Mask-Guided Region Attention Network for Person Re-Identification

Cong Zhou[1] and Han Yu[1,2(✉)]

[1] Nanjing University of Posts and Telecommunications School of Computer Science and Technology, Nanjing 210023, China
519658713@qq.com, han.yu@njupt.edu.cn
[2] Jiangsu Key Lab of Big Data Security and Intelligent Processing, Nanjing 210023, China

Abstract. Person re-identification (ReID) is an important and practical task which identifies pedestrians across non-overlapping surveillance cameras based on their visual features. In general, ReID is an extremely challenging task due to complex background clutters, large pose variations and severe occlusions. To improve its performance, a robust and discriminative feature extraction methodology is particularly crucial. Recently, the feature alignment technique driven by human pose estimation, that is, matching two person images with their corresponding parts, increases the effectiveness of ReID to a certain extent. However, we argue that there are still a few problems among these methods such as imprecise handcrafted segmentation of body parts, and some improvements can be further achieved. In this paper, we present a novel framework called Mask-Guided Region Attention Network (MGRAN) for person ReID. MGRAN consists of two major components: Mask-guided Region Attention (MRA) and Multi-feature Alignment (MA). MRA aims to generate spatial attention masks and meanwhile mask out the background clutters and occlusions. Moreover, the generated masks are utilized for region-level feature alignment in the MA module. We then evaluate the proposed method on three public datasets, including Market-1501, DukeMTMC-reID and CUHK03. Extensive experiments with ablation analysis show the effectiveness of this method.

Keywords: Person re-identification · Human pose estimation · Mask

1 Introduction

Person re-identification (ReID) aims to identify the same individual across multiple cameras. In general, it is considered as a sub-problem of image retrieval. Given a query image containing a target pedestrian, ReID is to rank the gallery images and search for the same pedestrian. It plays an important role in various surveillance applications, such as intelligent security and pedestrian tracking.

In the past years, many methods [1–4] have been proposed to address the ReID problem. However, it still remains as an incomplete task due to large pose variations, complex background clutters, various camera views, severe occlusions and uncontrollable illumination conditions. Recently, with the improvement of human pose

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12085, pp. 286–298, 2020.
https://doi.org/10.1007/978-3-030-47436-2_22

Fig. 1. Imprecise shapes of body parts set by handcraft, such as patches [1, 11] and rectangular RoIs [9, 12], include extensive noise.

estimation [5–7], some researches [8–10] utilize the estimation results as spatial attention maps to learn features from pedestrian body parts and then align them. These methods achieve great success and prove that extracting features exactly from body regions rather than background regions is helpful for ReID.

However, there are still notable problems in these methods concluded as follows. 1) As shown in Fig. 1, these methods tend to extract features from imprecise part shapes set by handcraft, such as patches [1, 11] and rectangular regions of interest (RoIs) [9, 12], which can introduce noise. 2) Part-level feature alignment which means matching two pedestrians with their heads, arms, legs, and other body parts is improper for ReID. 3) Feature representation is not accurate and comprehensive enough.

In the first problem, the main reason that the handcrafted shapes cannot precisely describe the silhouettes of body parts is that the shapes of body parts are irregular. Feature alignment based on these shapes can introduce noise from background clutters, occlusions and even adjacent parts as in Fig. 1, leading to inaccurate matching. To deal with this problem, we propose to use pedestrian masks as the spatial attention maps for masking out clutters and meanwhile obtaining the finer silhouettes of body parts both in pixel-level, as shown in Fig. 2(a). These silhouettes obtained by pedestrian masks should be more precise and closer to the reality of body shapes. For the second problem, the works mentioned above generally align features based on part-level and this is inappropriate for ReID. As walking is a dynamic process, and in this process, the moving arms and legs have huge morphological changes and often cause heavy self-occlusion, which implies that a body part will inevitably be occluded by other parts. For example, left legs are often occluded by right legs. Due to self-occlusion, it is difficult to align features based on part-level. Furthermore, each pedestrian has his own walking postures that are different from others', which means his head, upper body and lower body have their own morphological characters when walking. But the part-level alignment may discard these characters, as shown in Fig. 2(b). Meanwhile, the head, upper body and lower body are generally separate from each other in a walking pedestrian, which indicates there is no self-occlusion among these three parts as demonstrated in Fig. 2(b).

