ReID models. 2) Extra cues is usually extracted with estimation error, which harms the performance of ReID models; The ground-truth fine-grained extra cue can bring performance gains. But the ground-truth fine-grained extra cue is hard to obtain, which usually requires time-costing and labor-consuming annotations. In this paper, we find an easy but effective solution, which utilizes the imperfect extra cues in coarse manner and is robust to the granularity of extra cues and the estimation error. Meanwhile, the method can achieves the similar ReID performance as exploiting the fine-grained precise extra cue.

3.2. Overall Part-label based Framework

As shown in Fig. 3, our model is based on PCB [20]. The model mainly includes a Part Label Generator, a region visibility discriminator and a visibility-guided constraint, which aims to learn a robust mapping that maps keypoints to visibility scores of coarse stripes. The detail of our model is described below.

3.3. Part Label Generator with a Redundant Voting

To convert the keypoint cues to visibility information of person regions, we design a part label generator based on a redundant voting strategy, which is partially inspired by work [17]. But we further consider the information of multiple keypoints in each person stripe. Meanwhile, our purpose is quite different from [17], which has been discussed in related work. Following existing works [17, 4, 21], we use an off-the-shelf pose estimator to obtain pose keypoints of a person. In this paper, we choose the AlphaPose [2] as our pose estimator, which is pre-trained on the COCO dataset [15]. For each keypoint, the pose estimator predicts its coordinates and confidence score. For a person image, we can obtain totally K human keypoints, where K is 18 in this paper. The information of each keypoint can be denoted mathematically as $(cx_j, cy_j, s_j), j = 1, 2, \ldots, K$. cx_j, cy_j, s_j denote the horizontal and vertical coordinates, confidence score of jth keypoint respectively.

The original meaning of the confidence score is to represent the probability that the keypoint belongs to the human joint (e.g., head, hand, leg, etc.). It can also be used to represent a more coarse meaning that indicates the visible information of person part. As shown in Fig. 2, we can see a body stripe may contains multiple keypoints. So we design a redundant voting method to convert confidence score of multiple keypoints to visible information of person part, denoted as part label. We argue that each keypoint should has a voting weight to indicate whether or not a part should be considered visible or occluded. Meanwhile, we also know that when the confidence score of a keypoint is high, it is likely to be visible. Otherwise, it may be occluded. Thus we

