Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes

Lingxiao He* and Wu Liu*

JD AI Research, Beijing, P.R. China
https://air.jd.com
{helingxiao3, liuwu1}@jd.com

Abstract. Person Re-identification (Re-ID) in crowed scenes is a challenging problem, where people are frequently partially occluded by objects and other people. However, few studies have provided flexible solutions to re-identifying people in an image containing a partial occlusion body part. In this paper, we propose a simple occlusion-aware approach to address the problem. The proposed method first leverages a fully convolutional network to generate spatial features. And then we design a combination of a pose-guided and mask-guided layer to generate saliency heatmap to further guide discriminative feature learning. More importantly, we propose a new matching approach, called Guided Adaptive Spatial Matching (GASM), which expects that each spatial feature in the query can find the most similar spatial features of a person in a gallery to match. Especially, We use the saliency heatmap to guide the adaptive spatial matching by assigning the foreground human parts with larger weights adaptively. The effectiveness of the proposed GASM is demonstrated on two occluded person datasets: Crowd REID (51.52%) and Occluded REID (80.25%) and three benchmark person datasets: Market1501 (95.31%), DukeMTMC-reID (88.12%) and MSMT17 (79.52%). Additionally, GASM achieves good performance on cross-domain person Re-ID. The code and models are available at: https://github.com/JDAI-CV/fast-reid/blob/master/projects/CrowdReID.

Keywords: Person re-identification, Guided saliency feature learning, Guided adaptive spatial matching

1 Introduction

Person re-identification (Re-ID) has achieved significant improvement both in academic and industrial society in the past two years, making it widely used in many real-world scenarios, such as train station, airport, etc.. However, two urgent yet challenging problems of person Re-ID in crowded scenes are still not well addressed. One major issue that challenges this task is the ubiquitous pedestrian occlusion by each other. For instance, as shown in Fig. 1, the captured person is occluded by a partial body of the front person, making it difficult to track her

* Corresponding Author

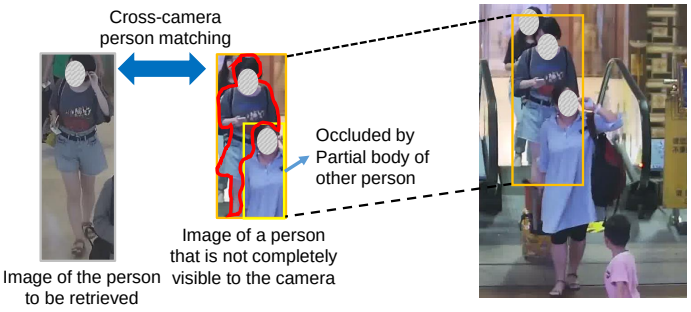

Fig. 1. Illustration of the person Re-ID in crowed scenes. Here, the Re-ID system aims to recognize the person within the red region. The captured person by the surveillance operator is occluded by the partial body of the front person.

movement. Another major issue that challenges this task is cross-camera person image matching because of unpredictable dynamic background-bias under different cameras. From this perspective, person Re-ID in crowded scenes has attracted significant research attention as the demand for identification using images captured by video surveillance systems has been rapidly growing.

Most existing person Re-ID [9], [11] approaches fail to identify a person when the body region is severely occluded by the partial body of the other person. To match an occluded person image, most of the attention-based approaches [10], [18], [25] (see in Fig. 2(a)) enforce the output features to mainly focus on the foreground human bodies. Our observations show that the attention-based approach can only eliminate the influence of background, but fail to remove the partial body of another person. Some other approaches [8], [15], [21] proposed a two-stream network as shown in Fig. 2(b), which consists of an appearance map extraction stream and a body part heatmap extraction stream. Following the two streams, a part-aligned feature map was obtained by a bilinear mapping of the corresponding local appearance and body part descriptors. But it inevitably requires more inference time to obtain the body part heatmap and fails to generate accurate body part heatmap under heavily occluded by the partial body of the other person in crowded scenes.

In this paper, we propose a towards accurate camera-aware person Re-ID framework that can re-identify a person occluded by a partial person's body as shown in Fig. 2(c). First, we utilize a fully convolutional neural network to produce spatial feature maps that contains the features of the person's body occlusion person body(partial body of another person) and background. To guarantee the extracted feature with less contamination of background and occlusion person body, we design a mask layer combined with a pose layer to predict the saliency heatmap. In particular, The predictive power of the saliency heatmap for the mask layer and pose layer are respectively learned from large human segmentation and pose estimation models. And then we obtain the discriminative feature by a bilinear mapping of the spatial feature map and the saliency heatmap. Finally, we develop an adaptive spatial matching method, which ex-

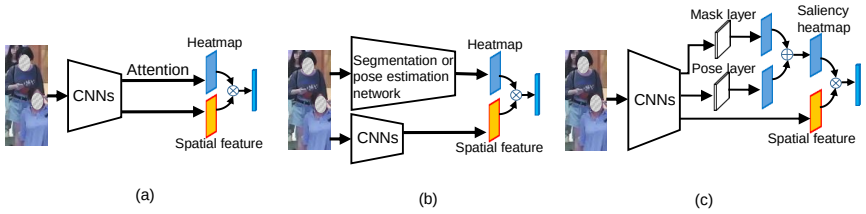

Fig. 2. (a) Attention-based approach (b) Two-stream approach (c) Our approach.

