Part-aware network: A simple but efficient method for occluded person re-identification

Peijun Ye, Haitang Zeng, Wei Zhang, Dihu Chen*
School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou, China

ABSTRACT

Person re-identification is to query person across cameras and occlusion is one of the difficulties. Previous works have proved that local feature extraction and alignment are critical for occluded person re-identification. However, directly horizontal partition causes mis-alignment and extra-cue methods highly depend on the quality of extra-cues. In this work, we propose a novel architecture including a weakly supervised mask generator without introducing extra-cues to create fine-grained semantic masks for local feature extraction and alignment, and a weight-shared fully connection to control the balance of local and global features. We also propose a general form of weighted pooling to improve gradient transfer, which gets rid of the probability explanation with softmax. Moreover, we unravel that there is a conflict between local branches and global branch, and a buffer convolution layer helps to fix this conflict. Extensive experiments show the effectiveness of our proposed method on occluded and holistic ReID tasks. Specially, we achieve 62.5% Rank-1 and 52.6% mAP (mean Average Precision) scores on the Occluded-Duke dataset.

Keywords: Person re-identification, deep learning, weakly supervised, local feature extraction, global-local balance

1. INTRODUCTION

Person re-identification (ReID) is to query a person across cameras, which is widely applied in video surveillance, security, and smart city. Recently, various person ReID methods[1-10] have been proposed and achieved a good performance on holistic cases. However, occlusion of a person happens when meeting obstacles (e.g., plants, cars, other pedestrians), so occlusion is one main challenge of ReID. Particularly, this problem caused by occlusion is called *Occluded Person Re-identification*[11,12].

Local features are proved to be helpful to solve occlusion problem[4,10,13], then new problems are how to segment local features reasonably and how to align local features precisely. Simply horizontally partition human features into several strips (e.g., PCB[4]) causes serious mis-alignment, some works introduce extra-cues (e.g., body key-points[10,12]] or human parsing[8,14]) for local feature extraction and achieve semantic alignment. However, extra-cue methods usually depend on an additional model, which greatly increases computation and parameters. What's more, the model is trained on datasets other than ReID datasets, thus leads to performance decrease on ReID datasets and influence local feature extraction and alignment.

Is there a simple but efficient end-to-end model? Concretely, a model does not introduce extra-cue, performs better than extra-cue methods, and is more light-weighted than extra-cue methods. Previous VPM[9] is an end-to-end model more light-weighted than extra-cue methods but performs worse than extra-cue methods. As previous works[4,15,16], we combine global and local branches, and further employ fully connection for global-local balance. In addition, previous works[9,12,16] usually apply a following softmax on masks, thus improperly restrict the representation of masks, we remove the softmax and propose a more general weighted pooling for masks without softmax. Finally, in this work, we contribute as follow:

- We propose a novel architecture including a weakly supervised mask generator to extract fine-grained semantic masks and a shared fully connection to control global-local balance;
- We propose a new form of weighted pooling which gets rid of probability explanation and performs better;
- We also discover a conflict between local and global feature extraction, thus a buffer layer to separate these two branches is useful;

*stscdh@mail.sysu.edu.cn

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122600M · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637388

• Extensive experiments and visualization of masks demonstrate that our method is effective. Especially in Occluded Duke, our method achieves 62.5% and 52.6% on Rank-1 and mAP scores.

2. THE PROPOSED METHOD

2.1 Local feature extraction

As in Figure 1, we adopt a mask-based manner for local feature extraction. Because most mask-based methods 9, 12, 16] have similar structures for local feature extraction, we would emphasise the difference of our method. In our work, the proposed Part-Aware Network (PAN) transforms pedestrian images into an embedded feature f_{cat} for discrimination, and this embedded feature f_{cat} is a concat of one global embedded feature and p local embedded features.

