Local-Global Extraction Unit for Person Re-identification

Peng Wang[1], Chunmei Qing[1(✉)], Xiangmin Xu[1], Bolun Cai[1], Jianxiu Jin[1],
and Jinchang Ren[2]

[1] School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, China
winper001@gmail.com, {qchm,xmxu,jxjin}@scut.edu.cn, caibolun@gmail.com
[2] Department of Electronic and Electrical Engineering, University of Strathclyde,
Glasgow, UK
jinchang.ren@strath.ac.uk

Abstract. The huge variance of human pose and inaccurate detection significantly increase the difficulty of person re-identification. Existing deep learning methods mostly focus on extraction of global feature and local feature, or combine them to learn a discriminative pedestrian descriptor. However, rare traditional methods have been exploited the association of the local and global features in convolutional neural networks (CNNs), and some important part-wise information is not captured sufficiently when training. In this paper, we propose a novel architecture called **Local-Global Extraction Unit (LGEU)**, which is able to adaptively re-calibrate part-wise information with integrating the channel-wise information. Extensive experiments on Market-1501, CUHK03, and DukeMTMC-reID datasets achieve competitive results with the state-of-the-art methods. On Market-1501, for instance, LGEU achieves 91.8% rank-1 accuracy and especially 88.0% mAP.

Keywords: Person re-identification · Local-Global Extraction Unit
Convolutional neural networks

1 Introduction

Person re-identification (ReID), matching two specified person images crossing non-overlapping camera views, has been receiving rapid attention in recent years. However, ReID remains a challenging problem due to following factors: human pose changes, background cluster, and local occlusion [24].

In the gallery, we aim at searching for images containing the same person in a cross-camera mode. To address this problem, two crucial facts must be considered. First, discriminative features are required to represent both the query

C. Qing—This work is supported by the National Natural Science Foundation of China (61401163, 61702192).

J. Ren et al. (Eds.): BICS 2018, LNAI 10989, pp. 402–411, 2018.
https://doi.org/10.1007/978-3-030-00563-4_39

and gallery images. Second, suitable distance metrics are inevitable to determine whether a gallery image contains the same person as the query image. Early ReID methods mainly focus on discriminative feature representation or robust distance measure. With the development of deep learning, CNN-based approaches concentrate on better feature representation for pedestrians, which can be roughly divided into three aspects.

- **Global descriptors** [2,3] pay more attention on global information, such as gender, stature, and body shape. However, they lose the explicit information or crucial details because of pose variations and person detection errors [9, 21,22,30].
- **Local descriptors** [1,20,23] directly divide the whole image into some fixed patches, and roughly feed them into the model. Therefore, they omit the fact that fixed-length strips are sensitive to the pose variance.
- **Global and local representation** [5,24] are combined to form a fusion descriptor, and achieve satisfactory performance. In [29], the global features is extracted for jointly learning local features. Cheng *et al.* [18] simply splits the convolutional maps into several parts, and fuses the local features with the global features. However, these method require more computation and additional storage space.

In this paper, we propose a novel architectural unit called **Local-Global Extraction Unit (LGEU)**, which exploits the selection of part-wise and channel-wise information to optimize global feature discriminative capabilities. LGEU embedded in backbone network replaces the original global pooling layer, which expresses the importance of each part and each channel for pedestrian images. Then, we fuse both part-wise and channel-wise information to get a global descriptor as the final feature. Largely different from existing methods, LGEU is not simply combines local and global information, but regards the local features as complementary information that benefits global features.

2 Local-Global Extraction Unit

During the CNN training for ReID, we expect the convolutional layer to learn integrated information by fusing part-wise and channel-wise information. Inspired by [7], we propose a novel unit called LGEU to address this problem. LGEU can do work on any networks, and we employ the ResNet50 as the backbone network in this paper. As shown in Fig. 1, the proposed architecture actually includes two part: local selection and global fusion.