pects that each spatial feature in the query can find the most similar spatial feature of a person in a gallery to match. To achieve the more accurate matching, We also use the saliency heatmap to guide the spatial matching by assigning the person body with larger weights and the occlusion body and background with smaller weights to overcome the occlusion and background problem. Remarkably, the model improves the performance of cross-camera Re-ID by eliminating background-bias. The main contributions of our work are summarized as follows:

- We propose a novel saliency feature learning network that integrates Re-ID feature learning, human segmentation and pose estimation in a unified framework, which can effectively address person re-identification under severe occlusion in crowded scenes. Our approach does not rely on any external cues during the inference term.
- We propose guide adaptive spatial matching for person Re-ID, which can address person image misalignment problem and is flexible to arbitrary-sized person images.
- The proposed saliency feature presentation framework effectively eliminates dynamic background-bias, which helps to improve cross-domain person Re-ID.
- Experimental results demonstrate that the proposed approach achieves impressive results on multiple occlusion datasets including new Crowd REID and Occluded REID. It exceeds some occluded Re-ID approaches by more than 15% in terms of rank-1 accuracy. Besides, the proposed method achieves competitive results on multiple benchmark datasets including Market1501 [27], DukeMTMC-reID [30], MSMT17 [29].

2 Related Work

Occluded Person Re-identification Occluded/Partial person Re-ID [16], [2] has become an emerging problem in video surveillance. To address this problem. Part-based models are considered as a solution to occluded person Re-ID. A local-to-local matching strategy is employed to handle occlusions and cases where the target is partially out of camera's view. Zheng *et al.* [28] proposed a local-level matching model called Ambiguity-sensitive Matching Classifier (AMC) based on the dictionary learning and introduced a local-to-global matching model called

Sliding Window Matching (SWM) that can provide complementary spatial layout information. To address the Re-ID problem with occlusion, *i.e.*, only part of the human body is visible to the camera, He *et al.* [4] proposed an alignment-free approach namely Deep Spatial feature Reconstruction (DSR) that uses a fully convolutional network to extract corresponding-sized spatial feature maps for the incomplete person images, and then exploits the reconstruction error based on sparse coding to avoid explicit alignment. Furthermore, a Foreground-aware Pyramid Reconstruction (FPR) [6] scheme also tries to remove the influence of background and scale various. Sun *et al.* introduce a Visibility-aware Part Model (VPM) [22] to extract stripe-level features, thus addressing the spatial misalignment in the incomplete images. Luo *et al.* [13] proposed STNReID that combines a spatial transformer network (STN) and a re-id network for partial re-id Besides, the Pose-Guided Feature Alignment (FGFA) [14] utilizes the pose landmarks to mine discriminative part information to address the occlusion noise. Although these methods mentioned above can solve the occlusion problem to some extent, it cannot deal with the situation where a person is partially occluded by a partial body of another person. To this end, we propose a guided saliency feature learning model that can effectively classify background and partial occlusion body, and then incorporate adaptive spatial matching to address occlusion problem in crowded scenes.

Person Re-identification The attention mechanism is usually utilized in many human-centric video analysis applications [5], [12], such as ReID, to extract more discriminative features.. Si *et al.* [18] proposed a dual attention mechanism, in which both intra-sequence and inter-sequence attention strategies are used for feature refinement and feature-pair alignment, respectively. Chen *et al.* proposed an Attentive but Diverse Network (ABD-Net) [1], which integrates spatial-channel attention modules and diversity regularizations throughout the entire network to extract more discriminative features. Zhou *et al.* [32] introduce a novel consistent attention regularizer between feature representation layers to learn foreground-aware feature maps. Besides, most of the person Re-ID approaches leverage external cues such as human mask and human pose estimation to enhance feature representation. In some mask-guided models, mask as an external cue helps to remove the background clutters in pixel-level and only contain body shape information. Song *et al.* [19] introduced the binary segmentation masks to construct the synthetic RGB-Mask pairs as inputs, then they design a mask-guided contrastive attention model (MGCAM) [19] to learn features separately from the body and background regions. Kalayeh *et al.* [8] proposed a person re-identification model that integrated human semantic parsing in person re-identification. Pose-guided models utilize pose estimation information as an external cue in person Re-ID to reduce the part misalignment problem. Each key part can be well located using person landmarks. Su *et al.* [20] proposed a Pose-driven Deep Convolutional (PDC) model to learn improved feature extractors and matching models from end-to-end, PDC can explicitly leverage the human part cues to alleviate the pose variations. Suh *et al.* [21] proposed a two-stream network that consisted of an appearance map extraction stream

