Do Not Disturb Me: Person Re-identification Under the Interference of Other Pedestrians

Shizhen Zhao[1][*], Changxin Gao[1][**], Jun Zhang[2], Hao Cheng[2], Chuchu Han[1], Xinyang Jiang[2], Xiaowei Guo[2], Wei-Shi Zheng[3], Nong Sang[1], Xing Sun[2]

[1]Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, [2]Tencent Youtu Lab, [3]Sun Yat-sen University
Email: {zhaosz, cgao}@hust.edu.cn

Abstract. In the conventional person Re-ID setting, it is widely assumed that cropped person images are for each individual. However, in a crowded scene, off-shelf-detectors may generate bounding boxes involving multiple people, where the large proportion of background pedestrians or human occlusion exists. The representation extracted from such cropped images, which contain both the target and the interference pedestrians, might include distractive information. This will lead to wrong retrieval results. To address this problem, this paper presents a novel deep network termed Pedestrian-Interference Suppression Network (PISNet). PISNet leverages a Query-Guided Attention Block (QGAB) to enhance the feature of the target in the gallery, under the guidance of the query. Furthermore, the involving Guidance Reversed Attention Module and the Multi-Person Separation Loss promote QGAB to suppress the interference of other pedestrians. Our method is evaluated on two new pedestrian-interference datasets and the results show that the proposed method performs favorably against existing Re-ID methods. Our project is available at https://github.com/X-BrainLab/PI-ReID.

Keywords: Person Re-identification; Pedestrian-Interference; Location Accuracy; Feature Distinctiveness; Query-Guided Attention

1 Introduction

Re-IDentification (Re-ID) aims to identify the same person across a set of images from nonoverlapping camera views, facilitating cross-camera tracking techniques used in video surveillance for public security and safety. In general, person Re-ID is considered to be the next high-level task after a pedestrian detection system. Therefore, as shown in Figure 1(a), the basic assumption of Re-ID is that the detection model can provide a precise and highly-aligned bounding box for each individual. However, in a crowded scene, off-shelf-detectors may draw a bounding box containing multiple people, as shown in Figure 1(b). This means the cropped

---

[*] This work was done when Shizhen Zhao was an intern at Tencent Youtu Lab.
[**] Corresponding author.

Fig. 1. Typical samples in the (a) traditional Re-ID, (b) Pedestrian-Interference Re-ID, and (c) Occluded Re-ID from Market-1501 [50], our constructed PI-CUHK-SYSU, and Occluded-DukeMTMC [27,54], respectively. (d) shows the query image (the first sample) and, the comparison of visualization results between the Occluded Re-ID method Foreground-aware Pyramid Reconstruction [7] (the second one) and our PISNet (the third one)

images contain both the target and the interference pedestrians. The interference pedestrian makes the feature ambiguous to identify the target person, which might lead to wrong retrieval results. We call this the Pedestrian-Interference person Re-IDentification (PI Re-ID) problem.

We observe that mutual occlusion of pedestrians often occurs in PI Re-ID. Recent works [7,23,41,3] have well studied the Occluded Re-ID problem. However, in their setting of Occluded Re-ID, the person images are mainly occluded by obstructions like cars, trees, or shelves. This is also reflected in the existing benchmarking Occluded Re-ID datasets, most of which consist of non-pedestrian occlusion, as shown in Figure 1(c). The performance of their approaches degrades if directly applied to PI Re-ID, as shown in the second sample of Figure 1(d). Because they only focus on reducing the influence caused by obstructions and do not specifically consider the interference between pedestrians in a cropped image. Moreover, they do not explicitly learn to draw a precise boundary between two overlapping people so that the extract features are corrupted by each other. As for PI Re-ID, it is different from Occluded Re-ID in two aspects: 1) PI Re-ID focuses on the pedestrian interference, which is more confusing than the non-pedestrian obstructions. 2) PI Re-ID aims to re-identify all the pedestrians appearing in a cropped image, which might be interfered with the background pedestrians or the pedestrian occlusion. Therefore, our setting is more challenging than Occluded Re-ID. Moreover, our setting is more practical in the crowded

situation (*e.g.*, airports, railway stations, malls, and hospitals), where people always share overlapping regions under cameras.