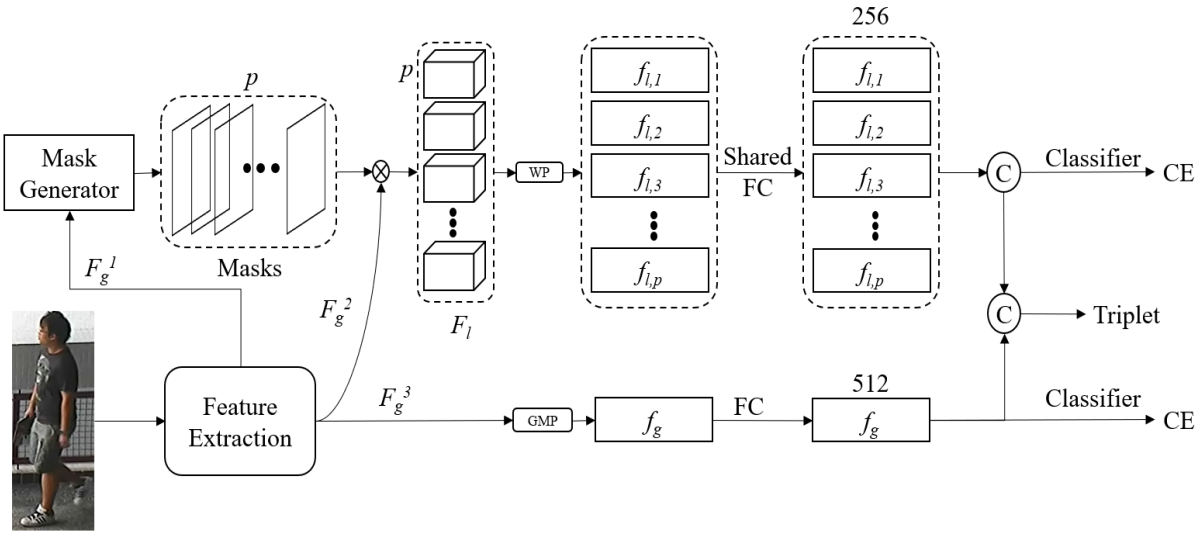


Figure 1. The Proposed Part-Aware Network. The feature extraction module is a backbone network (e.g., ResNet50) and the mask generator is a 1×1 convolution layer. "WP" is weighted pooling, "GMP" is global maximum pooling. "FC" is for dimension reduction and global-local balance, while "Shared FC" is a share-weight fully connection employed to all local features. $\otimes$ means element-wise matrix multiply, © denotes "concat".

Given a pedestrian image, we first resize and input it to PAN. Through a feature extraction module (a backbone network, e.g., ResNet 17] or OSNet 18]), PAN outputs **3 global features**:

- $F_g^1 \in \mathbb{R}^{c_1\times h\times w}$ for mask generation;
- $F_g^3 \in \mathbb{R}^{c_2\times h\times w}$ for mask-based local feature extraction;
- $F_g^3 \in \mathbb{R}^{c_3\times h\times w}$ for global feature extraction.

c, h, w are channel number, height and width of the feature respectively. Compared to VPM where these three tensors are the same tensor, we separate them to discuss whether an additional buffer layer helps to improve performance.

Then a mask generator is employed to F_g^1, which is formulated by

$$M = G(F_g^1) \tag{1}$$

where masks $M \in \mathbb{R}^{p\times h\times w}$, $G(\cdot)$ is the mask generator, which is a simple 1×1 convolution layer in our work. Though most mask-based works[9,12,19] employ a following softmax to masks for a probability explanation, our opinion is that this

operation affects the independence of masks and causes gradient vanishing. So we remove the softmax and develop a new weighted pooling method.

Based on the masks M, we extract local features from F_g^2, which is formulated by

$$F_{l,i} = F_g^2 \otimes m_i, \qquad i = 1, 2, ..., p \tag{2}$$

where $\otimes$ denotes element-wise multiply operation, the i^{th} mask $m_i \in \mathbb{R}^{1\times h\times w}$, and the i^{th} output local feature is $F_{l,i} \in \mathbb{R}^{c_2\times h\times w}$.