Based on the above analysis, it is concluded that region-level feature alignment based on head, upper body and lower body is more reasonable for ReID. Furthermore, apart from self-occlusion, pedestrians may have some carry-on items, such as

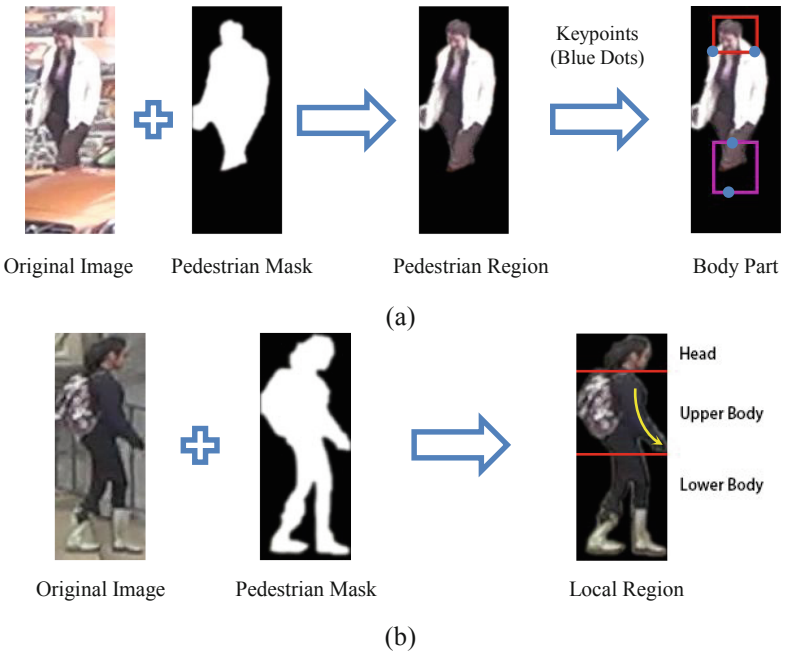

Fig. 2. (a): Pedestrian masks can be used to mask out clutters and obtain the finer silhouettes of pedestrian body parts. (b): Pedestrians' heads, upper bodies and lower bodies have their own morphological characters which can not be presented by a single body part. For example, the morphological characters of upper bodies are presented by arms and upper torsos, such as the amplitude of arm swing (Yellow Arrow). (Color figure online)

backpacks, handbags and caps. These items are definitely helpful for ReID and we can treat them as special parts of pedestrians, which should be included in the corresponding local region like in Fig. 2(b). In the third problem, these methods like [1, 9, 12] only align the part features, considered as local features, and the global feature of the whole pedestrian region is not considered. However, each pedestrian is intuitively associated with a global feature including body shape, walking posture and so on, which cannot be replaced by local features. Due to the neglect of global features, the final feature representation will not be comprehensive and robust enough. Meanwhile, previous works [13, 14] extract the global feature from the entire pedestrian image including background clutters and occlusions, which will introduce noise and lead to the inaccuracy of feature representation. Here, we utilize pedestrian masks to redesign the global features, removing clutters with masks firstly and then extracting the global features of pedestrians. After these operations, multi-feature fusion can be used to align features.

Based on above motivations, we propose a new Mask-Guided Region Attention Network for person re-identification. The contributions of our work can be summarized as follows:

- To make the better use of feature alignment technique for person re-identification, a unified framework called Mask-Guided Region Attention Network (MGRAN) is proposed.
- To further reduce the noise from background clutters and occlusions, we explore to utilize masks to separate pedestrians from them and obtain the finer silhouettes of pedestrian bodies.
- Region-level feature alignment, based on head, upper body and lower body, is introduced as a more appropriate method for ReID.
- We redesign the global feature and utilize multi-feature fusion to improve the accuracy and the completeness of feature representation.

2 Related Work

2.1 Person Re-Identification

Recently, person re-identification methods based on deep learning achieved great success [13, 15, 16]. In general, these methods can be classified into two categories, namely feature representation and distance metric learning. The first category [1, 3, 17, 18] often treats ReID as a classification problem. These methods dedicate to design view-invariant representations for pedestrians. The second category [19–21] mainly aims at measuring the similarity between pedestrian images by learning a robust distance metric.