To retrieve a person in the PI Re-ID setting, the extracted features should ensure 1) **location accuracy**: the strong activation on all the regions of targets, 2) **feature distinctiveness**: the trivial feature corruption by other pedestrians. To achieve this goal, we propose a novel deep network termed Pedestrian-Interference Suppression Network (PISNet), which consists of a backbone Fully Convolutional Network (FCN), a Query-Guided Attention Block (QGAB) and a Guidance Reversed Attention Module (GRAM). First, FCN is utilized to extract features for person images. Since the target feature, in a gallery image containing multi-person information, differs on the query, QGAB is designed to enhance the feature of the target in the gallery and suppress that of interference pedestrians, under the guidance of the query. On the one hand, as shown Figure 2(a), for encouraging the **location accuracy** of the attention, our motivation is that, if the attention well covers the regions of the target, the attention feature can be further utilized as the guidance to search the target in other multi-person images. Therefore, GRAM leverages the refined gallery features to guide other multi-person features to formulate attention for targets. On the other hand, as shown in Figure 2(b), to facilitate the **feature distinctiveness** of the attention learning, PISNet utilizes the Multi-Person Separation Loss (MPSL) to maximize the distance between the features, which are extracted from the same gallery but guided by different queries. In addition, as shown in the third sample of Figure 1(d), our PISNet is more capable of depressing the pedestrian interference than the Occluded Re-ID method.
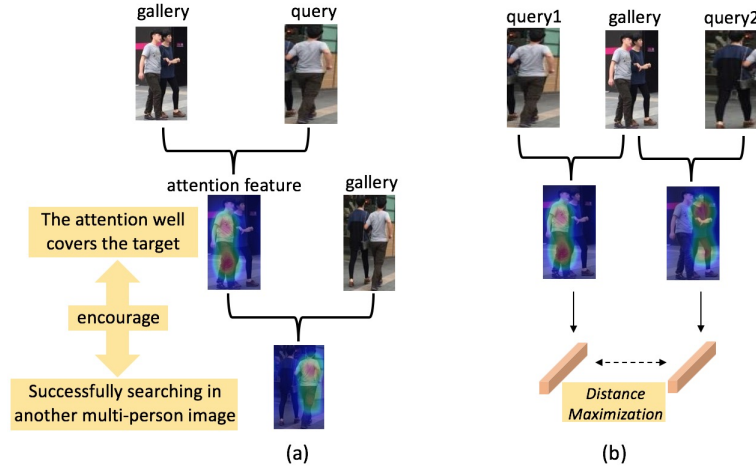
Fig. 2. (a) Enhance the location accuracy of the attention by GRAM. (b) Facilitate the feature distinctiveness of the attention by MPSL

Our **contributions** are listed as follows: 1) To the best of our knowledge, it is the first work that particularly addresses the problem of PI Re-ID. 2) We propose a Pedestrian-Interference Suppression Network (PISNet), which utilizes a query-guided approach to extract the target feature. The involving GRAM and MPSL further promote the location accuracy and the feature distinctiveness of the attention learning, respectively. 3) Since the existing benchmarks largely ignored this problem, we contribute two new datasets, which are specifically designed for this problem with a great deal more pedestrian-interference instances. Our experimental results on these two datasets show that the proposed model is effective in addressing the PI Re-ID problem, yielding significant improvement over representative Re-ID methods applied to the same problem. The experimental results also demonstrate the generalization ability of our method on the general Re-ID datasets (Market-1501 and DukeMTMC-ReID).

2   Related Work

2.1   Person Re-ID

Generally, Person Re-ID can be divided into two steps: calculating a feature embedding and performing feature matching under some distance metric [35,16,10]. We mainly review the former including both handcrafted feature [10,11,13,22] and learned feature [14,45,2,8,19,47] approaches.

In recent years, Re-ID has witnessed great progress owing to the prevailing success of convolutional neural networks (CNNs) in computer vision. However, simply applying CNNs to feature extraction may not yield ideal Re-ID performance due to many problem-specific challenges such as partial body, background perturbance, view point variation, as well as occlusion/misalignment. Combining the image-level information with the human-part information can enhance the robustness of Re-ID models. Moreover, many part-based approaches have achieved considerable improvement [4,24,17,21,53,2,29,34,49,55,30,44,43]. We refer readers to [51] for a more comprehensive review.

2.2   Attention Mechanisms in Person Re-ID

Several studies leverage attention mechanisms to address the misalignment problem in person Re-ID. For example, Chen et al. [1] propose an attentive but diverse network which consists of a pair of complementary attention modules, focusing on channel aggregation and position awareness, respectively. Si et al. [28] use an inter-class and an intra-class attention module to capture the context information for person Re-ID in video sequences. Li et al. [15] leverage hard region-level and soft pixel-level attention, which can jointly produce more discriminative feature representations. Xu et al. [39] utilize pose information to learn attention masks and then combine the global with the part features as feature embeddings.

Previous methods [1,37,39] leverage attention mechanisms to enhance the feature of human bodies. In contrast, in our proposed setting, images contain

other pedestrians, which severely corrupt the feature of a target. Since they cannot distinguish between the target and interference pedestrians, directly applying their approaches will cause the severe corruption of the target feature. .