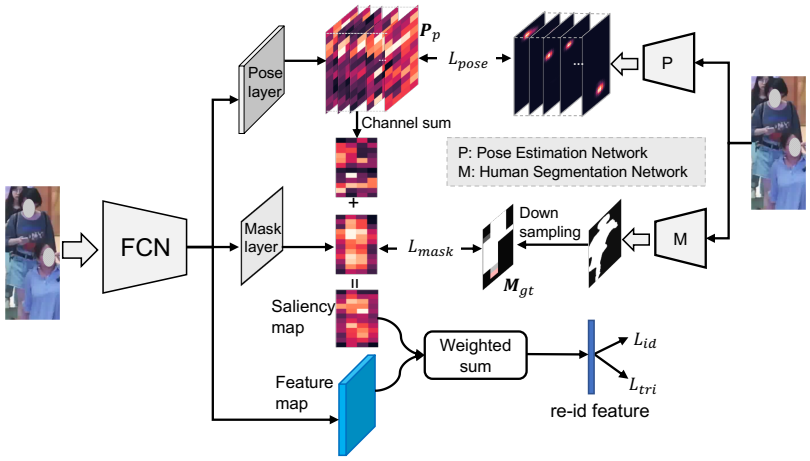

Fig. 3. Architecture of the proposed guided saliency feature learning. The backbone consists of a fully convolutional network (FCN), a mask layer and a pose layer. FCN is used to extract spatial features. The mask layer and pose layer are used to predict the mask-guided heatmap and pose-guided heatmap. And then saliency features are obtained by a bilinear mapping of the feature map and the saliency heatmap.

and body part map extraction stream. And then a part-aligned feature map is obtained by a bilinear mapping of the corresponding local appearance and body part descriptors. Although mask-guided and pose-guided approaches can achieve satisfying performance, they extremely rely on accurate segmentation model and pose estimation model. Also, these existing approaches are two-stream structure so that the computation cost of these approaches is rather extensive during the inference term.

3 Proposed Approach

In this section, we elaborate on the proposed guided saliency feature learning for the occluded person re-identification approach. We first introduce the entire network architecture. After that, we will introduce the training of our model. Finally, we will explain saliency adaptive spatial matching.

3.1 Architecture of the Proposed Model

The architecture of the proposed Re-ID model is shown in Fig. 3. Structurally, it consists of a full convolutional neural network (FCN), a mask layer and a pose layer. We now introduce them one by one.

FCN In the crowded scenes, the target person to re-identify is provided with person detection bounding boxes by a human detector. As shown in Fig. 1, the detected person bounding boxes are coarse, often containing background and

partial body part of occlusion person. Conventional neural networks involving a feature aggregation layer like the fully connected layer or average pooling layer would output global features contaminated by background and occlusion. To obtain better representation that only focuses on the person to be re-identified, we use a fully convolutional network (FCN) as the backbone for spatial feature extraction. FCN can still retain spatial coordinate so that the background, occlusion features, and person features are extracted without interference with each other. We discard the last average pooling layer based on ResNet-50 to implement FCN, and the last Resblock outputs the spatial feature map.

Mask layer The extracted spatial features are often contaminated by the background and occlusion features. To guarantee the following person spatial feature are less contamination from the background, we design a mask layer to obtain the foreground probability of spatial features. The designed mask layer consists of a convolution layer with a $1 \times 1 \times d$ (d denotes the number of the channel of the output feature map, and d = 2048 in our model) convolution kernel. The output spatial features pass the mask layer to generate a correspondingly-size mask-guided heatmap, and the size of each mask-guided heatmap is $W \times H$, where W and H denote the width and the height of output feature map, respectively.

Pose layer The detected person may be occluded by other partial bodies of other persons, resulting in that only mask-guided heatmap unable to distinguish between persons to be identified or occlusion persons. Therefore, we also design a pose layer to predict the C human keypoints heatmap (C= 13)[1] that can only focus on the identified person. The pose layer consists of a convolution layer with 13 $1 \times 1 \times 2048$ kernels. Likewise, the output spatial features pass the pose layer to generate a correspondingly-size pose-guided heatmap, and the size of each pose-guided heatmap is $W \times H \times C$.

Pose-guided heatmap only focuses on human keypoints, the weights around person keypoints are relatively large while the weights beyond person keypoints are relatively small. As for Mask-guided heatmap, it only focuses on the person part, but it cannot address partial occlusion body contamination. Therefore, pose-guided heatmap combined with mask-guided heatmap can output a more accurate saliency heatmap for further guiding following feature learning and adaptive spatial matching.

3.2 Guided Saliency Feature Learning

In this section, we now explain the training strategy for learning saliency features. As shown in Fig. 3, four loss functions are used to optimize the whole Re-ID model.

Mask-guided loss We design a mask layer to classify the spatial features of the background and foreground person. We treat this problem as a probabilistic prediction problem. To generate a mask-guided heatmap, we use the semantic segmentation model PSPNet [26] trained with COCO dataset to guide the mask

[1] Head, Left-shoulder, Right-shoulder, Left-elbow, Right-elbow, Left-wrist, Right-wrist, Left-hip, Right-hip, Left-knee, Right-knee, Left-ankle, Right-ankle.