To squeeze the space information of local features, a proposed mask-based weighted pooling is employed to the local features, as in

$$f_{l,i} = \frac{\sum_j^{h\times w} F_{l,i}}{\sum_j^{h\times w} |\, m_i \,|} \tag{3}$$

the output local embedded feature $f_{l,i} \in \mathbb{R}^{c_2}$. As elements x_j of masks with a softmax satisfy $x_j \in [0,1]$, the elements of our masks are $x_j \in (-\infty, \infty)$, thus the sum of a mask may equal to zero, so we sum up the absolution of the mask. Our weighted pooling is a general form of the weighted pooling in VPM 9] and PAT 16], which gets rid of probability explanation and improves gradient transfer.

2.2 Global-local balance strategy

As we combine global feature and local features to a discriminative embedded feature, there is a balance between the global feature and the local features. Previous methods balance global and local branches on loss weight rather than on embedded feature dimensions, thus the range of global features and local features may be different, which means redundancy on the final embedded feature. We adopt fully connection layers to reduce dimension of embedded feature and achieve global-local balance. Concretely, we apply a fully connection layer for the global feature (dimension $c_3 \to N_g$) and a share-weight fully connection layer ("Shared FC" in Figure 1) for local features (dimension $c_2 \to n_l$). The shared weight manner on all local branches restricts that the difference of local features is only from mask-based local feature extraction.

We define the balance of global and local branches with the ratio of local features versus global γ

$$\gamma = \frac{N_l}{N_g} = \frac{pn_l}{N_g} \tag{4}$$

where n_l is the dimension of each local embedded feature (256 as default). Here we exploit a weight shared fully connection layer to adjust local feature size, thus change the ratio γ. Since γ means the importance ratio of local features vs. global feature, we expect our model reach the best performance when $1 < \gamma < n_l$. Because local features are more important than the global feature, while the global feature contains more information than a single local feature.

2.3 Training loss

We utilize identity loss and triplet loss for our model:

$$L = L_{id} + L_{tri} = \frac{1}{2}(L_{g,id} + L_{l,id}) + L_{tri} \tag{5}$$

while identity loss L_{id} is cross entropy loss with label smoothing[19], and triplet loss L_{tri} is Soft Margin Hard Mining Triplet Loss[20].

3. EXPERIMENT

3.1 Experiment setting

All models are trained and tested on Ubuntu 18.04 with 2 GTX 1080Ti GPUs. And we employ random flip and random erasing 16] as our data augmentation strategy. Samples are re-scaled to 256×128. We employ Adam optimizer and the base learning rate is 3.5×10^{-5}. On the training, there are 20 linear warm-up epochs for learning rate going to 3.5×10^{-4}, then on milestones 60 and 90 epoch, learning rate multiply 0.1 respectively.

3.2 Datasets

Market1501[1] and DukeMTMC-reID[2] are 2 common holistic ReID datasets. Occluded Duke[12] is made from DukeMTMC-reID. Its training, query, and gallery set contain 9%/100%/10% occluded images (14%/15%/10% for DukeMTMC-reID). In other words, the training set of Occluded Duke contains fewer occluded images than DukeMTMC-reID, and only occluded images are in its query set.

3.3 Evaluation protocol

For performance evaluation, we employ the common metrics of person re-identification, Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP).

4. PERFORMANCE COMPARISON

In Table 1, the 1st group is CNN-based occluded methods and the 2nd group is our baseline and the proposed method. The proposed method achieves an improvement on occluded dataset (Occluded Duke +10.5% R1/+8.9% mAP) and holistic datasets (Market1501 +1.5% R1/+2.7% mAP, DukeMTMC-reID +2.2% R1/+3.0% mAP) compared to our baseline part-based method. When we replace ResNet with OSNet, a multiscale network, performance have an additional increase on all datasets (Occluded Duke +4.7% R1/+4.4% mAP, Market1501+0.7% R1/+1.4% mAP, DukeMTMC-reID+0.7% R1/+2.5% mAP). The computation and parameter number of our method (ResNet50 backbone) are 5.98 GFLOPS and 46.9 M, while the computation and parameter number of the baseline method PCB are 5.97 GFLOPS and 42.7 M. It is safe to say we achieve our design goal.

Table 1. Performance comparison on public datasets.