2.3   Occluded Re-ID

Some related works for the Occluded Re-ID have been well studied. Zheng et al. [36] propose an Ambiguity sensitive Matching Classifier (AMC) and a Sliding Window Matching (SWM) model for the local patch-level matching and the part-level matching, respectively. He et al. [6] propose a Deep Spatial Feature Reconstruction (DSR) model for the alignment-free matching, which can sparsely reconstruct the spatial probe maps from spatial maps of gallery images. He et al. [7] further present a Spatial Feature Reconstruction (SFR) method to match different sized feature maps for the Partial Re-ID. Miao et al. [23] propose the Pose-Guided Feature Alignment (PGFA), which introduces the pose estimation algorithm to enhance the human part feature in an occlusion image. Sun et al. [41] propose a self-supervision model called Visibility-aware Part Model (VPM), which can perceive the visibility of regions. Fan et al. [3] propose a spatial-channel parallelism network (SCPNet), which enhances the feature of a given spatial part of the body in each channel of the feature map.

These methods ignore the interference between pedestrians within a cropped image. Therefore, they cannot well address the PI Re-ID problem, where the large proportion of the pedestrian interference exists. In contrast, in this paper, we focus on suppressing the pedestrian interference, by learning the query-guided attention with the location accuracy and the feature distinctiveness.

3   Pedestrian-Interference Suppression Network

In this work, we assume that in PI Re-ID a query image contains only a single person and the task is to match this query with a gallery consisting of the pedestrian interference. This is based on a practical scenario where a human operator has manually cropped the human body and sent a query to a Re-ID system to search for the same person in another camera view. In this section, we first give an overview of our framework, and then describe more details for each component individually.

3.1   Overview

As shown in Figure 3, PISNet consists of (1) a backbone Fully Convolutional Network (FCN), (2) a Query-Guided Attention Block (QGAB), and (3) a Guidance Reversed Attention Module (GRAM). For each forward propagation in the training stage, we pair a gallery image with a query image. FCN can extract features for the query and the gallery. QGAB finds the common regions between the query and gallery feature maps, and then enhances the common feature in the
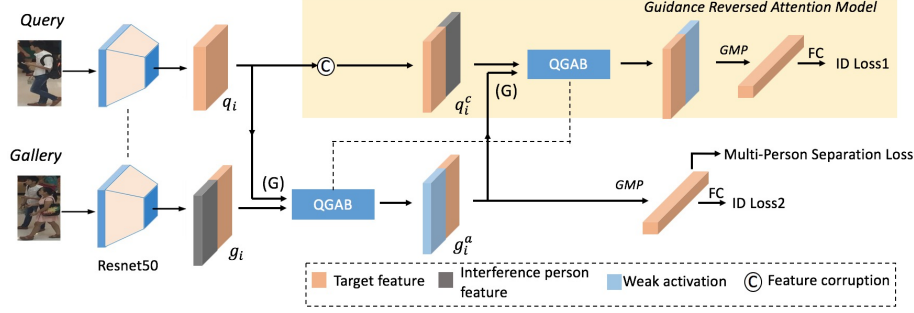
Fig. 3. Illustration of our Pedestrian-Interference Suppression Network (PISNet). For further clarity, the target feature represents the same ID information to the query. PISNet consists of (1) a backbone Fully Convolutional Network (FCN), (2) a Query-Guided Attention Block (QGAB), and (3) a Guidance Reversed Attention Module (GRAM). FCN can extract features for the query and the gallery. QGAB leverages the query feature as the guiding feature (single-person) to formulate attention on the gallery feature (with pedestrian interference). GRAM plays a role in encouraging QGAB to enhance the feature on the regions of a target. The Multi-person Separation Loss promotes the attention to draw a more precise boundary for overlapping instances. g_i and q_i denote the feature map of a gallery and a query, respectively. g_i^a and q_i^c are the refined gallery feature and the corrupted query feature, respectively. GMP denotes the Global Max Pooling. QGABs share the same parameters. (G) denotes the feature as the guidance to QGAB. GRAM is only used in the training stage

gallery feature. For encouraging the location accuracy of the attention, GRAM aims to guarantee that the refined gallery feature has strong attention on all the regions of the target. For the feature distinctiveness of the attention, the Multi-person Separation Loss (MPSL) magnifies distance of the features from the same gallery but guided by different queries. In addition, GRAM is ignored in the testing stage.

