Robust Pedestrian Detection for Semi-automatic Construction of a Crowded Person Re-Identification Dataset

Zengxi Huang[1,2], Zhen-Hua Feng[2(✉)], Fei Yan[2], Josef Kittler[2], and Xiao-Jun Wu[3]

[1] School of Computer and Software Engineering, Xihua University, Chengdu, China
huangzx@mail.xhu.edu.cn
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
{z.feng,f.yan,j.kittler}@surrey.ac.uk
[3] Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Machine Intelligence, Jiangnan University, Wuxi, China
wu_xiaojun@jiangnan.edu.cn

Abstract. The problem of re-identification of people in a crowd commonly arises in real application scenarios, yet it has received less attention than it deserves. To facilitate research focusing on this problem, we have embarked on constructing a new person re-identification dataset with many instances of crowded indoor and outdoor scenes. This paper proposes a two-stage robust method for pedestrian detection in a complex crowded background to provide bounding box annotations. The first stage is to generate pedestrian proposals using Faster R-CNN and locate each pedestrian using Non-maximum Suppression (NMS). Candidates in dense proposal regions are merged to identify *crowd patches*. We then apply a bottom-up human pose estimation method to detect individual pedestrians in the crowd patches. The locations of all subjects are achieved based on the bounding boxes from the two stages. The identity of the detected subjects throughout each video is then automatically annotated using multiple features and spatial-temporal clues. The experimental results on a crowded pedestrians dataset demonstrate the effectiveness and efficiency of the proposed method.

Keywords: Person Re-identification · Pedestrian detection
Faster R-CNN · Human pose estimation

1 Introduction

Person re-identification (ReID) is concerned with person tracking across non overlapping cameras using soft biometrics such as clothes, hairstyle and accessories [21]. The focus on general appearance stems from the lack of availability of

F. J. Perales and J. Kittler (Eds.): AMDO 2018, LNCS 10945, pp. 63–72, 2018.
https://doi.org/10.1007/978-3-319-94544-6_7

good quality hard biometrics in surveillance video footage due to poor resolution of face data and challenging poses, which preclude the use of face recognition technology.

In person ReID, a query image enclosing a subject is compared against a gallery of candidates. There are a number of benchmarking databases used for the development of ReID algorithms, but they share the same two limitations: (i) Their query and gallery images contain one subject each. This makes the training of ReID systems and their evaluation straightforward. (ii) The image data has been collected at a single point in time.

In order to address these two issues, a new ReID database of surveillance videos has been collected over an extended period of time, where the subjects were free to change their apparel. They were also encouraged to move in groups or crowded areas, creating occlusion and overlap. The ultimate aim is to release this database with ground truth annotation to the computer vision community. In the ground truth annotation task, the key initial challenge is to perform pedestrian detection, which can cope with occlusions exhibited in crowd scenes.

Compared to the traditional pedestrian detectors that rely on hand-crafted features, such as Aggregate Channel Features (ACF) [2] and Locally Decorrelated Channel Features (LDCF) [10], recently deep models, such as Faster Region-based Convolutional Neural Network (Faster R-CNN) [12,17,18], have become the defacto standard detector approaches. However, the crowd occlusion, remains a significant challenge in pedestrian detection. Wang et al. [15] argued that 26.4% of all pedestrians in CityPersons [18] have a considerable overlap with another pedestrian, severely harming the performance of pedestrian detectors.

Some effort in deep model research has been directed towards addressing the crowd occlusion issue, including Hosang et al. [7] and Wang et al. [15]. However, more promising appear to be part-based models. Ouyang et al. [11] introduced a deep model for learning the visibility relationship among overlapping parts at multiple layers, which can be viewed as a general post-processing of part-detection results. Tian et al. [14] proposed a DeepParts framework that consists of lots of complementary part-based models selected in a data driven manner.

In this paper, we present a two-stage pedestrian detection method that is robust to crowd occlusion. It employs a human body model composed of joints and important components, called keypoints, as local parts to facilitate global body detection. In the first stage, we use the Faster R-CNN to create pedestrian proposals. In sparse proposal regions the subjects are located using NMS as a part of our detection result. In the dense proposal regions, proposals are merged to create crowd patches. In the second stage, we apply a bottom-up human articulated pose estimation method using a part-based model to detect individual pedestrians in the crowd patches. Using the resultant accurate bounding boxes, detected frame by frame, we introduce a simple but effective method to associate them and then crop the track of sequential samples belonging to each specific person. In this method, Histogram of Oriented Gradients (HOG) and colour histogram are integrated at score level for person identity classification. Experimental results are presented to demonstrate the effectiveness and efficiency of the proposed approach.

Table 1. A statistical summary of widely-used person ReID benchmarks.