Method	Occluded Duke		Market1501		DukeMTMC	
	R1	mAP	R1	mAP	R1	mAP
PCB 4]	42.6	33.7	92.3	77.4	81.8	66.1
PGFA 12]	51.4	37.3	91.2	76.8	82.6	65.5
HOReID 10]	55.1	43.8	94.2	84.9	86.9	75.6
ISP 21]	62.8	52.3	95.3	88.6	89.6	80.0
PCB (baseline)	52.0	43.7	93.4	84.9	87.3	75.2
PAN-ResNet (ours)	62.5	52.6	94.9	87.6	89.5	78.2
PAN-OSNet (ours)	**67.2**	**57.0**	**95.6**	**89.0**	**90.2**	**80.7**

5. ABLATION STUDY

5.1 Share-weight fully connection

Compared shared and not shared in Table 2, shared fully connection has little influence on holistic datasets, but shows consistent improvement on occluded dataset in spite of backbones (ResNet50 +1.0% R1/+0.9% mAP, OSNet+4.5% R1/+3.4% mAP). So the restriction on generating local features is helpful for occluded cases, perhaps leads to more precise local feature alignment.

Table 2. Comparison of shared and not shared fully connection for local features.

Shared	Backbone	Occluded Duke		Market1501		DukeMTMC	
		R1	mAP	R1	mAP	R1	mAP
√	ResNet50	**62.5**	**52.6**	**94.9**	**87.6**	**89.5**	**78.2**
	ResNet50	61.5	51.7	94.8	87.2	89.0	78.0
√	OSNet	**67.2**	**57.0**	95.6	89.0	90.2	80.7
	OSNet	62.7	53.6	**95.8**	**89.2**	**90.8**	**80.7**

5.2 Global-local balance

We vary local-global feature ratio γ in equation (4) by changing shared FC output channels in Figure 2. There are two peaks on the curve, the first peak is at $\gamma = 1.75$, the second peak is at $\gamma = 7$. When γ increase from 0.875 to 1.75, performance increase 1.4% Rank-1 and 1.4% mAP scores, thus indicates local branch is necessary for discrimination. When γ increases from 1.75 to 7, a valley is at $\gamma = 3.5$, so there may be a competition between global and local branches. And when γ increases from 7 to 28, performance continuously drops down, which means global feature cannot be replaced by local features. Finally, we choose $\gamma = 7$ ($n_l = 256$) for our model.

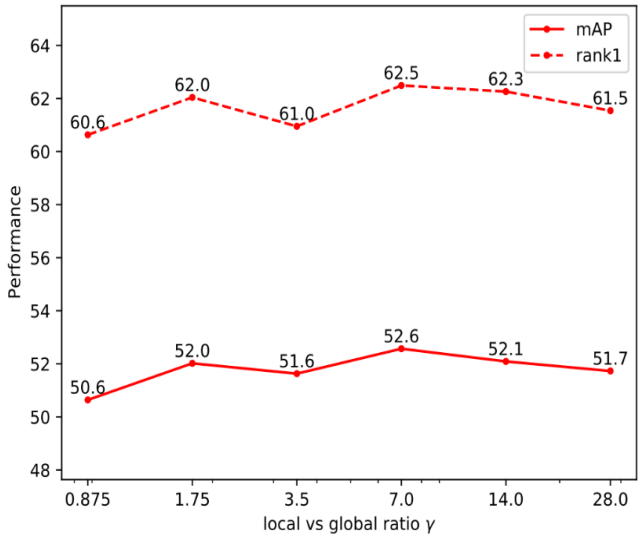


Figure 2. Global-local balance.

5.3 The usage of mask generator

In Table 3, we discuss the usage of mask generator, including using which tensor as F_g^1 to generate masks and employing masks to which tensor as F_g^2. Compared paired experiments (1, 2), (3, 4), and (5, 6), using the output of the separated buffer layer as F_g^2 can improve performance, which indicates there is a conflict between global branch and local branches, and global branch and local branches should not share the same global feature for feature extraction. And compared experiments (1, 3) and (2, 4), using the output of the separated buffer layer "Conv41" as F_g^1 performs worse. Compared paired experiments (2, 6) and (1, 5), it is better to use the final output of feature extraction module for mask generation. Finally, we choose the output of "Conv50", the final output of global branch, as F_g^1 for mask generation, and the output of "Conv51", the output of the separated buffer layer, as F_g^2 to extract local features.