2.1 Local Selection

Differently from the previous works, the proposed architecture embeds local selection to directly train the network as final descriptor. Local selection pays more attention to part-wise and channel-wise information. In some ways, local

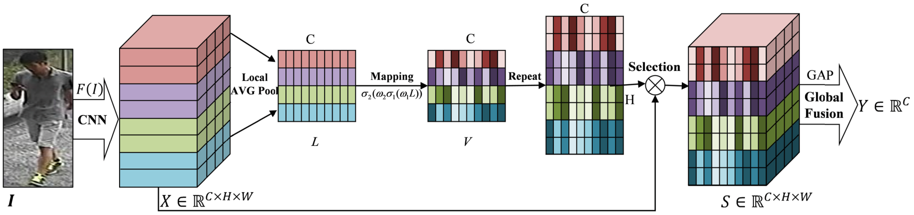
I
F(I)
CNN
$X \in \mathbb{R}^{C\times H\times W}$
Local
AVG Pool
C
L
Mapping
$\sigma_2(\omega_2\sigma_1(\omega_1 L))$
C
V
Repeat
C
H
Selection
$S \in \mathbb{R}^{C\times H\times W}$
GAP
Global
Fusion
$Y \in \mathbb{R}^C$

Fig. 1. The architecture of LGEU. We feed the raw image I into the backbone network (ResNet50) $\mathcal{F}$, and define the output $X \in \mathbb{R}^{C\times H\times W}$ of the last convolutional layer as $X = \mathcal{F}(I)$. For the sake of understanding, the LGEU can be shallowly represented as $f : X \rightarrow Y$, where $Y \in \mathbb{R}^C$ is integration of part-wise and channel-wise information.

selection works like activation functions [6,7] for feature re-weighting, which would be discussed detailedly in Sect. 2.3.

Along the vertical direction, a pedestrian body can be divided into different parts, such as head, upper clothes, lower clothes, and shoes. As is shown in Fig. 1, we integrate the part-wise feature $L \in \mathbb{R}^{C\times K}$ by local average pooling with $(H/K) \times W$ receptive fields. In this paper, the input feature $X \in \mathbb{R}^{C\times H\times W}$ is divided into $K = 4$ pieces, and each element is calculated by

$$L^{c,k} = \frac{1}{(WH)/K} \sum_{i=(k-1)\Delta+1}^{k\Delta} \sum_{j=1}^{W} X^{c,i,j}, \tag{1}$$

where $\Delta = H/K$.

Then we take these part-wise features L into K branches, separately. Each branch is sequentially composed by weight, ReLU, weight, and Sigmoid function. The non-linear mapping aims to learn the importance feature of each part. We concatenate them to single tensor $V \in \mathbb{R}^{C\times K}$, whose elements can be expressed by

$$V^{c,k} = \sigma_2(\omega_2^k \sigma_1(\omega_1^k L^{c,k})), \tag{2}$$

where σ_1, σ_2 refer to the ReLU function and sigmoid function, respectively. The size of ω_1^k is $R \times C$, and ω_2^k is $C \times R$. Here R is the reduction dimension and set to 16 in the paper.

If we let $K = 1$, it means that the whole pedestrian discriminator is fed into the unit. As pointed in Sect. 2.3, the block is a little similar to SE unit. However, LGEU has many differences in mechanism or results. The results of LGEU on these public datasets would show its powerful ability, which is explained in Sect. 3.3.

Finally, we select the channel-wise information in each part-wise branch. Each element of V along the vertical direction can be repeated by H/K times to have the same height as X. The result of local selection can be written as follow:

$$S^{c,i,j} = X^{c,i,j} \otimes V^{c,\lceil i/\Delta \rceil}, \tag{3}$$

where $\otimes$ denotes point-wise multiplication, and $\lceil\rceil$ means the operation of rounded up.

2.2 Global Fusion

As described above, local selection takes account of both part-wise and channel-wise information to re-weight the input X. Similarly to many CNN-based frameworks, we straightforward embed global average pooling for feature fusion, which can be expressed by

$$Y^c = \frac{1}{WH}\sum_{i=1}^{H}\sum_{j=1}^{W} S^{c,i,j}, \tag{4}$$

where Y^c is the mean value responsing to each feature map. Indeed, the global feature contains more high-level information, such as shape and pose, and it benefits to avoid overfitting on small training sets. Then a Softmax loss function is appended for global feature learning.

2.3 Comparison

In this section, we will discuss the mechanism of LGEU, and compare it with ReLU [6] and SE unit [7].