Among these methods, many of them [9, 12] achieved the success by feature alignment. Numerous studies proved the importance of feature alignment for ReID. For example, Su et al. [5] proposed a Pose-driven Deep Convolutional model (PDC) that used Spatial Transformer Network (STN) to crop body regions based on pre-defined centers. Xu et al. [9] achieved the more precise feature alignment based on their proposed network called Attention-Aware Compositional Network (AACN) and further improved the performance of identification. However, these methods align the part features based on the body shapes set by handcraft, which is usually imprecise. In our model, we utilize pedestrian masks in pixel-level to align features, intending to obtain more precise information of body parts.

2.2 Instance Segmentation and Human Pose Estimation

With the rapid development of instance segmentation based on deep learning methods such as Mask R-CNN [22] and the Fully Convolutional Networks (FCN) [23], now we can easily obtain high-quality pedestrian masks which can be used in person re-identification. Furthermore, these instance segmentation methods can be naturally extended to human pose estimation by modeling keypoint locations as one-hot masks. We can further improve the performance of person re-identification by integrating the results of instance segmentation and human pose estimation.

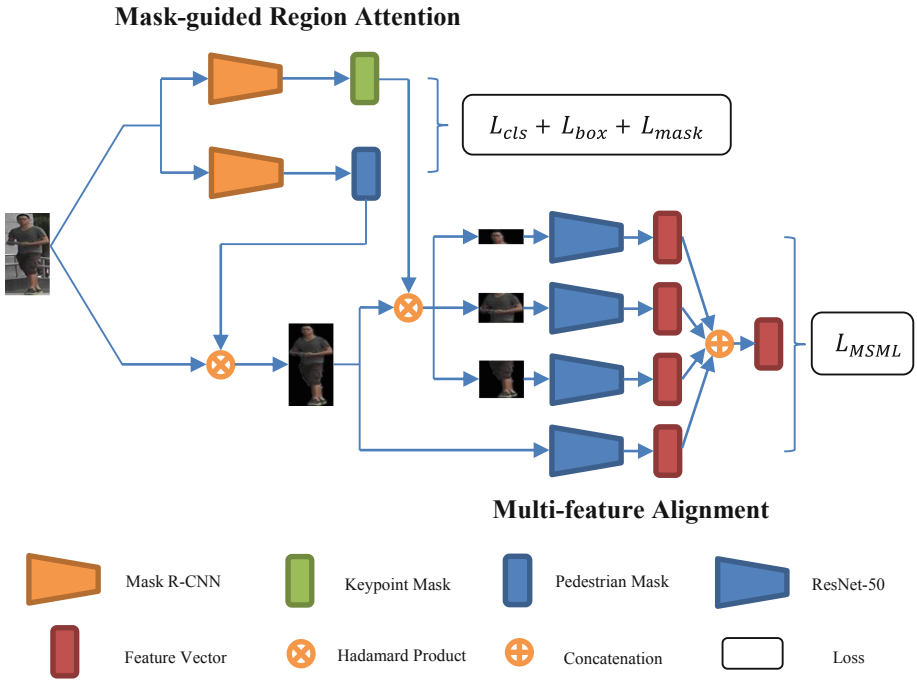

Fig. 3. Mask-Guided Region Attention Network (MGRAN). Our proposed MGRAN consists of two main components: Mask-guided Region Attention (MRA) and Multi-feature Alignment (MA). MRA aims to generate two types of attention maps: pedestrian masks and human body keypoint masks. MA utilizes the attention maps generated by MRA to obtain the pedestrian region and the three associated local regions. Then the global feature and local features are extracted and multi-feature fusion is used to align them.

2.3 Spatial Attention Mechanism

Spatial attention mechanism has achieved great success in understanding images and it has been widely used in various tasks, such as semantic segmentation [24], object detection [25] and person re-identification [26]. For example, Chu et al. [6] proposed a multi-context attention model for pose estimation. Inspired by these methods, we use spatial attention maps to remove the undesirable clutters in pedestrian images. However, different from them, we use binary pedestrian masks as spatial attention maps to obtain more precise information of pedestrian bodies.