Dataset	Release time	Num. of identities	Num. of cameras	Num. of images	Labelling method	Frames available
VIPeR [5]	2007	632	2	1264	Hand	No
GRID [9]	2009	1025	8	1275	Hand	Yes
PRID2011 [6]	2011	934	2	24541	Hand	No
CUHK03 [8]	2014	1467	10	13164	Hand/DPM	No
Market1501 [20]	2015	1501	6	32217	Hand/DPM	No
MARS [19]	2016	1261	6	1191003	DPM+GMMCP	No
DukeMTMC-reID [22]	2017	1812	8	36441	Hand	Yes
DukeMTMC4ReID [4]	2017	1852	8	46261	Doppia	Yes

The paper is organized as follows. In Sect. 2, the new ReID dataset facilitating research in ReID over an extended period of time is introduced. The pedestrian detection method developed to meet the required specification is presented in Sect. 3. The automatic data association technique is employed to link individual frame detections in Sect. 4. The pedestrian detection algorithm is evaluated in Sect. 5. The conclusions are drawn in Sect. 6.

2 The JNU Dataset

Table 1 provides a statistical summary of widely-used person ReID benchmarks, including their scales and annotating methods. We can see from the table that over the last decade ReID benchmarks have grown significantly in size. However, to date none of them has touched the two issues raised in the introduction.

The Jiangnan University (JNU) dataset collection is an on-going project conducted by the Jiangnan University and University of Surrey. The aim of the project is to create a new large-scale benchmark with temporal characteristics. The highlights of the JNU dataset can be summarised as follows: (1) most videos were captured in crowd scenarios; and (2) each subject had been captured at least twice with at least one day interval to allow apparel changes. In such a recording script, there is a high probability that a volunteer may change his/her cloths, which introduces new challenges to the person ReID problem. Note that we suggested the volunteers to dress as usual and did not force them to change their clothing when they came back. The new dataset can be used as a benchmark for both pedestrian detection and person ReID. The aim is to encourage and facilitate research in this subject area that considers apparel variations.

All the videos were captured at four different scenes in the Jiangnan University, two outdoor and two indoor, using four different video recording devices including a controlled Canon SLR camera (1080p), a Sony hand-held video recorder (1080p) and two mobile phones (720p). Some example frames of the videos captured under different camera settings are shown in Fig. 1. 274 subjects participated in the first data collection session recorded in the summer.

Fig. 1. Some example frames of different camera settings of the JNU dataset.

The final target is to include more than 1000 subjects. Other sessions will be collected in different seasons. More information will be given along with the release of JNU dataset.

3 Automatic Pedestrian Detection

NMS is generally used as a post-processing step in deep model-based object detection frameworks. Given a predetermined IoU, NMS selects high scoring proposals and deletes close-by less confident neighbours assuming they are likely to cover the same object. Its choice is often problematic and leads to detection errors in crowd scenes. In a crowd scene, pedestrians with similar appearance features yield many overlapping proposals. We summarise the issues caused by crowd occlusion into three categories: miss detection, false alarm, and localisation error. A too low IoU threshold is likely to cause miss detection, while an excessively high one may lead to false alarm. Both [7] and [15] were motivated by the NMS' vulnerability to crowd occlusion and made efforts to alleviate it.

Localisation error refers to an imprecise bounding box that contains excessive background or misses some important body parts. But, such a bounding box is still counted as correct according to the commonly used criteria such as 0.5 or 0.75 IoU. There is no doubt that such localisation errors degrade person ReID accuracy. Localisation error is of particular concern in pedestrian annotation for benchmarking datasets. In Fig. 2, the first row of experimental results shows that either a too low or too high IoU threshold of NMS is likely to introduce localisation error in a crowd, while the second row tells that it is inevitable to avoid miss detection and localisation error by tuning the IoU threshold. Therefore, we argue that the post-processing algorithms like NMS can not handle sufficiently with the problem of pedestrian detection in complex crowd scenes.

In contrast to global pedestrian detection algorithms, human articulated pose estimation is a fine-grained human body perception task. Human pose estimation is defined as the localisation of human joints or keypoints on the arms, legs, torso,