ReLU. ReLU [6] sparsely selects which element in the feature maps from previous layer can be fed into the next layer. As illustrated in Fig. 2, ReLU, as a special form of LGEU, can be regarded as a point-wise multiplication of an linear mapping and a gate function. LGEU is capable of learning how these patches and channels are selected, while ReLU cannot do so. Moreover, the derivative of ReLU switch function is zero, which causes that the training error cannot be back-propagated though the switch during the training process. LGEU uses Sigmoid as gate function, which is able to optimize the whole selection control as training error can be back-propagated through itself.

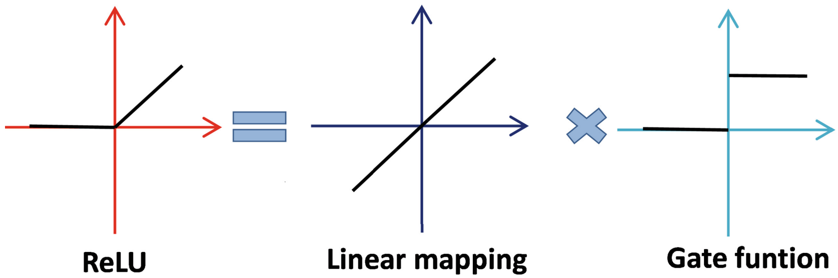

Fig. 2. Mechanism of ReLU. ReLU can be re-defined as a point-wise multiplication of an identity mapping and a switch.

SE Unit. In some ways, LGEU could be compared with SE unit [7] in architecture. However, they have many differences, especially in function. SE unit fuses the global spatial information into channel descriptor, aiming to fully capture channel-wise dependencies. While LGEU not justly consider channel-wise information, but lay more emphasis on part-wise. In addition, SE squeezes the each feature map into a scalar, which is used for re-weighting the channel, but we do not do so. In short, LGEU focus on more information than SE unit when training the network.

3 Experiments

3.1 Datasets and Evaluation Protocol

Datasets. We conduct the proposed method on three widely used datasets: Martket-1501 [8], CUHK03 [4], and DukeMTMC-reID [10,11]. **Market-1501** dataset contains 1501 identities observed by 6 cameras, 19732 gallery images, and 12936 training images. We split the dataset with standard protocol: 751 identities for training and 750 identities for testing. **CUHK03** dataset includes 1467 identities and 13164 images, and offers both hand-labeled and DPM-detected bounding boxes. In this work, we use latter, and employ the new training/testing protocol introduced in [12]. **DukeMTMC-reID** dataset, which contains 1404 identities (702/702 for training and testing), 2228 queries, 17661 gallery images, and 16552 training images, is captured with 8 cameras. The single-query setting is used in all our experiment.

Evaluation Protocol. For the distance measure, we compute the cosine distance between a query and a gallery. Our experiments adopt the cumulative matching characteristic (CMC) and mean average precision (mAP) as performance measure. Moreover, we also employ re-ranking approach [12] to improve ReID accuracy, which combines original distance and Jaccard distance.

3.2 Implementation Details

We implemented the proposed method in *Pytorch* package. For model learning, we use the ResNet50 fine-tuned from ImageNet as our baseline. The model is trained for 60 epochs with SGD optimizer, whose starting learning rate initialized at 0.1 and decayed to 0.01 after 40 epochs. We set the size of training inputs to be 256×128, and set the batchsize to 32. All the images are random horizontal flipped and cropped for data augmentation.

3.3 Performance Evaluation

The Effectiveness of LGEU. As is shown in Fig. 3, the backbone network inserted LGEU achieves a great promotion in all these datasets. Without re-ranking approach in test, mAP on three datasets increases from 69.8%, 41.6%, 57.4% to **74.5% (+4.7%), 47.4% (+5.8%), 62.3% (+4.9%)**, respectively. And rank-1 increases from 86.5%, 46.5%, 73.9% to **90.0% (+3.5%), 50.7% (+4.2%), 77.4% (+3.5%)**, respectively. On the other hand, LGEU is more superior than SE unit [7], which is shown in Table 1. The significant raise indicates that the combination of part-wise and channel-wise information enhances the discriminative ability of the network.