3 Mask-Guided Region Attention Network

3.1 Overall Architecture

The overall framework of our Mask-Guided Region Attention Network (MGRAN) is illustrated in Fig. 3. MGRAN consists of two main components: Mask-guided Region Attention (MRA) and Multi-feature Alignment (MA).

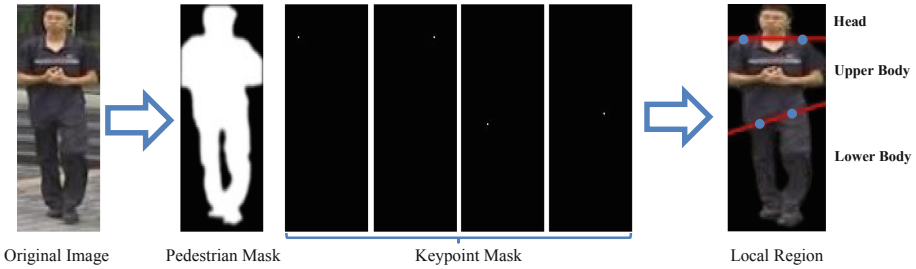

Fig. 4. Two types of masks: pedestrian masks and keypoint masks. In this paper, we define four keypoints (Blue Dots). By connecting two adjacent keypoints, we can divide the pedestrian region into three local regions: the head, the upper body and the lower body. (Color figure online)

The MRA module aims to generate two kinds of attention maps: pedestrian masks and human body keypoints. It is constructed by a two-branch neural network, which predicts the attention maps of the pedestrians and their keypoints, respectively.

The MA module is constructed by a four-branch neural network. It utilizes the estimated attention maps to extract global features and local features. A series of extracted features are then fused for multi-feature alignment.

3.2 Mask-Guided Region Attention

Different from other works, we use binary masks as attention maps for highlighting specific regions of human body in the image. With the rapid development of instance segmentation, there are many alternative methods to generate pedestrian masks. In this paper, we choose Mask R-CNN [22] to predict the masks due to its high accuracy and flexibility. As shown in Fig. 4, there are two types of masks: pedestrian masks and keypoint masks. They are simultaneously learned in a unified form through our proposed Mask-guided Region Attention module.

Pedestrian Masks $\boldsymbol{P}$. A pedestrian mask $\boldsymbol{P}$ is the encoding of an input image's spatial layout. It is a binary encoding which means that the pixels of pedestrian region are encoded as number 1 and the others are encoded as number 0. Following the original article of Mask R-CNN, we set hyper-parameters as suggested by existing Faster R-CNN work [27] and define the loss $\boldsymbol{L_{mask}(P)}$ on each sampled RoI in Mask R-CNN as the average binary cross-entropy loss,

$$L_{mask}(\boldsymbol{P}) = -\frac{1}{N}\sum\nolimits_{i=1}^{N} \boldsymbol{P_i^*} \cdot \log(\sigma(\boldsymbol{P_i})) + \left(1 - \boldsymbol{P_i^*}\right) \cdot \log(1 - \sigma(\boldsymbol{P_i})), \quad (1)$$

where N is the number of pixels in a predicted mask, σ denotes the sigmoid function, $\boldsymbol{P_i}$ is a single pixel in the mask, and $\boldsymbol{P_i^*}$ is the corresponding ground truth pixel. Furthermore, the classification loss L_{cls} and the bounding-box loss L_{box} of each sampled RoI are set as indicated in [21].

Keypoint Masks $\boldsymbol{K}$. Mask R-CNN can easily be extended to keypoints detection. We model a keypoint's location as a one-hot mask and use Mask R-CNN to predict four

masks, one for each of the four keypoints as shown in Fig. 4. Following the original article of Mask R-CNN, during training, we minimize the cross-entropy loss over an $\boldsymbol{m}^2$-way softmax output for each visible ground-truth keypoint, which encourages a single point to be detected.

3.3 Multi-feature Alignment

Based on the attention masks generated by Mask-guided Region Attention module, we propose a Multi-feature Alignment (MA) module to align the global feature and local features. MA consists of two main stages called Space Alignment (SA) and Multi-feature Fusion (MF). The complete structure of MA is shown in Fig. 3.