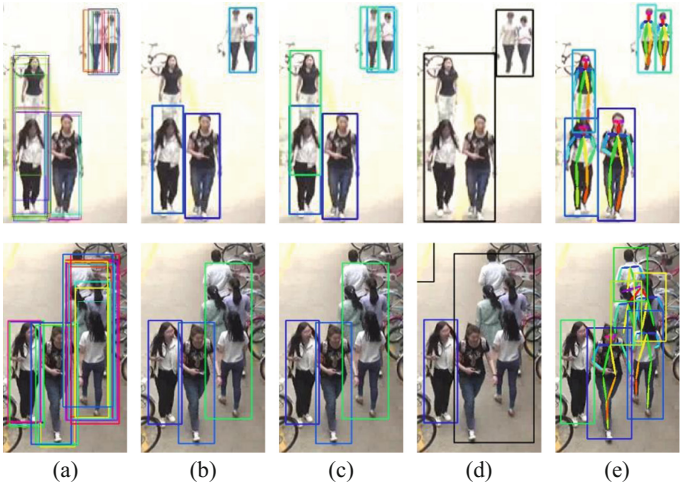

Fig. 2. Three visualization examples of detection error caused in crowd by Faster R-CNN, and our results. (a) Pedestrian proposals; (b) NMS with 0.2 IoU threshold; (c) NMS with 0.5 IoU threshold; (d) Crowd patches and single pedestrian; (e) Our detection results.

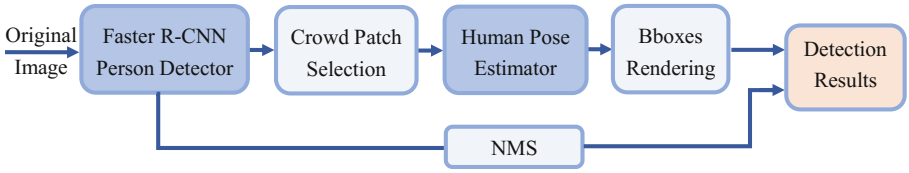

Fig. 3. Overview of the proposed pedestrian detection framework.

and face. A bottom-up human pose estimation approach firstly detects keypoints and then groups them to form a person instance.

Recently, the human keypoint detection approach using Convolutional Pose Machines (CPMs) was shown to exhibit encouraging performance for robust detection in crowd scenes [16]. OpenPose introduces a very effective and efficient parsing method to connect the keypoints of multiple persons by learning part affinity fields [1]. Intuitively, human pose estimation methods can be directly used for pedestrian detection with decent robustness to crowd occlusion. However, compared with Faster R-CNN, they have some evident disadvantages, including being computationally demanding, inferior keypoint detection in low resolution subjects, and tendency to generate false keypoint groups in a complex background clutter. Motivated by these observations, we propose to integrate Faster R-CNN and human pose estimator to construct a robust and efficient pedestrian detection method.

As shown in Fig. 3, the proposed method is a two-stage method. In the first stage, we use Faster R-CNN to generate pedestrian proposals and locate each pedestrian using NMS. Meanwhile, candidates in dense proposal regions are merged to identify crowd patches. In the second stage, a bottom-up human pose estimator is applied to detect the human keypoints in the crowd patches and cluster them into groups. Based on the keypoint groups, we heuristically infer the bounding box for each individual. The bounding boxes derived inside and outside a crowd from both stages constitute our detection result.

In the remaining part of this section, we briefly introduce the crowd patch selection and bounding box rendering modules in the proposed pipeline.

Crowd Patch Selection. We use a very low confidence score threshold for Faster R-CNN to retain as many pedestrian proposals as possible in order to avoid missing any subjects. Following the R-CNN detection step, we select the locations of pedestrians based on NMS using a relatively high IoU threshold, referred to as NMS bounding boxes. Simultaneously, the R-CNN proposals are simply merged using a very low IoU threshold, circa 0.15, to create larger bounding boxes. We term these big bounding boxes as crowd patches if one contains two or more NMS bounding boxes. Only the crowd patches are fed into the human pose estimator for detection speed. As shown in Fig. 2(d), crowd patches are outlined by black bold boxes, while a single NMS bounding box is marked by a blue box.

Bounding Box Rendering. The human pose estimator detects human keypoints and parses their affinity relations, shown by keypoint connections in Fig. 2. Some keypoints of an individual may not be found owing to the crowd occlusion or low image resolution. In such cases, we infer the actual pedestrian bounding box based on the available human limbs or keypoint connections. The six limbs like neck-left shoulder, neck-right shoulder, neck-left hip, neck-right hip, left shoulder-left elbow, and right shoulder-right elbow are considered most important in our bounding box inference. It should be noted that in Fig. 2 we do not infer the lower body if no relevant keypoints are found. An exception is made only in the case of annotations required for the detection evaluation in Sect. 5.

4 Automatic Data Association

Without obvious crowd occlusion, we can easily associate all the detected persons to a unique identity based on spatial-temporal information throughout the entire video, using the bounding boxes obtained in each frame by the proposed pedestrian detection algorithm. However, in our videos, more than half of the volunteers were occluded with others more or less. Furthermore, to provide sequential samples for each subject, we opted to keep as many bounding boxes as possible. This policy makes our approach inevitably introduce many false detections and distracts the subsequent data association. To address this issue, we introduce a simple but effective bounding box association method. This method to associate and crop consecutive samples for each subject is based on similar principles to

those used in object tracking. We integrate HOG and colour histogram features at score level for person identity classification.