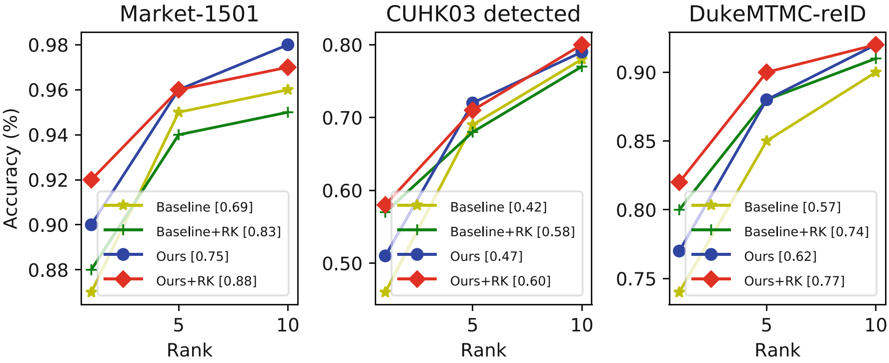

Fig. 3. The comparison of the baseline of and our methods. Note that the score in the parenthesis means the mAP corresponding to each method. RK stands for re-ranking approach.

Comparison with Other Methods. The comparisons of the proposed methods on **Market-1501** in terms of rank-1, rank-5, rank-10 matching rate are given in Table 2. Related methods are roughly divided into three groups: (1) hand-crafted methods, such as [8,13,14]. (2) CNN-based methods with employing global features [15,16]. (3) CNN-based methods with employing local features [19,23,24]. GLAD [24] gets the best rank-1 accuracy (89.9%) and mAP (73.9%), which is surpassing the other methods. However, our result is 0.1% higher than its in rank-1 accuracy, and 0.6% higher in mAP. When using re-ranking, we can obtain 91.8% rank-1 accuracy, and the mAP further surpass GLAD.

For **CUHK03-detected** dataset, we use the training/testing protocol as [12]. The comparison is summarized in Table 3. SVDNet [26] is comparatively higher than other works either rank-1 accuracy or mAP, while we get better result than it. Our method obtains 50.7% accuracy and 47.4% mAP. After using re-ranking, the rank-1 accuracy increases to 57.7% (+10%), and the mAP increases to 60.0% (+12.6%).

We also show our result for **DukeMTMC-reID**. The comparison with related methods is depicted in Table 4. Our method outperforms most the state-of-the-art approaches, but slightly lower the DPFL [1] in rank-1, which takes multi-scale images as input and calculates more complicatedly than ours. However, our method gets the best mAP (62.35%) among the all approaches. When employing the re-ranking, our performance is improved enormously from 77.40% to 81.96% in accuracy, from 62.35% to 77.17% in mAP.

Table 1. Comparison of our method with the baseline and Se-Unit, respectively. The rank-1 accuracy (%) and mAP (%) are shown. All the methods are conducted without re-ranking approach.

Methods	Datasets					
	Market-1501		CUHK03		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	86.5	69.8	46.5	41.6	73.9	57.4
Se-Unit	89.8	74.3	47.6	43.9	74.1	57.5
Ours	**90.0**	**74.5**	**50.7**	**47.4**	**77.4**	**62.3**

Table 2. Comparison of our method with the state of the art methods on Market-1501. The rank-1, rank-5, rank-10 accuracy (%) and mAP (%) are shown.

Methods	R-1	R-5	R-10	mAP
Bow+kissme [8]	44.4	63.5	72.2	20.8
WARCA [13]	45.2	68.1	76.0	-
KLFDA [14]	46.5	71.1	79.9	-
SOMAnet [15]	73.9	-	-	47.9
SVDNet [16]	82.3	92.3	95.2	62.1
PAN [17]	82.8	-	-	63.4
PAR [19]	81.0	92.0	94.7	63.7
MultiLoss [20]	83.9	-	-	64.4
PartLoss [23]	88.2	-	-	69.3
DPFL [1]	88.9	-	-	73.1
GLAD [24]	89.9	-	-	79.3
Ours	90.0	**96.2**	**97.2**	74.5
Ours+RK	**91.8**	95.7	96.9	**88.0**

Table 3. Comparison with state of the art methods on CUHK03-detected. rank-1 accuracy (%) and mAP (%) are shown.

Methods	R-1	R-5	R-10	mAP
Bow+kissme [8]	6.4	-	-	6.4
LOMO+XQDA [25]	12.8	-	-	11.5
SVDNet [16]	41.5	-	-	37.5
PAN [17]	36.3	-	-	34.0
DPFL [1]	40.7	-	-	37.0
SVDNet+Era [26]	48.7	-	-	43.7
Ours	50.7	71.6	78.9	47.4
Ours+RK	**57.7**	**70.8**	**79.8**	**60.0**

Table 4. The result of DukeMTMC-reID. The rank-1 accuracy (%) and mAP (%) are shown.