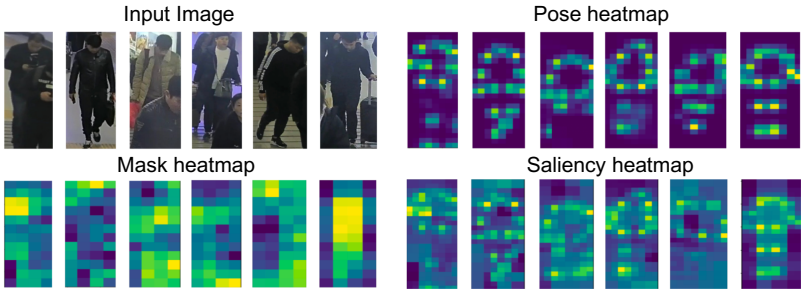

Fig. 4. Example results of source image, pose heatmap, mask heatmap and saliency heatmap.

layer to predict the mask-guided heatmap $\mathbf{M}_p \in \mathbb{R}^{W \times H}$. Let $\mathbf{M}_{gt}$ generated by PSPNet with downsampling operation be the ground truth human mask of width W and height H. Our aim is to regress the $\mathbf{M}_p$ with regard to $\mathbf{M}_{gt}$. Therefore, the training objective of mask layer is a pixel-wise regression:

$$\mathcal{L}_{mask} = ||\mathbf{M}_p - \mathbf{M}_{gt}||_F \tag{1}$$

Pose-guided loss The predicted mask-guided heatmap can only distinguish between backgrounds and persons, but is unable to distinguish between persons to be identified and partial body of other persons. Therefore, we want to design a pose layer to classify the spatial features of the person and partial body. The human pose estimation model CenterNet [33] (visualized human joints are shown in Fig. 4) is used to guide the pose layer to predict the pose-guided heatmap. Keypoint types included 13 human joints in human pose estimation. Let $\mathbf{P}_p \in \mathbb{R}^{W \times H \times 13}$ and $\mathbf{P}_gt \in \mathbb{R}^{W \times H \times 13}$ be predicted pose-guided heatmap by the pose layer and the generated ground truth pose heatmap by CenterNet with downsampling operation, respectively. Our aim is to use $\mathbf{P}_p$ to regress $\mathbf{P}_{gt}$. Therefore, the training objective of the pose layer is a pixel-wise regression:

$$\mathcal{L}_{pose} = ||\mathbf{P}_p - \mathbf{P}_{gt}||_F \tag{2}$$

And then, we use channel sum operation to obtain the final pose-guided heatmap $\mathbf{P}_p$.

The two complementary mask-guided heatmap and pose-guided heatmap are fused to obtain saliency heatmap to predict more accurate saliency person coordinate information.

Triplet Loss and Identification Loss Our person re-identification model simply aggregates the output spatial feature map using the saliency heatmap. By using the weighted sum operation, the network finally generates a 2048-D global saliency feature representation. To train the network, we pass it to a multi-class classification objective with softmax cross-entropy loss L_{id}. To obtain a more compact feature representation, we also use a metric learning objective with triplet loss L_{tri} to train the network.

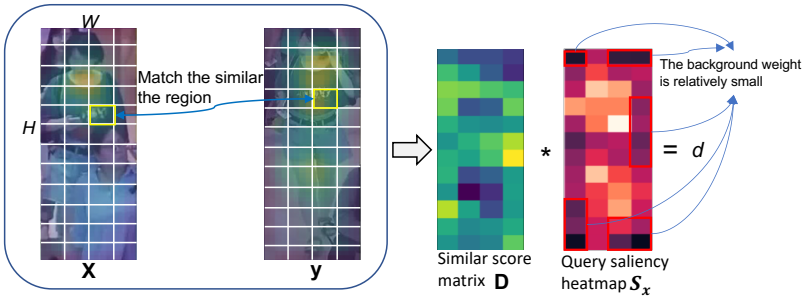

Fig. 5. Illustration of guided adaptive spatial matching. Here, each spatial feature in the query can find the most similar spatial feature of a person in a gallery to match. The saliency heatmap can refine the final matching result.

The final loss function is a weighted sum of all the losses defined above: L_{mask} and L_{pose} are complementary to predict the saliency person heatmap. L_{id} and L_{tri} to preserve identity discrimination.

$$\mathcal{L} = \lambda_{mask}\mathcal{L}_{mask} + \lambda_{pose}\mathcal{L}_{pose} + \mathcal{L}_{id} + \mathcal{L}_{tri} \tag{3}$$

We set $\lambda_{mask} = 0.05$ and $\lambda_{pose} = 0.1$ in our experiments.

3.3 Guided Adaptive Spatial Matching

At present, Euclidean distance is used to match global features. However, such coarse matching is hard to solve the misalignment problem. The saliency adaptive spatial matching (SASM) is a more refined matching method, which can solve the misalignment to some extent. Given a pair of person images I_x (query) and I_y (gallery), Corresponding spatial feature maps $\mathbf{x} \in \mathbb{R}^{W\times H\times d}$ and $\mathbf{y} \in \mathbb{R}^{W\times H\times d}$ are inferred by FCN, where W, H and d denote the height, the width and the number of the channel of spatial feature map. $\mathbf{x}_{h,w} \in \mathbb{R}^{1\times 1\times d}$ represents a local spatial feature corresponding to a local region of a source image. As shown in Fig. 5, $\mathbf{x}_{w,h}$ attempt to search similar local spatial feature in $\mathbf{y}$ to match, and then outputs the similar score $d_{w,h}$. For $W \times H$ spatial features in $\mathbf{x}$, we can calculate the similarity score of each spatial feature $\mathbf{x}_{w=1,h=1}^{W,H}$ with regard to $\mathbf{y}$. The similar score matrix can be defined as $\mathbf{D}$.