3.2   Query-Guided Attention Block

QGAB is depicted in Figure 4. The main goal of this block is to search for spatial similarity between the query and the multi-person gallery. The inputs of QGAB are the query and gallery feature maps. The query is used as the guidance. The output is the spatially enhanced gallery feature. The spatial similarity calculates the inner product of the features from gallery and query branch first, after which Global Max Pooling (GMP) in the channel dimension is applied to formulate a pixel-wise attention matrix. This matrix then is multiplied with the gallery feature in order to enforce a spatial similarity search between the query and gallery feature maps. The overall process of this feature enhancement is formulated as:

$$QGAB(g_i, q_i) = GMP\big(Softmax(c_1(g_i)^T \times c_2(q_i))\big) \times g_i + g_i, \qquad (1)$$
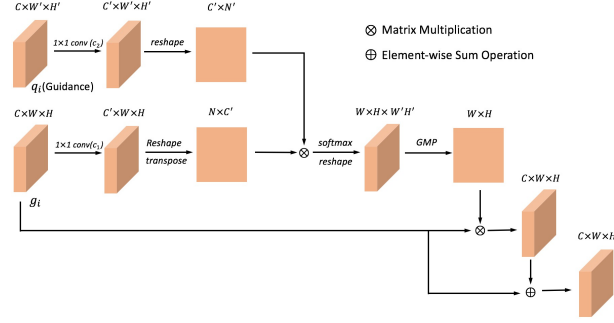
Fig. 4. Illustration of our proposed Query-Guided Attention Block (QGAB). g_i denotes the feature map of a gallery. q_i is the feature map of a query. The feature maps are shown as the shape of their tensors. W and H are the width and height of the gallery feature map. $W^{'}$ and $H^{'}$ are the width and height of the query feature map. C is the number of channels after the backbone. GMP denotes the Global Max Pooling

where g_i is the multi-person feature (gallery), q_i is the single-person feature (query), c_1 and c_2 are convolutional layers, GMP is the Global Max Pooling in the channel dimension and $\times$ denotes matrix multiplication.
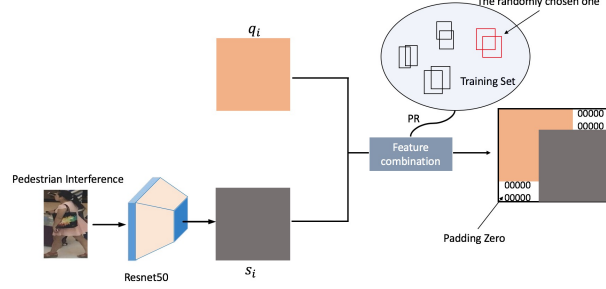
Fig. 5. Illustration of the feature corruption. q_i denotes the query feature map. s_i denotes the feature map of the sampled single-person image. PR denotes the function that extracts the relative position relationship of ground-truth boxes from a multi-person image. Two features are combined following the relative position relationship

3.3   Guidance Reversed Attention Module

GRAM aims to guarantee that the refined gallery feature has the strong attention on all the regions of the target. As shown in Figure 2(a), our motivation is that a well-refined gallery feature can be used as the guidance to formulate the attention

for another gallery feature containing the pedestrian interference. For example, if a gallery contains IDs of A and B. Using the query image of A, the well-refined gallery feature will have the strong activation on regions of A while the feature of B is suppressed. Therefore, the refined feature should be capable of serving as the guidance to formulate attention for another gallery containing person A. In this new attention mask, we expect the feature of A is still enhanced and the feature of another pedestrian is suppressed. The attention formulation for person A in these two feature maps is encouraged by each other.

Therefore, as shown in Figure 3, we utilize the feature, which is refined by our query-guided attention operation, as the guidance feature to formulate the spatial attention on another gallery feature map. In order to reduce the labour of the data collection, we construct the new gallery by a feature corruption operation. Specifically, we randomly select a single-person image and a gallery image with the pedestrian interference. We can extract the single-person feature from the former and the relative position relationship of the involving ground-truth bounding boxes from the latter. Then we corrupt the query feature by combining it with the single-person feature. Specifically, as shown in Figure 5, following the relative position relationship, we put two feature maps on the corresponding positions to generate a multi-person feature map and we pad the remaining regions with zero. The process of this feature corruption is formulated as:

$$FC(q_i, s_i, m) = Combine(q_i, s_i, PR(m)), \qquad (2)$$

where s_i is the feature map of the sampled single person image, m denotes the image with pedestrian interference, PR denotes the function that extracts the relative position relationship of ground-truth boxes from m, and $Combine$ denotes the function that can combine features depending on the relative position relationship of bounding boxes.

Then we input the corrupted query feature and the refined gallery feature into QGAB. In contrast to the last QGAB operation, we reverse the roles of two features. The refined gallery feature is served as a query feature that can guide QGAB to enhance the target feature in the corrupted query features. The overall process of reversed feature enhancement is formulated as:

$$QGAB^r(g_i^a, q_i^c) = GMP\big(Softmax(c_1(q_i^c)^T \times c_2(g_i^a))\big) \times q_i^c + q_i^c, \qquad (3)$$

where g_i^a is the refined gallery feature, q_i^c is the corrupted query feature, and c_1 and c_2 share parameters with the last QGAB.