Space Alignment (SA). Space Alignment aims to obtain the pedestrian region and the three local regions. Based on the attention masks generated by MRA module, we propose a simple and effective approach to obtain them. Specifically, we firstly apply Hadamard Product between the original image $\boldsymbol{M}$ and the corresponding pedestrian mask $\boldsymbol{P}$ to obtain the pedestrian region, as follows:

$$\boldsymbol{M}^* = \boldsymbol{M} \circ \boldsymbol{P}, \tag{2}$$

where, $\circ$ denotes the Hadamard Product operator which performs element-wise product on two matrices or tensors and $\boldsymbol{M}^*$ denotes the pedestrian region. It is worth noting that we use Hadamard Product on the original image to guarantee the accuracy of features. Some works [9, 12] use spatial attention maps on processed data such as data processed by convolution, which will introduce noise into the attention region from other regions in the image. Secondly, based on the obtained pedestrian body region, we utilize the four keypoint masks to obtain the three local regions by connecting two adjacent keypoints and segmenting the pedestrian region, as shown in Fig. 4.

Multi-feature Fusion (MF). In this module, we use four ResNet-50 networks [28] to extract the features of the four regions generated by SA module, respectively. Then feature fusion is used to align features, as follows:

$$\boldsymbol{F} = Concat\left(\left\{\boldsymbol{f}_g\ \boldsymbol{f}_l^1\ \boldsymbol{f}_l^2\ \boldsymbol{f}_l^3\right\}\right), \tag{3}$$

where $Concat(\cdot)$ denotes the concatenation operation on feature vectors, $\boldsymbol{f}_g$ represents the global feature of the whole pedestrian body region, $\boldsymbol{f}_l^1$, $\boldsymbol{f}_l^2$ and $\boldsymbol{f}_l^3$ denote the features of the three local regions respectively, and $\boldsymbol{F}$ is the final feature vector for the input pedestrian image.

Overall, our framework integrated the MRA and MA to extract features for input pedestrian images.

3.4 Implementation Details

We construct the Mask R-CNN model with a ResNet-50-FPN backbone and use the annotated person images in the COCO dataset [29] to train it. Furthermore, the floating-number mask output is binarized at a threshold of 0.5. In MF, the four ResNet-50

Table 1. The details of three public datasets used in experiments.

Datasets	# IDs	# cameras	# resolution
Market-1501 [31]	1501	6	64 × 128
DukeMTMC-reID [32]	1812	8	Vary
CUHK03 [1]	1467	2	Vary

networks share the same parameters and we use the Margin Sample Mining Loss (MSML) [30] to conduct distance metric learning based on the four features extracted by ResNet-50. We scale the all images input into Mask R-CNN and ResNet-50 with a factor of 1/256. Finally, MRA and MA are trained independently.

4 Experiments

In this section, the performance of Mask-Guided Region Attention Network (MGRAN) is compared with several state-of-the-art methods on three public datasets. Furthermore, detailed ablation analysis is conducted to validate the effectiveness of MGRAN components.

4.1 Datasets and Protocols

We evaluate our method on three large-scale public person ReID datasets, including Market-1501 [31], DukeMTMC-reID [32] and CUHK03 [1], details of them are shown in Table 1. For fair comparison, we follow the official evaluation protocols of each dataset. For Market-1501 and DukeMTMC-reID, rank-1 identification rate (%) and mean Average Precision (mAP) (%) are used. For CUHK03, Cumulated Matching Characteristics (CMC) at rank-1 (%) and rank-5 (%) are adopted.

4.2 Comparison with the State-of-the-Art Methods

We choose 13 methods in total with state-of-the-art performance for comparisons with our proposed framework MGRAN. These methods can be categorized into two classes according to whether human pose information is used. The Spindle-Net (Spindle) [12], Deeply-Learned Part-Aligned Representations (DLPAR) [10], MSCAN [33], and the Attention-Aware Compositional Network (AACN) [9] are pose-relevant. The Online Instance Matching (OIM) [14], Re-ranking [34], the deep transfer learning method (Transfer) [35], the SVDNet [15], the pedestrian alignment network (PAN) [36], the Part-Aligned Representation (PAR) [10], the Deep Pyramid Feature Learning (DPFL) [13], DaF [37] and the null space semi-supervised learning method (NFST) [38] are pose-irrelevant. The experimental results are presented in Table 2, 3 and 4.