Methods	R-1	R-5	R-10	mAP
Bow+kissme [8]	25.13	-	-	12.17
LOMO+XQDA [25]	30.75	-	-	17.04
LSRO [10]	67.68	-	-	47.13
AttIDNet [27]	70.69	-	-	51.88
PAN [17]	71.59	-	-	51.59
ACRN [28]	72.58	-	-	51.98
SVDNet [16]	76.70	-	-	56.80
DPFL [1]	<u>79.20</u>	-	-	<u>60.60</u>
Ours	77.4	88.11	92.10	62.35
Ours+RK	**81.96**	**89.45**	**92.37**	**77.17**

4 Conclusion

To obtain discriminative features, we proposed a novel architecture named Local-Global Extraction Unit (LGEU) to exploit both part-wise and channel-wise information aiming to optimize global feature discriminative ability. LGEU is largely different from tradition methods, which either focus on local and global features, or combination of them. LGEU work more like the mechanism of activation function, such as ReLU, but it can automatically select discriminative information. Moreover, LGEU is a little bit like the SE units, which only focus on channel-wise information. While LGEU pays attention to part-based information too. Abundant experiments with state-of-the-art methods on the challenging datasets demonstrated that our method achieves favorable results in terms of rank-1 accuracy and mAP. In the next work, we would like to combine other mechanism to select crucial information for person re-identification.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (61401163, 61702192).

References

1. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2590–2600. IEEE, Hawaii (2017)
2. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 384–393. IEEE, Hawaii (2017)
3. Chen, W., Chen, X., Zhang, J., Huang, K.: A multi-task deep network for person re-identification. In: AAAI, San Francisco, vol. 1, p. 3 (2017)

4. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159. IEEE, Hawaii (2014)
5. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3980–3989. IEEE, Venice (2017)
6. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on machine learning (ICML-2010), Haifa, pp. 807–814 (2010)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, Venice (2017)
8. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person reidentification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124. IEEE, Santiago (2015)
9. Wang, Z., Ren, J., Zhang, D., Sun, M., Jiang, J.: A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing **287**, 68–83 (2018)
10. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3774–3782. IEEE, Venice (2017)
11. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
12. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Computer Vision and Pattern Recognition (CVPR), pp. 3652–3661. IEEE, Hawaii (2017)
13. Jose, C., Fleuret, F.: Scalable metric learning via weighted approximate rank component analysis. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 875–890. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_53
14. Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., Radke, R.J.: A comprehensive evaluation and benchmark for person re-identification: features, metrics, and datasets. arXiv preprint arXiv:1605.09653 (2016)
15. Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T.: Looking beyond appearances: synthetic training data for deep CNNs in re-identification. arXiv preprint arXiv:1701.03153, 2017
16. Sun, Y., Zheng, L., Deng, W., Wang, S.: SVDNet for pedestrian retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Hawaii (2017)
17. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV. IEEE, Venice (2017)
18. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344. IEEE, Nevada (2016)
19. Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice (2017)

20. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multiloss classification. In: International Joint Conference on Artificial Intelligence (2017)
21. Zaihidee, E.M., Ghazali, K.H., Ren, J., Salleh, M.Z.: A hybrid thermal-visible fusion for outdoor human detection. J. Telecommun. Electron. Comput. Eng. (JTEC) **10**(1–4), 79–83 (2018)
22. Ren, J., Jiang, J., Wang, D., Ipson, S.: Fusion of intensity and inter-component chromatic difference for effective and robust colour edge detection. IET Image Process. **4**(4), 294–301 (2010)
23. Yao, H., Zhang, S., Zhang, Y., Li, J., Tian, Q.: Deep representation learning with part loss for person re-identification. arXiv preprint arXiv:1707.00798 (2017)
24. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 420–428. ACM (2017)
25. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206. IEEE, Boston (2015)
26. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
27. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220 (2017)
28. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute complementary information. In: Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1435–1443. IEEE, Hawaii (2017)
29. Zhang, X., et al.: AlignedReID: surpassing human-level performance in person reidentification. arXiv preprint arXiv:1711.08184 (2017)
30. Yan, Y., et al.: Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cogn. Comput. **10**(1), 94–104 (2018)