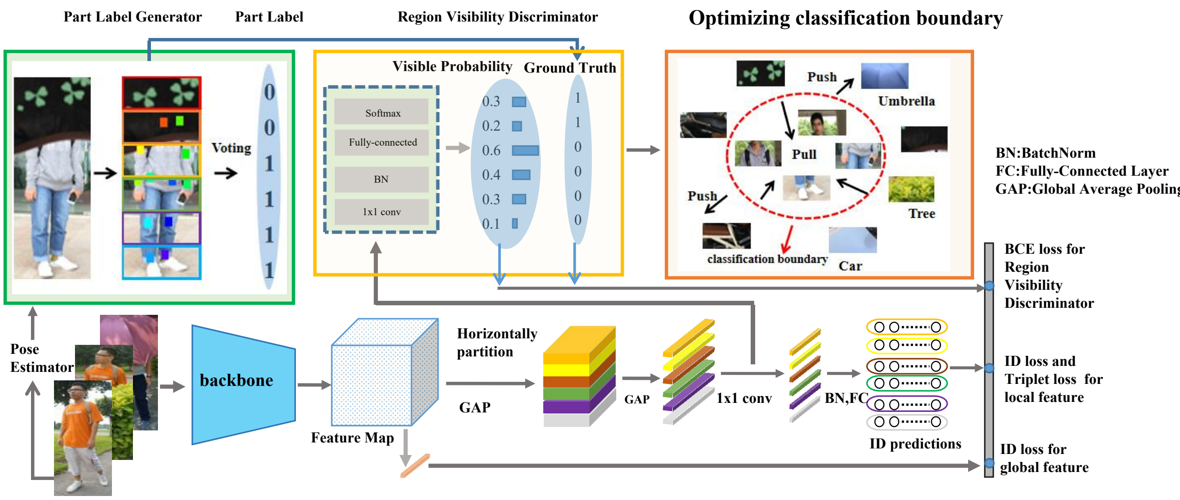

Figure 3. The overall architecture of our proposed method. We use a CNN backbone for the feature extraction. Similar to PCB [20], the feature is learnt in both global and local view. We design a region visibility discriminator in the local branch to predict visibility of coarse stripes, as illustrated in the yellow rectangle. The ground-truth part label is generated by the part label generator based on the keypoint predictions, as illustrated in the green rectangle. The region visibility discriminator is supervised by part labels to learn a compact classification boundary by pulling visible body part(e.g., head, the upper body,etc.) together and pushing the occluded body part(e.g., umbrella, tree, car) together in training stage, as illustrated in the orange rectangle.

set a threshold λ to assign different voting weight to different keypoints. The voting weight can denote whether vote of a keypoint is valid or not. When the $s_j \geq \lambda$, the voting weight of keypoint is 1, which indicates that the keypoint can produce a vaild vote. When the $s_j < \lambda$, the voting weight of keypoint is 0. Finally, we sum the total voting weight of all keypoints in each stripe, which is denoted as T. If the total voting weight of a body part reaches a certain level, this part can be regarded as visible. Thus we set a voting weight threshold W to judge whether a part should be considered visible or occluded. When $T \geq W$, the stripe is seen as visible and set its part label to 1. Otherwise, the stripe is seen as occluded and set its part label to 0. Thus, for each part $i = 1, \ldots, N$, its part label l_i is written as follows:

$$\begin{aligned} w_j &= f(s_j - \lambda) \quad (j = 1, \ldots, K), \\ T_i &= \sum_{j=1}^{K} w_j \text{ if } \exists c y_j \in \left[\frac{i-1}{N}H, \frac{i}{N}H\right), \\ l_i &= f(T_i - W) \quad (i = 1, \ldots, N), \end{aligned} \tag{1}$$

where $f(x) = 1$ if $x \geq 0$, otherwise, $f(x) = 0$. H denotes the height of the original image. Thus, a person horizontally partitioned into N body regions, each region has a ground truth label $l_i \in \{0, 1\}$. $l_i = 0$ or 1 means occluded or visible, respectively.

3.4. Region Visibility Discriminator

Person ReID aims to learn a distance that images of the same identity should be as close as possible. As shown in Fig. 4 (a), if we directly calculate the distance between occluded and holistic person, the final distance would be obvious inaccurate. An intuitive idea is to rough visibility information of person images by utilize extra pose cues. But the original pose cues cannot be directly quantified and used in calculating the effective distance between occluded and holistic person images. Thus we design a region visibility discriminator to learning a robust mapping from the coarse pose cues to the visibility scores of person parts.

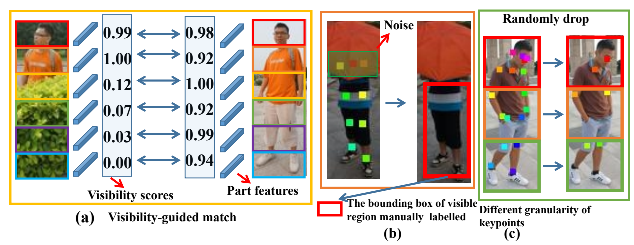

Figure 4. (a) An illustration of visibility scores for parts, which help avoid disturbances from occlusions. (b) This red box of manual annotation can be seen a precise extra cues without any noise, which can used to explore the effect of noise in PGFA [17] in Sec. 3.1. (c) An illustration of different granularity of pose cue.

We introduce the region visibility discriminator formally below. Given a person image, we denote it as I, whose feature map extracted by the backbone model is denoted as F. $F_i \in \mathbb{R}^{h \times w \times c}$, in which h, w, c denote the height, width and channel number, respectively. The feature map F is horizontally partitioned into N regions. Then each region is processed by average pooling and each region feature is represented as $x_i \in \mathbb{R}^c, i = 1, \ldots, N$. c is the dimension of feature. Our region visibility discriminator takes each region feature x_i as input to estimate its visibility scores. The ground-truth part label, generated by our part label generator, is used to optimize the region visibility discriminator in an end-to-end manner.