3.4   Multi-Person Separation Loss

In a pedestrian-interference image, people always share an overlapping area of their body. This is the key reason that causes the failure detection. Moreover, it also improves the difficulty of the attention learning. Therefore, we conduct
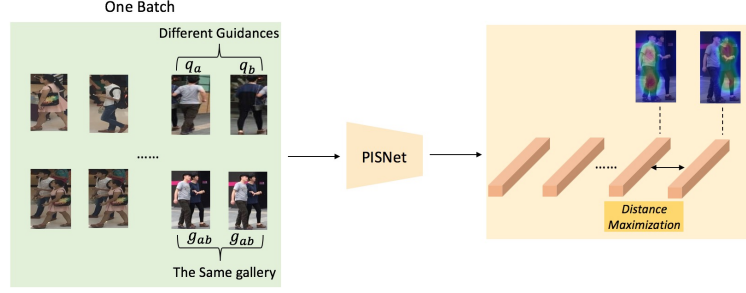
Fig. 6. The computation process of the Multi-Person Separation Loss. q_a and q_b denote the queries. g_{ab} is the gallery. The subscripts represent the IDs that appear in the image. In one batch, we pair a multi-person gallery image with different query images as the guidances. To promote the feature distinctiveness of the attention, the distance between the refined features guided by different query images should be maximized

the feature distinctiveness enhancement by the Multi-Person Separation Loss for further guaranteeing the purity of refined features.

As shown in Figure 2(b), we expect that the refined feature should have a large distance to the feature guided by another query image with a different person ID. For example, if a gallery image contains A and B, given the query image of A, we expect to extract the pure feature of A while suppressing the feature of B. In contrast, given the query image of person B, the pure feature of B should be extracted. In order to achieve this goal, as shown in Figure 6, we first construct the image batch for training, where a multi-person gallery image is paired with different query images as the guidances. Then, the distances can be minimized and maximized, respectively, by the Multi-Person Separation Loss, which is given by,

$$L_m = max(0, c + dist(QGAB(g_{ab}, q_a), QGAB(g_{ab}, q_b)) - dist(QGAB(g_{ab}, q_a), q_a)), \tag{4}$$

where $dist$ is the cosine distance and c denotes the margin coefficient. We should maximize the distance between $QGAB(g_{ab}, q_a)$ and $QGAB(g_{ab}, q_b)$ and meanwhile minimize the distance between $QGAB(g_{ab}, q_a)$ and q_a, where the subscripts represent the IDs that appear in the image.

3.5   Overall Objective Function

We utilize the cross entropy loss for both the gallery branch and GRAM, which is denoted as L_g and L_q, which is corresponding to $ID\ Loss_1$ and $ID\ Loss_2$, respectively, in Figure 3.

$$L_g = CE(\hat{y}, y), \tag{5}$$

$$L_q = CE(\bar{y}, y), \tag{6}$$

where CE denotes the cross-entropy loss, $\hat{y}$ and $\bar{y}$ denote the prediction ID in the gallery branch and GRAM, respectively, and y is the ground-truth ID. By combining with the Multi-Person Separation Loss, the final loss for the network is formulated as

$$L_{final} = L_g + \alpha * L_q + \beta * L_m, \tag{7}$$

where α and β are the coefficients to balance the contributions from the latter two losses.

3.6   Implementation Details

To implement our proposed model, we adopt Resnet-50 [5] as our basic CNN for feature extraction, which is pretrained on ImageNet. We first train the backbone on the single-person images using all training tricks in the strong baseline [20]. Then we add QGAB on the top of the Siamese Network. Both c_1 and c_2 are 1×1 convolutional layers with 1024 channels. Then we freeze the backbone network and train QGAB by pairing the multi-person images with single person ones. The batch size of samples for training is 64. The SGD optimizer is applied, with a learning rate of 0.00035. They are decayed by 0.1 after 20 epochs, and the training stops at 60 epochs. Parameters for the final loss function are $\alpha = 1.0$ and $\beta = 0.5$.

4   Experiments

4.1   Datasets and Evaluation Metrics

To demonstrate the effectiveness of our model on the Person-Interference Re-ID problem, we carry out the experiment on our constructed PI-PRW and PI-CUHK-SYSU dataset. Besides, in order to prove the generalization ability of our method on single-person images, we also evaluate the proposed PISNet on the another two datasets: Market-1501 [50] and DukeMTMC-ReID [27,54].

PI-PRW is derived from the PRW [52] dataset. We use the off-the-shelf detector Faster R-CNN [26] to perform pedestrian detection. Then we select the bounding boxes with multiple pedestrians. The selection criterion is: 1) At least 70% area of each ground-truth bounding box should be contained in the multi-person boxes. 2) The contained part of bounding boxes is at least 0.3 times the size of multi-person boxes in order to ensure the degree of the person interference. 3) Each bounding box has the overlapping area with any other ones. We get 1792 multi-person images with 273 IDs for training and 1258 multi-person gallery images and 211 single person query images for testing. Besides, in order to get closer to the actual scene, we add another 10000 single-person images in the test set as gallery images.