However, it suffers from an obvious limitation: since $\mathbf{x}$ contains the background and partial occlusion body features, the spatial matching scores are inaccurate. To alleviate the problem, we want to reduce the influence of background and partial occlusion body by assigning their matching scores small weights, while enhancing the effect of foreground person by assigning these matching scores large weights adaptively. Therefore, we use the generated saliency heatmap to guide adaptive spatial matching for further improving the Re-ID performance. Given a query image I_x, the pose layer and mask layer as introduced above output the saliency heatmap $\mathbf{S}_x$. Then the saliency heatmap $\mathbf{S}_x$ can be used to

guide the adaptive spatial matching. We perform weight sum operation over the spatial matching matrix $\mathbf{D}$ and saliency heatmap $\mathbf{S}_x$. Then the final guided adaptive spatial matching (GASM) distance d of I_x and I_y can be defined as

$$d = \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{D}^{h,w} * \mathbf{S}_x^{h,w} \tag{4}$$

Also, the global matching distance provides the complementary matching information to GASM by average weighting operation.

4 Experiments

In this section, we first verify the effectiveness of our proposed approach for the task of occluded person Re-ID on two occluded person Re-ID dataset: Crowd REID and Occluded REID [7], and then experiment on three person Re-ID benchmarks: Market1501 [27], DukeMTMC-reID [30], and MSMT17 [24] to show its generalizability. Also, we verify the advantage of our proposed approach for the task of cross-domain person Re-ID. Finally, we perform the parameter analysis to investigate the influence of the separated pose layer and mask layer.

4.1 Experiment Settings

Implementation Details. Our implementation is based on the publicly available code of PyTorch. All models are trained and tested on Linux with P40 GPUs. In the training term, all training images are re-scaled to 384×128. For batch hard triplet loss function, one batch consists of 16 subjects, and each subject has 4 different images. We train the overall network with 120 epochs, We use the Adam optimizer with the base learning rate initialized to 10^{-4}, then decayed to 10^{-5}, 10^{-6} after 40, 90 epochs, respectively.
Evaluation Protocol We employ the standard metrics as in most person Re-ID literature namely the cumulative matching curve (CMC) and the mean Average Precision (mAP).

4.2 Datasets

Crowd REID is a newly created, which is specifically designed for person Re-ID in crowded scenes. The dataset is collected in a train station by 11 surveillance cameras. The dataset is very challenging because most of the person images in these images to be identified are occluded by the partial body of other persons. The examples of occlusion persons in the Crowd REID dataset are shown in Fig. 6(a). To evaluate this dataset, 2,412 images of 835 identities are used as the gallery set and 845 images of 605 identities are used as the gallery set.
Occluded REID is an occluded person dataset captured by mobile cameras, consisting of 2,000 images of 200 occluded persons (see Fig. 6(b)). Each identity has 5 full-body person images and 5 occluded person images with different viewpoints, backgrounds and different types of severe occlusions.

Fig. 6. (a) Example of person images in the crowd REID dataset, the person to be identified are occluded by partial bodies of other persons such as head-shoulder or half body. (b) Example of person images in the Occluded dataset, the person images are occluded by static objects such as tree, car, and umbrella, etc..

Market1501 has 12,936 training and 19.732 testing images with 1,501 identities in total from 6 cameras. Deformable Part Model (DPM) is used as the person detector. We follow the standard training and evaluation protocols in [27] where 751 identities are used for training and the remaining 750 identities for testing.

DukeMTMC-reID is the subset of Duke Dataset [17], which consists of 16,522 training images from 702 identities, 2,228 query images and 1,7,661 gallery images from other identities. It provides manually labeled person bounding boxes. Here, we follow the setup in [30].

MSMT17 is the current largest available person Re-ID dataset, consisting of 126,441 images of 4,101 identities captured by 12 outdoor cameras and 3 indoor cameras. We follow the training-testing split of [24]. MSMT17 is significantly more challenging than market1501 and DukeMTMC-reID due to more complex collected scenes, different weather conditions (morning, noon, afternoon).

Table 1. Performance comparison on Crowd REID and Occluded REID using Market1501, DukeMTMC and MSMT17 training datasets, respectively.