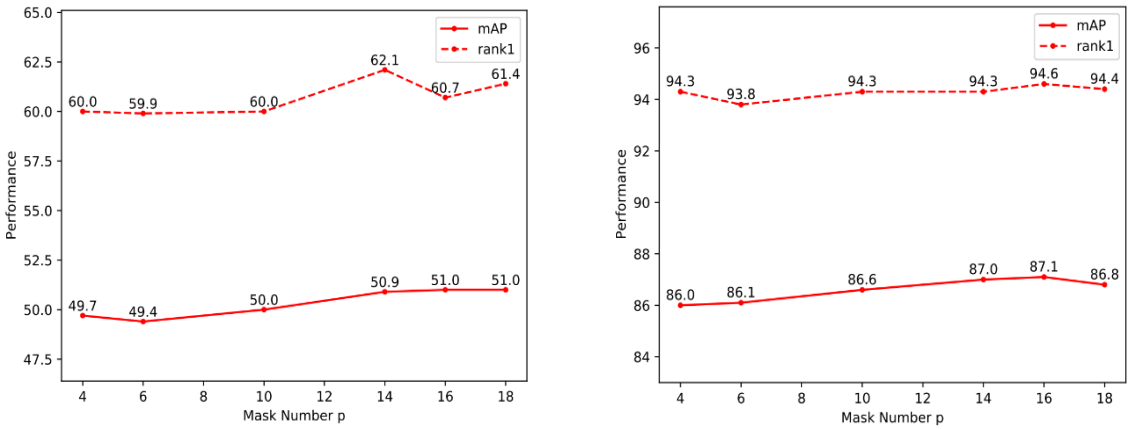
Table 3. The usage of mask generator.

No.	F_g^1	F_g^2	Occluded Duke	
			R1	mAP
1	Conv40	Conv50	60.4	50.6
2	Conv40	Conv51	61.8	51.7
3	Conv41	Conv50	59.0	49.6
4	Conv41	Conv51	59.8	50.4
5	Conv50	Conv50	61.8	51.3
6	Conv50	Conv51	62.5	52.6

Note: "Conv40" and "Conv50" are the last two ResNet blocks; "Conv41" is a copied block of "Conv40"; "Conv51" is a copied block of "Conv50".

5.4 Mask number

We further study the influence of mask number (or parts), as shown in Figure 3. In our model, best mask number is 14 on Occluded Duke, and 16 on Market1501, which means fine-grained masks is helpful not only for occluded cases but also for holistic cases. The observation is different from PAT[21] on holistic cases, because PAT employs a softmax to masks and adopts a loss to supervise the diversity of local features. However, when we employ softmax to masks, Rank-1/mAP for Occluded Duke decreases from 62.5%/52.6% to 60.6%/51.5%, thus the probability explanation (softmax operation) is not necessary for masks.



(a) (b)

Figure 3. The influence of mask number: (a) is experiments on occluded dataset Occluded Duke, (b) is experiments on holistic dataset Market1501.

5.5 Weighted pooling

In Table 4, we compare commonly used weighted pooling (use softmax to provide probability explanation to masks) and the proposed weighted pooling without softmax. Our proposed method shows consistent improvement on most datasets, especially on Occluded Duke (Rank-1+1.9%/mAP+1.1%). The proposed weighted pooling without softmax improves the gradient transfer, so performance commonly increase. For occlusion cases, our opinion is that without softmax, different local features freely combine to discriminative features (Figure 4), thus improve occlusion performance.

Table 4. The comparison of weighted pooling.

Softmax	Market1501		DukeMTMC		Occluded Duke	
	R1	mAP	R1	mAP	R1	mAP
w/	94.6	87.3	89.1	**79.2**	60.6	51.5
w/o (ours)	**94.7**	**87.7**	**89.5**	78.4	**62.5**	**52.6**

5.6 Visualization of masks

12 different persons' masks are shown on the visualization of Figure 4. The 1st row is original images; the 2nd row is the fusion of all masks, they are all concerned on bodies; the rows from 4th to 6th are the 1st to 3rd masks, they are combination of different parts, as mentioned on Section 5.5. Visualization of masks shows that our method achieves good performance for concerning on human bodies, but mis-alignment also exists (e.g., the 1st mask of the last person), this might be caused by the lack of restriction of masks.