Based on the experimental results, it is obvious that our MGRAN framework outperforms the compared methods, showing the advantages of our approach. To be specific, compared with the second best method on each dataset, our framework achieves 6.10%, 1.89%, 1.28%, 7.62% and 6.57% rank-1 accuracy improvement on

Table 2. Comparison results on Market-1501 dataset.

Market-1501	Single query		Multiplequery	
	Rank-1	mAP	Rank-1	mAP
Spindle [12]	76.90	–	–	–
DLPAR [10]	81.00	63.40	–	–
MSCAN [33]	80.31	57.53	–	–
SVDNet [15]	82.30	62.10	–	–
PAN [36]	82.81	63.35	88.18	71.72
Re-ranking [34]	77.11	63.63	–	–
NFST [38]	61.02	35.68	71.56	46.03
MGRAN (Ours)	88.91	78.03	90.07	81.30

Table 3. Comparison results on DukeMTMC-reID dataset.

DukeMTMC-reID	Rank-1	mAP
SVDNet [15]	76.70	56.80
OIM [14]	68.10	–
PAN [36]	71.59	51.51
AACN [9]	76.84	59.25
MGRAN (Ours)	78.12	63.57

Table 4. Comparison results on CUHK03 dataset.

CUHK03	Labeled		Detected	
	Rank-1	Rank-5	Rank-1	Rank-5
PAR [10]	85.40	97.60	81.60	97.30
NFST [38]	62.55	90.05	54.70	84.75
SVDNet [15]	81.80	–	–	–
DPFL [13]	43.00	–	40.70	–
DaF [37]	27.50	–	26.40	–
Transfer [35]	85.40	–	84.10	–
MGRAN (Ours)	93.02	98.94	90.67	98.21

Market-1501 (Single Query), Market-1501 (Multiple Query), DukeMTMC-reID, CUHK03 (Labeled) and CUHK03 (Detected), respectively. Furthermore, compared with the second best method on Market-1501 and DukeMTMC-reID, 14.40%, 9.58% and 4.32% mAP improvement on Market-1501 (Single Query), Market-1501 (Multiple Query) and DukeMTMC-reID are achieved, respectively.

Table 5. Effectiveness of MF. MGRAN − GF means removing global features in final feature vectors.

Ablation Analysis	Market-1501				DukeMTMC-reID	
	Single Query		Multiple Query		Rank-1	mAP
	Rank-1	mAP	Rank-1	mAP		
MGRAN − GF	86.30	74.26	87.82	80.53	77.31	60.10
MGRAN	88.91	78.03	90.07	81.30	78.12	63.57

Table 6. Effectiveness of RFA. MGRAN-PL means aligning features based on part-level. MGRAN-RL means aligning features based on region-level.

Ablation Analysis	CUHK03			
	Labeled		Detected	
	Rank-1	Rank-5	Rank-1	Rank-5
MGRAN-PL	91.83	97.41	89.35	96.75
MGRAN-RL	93.02	98.94	90.67	98.21

4.3 Ablation Analysis

In this section, we evaluate the effect of our proposed multi-feature fusion and region-level feature alignment by ablation analysis.

Multi-feature Fusion (MF). We verify the effectiveness of MF on Market-1501 and DukeMTMC-reID dataset by removing global features in final feature vectors. As shown in Table 5, MF increases the rank-1 accuracy by 2.61%, 2.25% and 0.81% on Market-1501 (Single Query), Market-1501 (Multiple Query) and DukeMTMC-reID. Furthermore, 3.77%, 0.77% and 3.47% mAP improvement on Market-1501 (Single Query), Market-1501 (Multiple Query) and DukeMTMC-reID are achieved based on MF.