We implement the Region Visibility Discriminator via a tiny network which consists of a Convolutional layer of 1×1 filter, a BatchNorm layer, a Fully-connected layer,and a Sigmoid activation layer. The output of the region visibility discriminator is visibility scores of person parts, and its range is $(0, 1)$. Further the visibility scores are introduced into the ReID triplet loss to obtain a more accurate distance (see Sec. 3.5).

3.5. Visibility-Guided Constraint Loss

In the ReID task, the fundamental learning process of human appearance consistency is usually accomplished by common metrics such as the contrastive loss and the triplet loss. Typically, the batch hard triplet loss [12] is exploited in this paper and formulated as follows:

$$L_{triplet} = \max(D(f_a, f_p) - D(f_a, f_n) + \mathrm{m}, 0), \quad (2)$$

where f_a,f_p,f_n are features extracted from the anchor, positive and negative samples, respectively, and m is the margin of the triplet loss. Traditionally, $D(\cdot)$ is the L2-norm distance or cosine distance. However, as shown in Fig. 4 (a), the occluded region is obviously obstacle for computing the effective distance. Thus only consider the global distance is inaccurate and we need to alleviate the negative influence of occluded part by part labels. Therefore, borrowed from the similar matching strategy in [17], we use Equation (3) as the new distance function of measuring image pairs. But different from [17], we integrate the matching strategy into triplet loss, which can optimize the model in end-to-end manner.

$$\text{dist} = \frac{\sum_{i=1}^{N} (l_i^q \cdot l_i^g) \cdot D(x_i^q, x_i^g) + D(F^q, F^g)}{\sum_{i=1}^{N} l_i^q \cdot l_i^g + 1}. \quad (3)$$

Specifically, the l_i^q and l_i^g denote the visibility scores of ith part of query and gallery images respectively. $\cdot$ denotes multiplication. $D(x_i^q, x_i^g)$ is original distance of each corresponding feature pairs. As shown in Fig. 4(a), when there exists occlusion, the $l_i^q \cdot l_i^g$ would be small (close to 0), which can be seen as a weight that reduce the influence of the occlusion on the final distance. When a pair of features are both visible, the $l_i^q \cdot l_i^g$ would be high(close to 1). Then $D(F^q, F^g)$ are utilized to measure distance of global feature. Finally, we use Equation (3) to replace the conventional $D(\cdot)$ as more representative measurement metric in hard triplet loss. The visibility-guided hard triplet loss is denoted as $\mathcal{L}_{VGTri}$.

Meanwhile, in the test stage, we also use Equation (3) as the distance measurement metric between query and gallery images. Note that the visibility scores of person part are generated by region visibility discriminator instead of pose detectors in test stage. It means we do not need to rely on any extra cues in test stage, which can relax the constraints of expensive resource overhead when deploying re-id models.

3.6. Objective Function

Except for the Visibility-Guided Hard Triplet loss mentioned above, we employ two type loss function to optimize our proposed model, including the Cross-Entropy loss and the hard triplet loss in training phase.

ID loss. For basic discrimination learning, we regard the identification task as a multi-class classification problem, the ID loss can be formulated as Equation (4)

$$\mathcal{L} = CE(\hat{y}, y), \quad (4)$$

where CE denotes the cross-entropy loss, $\hat{y}$ denotes the prediction and y denotes the ground truth identity. In the training phase, global feature, part-level feature and region visibility discriminator are supervised by cross-entropy loss which are denoted as $\mathcal{L}_{gID}$, $\mathcal{L}_{pID}$, $\mathcal{L}_{BCE}$.

Triplet loss. For the Sec. 3.5, we propose to use visibility-guided hard triplet loss as the new distance metric of global feature. For part-level feature, we still use batch hard triplet loss to further learn discriminative part-level representation, which is be denoted as $\mathcal{L}_{pTri}$.

Therefore, the overall objective function is:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{pID} + (1-\alpha)\mathcal{L}_{gID} + \mathcal{L}_{pTri} + \mathcal{L}_{BCE} + \mathcal{L}_{VGTri}. \quad (5)$$

where α is balanced factor.