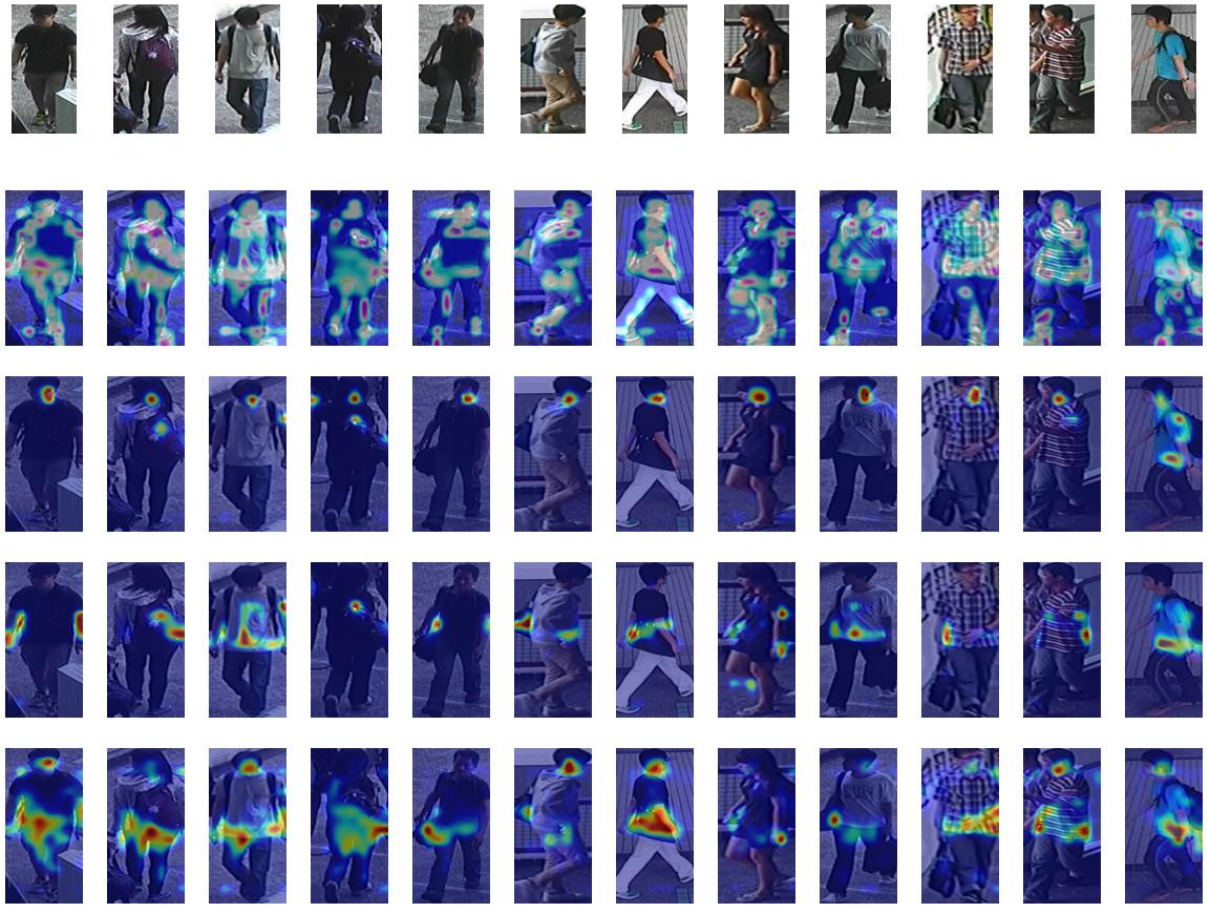


Figure 4. Visualization of masks. The 1st row is original images, the 2nd row is fusion masks of all masks, the 3rd, 4th, and 5th row are masks 1, 2, 3 of each person.