PI-CUHK-SYSU is derived from the CUHK-SYSU [38] dataset. We get multi-person cropped images following the same procedure in PI-PRW, resulting 3600 multi-person images for training with 1485 IDs and 3018 multi-person gallery images and 1479 single person query images for testing. We also add another 10000 single-person images in the test set as gallery images. More details of PI-PRW and PI-CUHK-SYSU can be referred to our supplementary material.
Evaluation Metrics. We use Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP) to evaluate the quality of different Re-ID models. All the experiments are performed in a single query setting.

Table 1. Comparison results (%) on PI-PRW and PI-CUHK-SYSU dataset at 4 evaluation metrics: *rank 1*, *rank 5*, *rank 10*, *mAP* where the bold font denotes the best method. The methods in the 1st group are proposed for the traditional Re-ID problem. The methods in the 2nd group are proposed for the multi-label learning. The 3rd group is the methods of Occluded Re-ID. The 4th group is our method

| Method	PI-PRW				PI-CUHK-SYSU			
	rank1	*rank 5*	*rank 10*	*mAP*	*rank1*	*rank 5*	*rank 10*	*mAP*
HA-CNN [15]	32.4	56.9	68.0	32.0	71.3	82.0	87.5	65.3
PCB [32]	31.3	55.1	67.5	30.2	70.1	80.4	86.9	63.1
Strong Baseline [20]	34.7	59.4	70.3	36.0	72.5	83.9	88.2	70.1
PyramidNet [48]	35.9	60.2	70.1	37.0	73.1	83.5	87.9	70.5
ABD-Net [1]	35.4	59.9	69.7	36.3	72.9	82.6	87.5	70.4
QAConv [18]	36.0	61.2	70.9	38.2	73.2	84.7	88.3	70.9
HCP [42]	30.2	49.7	61.2	29.6	67.2	75.3	83.5	61.9
LIMOC [40]	32.9	52.4	63.3	32.6	69.1	78.2	85.3	65.2
FPR [7]	36.3	60.7	70.4	37.9	73.7	85.0	89.1	71.2
AFPB [57]	34.1	58.2	67.2	35.1	70.7	83.2	87.3	68.3
Ours	**42.7**	**67.4**	**76.2**	**43.2**	**79.1**	**88.4**	**91.9**	**76.5**

4.2   Results Comparison

Results on PI-PRW and PI-CUHK-SYSU. We first compare the proposed approach with the existing methods on the two proposed PI Re-ID datasets. Table 1 shows the result of our method and previous works. The compared methods (including six existing representative Re-ID models, two multi-label learning approaches, and two Occluded Re-ID methods) are listed in the table. These results show: (1) Among existing methods, the Occluded Re-ID model FPR is superior. For example, FPR achieves 36.3% Rank-1 accuracy and 37.9% mAP on PI-PRW, which outperforms all the previous Re-ID methods. This is because, similar to our method, FPR [7] leverage query feature maps as multi-kernels to calculate the spatial affinity with the gallery feature maps, and then enhance the common features in the gallery features. (2) The performance of

LIMOC [40] and HCP [42] proposed for the multi-label learning is ordinary. For example, compared to the strong baseline [20], HCP [42] is less by -4.5% Rank-1 accuracy and -3.5% mAP on PI-PRW. (3) Our new model PISNet outperforms all competitors by significant margins. For example, PISNet achieves 42.7% Rank-1 accuracy and 43.2% mAP on PI-PRW and 79.1% Rank-1 accuracy and 76.5% mAP on PI-CUHK-SYSU. This is because our proposed method explicitly utilizes query information and learn a more precise boundary between pedestrians by GRAM and MPSL.

Table 2. Component analysis of the proposed method on the PI-PRW and PI-CUHK-SYSU datasets (%)

Method	PI-PRW				PI-CUHK-SYSU			
	rank1	*rank 5*	*rank 10*	*mAP*	*rank1*	*rank 5*	*rank 10*	*mAP*
Baseline	34.7	59.4	70.3	36.0	72.5	83.9	88.2	70.1
Baseline + QGAB	38.9	61.5	72.4	38.0	73.9	85.0	89.1	72.3
Baseline + QGAB + MPSL	39.7	63.2	74.1	40.1	76.2	87.1	91.4	74.2
Baseline + QGAB + GRAM	41.8	66.1	75.2	42.4	77.9	87.5	91.0	75.0
Baseline + QGAB + GRAM + MPSL	42.7	67.4	76.2	43.2	79.1	88.4	91.9	76.5
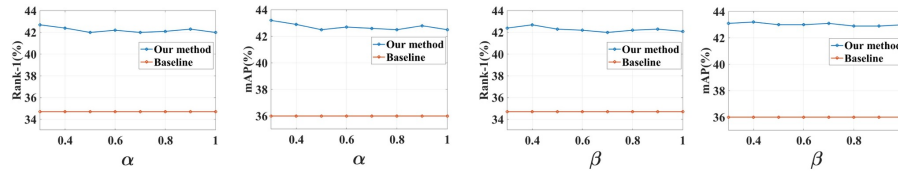
Fig. 7. Evaluation of different parameters of PISNet using Rank-1 and mAP accuracy on the PI-PRW dataset (%)