4. Experiments

4.1. Datasets and Evaluation Protocols

To illustrate the superiority of our proposed method in ovecoming the occlusion problem in person ReID, we conduct our experiments on four datasets. They are Occluded-DukeMTMC [17], Partial-REID [28], Partial-iLIDS [27], Occluded-ReID [31], respectively.

The Occluded-DukeMTMC dataset is proposed recently by work [17]. Its characteristic is that the query images are all occluded images. It contains 15618 training images, 2210 occluded query images and 17661 gallery images, which is the most largest dataset in occluded person ReID. The Occluded-ReID dataset contains 2000 annotated person images of 200 occluded persons. Each person has 5 full-body person images and 5 occluded person. The Partial-REID dataset includes 600 images of 60 persons. 5 occluded person images and 5 holistic person images are collected for each identity. The Partial-iLIDS is based on the iLIDS dataset [27] dataset contains 119 people with a total 238 person images. Following the test setting of previous work [19, 4, 11, 17, 9, 21], Occluded-ReID dataset , Partial-REID dataset and Partial-iLIDS dataset are only used as test set. When conducting our experiments on these three public datasets, we train our proposed model on the training set of Market-1501 [25]. In this work, for Partial-iLIDS and Partial-REID, we only use the full-body and occluded person images for evaluation. For performance evaluation, we used the cumulative matching characteristic (CMC) and the mean Average Precision (mAP).

Methods	Source	Partial-REID		Partial-iLIDS	
		Rank-1	Rank-3	Rank-1	Rank-3
DSR [9]	CVPR18	50.7	70.0	58.8	67.2
SFR [10]	arXiv18	56.9	78.5	63.9	74.8
AFPB [31]	ICME18	78.5	-	-	-
FPR [11]	ICCV19	81.0	-	68.1	-
TCSDO [32]	arXiv19	82.7	-	-	-
VPM [19]	CVPR19	67.7	81.9	65.5	74.8
PGFA [17]	ICCV19	68.0	80.0	69.1	80.9
PVPM+Aug[4]	CVPR20	78.5	-	-	-
HOReID [21]	CVPR20	85.3	91.0	72.6	86.4
Ours	-	**85.7**	**93.7**	**80.7**	**88.2**

Table 3. Comparison with state-of-the-arts on two partial datasets, *i.e.* Partial-REID [28] and Partial-iLIDS [9] datasets.

4.2. Implementation Details

For the model architectures, we adopt the same modifications for ResNet50 [8] with [16, 20]. The modified ResNet50 is used as our backbone to extract feature maps from input images. For each PCB classifier and the classifier of the region visibility discriminator, following [16], we adopt 1×1 conv, a batch normalization layer [14] and a fully-connected layer followed by a softmax function. For data preprocessing, the input images are resize to 384x128 and augmented with we adopt random horizontal flipping, padding ten pixels, random cropping, and random erasing [29]. When conduct the test on Partial-REID, Partial-iLIDS and Occluded-ReID datasets, followed by [21], we adopt color jitter augmentation to avoid domain variance. For optimization, we adopt standard Stochastic Gradient Descent (SGD) optimization strategy to train our model. The total training epoch number is 80. We use cosine annealing strategy to adjust the learning rate. For Market-1501 and Occluded-DukeMTMC, we set the initial learning rate of backbone to 0.02 and 0.05, respectively. The hyper-parameter α, N are set to 0.9,6 and 0.8,4 ,respectively. For the batch hard triplet loss and visibility-guided hard triplet loss, the batch size is set to 128 with 4 images per person. The threshold λ, which assigns voting weight to different key-points, is set to 0.2. The voting weight threshold $W = 1$.

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

To validate the effectiveness of our proposed method on the occluded person ReID problem, we compare our presented method with several recent state-of-the-art methods on the above mentioned four datasets. Meanwhile, we also compare our method with the state-of-the-art methods designed for holistic person.