Method	Crowd REID rank-1 (mAP)			Occluded REID rank-1 (mAP)		
	Market1501	DukeMTMC	MSMT17	Market1501	DukeMTMC	MSMT17
PCB [23]	14.1 (13.4)	20.3 (17.5)	32.7 (26.6)	41.3 (38.9)	47.5 (42.4)	56.2 (47.5)
MaskReID [15]	14.3 (13.7)	21.7 (19.8)	37.2 (31.2)	26.8 (25.0)	44.2 (38.1)	54.2 (51.2)
PABR [21]	24.3 (21.7)	26.5 (23.7)	41.6 (36.8)	67.1 (55.4)	-	-
DSR [4]	28.8 (26.7)	30.3 (27.3)	48.7 (43.3)	72.8 (62.8)	-	-
FPR [6]	30.2 (29.9)	32.6 (30.3)	50.7 (45.8)	78.3 (68.0)	-	-
self-attention	22.3 (23.2)	-	-	67.4 (55.8)	-	-
Baseline	16.7 (16.4)	26.6 (24.3)	44.0 (41.2)	60.2 (57.0)	66.1 (60.2)	74.6 (69.2)
GASM	32.4 (31.1)	35.8 (31.8)	51.5 (47.1)	74.5 (65.6)	76.5 (67.8)	80.3 (73.2)

4.3 Occluded Person Re-identification

The designed person Re-ID model is respectively trained with Market1501, DukeMTMC-reID, and MSMT17 datasets, and tested on Crowd REID and Occluded REID datasets. Therefore, it is a cross-domain setting. Firstly, we compare our GASM against existing person Re-ID approaches including a part-based model PCB [23], a mask-guided Re-ID model MaskReID [15] and pose-guided Re-ID model PABR [21]. Besides, we also compare our GASM against Partial partial person Re-ID approaches DSR [4] and FPR [6]. For PCB, MaskReID, PABR, DSR, and FPR, we follow their original parameter settings. Besides, we compare our proposed GASM against the baseline model based on ResNet-50.

Table 1 respectively shows the experimental results on Crowd REID and Occluded REID datasets. It is noted that: (1) The proposed GSAM achieves stable results regardless of different training datasets and different testing datasets. Clear gaps are shown between our proposed GASM and these state-of-the-art, which suggests that GSAM is very solid to address the occlusion problem; (2) The proposed GASM outperforms AMC-SWM, PCB, MaskReID, SPReID, PABR, DSR and FPR on Crowd REID dataset. Such a result justifies the fact that GASM can reduce the influence of the partial body of other persons, making the network only focus on foreground person part; (3) The gap between GASM and AMC-SWM, PCB are significant on the two datasets. Compared to PCB: GASM increases from 14.1% to 32.4% and from 41.3% to 74.5% at rank-1 accuracy on Crowd REID and Occluded REID when we use Market1501 dataset to train the model. The results suggest that it is difficult to address occlusion problem since it fuses both occlusion/background part feature and human part feature to the final feature.; (4) PABR and FPR achieve comparable results (24.3%, 67.1% and 30.2%, 78.3% at rank-1 accuracy) on the two datasets because pose estimation heatmaps used in PABR and foreground-aware network can alleviate the influence the background; (5) GASM performs better than MaskReID due to eliminating the background-bias is not conducive to person Re-ID; (6) Although MaskReID and PABR are well suited for addressing occlusion problem, they depend on external models such segmentation network and pose estimation network during the inference. Therefore, the forward inference is inefficient.

As discussed above, GASM is an occlusion-aware approach that can both address background-bias, occlusion object, and occlusion partial body by assigning them small weights. Furthermore, GSAM is an adaptive local matching approach that can address misalignment person image.

4.4 Non-Occluded Person Re-identification

We also experiment on non-occluded person datasets: Market1501, DukeMTMC-reID, and MSMT17 to test the generalizability of our proposed approach.
Results on Market1501 Comparisons between GSAM and 11 state-of-the-art: PCB [23], MasKReID [15], SPReID [8], PABR [21], DSR [4] and FPR [6], OSNet [31], ABDNet [1], CAR [32] and P^2-Net [3]. We only conduct a single query experiment. The results are shown in Table 2, which suggests that the proposed

Table 2. Performance comparison on Market1501, DukeMTMC and MSMT17 datasets.

Method	Market1501	DukeMTMC	MSMT17
PCB [23]	92.3 (77.4)	81.8 (66.1)	-
PCB+RPP [23]	93.8 (81.6)	83.3 (69.2)	-
SPReID [8]	92.5 (81.3)	83.3 (68.8)	-
MaskReID [15]	90.0 (75.3)	-	-
PABR [21]	90.2 (76.0)	82.1 (64.2)	-
OSNet [31]	94.8 (84.9)	88.6 (73.5)	78.7 (52.9)
ABDNet [21]	95.6 (88.3)	89.0 (78.6)	82.3 (60.8)
CAR [32]	96.1 (84.7)	86.3 (73.1)	-
P^2-Net [3]	95.2 (85.6)	86.5 (73.1)	-
DSR [4]	94.7 (85.8)	88.1 (77.1)	-
FPR [6]	95.4 (86.6)	88.6 (78.4)	-
Baseline	94.8 (83.4)	87.3 (73.7)	77.3 (49.4)
GASM	95.3 (84.7)	88.3 (74.4)	79.5 (52.5)

GSAM achieves competitive performance on all evaluation criteria. It is noted that: (1) The gaps between our results and baseline model (ResNet-50+Triplet loss + softmax cross-entropy loss) are significant: GASM increases from 94.8% to 95.3% at rank-1 accuracy and from 83.4% to 84.7% at mAP accuracy, which fully suggests that guided adaptive spatial matching is more effective than only using global feature matching. GASM can reduce the influence of background for cross-camera person Re-ID; (2) Benefit from guided saliency feature learning based on external human segmentation model and pose estimation model, GSAM can reduce the influence of background. Besides, adaptive spatial matching effectively addresses the misalignment problem; (3) Contributed by exacting human semantic parsing and pose estimation, SPReID, PABR, and P^2-Net achieves the competitive accuracy. However, SPReID, PABR, and P^2-Net rely on excellent human segmentation model and pose estimation during the inference term. GASM place a huge advantage over these mask-guided and pose-guided approaches since it does not depend on external cue models during the inference term; (4) Although attention-based approaches OSNet, ABDNet and CAR improve the performance of person Re-ID, only using self-attention mechanism easily result in unstable of these methods due to partial body occlusion.