4.3   Further Analysis

Contributions of Individual Components. In Table 2, we evaluate the three components on how they contribute to the full model. The results show that all of them are effective on their own (each outperforms all the compared methods). Moreover, when combined, the best performance is achieved. This validates our design consideration in that they are complementary and should be combined.
**Does PISNet Perform well on the General Re-ID Dataset?** We also apply our method on the general Re-ID datasets, Market-1501 and DukeMTMC-ReID. The compared methods (including fifteen existing representative Re-ID models, and two state-of-the-art Occluded Re-ID methods) are listed in Table 3. The results show that: 1) Compared to existing representative Re-ID

Table 3. Comparison results (%) on the Market-1501 and DukeMTMC-ReID datasets. N_f is the number of features used in the inference stage. The methods in the 1st group are proposed for the traditional Re-ID problem. The 2nd group is the state-of-the-art methods of Occluded Re-ID. The 3rd group is our method

Method	N_f	Market1501 r = 1	mAP	DukeMTMC-ReID r = 1	mAP
PIE [49]	3	87.7	69.0	79.8	62.0
SPReID [9]	5	92.5	81.3	84.4	71.0
MaskReID [12]	3	90.0	75.3	78.8	61.9
MGN [33]	1	95.7	86.9	88.7	78.4
SCPNet [3]	1	91.2	75.2	80.3	62.6
PCB [32]	6	93.8	81.6	83.3	69.2
Pyramid [48]	1	92.8	82.1	-	-
Pyramid [48]	21	95.7	88.2	89.0	79.0
HA-CNN [15]	4	91.2	75.7	80.5	63.8
ABD-Net [1]	1	95.6	88.3	89.0	78.6
Camstyle [56]	1	88.1	68.7	75.3	53.5
PN-GAN [25]	9	89.4	72.6	73.6	53.2
IDE [46]	1	79.5	59.9	-	-
SVDNet [31]	1	82.3	62.1	76.7	56.8
TriNet [8]	1	84.9	69.1	-	-
SONA [37]	1	95.6	88.8	89.6	78.2
FPR [7]	1	95.4	86.5	88.6	78.2
PGFA [23]	1	91.2	76.8	82.6	65.5
Baseline	1	94.5	85.9	86.4	76.4
Ours	1	95.6	87.1	88.8	78.7

models, our method achieves comparable performances with state-of-the-art on both datasets. These models leverage complicated attention mechanisms or local-based methods to achieve the results, while our PISNet is specifically designed for PI Re-ID. 2) Our method outperforms the existing Occluded Re-ID models on both general Re-ID datasets. Specifically, our method can reach 95.6% rank-1 accuracy and 87.1% mAP on Market1501, and 88.8% rank-1 accuracy and 78.7% mAP on DukeMTMC-ReID. The results prove the generalization ability of PISNet on the general Re-ID datasets.

Influence of Parameters. We evaluate two key parameters in our modelling, the loss weights α and β in Eq.(7). The two parameters would influence the performance of the proposed method. As shown in Figure 7, when α and β are set between 0.3 and 1.0, and 0.4 and 1.0, respectively, the performance does not change dramatically, which indicates that PISNet is not sensitive to the α and β in the value ranges.
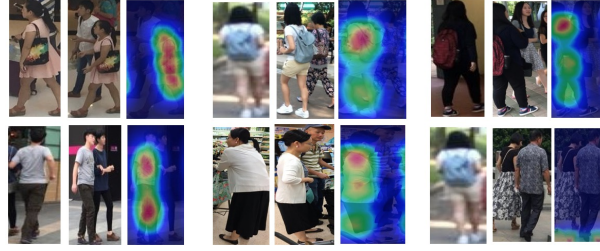
Fig. 8. Visualisation of our query-guided attention for multi-person images in PI-CUHK-SYSU. In each group, from left to right, (1) the single-person query, (2) the multi-person gallery and (3) the masked feature map. In the heat map, the response increases from blue to red. Best viewed in color

4.4   Attention Visualisation

We visualise our query-guided attention for multi-person images in the PI-CUHK-SYSU dataset. Figure 8 shows that: (1) The attention mask filters out other pedestrians in multi-person images, (2) When the multi-person gallery does not include the query, the attention is weak for the whole image (the third group in the second rows). The visualisation results further prove that our method can suppress the pedestrian interference effectively.