Compared to state-of-the-art methods designed for Occluded person ReID. Compared with PGFA [17], the rank1 accuracy of our model significantly surpass this method by 17.7%, 11.6% on Partial-REID and Partial-iLIDS, respectively. Meanwhile, on the Occluded-Duke dataset, our model outperforms the method by 10.8%/9.0% in rank1/mAP. The main reason is that PGFA directly depends on the pre-trained pose estimator to generate part labels, it is easily affected by the accuracy of the pose estimator. However, our superiority lies in the design of the region visibility discriminator, which predicts the visibility scores of stripes and can conveniently be integrated into the feature learning model in an end-to-end scheme.

Methods	Source	Occluded-Duke		Occluded-REID	
		Rank-1	mAP	Rank-1	mAP
Part-Aligned [23]	ICCV17	28.8	20.2	-	-
PCB [20]	ECCV18	42.6	33.7	41.3	38.9
Part Bilinear [18]	ECCV18	36.9	-	-	-
FD-GAN [5]		40.8	-	-	-
AMC+SWM [28]	ICCV15	-	-	31.2	27.3
DSR [9]	CVPR18	40.8	30.4	72.8	62.8
SFR [10]	arXiv18	42.3	32	-	-
Ad-Occluded [13]	CVPR18	44.5	32.2	-	-
TCSDO [32]	arXiv19	-	-	73.7	77.9
FPR [11]	ICCV19	-	-	78.3	68.0
PGFA [17]	ICCV19	51.4	37.3	-	-
PVPM+Aug[4]	CVPR20	-	-	70.4	61.2
HOReID [21]	CVPR20	55.1	43.8	80.3	70.2
ISP [30]	ECCV20	62.8	52.3	-	-
Ours	-	**62.2**	**46.3**	**81.0**	**71.0**

Table 4. Comparison with state-of-the-arts on two occluded datasets, *i.e.* Occluded-Duke [17] and Occluded-REID [31].

Compared with PVPM [4], PVPM extracts heatmaps of the keypoints to produce pose-guided attention (PGA) for mining visible parts. Our method surpasses the PVPM by 7.2% rank-1 accuracy on Partial-REID. And also, 11.5%/10.6% rank1/mAP on Occluded-REID. HOReID [21] utilizes graph conventional network to model relations of keypoints. Our approach outperforms HOReID by 7.1%/2.5% rank-1/mAP on Occluded-Duke. Note that both PVPM [4] and HOReID [21] heavily rely on dense and accurate pose keypoints, which is not always accessible as the pose estimator are not optimal. On the contrary, by transforming keypoints to visibility scores, our model is less sensitive to pose estimation error, thus can better locate occluded regions. Compared to ISP [30], we don't need to iteratively learn and classify feature maps on pixel-by-pixel level, which is time-consuming in training and testing stage. Furthermore, compared with the complicated training scheme of ISP [30], our method can be embedded into most person ReID methods easily.

Compared to state-of-the-art methods designed for holistic person. We reproduce and evaluate the two strongest state-of-the-art methods [16, 24] designed for holistic person on the four occluded ReID datasets. As shown in Table 6, our model significantly surpasses these methods because they don't explicitly handle occlusion.

4.4. Ablation Study

The effect of Region Visibility Discriminator. To verify the necessity of region visibility discriminator, we straightly set all the part labels for each image to 1 in testing stage, which means the model can't judge the visibility of each part. From the 1-st row and 2-nd row of Table 5, we can see disabling the function of the region visibility discriminator significantly degrades ReID performance by 8.8%/8.3%

VGTri	VRD	Partial-REID		Partial-iLIDS		Occluded-REID		Occluded-Duke	
		Rank-1	Rank-3	Rank-1	Rank-3	Rank-1	mAP	Rank-1	mAP
×	×	79.7	89.0	78.9	86.5	71.8	62.7	41.7	34.3
×	✓	85.0	93.0	79.8	87.3	80.6	71.0	61.3	46.0
✓	×	81.6	91.3	78.9	88.2	74.7	64.7	47.0	38.8
✓	✓	**85.7**	**93.7**	**80.7**	**88.2**	**81.0**	**71.0**	**62.2**	**46.3**

Table 5. Analysis of Visibility-Guided Hard Triplet loss (VGTri), Region Visibility Discriminator (VRD). **The results of first row can be seen as baseline.** The experimental results show the effectiveness of our proposed methods.