6. CONCLUSION

In this work, we propose a simple but efficient architecture for occluded person re-identification with a weakly supervised mask generator and a share-weight fully connection layer, and extensive experiments show the effectiveness

of our architecture. We also discover that a conflict of global and local branches, and a separated buffer layer is helpful to fix the conflict.

ACKNOWLEDGMENTS

This work was supported in part by the Science and Technology Program of Guangdong Province under Grant 2021B1101270007.

REFERENCES

[1] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q., "Scalable person re-identification: A benchmark," 2015 IEEE Inter. Conf. on Computer Vision (ICCV), 1116-1124 (2015).
[2] Ristani, E., Solera, F., Zou, R., Cucchiara, R. and Tomasi, C., "Performance measures and a data set for multi-target, multi-camera tracking," Computer Vision—ECCV 2016 Work., 17-35 (2016).
[3] Li, W., Zhao, R., Xiao, T. and Wang, X., "DeepReID: Deep filter pairing neural network for person re-identification," 2014 IEEE Conf. on Computer Vision and Pattern Recognition, 152-159 (2014).
[4] Sun, Y., Zheng, L., Yang, Y., Tian, Q. and Wang, S., "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," Computer Vision—ECCV 2018, 501-518 (2018).
[5] Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C. and Sun, J., "AlignedReID: Surpassing human-level performance in person re-identification," IEEE Conf. on Computer Vision and Pattern Recognition, (2017).
[6] Luo, H., Gu, Y., Liao, X., Lai, S. and Jiang, W., "Bag of tricks and a strong baseline for deep person re-identification," 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Work. (CVPRW), 1487-1495 (2019).
[7] He, L., Liang, J., Li, H. and Sun, Z., "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 7073-7082 (2018).
[8] He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z. and Feng, J., "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," 2019 IEEE/CVF Inter. Conf. on Computer Vision (ICCV), (2019).
[9] Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S. and Sun, J., "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 393-402 (2019).
[10] Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E. and Sun, J., "High-order information matters: Learning relation and topology for occluded person re-identification," 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 6448-6457 (2020).
[11] Zhuo, J., Chen, Z., Lai, J. and Wang, G., "Occluded person re-identification," 2018 IEEE Inter. Conf. on Multimedia and Expo (ICME), 1-6 (2018).
[12] Miao, J., Wu, Y., Liu, P., Ding, Y. and Yang, Y., "Pose-guided feature alignment for occluded person re-identification," 2019 IEEE/CVF Inter. Conf. on Computer Vision (ICCV), 542-551 (2019).
[13] Fan, X., Luo, H., Zhang, X., He, L., Zhang, C. and Jiang, W., "SCPNet: Spatial-channel parallelism network for joint holistic and partial person re-identification," Computer Vision—ACCV 2018, 19-34 (2019).
[14] Qi, L., Huo, J., Wang, L., Shi, Y. and Gao, Y., "MaskReID: A mask based deep ranking neural network for person re-identification," (2019). preprint arxiv/1804.03864
[15] Wang, G., Chen, X., Gao, J., Zhou, X. and Ge, S., "Self-guided body part alignment with relation transformers for occluded person re-identification," IEEE Signal Processing Letters, 28, 1155-1159 (2021).
[16] Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y. and Wu, F., "Diverse part discovery: Occluded person re-identification with part-aware transformer," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2898-2907 (2021).
[17] He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition," 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 770-778 (2016).

[18] Zhou, K., Yang, Y., Cavallaro, A. and Xiang, T., "Omni-scale feature learning for person re-identification," 2019 IEEE/CVF Inter. Conf. on Computer Vision (ICCV), 3701-3711 (2019).
[19] Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y., "Random erasing data augmentation," (2017). preprint arxiv/1708.04896
[20] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., "Rethinking the inception architecture for computer vision," 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2818-2826 (2016).
[21] Zhu, K., Guo, H., Liu, Z., Tang, M. and Wang, J., "Identity-guided human semantic parsing for person re-identification," Computer Vision—ECCV 2020, 346-363 (2020).