5   Conclusions

We have considered a new and more realistic person Re-ID challenge: pedestrian-interference person re-identification problem. To address the particular challenges associated with this new Re-ID problem, we propose a novel query-guided framework PISNet with a Guidance Reversed Attention Module and the Multi-Person Separation Loss. Both are specifically designed to address the person interference problem. The effectiveness of our model has been demonstrated by extensive experiments on two new pedestrian-interference Re-ID datasets introduced in this paper. In our future work, we will extend this work to handle more kinds of hard cases caused by a non-perfect detector.

Acknowledgment

References

1. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
2. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
3. Fan, X., Luo, H., Zhang, X., He, L., Zhang, C., Jiang, W.: Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. Lecture Notes in Computer Science (2019)
4. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision (ECCV) (2008)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR) (2016)
6. He, L., Liang, J., Li, H., Sun, Z.: Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
7. He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., Feng, J.: Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
8. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
9. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
10. Khamis, S., Kuo, C.H., Singh, V.K., Shet, V.D., Davis, L.S.: Joint learning for attribute-consistent person re-identification. In: European Conference on Computer Vision (ECCV) (2014)
11. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
12. Lei, Q., Jing, H., Lei, W., Yinghuan, S., Yang, G.: Maskreid: A mask based deep ranking neural network for person re-identification (2019)
13. Li, W., Wang, X.: Locally aligned feature transforms across views. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
14. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
15. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
17. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: The IEEE conference on computer vision and pattern recognition (CVPR) (2015)
18. Liao, S., Shao, L.: Interpretable and generalizable deep image matching with adaptive convolutions (2019)

19. Liao, W., Yang, M.Y., Zhan, N., Rosenhahn, B.: Triplet-based deep similarity learning for person re-identification. In: The 2017 IEEE International Conference on Computer Vision Workshop (ICCVW) (2017)
20. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification (2019)
21. Ma, A.J., Yuen, P.C., Li, J.: Domain transfer support vector ranking for person re-identification without target camera label information. In: The IEEE International Conference on Computer Vision (CVPR) (2013)
22. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: British Machive Vision Conference (BMVC) (2012)
23. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
24. Prosser, B.J., Zheng, W.S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: British Machive Vision Conference (BMVC) (2010)
25. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: European Conference on Computer Vision (ECCV) (2018)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2015)
27. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking (2016)
28. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. arXiv preprint arXiv:1803.09937 (2018)
29. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
30. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. arXiv preprint arXiv:1804.07094 (2018)
31. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
32. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: European Conference on Computer Vision (ECCV) (2018)
33. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM Multimedia Conference on Multimedia Conference (ACM MM) (2018)
34. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: global-local-alignment descriptor for pedestrian retrieval. In: ACM on Multimedia Conference (ACM MM) (2017)
35. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in neural information processing systems (NIPS) (2006)
36. Weishi, Z., Xiang, L., Tao, X., Shengcai, L., Jianhuang, L., Shaogang, G.: Partial person re-identification. In: The IEEE International Conference on Computer Vision, ICCV (2015)

37. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
38. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
39. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. arXiv preprint arXiv:1805.03344 (2018)
40. Yang, H., Tianyi Zhou, J., Zhang, Y., Gao, B.B., Wu, J., Cai, J.: Exploit bounding box annotations for multi-label object recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
41. Yifan, S., Qin, X., Yali, L., Chi, Z., Yikang, L., Shengjin, W., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
42. Yunchao, W., Wei, X., Min, L., Junshi, H., Bingbing, N., Jian, D., Yao, Z., Shuicheng, Y.: Hcp: A flexible cnn framework for multi-label image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2016)
43. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
44. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
45. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
46. Zhedong, Z., Liang, Z., Yi, Y.: A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2018)
47. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
48. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
49. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. arXiv preprint arXiv:1701.07732 (2017)
50. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
51. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
52. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
53. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE transactions on pattern analysis and machine intelligence (PAMI) (2013)
54. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. arXiv preprint arXiv:1701.07717 (2017)

55. Zhu, F., Kong, X., Zheng, L., Fu, H., Tian, Q.: Part-based deep hashing for large-scale person re-identification. IEEE Transactions on Image Processing (TIP) (2017)
56. Zhun, Z., Liang, Z., Zhedong, Z., Shaozi, L., Yi, Y.: Camstyle: A novel data augmentation method for person re-identification. IEEE Transactions on Image Processing (TIP) (2019)
57. Zhuo, J., Chen, Z., Lai, J., Wang, G.: Occluded person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME) (2018)