Results on DukeMTMC-reID Person Re-ID results on DukeMTMC-reID are given in Table 2. This dataset is challenging because the person bounding box size varies drastically across different camera views, which naturally suits the proposed GASM. We report the results of PCB, MasKReID, SPReID, PABR, DSR, FPR, OSNet, ABDNet, CAR and P^2-Net. The results show that GASM achieves 88.1% at rank-1 accuracy and 74.4% mAP accuracy. Besides, GASM beats the spatial feature reconstruction method DSR by 0.3% at the rank-1 accuracy. However, GASM performs worse than FPR due to pyramid pooling used in FPR can cope with scale variations to some extent.

Results on MSMT17 Very few methods report the result on this dataset since it is recently released. Except for PCB, MasKReID, SPReID, PABR, other comparison methods have reported the result on MSMT17. The results show that the proposed GASM yields comparable results both at the rank-1 and mAP metrics.

Table 3. Performance comparison on Cross-domain person Re-ID.

Method	M→D	M→MS	D→M	D→MS	MS→M	MS→D
Baseline	54.3 (34.8)	22.2 (7.5)	61.6 (29.6)	26.8 (8.7)	64.7 (43.3)	64.6 (33.8)
GASM	56.4 (36.5)	27.8 (9.5)	67.2 (34.8)	32.8 (10.4)	68.4 (49.1)	68.3 (35.8)

As discussed above, GASM can achieve competitive performance on the three benchmark datasets. Referring to the results of occluded person Re-ID experimental result, our proposed GASM can both address occluded person Re-ID and non-occluded person Re-ID. Since these three datasets are collected in non-crowded scenes, it cannot show its superiority.

4.5 Cross-domain Person Re-IDentification

To verify the effectiveness of GASM in cross-domain person Re-ID, we conduct several cross-domain Re-ID experiments including M→D, M→MS, D→M, D→MS, MS→M and MS→D (M→D means that train model with Market1501 and test model with DukeMTMC-reID). More remarkable, we train a model using source-domain dataset, and then directly test on target-domain dataset without extra processing like finetuning and GAN on target-domain dataset. We only compare our proposed GASM against a baseline model (the baseline model is based on ResNet-50). Table 3 shows the results of cross-domain Re-ID. It is noted that GASM has lager superiority for cross-domain testing, improving rank-1 accuracy by 2.1%, 5.6%, 5.6%, 6.0%, 3.7% and 3.7% for M→D, M→MS, D→M, D→MS, MS→M and Ms→D, respectively. Such results also suggest that GASM improves the performance of cross-domain Re-ID by guided saliency feature learning. It is well known that the differences between different datasets mainly come from background, content, and camera parameters, GASM is capable of eliminating the background-bais to further improve the performance of person Re-ID.

In summary, we provide a good baseline for cross-domain person Re-ID and unsupervised person Re-ID.

4.6 Ablation Study

In this subsection, we aim to thoroughly analyze the effectiveness of each component in our Guided Saliency Feature Learning framework. Four different networks including FCN + average pooling (baseline), FCN + pose layer, FCN + mask layer, and FCN + pose layer + mask + layer are designed, and test on occluded datasets and three benchmark datasets, respectively. Besides, we also conduct cross-domain person Re-ID experiments using the four networks.

Table 4 shows the experimental results using Market1501, DukeMTMC-reID and MSMT17. We find the results on three datasets are similar, FCN + pose layer + mask layer outperforms better than other three designed networks on different tasks because pose-guided heatmap combined with mask-guided heatmap can output more accurate saliency heatmap. Pose-guided heatmap only focuses on human keypoints, the weights around person keypoints are relatively large

Table 4. The performance of person Re-ID using different components.