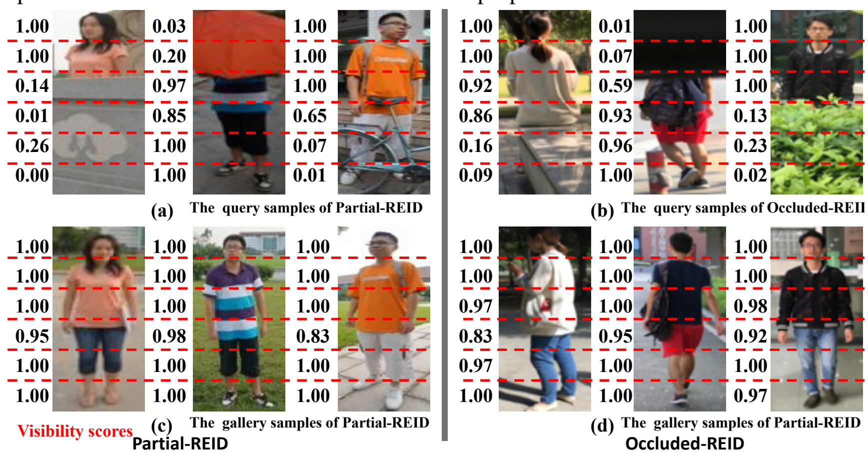

Figure 5. A visualization of visibility scores predicted by region visibility discriminator in testing stage on Partial-REID and Occluded-REID datasets. The six black values on left of image from top to bottom denote the visible probability of each part region in turn.

Methods	Partial-REID		Partial-iLDS		Occluded-REID		Occluded-Duke	
	Rank1	Rank3	Rank1	Rank3	Rank1	mAP	Rank1	mAP
Pyramid [24]	64.6	77.6	75.6	88.2	61.5	58.5	52.8	46.1
Strong Baseline [16]	75.3	85.0	79.8	86.6	64.2	58.5	49.9	43.4
Our	**85.7**	**93.7**	**80.7**	**88.2**	**81.0**	**71.0**	**62.2**	**46.3**

Table 6. Comparing the performance of our method and other two state-of-the-art methods on four datasets.

Methods	key num	Partial-REID Rank1	Partial-iLDS Rank1	Occluded-REID Rank1	Occluded-Duke Rank1
Our v1	6	83.33	76.5	77.7	60.7
Our v2	9	84.0	78.2	76.0	60.5
Our v3	12	84.7	78.2	76.0	60.5
Our v4	15	84.3	79.9	76.0	60.4
Ours	18	**85.7**	**80.7**	**81.0**	**62.2**

Table 7. The number of keypoints changes from 18,15,12,9 to 6. They are denoted as "Ours","Ours v4", "Ours v3","Ours v2,"Ours v1", respectively. The 18 keypoints is the original setup of our paper. The experimental results show the robustness of our part label strategy for utilizing pose information.

Methods	Partial REID	Partial iLDS	Occluded REID	Occluded Duke
	Rank1	Rank1	Rank1	Rank1
PGlabel	84.0	65.6	79.2	48.1
GT	88.0	80.7	81.0	62.2
Ours	**85.7**	**80.7**	**81.0**	**62.2**

Table 8. Analysis of three different methods to obtain part label. The experimental results show the superiority of our proposed method.

Methods	Partial-REID		Partial-iLDS		Occluded-REID		Occluded-Duke	
	Rank1	Rank3	Rank1	Rank3	Rank1	mAP	Rank1	mAP
HTri	85.0	93.0	79.8	87.4	80.6	70.9	61.2	46.0
VGTri	**85.7**	**93.7**	**80.7**	**88.2**	**81.0**	**71.0**	**62.2**	**46.3**

Table 9. Comparing the effect of our visibility-guided hard triplet loss with traditional hard triplet loss on four datasets. The experimental results show the superiority of our methods to traditional hard triplet loss.

and 19.6%/11.7% rank1/mAP on the two largest datasets, namely Occluded-REID and Occluded-Duke. Similarly, for the 3-rd row and 4-th row of Table 5, the accuracy of rank1/mAP drops 6.3%/ 6.3% and 15.2%/7.5% respectively on the same two datasets. Hence, the experimental results show the effectiveness of our proposed region visibility discriminator module. Meanwhile, the consistent significant improvements prove the solid generality of the region visibility discriminator module.