Region-Level Feature Alignment (RFA). We align features based on part-level and region-level respectively to verify the effectiveness of our proposed region-level feature alignment. Specifically, we replace region-level feature alignment in MGRAN with part-level feature alignment and keep the other parts unchanged. As shown in Table 6, RFA increases the rank-1 accuracy by 1.19% and 1.32% on CUHK03 (Labeled) and CUHK03 (Detected). Meanwhile, RFA increases the rank-5 accuracy by 1.53% and 1.46% on CUHK03 (Labeled) and CUHK03 (Detected). The experimental results show the usefulness of our proposed RFA.

5 Conclusion

In this paper, we propose a novel Mask-Guided Region Attention Network (MGRAN) for person re-identification to deal with the clutter and misalignment problem. MGRAN consists of two main components: Mask-guided Region Attention (MRA) and Multi-feature Alignment (MA). MRA generates spatial attention maps to mask out undesirable clutters and obtain finer silhouettes of pedestrian bodies. MA aims to align features based on region-level which is more appropriate for ReID. Our method has achieved some success, but with the rapid development of science, a great number of excellent technologies have been created, such as GAN, and in the future work, we propose to use these technologies to further improve the performance of ReID.

Acknowledgements. We are grateful for the financial support of the China Postdoctoral Science Foundation (grant no. 2018T110531), the National Natural Science Foundation of China (grant no. 11501302), and the Natural Science Foundation of Nanjing University of Posts and Telecommunications NUPTSF (grant no. NY219080).

References

1. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: CVPR, pp. 152–159 (2014)
2. Chen, S.Z., Guo, C.C., Lai, J.H.: Deep ranking for person re-identification via joint representation learning. IEEE Trans. Image Process. **25**(5), 2353–2367 (2016)
3. 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: CVPR, pp. 1335–1344 (2016)
4. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 475–491. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_30
5. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV, pp. 3960–3969 (2017)
6. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: CVPR, pp. 1831–1840. (2017)
7. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
8. Kumar, V., Namboodiri, A., Paluri, M., Jawahar, C.V.: Pose-aware person recognition. In: CVPR, pp. 6223–6232 (2017)
9. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: CVPR, pp. 2119–2128 (2018)
10. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: ICCV, pp. 3219–3228 (2017)
11. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: ICCV, pp. 2528–2535 (2013)

12. Zhao, H., et al.: Spindle Net: person re-identification with human body region guided feature decomposition and fusion. In: CVPR, pp. 1077–1085 (2017)
13. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: ICCV, pp. 2590–2600 (2017)
14. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR, pp. 3415–3424 (2017)
15. Sun, Y., Zheng, L., Deng, W., Wang, S.: SVDNet for pedestrian retrieval. In: ICCV, pp. 3800–3808 (2017)
16. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV, pp. 994–1002 (2017)
17. Wu, L., Shen, C., Hengel, A.V.D.: PersonNet: person re-identification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255 (2016)
18. Shi, H., et al.: Embedding deep metric for person re-identification: a study against large variations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 732–748. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_44
19. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
20. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognit. **48**(10), 2993–3003 (2015)
21. Girshick, R.: Fast R-CNN. In: ICCV, pp. 1440–1448 (2015)
22. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2961–2969 (2017)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
24. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: scale-aware semantic image segmentation. In: CVPR, pp. 3640–3649 (2016)
25. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: CVPR, pp. 5659–5667 (2017)
26. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Trans. Image Process. **26**(7), 3492–3506 (2017)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, pp. 91–99 (2015)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
29. Lin, T.-Y., et al.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
30. Xiao, Q., Luo, H., Zhang, C.: Margin sample mining loss: a deep learning based method for person re-identification. arXiv preprint arXiv:1710.00478 (2017)
31. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV, pp. 1116–1124 (2015)
32. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: ICCV, pp. 3754–3762 (2017)
33. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR, pp. 384–393 (2017)
34. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR, pp. 1318–1327 (2017)
35. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244 (2016)

36. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. IEEE Trans. Circ. Syst. Video Technol. **29**(10), 3037–3045 (2018)
37. Yu, R., Zhou, Z., Bai, S., Bai, X.: Divide and fuse: a re-ranking approach for person re-identification. arXiv preprint arXiv:1708.04169 (2017)
38. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR, pp. 1239–1248 (2016)