Market1501 train	Market1501	DukeMTMC	MSMT17	Crowd REID	Occluded REID
Baseline	94.8 (83.8)	54.3 (34.8)	22.18 (7.5)	16.7 (16.4)	60.2 (57.0)
+mask	94.5 (84.1)	55.3 (35.2)	27.1 (9.3)	32.2 (30.4)	74.1 (64.3)
+pose	95.1 (84.3)	56.1 (35.7)	26.1 (9.0)	30.0 (30.4)	74.5 (65.6)
+pose, +mask	95.3 (84.7)	56.4 (36.5)	79.5 (52.5)	32.4 (31.1)	74.5 (56.6)
DukeMTMC train	Market1501	DukeMTMC	MSMT17	Crowd REID	Occluded REID
Baseline	61.6 (29.6)	87.3 (73.68)	26.8 (8.7)	26.6 (24.3)	66.1 (60.2)
+mask	65.6 (32.8)	87.8 (72.4)	31.5 (9.9)	32.2 (28.7)	75.5 (68.5)
+pose	67.2 (33.8)	87.9 (73.9)	32.5 (10.2)	34.3 (31.1)	75.4 (67.4)
+pose, +mask	67.2 (34.2)	88.1 (74.2)	32.8 (10.4)	35.8 (31.8)	76.5 (67.8)
MSMT17 train	Market1501	DukeMTMC	MSMT17	Crowd REID	Occluded REID
Baseline	64.6 (33.8)	64.7 (43.3)	77.3 (49.4)	44.0 (41.2)	74.6 (69.2)
+mask	67.9 (35.5)	67.9 (47.7)	79.3 (51.7)	50.9 (47.4)	78.9 (72.0)
+pose	67.2 (34.5)	67.7 (47.5)	78.5 (51.7)	50.4 (47.0)	78.4 (72.9)
+pose, +mask	68.2 (35.9)	68.3 (49.1)	79.7 (52.5)	51.5 (47.1)	80.3 (73.1)

while the weights beyond person keypoints are relatively small. As for Mask-guided heatmap, it only focuses on the person part, but it cannot address partial occlusion body contamination. It is also noted that: (1) FCN + average pooling performs worse than other three networks on different tasks due to failing to address occlusion and background; (2) Mask-guided heatmap and pose-guided heatmap are complementary, mask-guided heatmap helps to find the person part but cannot address partial body occlusion; pose-guided heatmap helps to find the person to be identified but fails to address every part of person image. In summary, the two modules work well together to occluded person Re-ID, non-occluded person Re-ID and cross-domain person Re-ID.

5 Conclusion

We have proposed a novel approach called Guided Adaptive Spatial Matching to person Re-ID in crowded scenes. The proposed method design a pose-guided layer and mask-guide layer to learn the saliency heatmap to further guide feature learning. Besides, we proposed a novel matching approach called Guided Adaptive Spatial Matching (GASM) where each spatial feature in the query can find the most similar spatial features of a person in a gallery to match. Remarkably, saliency heatmap used in spatial matching can solve partial body occlusion and eliminate background-bais. Experimental results on Crowd REID and Occluded REID dataset validate that GSAM can address the occlusion problem. Besides, GSAM also shows its advantage in cross-domain by eliminating background-bais. Additionally, the proposed method is also competitive on the benchmark person datasets.

Acknowledgement Special thanks to Xingyu Liao who support our experiments, and thanks to FastReID: https://github.com/JDAI-CV/fast-reid [5] that provides with codebase for GASM. This work is partially supported by Beijing Academy of Artificial Intelligence (BAAI).

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: Proceedings of the IEEE International Conference on Computer Vision (CVPR) (2019)
2. 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. In: Asian Conference on Computer Vision (ACCV) (2018)
3. Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J.G., Han, K.: Beyond human parts: Dual part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
4. He, L., Liang, J., Li, H., Sun, Z.: Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7073–7082 (2018)
5. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: A pytorch toolbox for general instance re-identification. arXiv preprint arXiv:2006.02631 (2020)
6. He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z., Feng, J.: Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
7. Jiaxuan Zhuo, Zeyu Chen, J.L.G.W.: Occluded person re-identification. In: IEEE International Conference on Multimedia and Expo (ICME) (2018)
8. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1062–1071 (2018)
9. Li, S., Liu, X., Liu, W., Ma, H., Zhang, H.: A discriminative null space based deep learning approach for person re-identification. In: International Conference on Cloud Computing and Intelligence Systems (CCIS) (2016)
10. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
11. Liu, W., Zhang, C., Ma, H., Li, S.: Learning efficient spatial-temporal gait features with deep learning for human identification. Neuroinformatics **16**(3-4), 457–471 (2018)
12. Long, X., Gan, C., De Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: Purely attention based local feature integration for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2018)
13. Luo, H., Jiang, W., Fan, X., Zhang, C.: Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Transactions on Multimedia (T-MM) (2020)
14. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
15. Qi, L., Huo, J., Wang, L., Shi, Y., Gao, Y.: Maskreid: A mask based deep ranking neural network for person re-identification. arXiv preprint arXiv:1804.03864 (2018)
16. Ren, M., He, L., Li, H., Liu, Y., Sun, Z., Tan, T.: Robust partial person re-identification based on similarity-guided sparse representation. In: Chinese Conference on Biometric Recognition (CCBR) (2017)
17. 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. pp. 17–35. Springer (2016)

18. 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)
19. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1179–1188 (2018)
20. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: IEEE International Conference on Computer Vision (ICCV). pp. 3980–3989 (2017)
21. 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)
22. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: 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 (CVPR) (2019)
23. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling. arXiv preprint arXiv:1711.09349 (2017)
24. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2018)
25. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
26. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2017)
27. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
28. Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: IEEE International Conference on Computer Vision (ICCV) (2015)
29. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. arXiv preprint arXiv:1707.00408 (2017)
30. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3754–3762 (2017)
31. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
32. Zhou, S., Wang, F., Huang, Z., Wang, J.: Discriminative feature learning with consistent attention regularization for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
33. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)