The Effect of Visibility-Guided Hard Triplet loss. Similar to the analysis of the region visibility discriminator, the comparison results from the 1-st row and 3-rd row, 2-nd row and 4-th row of Table 5 depict that the visibility-guided hard triplet loss is effective. Meanwhile, for comparing visibility-guided hard triplet loss(VGTri) and the traditional hard Triplet loss(HTri) [12], we keep all experimental settings same except for the type of the triplet loss and further conduct two experiments. Traditional hard triplet loss and our visibility-guided hard triplet loss are denoted as "'HTri' and "VGTri" respectively. The experimental results are shown in Table 9. "VGTri" fully outperforms traditional hard Triplet loss on four datasets, which further illustrates the superiority of visibility-guided hard triplet loss compared to traditional hard triplet loss.

4.5. Model Analysis

The Superiority of Our Part Label Discretization Strategy. For verifying the superiority of our part label discretization strategy, we compare the other two strategies to get the part label. One is using the off-the-self pose detector to obtain the part label, which is denoted as "PGlabel". The other is using bounding box annotation to obtain part labels, which is denoted as "GT"(e.g., as shown in Fig. 4 (b)). We can regard it as a precise enough extra cue. For fair compar-

Methods	Partial-REID		Partial-iLDS		Occluded-REID		Occluded-Duke	
	Rank1	Rank3	Rank1	Rank3	Rank1	mAP	Rank1	mAP
Mask R-CNN [7]	86.3	91.0	79.0	86.6	81.6	71.6	62.9	45.8
OpenPose [1]	86.0	91.3	80.7	86.6	79.0	69.9	61.5	46.0
AlphaPose [2]	**85.7**	**93.7**	**80.7**	**88.2**	**81.0**	**71.0**	**62.2**	**46.3**

Table 10. Comparing the effect of using different pose detectors on four datasets. AlphaPose is used in our original paper.

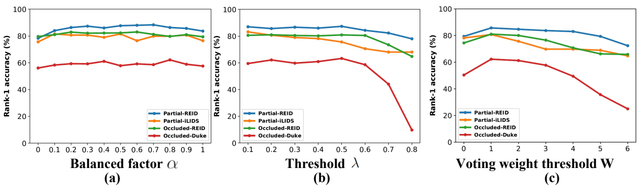

Figure 6. (a) The impact of the loss weight α. (b) The impact of confidence threshold λ. (c) The impact of Voting weight threshold W.

ison, we fix all setup and use the same trained model, only change the way of obtaining part label in inference stage.

The experimental results are presented on Table 8, which suggests the performance of our method fully outperforms the ”PGlabel” on four datasets. Specifically, the rank-1 of our method dramatically surpasses “PGlabel” by 14.1% on Occluded-Duke. The reason may is that the “PGlabel” is based on pose information, only obtaining hard part label (0 or 1). Moreover, “PGlabel” is susceptible to the performance of pose estimator itself. But our part label predicted by region visibility discriminator is in a soft form, whose range is from 0 to 1 (e.g., as shown in Fig. 5). Thus our soft part label is continuous, which can reflect the different degrees of occlusion. Meanwhile, under the guidance of part labels, our method can learn consistent feature in an end-to-end manner. On the other hand, we also can see the performance of our method and “GT” is basically same, which shows that our method can achieve the same performance as exploiting ideally precise extra cues. But compared with “GT”, our method don’t require time-costing and labor-consuming annotations.

The Robustness of Our Part Label Discretization Strategy. To validate the robustness of our part label strategy, we keep our experimental condition fixed and only change the granularity of pose information. And the Table 7 shows that with the variation of pose information, namely the variation of the number of keypoints , the performance of our model is nearly unchanged, which demonstrates that our part label strategy is robust to granularity of pose cues.

The Impact of the Pose Estimation Algorithm. To investigate the impact of different pose keypoint detectors, we use three different pose estimation algorithms, AlphaPose [2], Mask RCNN [7] and OpenPose [1]. As shown in Table 10, we find three pose estimation algorithms achieves similar performance on the four datasets, which proves that our model is robust to different keypoint detectors.

4.6. Hyper-parameter Analysis

The Impact of the balanced factor α. We analyze the effect of the balanced factor α parameters on four datasets. Results in figure 6 (a) indicate that our method is not sensitive to the α in range [0.1,0.9].

The Impact of the Threshold λ. We set a threshold λ (0.2 in this paper) to assign voting weight to different keypoints. We further evaluate the impact of λ. As shown in Figure 6 (b), when λ is set between 0.1 and 0.5, the performance changes smoothly, which indicates that our model is not sensitive to the λ of this range. When λ is too large, the model suffers a performance drop, because the voting weight of few keypoints are set to 1 and further lead to generating abundant incorrect part labels.

The Impact of the voting weight threshold W. We set a voting weight threshold W (1 in this paper) to assign part labels to different body parts. We further evaluate the impact of λ. As shown in Figure 6 (c), when $W = 1$, the model achieves the best performance. The phenomenon explains why our method is robust to granularity of keypoints. When W is set to 1, only needing a small number of key points that can produce valid votes can meet the condition. Thus, as shown in Table 7, we can see that even when the granularity is extreme sparse, our method is still competitive to the current state-of-the-art methods.

4.7. Visualization

we visualise the visibility scores predicted by region visibility discriminator in testing stage. From the Fig. 5, we find although the image region has a variety of diversified appearances, the completely occluded region and the completely visible region both can be predicted a almost ideal visible probability. Moreover, we can see the partially occluded region(e.g., as shown in the third person image on the Fig. 5 (a)), whose visible score is proper and continuously reflects the change of occlusion. It shows our soft part label based on our region visibility discriminator is superiority to hard part label based on pose information.

5. Conclusion

In this paper, by changing the granularity of the pose cue for current state-of-the-art methods, and adding annotation bounding boxes by hands, we conduct some experiments and observe following facts: 1) these method is sensitive to the granularity of the pose cue and prefers fine-grained and accurate pose cue; 2) the pose cue usually contains noise in real scenarios, increasing the difficulty of robust modeling. Then, we propose a simple but effective part-label based algorithm to discretely utilize coarse pose information while maintain good robustness and generality. The experimental results show the superiority of our algorithm.

6. Acknowledgement

This work was supported partially by the NSFC(U1911401,U1811461), Guangdong NSF Project (No. 2020B1515120085, 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03), and the Key-Area Research and Development Program of Guangzhou (202007030004).

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 8

[2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3, 8

[3] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302, 2019. 2

[4] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11744–11752, 2020. 1, 2, 3, 5, 6

[5] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in neural information processing systems*, pages 1222–1233, 2018. 6

[6] Shaogang Gong, Marco Cristani, Shuicheng Yan, Chen Change Loy, and Person Re-Identification. Springer publishing company. *Incorporated*, 1447162951:9781447162957, 2014. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 8

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[9] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018. 5, 6

[10] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018. 6

[11] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8450–8459, 2019. 1, 2, 5, 6

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5, 7

[13] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018. 6

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 6

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[16] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6, 7

[17] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 542–551, 2019. 1, 2, 3, 4, 5, 6

[18] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018. 6

[19] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2019. 5, 6

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 2, 3, 4, 6

[21] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6449–6458, 2020. 1, 2, 3, 5, 6

[22] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019. 2

[23] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3219–3228, 2017. 6

[24] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8514–8522, 2019. 2, 6, 7

[25] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 5

[26] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[27] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR 2011*, pages 649–656. IEEE, 2011. 5

[28] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015. 5, 6

[29] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 6

[30] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer, 2020. 6

[31] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 1, 2, 5, 6

[32] Jiaxuan Zhuo, Jianhuang Lai, and Peijia Chen. A novel teacher-student learning framework for occluded person re-identification. *arXiv preprint arXiv:1907.03253*, 2019. 1, 2, 6