We denote K persons in a video as $P_k(\mathbf{b}, s, \mathbf{f}^h, \mathbf{f}^c, n, d)_{k=1}^K$, where $\mathbf{b} = [x, y, w, h]^T$ is the bounding box of a pedestrian, s is the confidence score, $\mathbf{f}^h$ and $\mathbf{f}^c$ are the feature templates of HOG features and colour histograms. n is the index of the last frame where P_k is found. d stands for the duration of P_k in the video, which records all occurrences but will receive a specified penalty once P_k is lost. We use $Q(\mathbf{b}, s, \mathbf{f}^h, \mathbf{f}^c)$ to represent the subjects whose bounding box is detected in the next frame. The proposed data association method is briefly summarised as follows:

1. Find the bounding boxes that mutually overlap in the next frame. For simplicity, we assume that there are two subjects in a crowd, denoted by $Q_u(\mathbf{b}, s, \mathbf{f}^h, \mathbf{f}^c)$ and $Q_v(\mathbf{b}, s, \mathbf{f}^h, \mathbf{f}^c)$.
2. Find the person whose bounding box $\mathbf{b}_k$ has a certain IoU with $\mathbf{b}_u$ or $\mathbf{b}_v$.
3. Assume that the two persons P_i and P_j are matched and calculate their feature similarities with Q_u and Q_v, *i.e.* $S_{i,u}^h(\mathbf{f}_i^h, \mathbf{f}_u^h)$ and $S_{i,u}^c(\mathbf{f}_i^c, \mathbf{f}_u^c)$ for P_i and Q_u.
4. Normalise similarity scores of the two types into common domain, then combine them with a weighted-sum rule, like $S_{i,u} = \lambda \cdot S_{i,u}^h(\mathbf{f}_i^h, \mathbf{f}_u^h) + (1 - \lambda) \cdot S_{i,u}^c(\mathbf{f}_i^c, \mathbf{f}_u^c)$ for P_i and Q_u.
5. Based on the fused similarities, label Q_u and Q_v with the identities of P_i or P_j, otherwise, produce a new identity with Q's information.
6. If P_k matches a Q_k subject and its d is higher than a predetermined threshold, update its information template and crop an image sample within the bounding box of Q_k for this identity. If it gets lost continuously for a certain number of frames or if its d is below a certain threshold, we terminate its tracking to avoid its distraction in the subsequent data association. If not sure, we assign multiple identities to a person, rather than mistakenly assigning the same identity to different persons. The latter strategy requires much more manual effort in the final manual correction step.

5 Evaluation

To evaluate the performance of the proposed approach, we selected a subset with crowd scenes from the first 20 K frames of training/validation set of camera 8 in DukeMTMC [13]. To be more specific, we selected the frames that contain at least one ground truth bounding box with 40% crowd occlusion but without any bounding box with crowd occlusion over 70%. The selected subset has 2529 images with 11837 bounding boxes.

In our method, one major aim of the use of a global pedestrian detection algorithm, *i.e.* Faster R-CNN, is to keep as many pedestrian proposals as possible for crowd patch selection. The subject size in DukeMTMC is generally rather substantial and the resolution requirement of crowd patch fed into the human pose estimator can be guaranteed. Therefore, in this experiment, we directly used

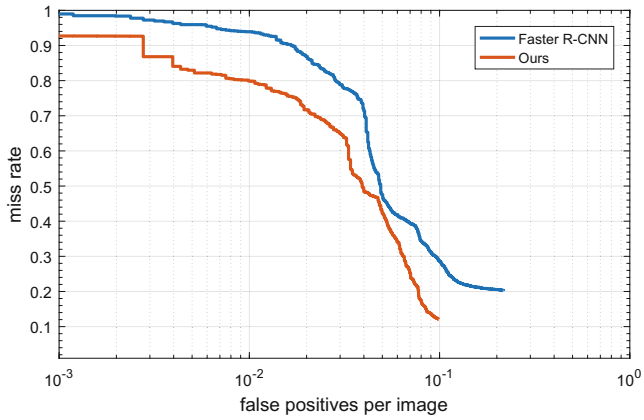

Fig. 4. Comparison of detection performance on a crowd subset of DukeMTMC.

Table 2. Comparison of runtime performance.

	Faster R-CNN	OpenPose	Our method
Average (sec per image)	0.951	9.215	2.293
Standard deviation	0.009	0.783	0.62

the off-the-shelf Faster R-CNN object detector [12] and OpenPose [1] human pose estimator as the global and part-based pedestrian detectors.

We compared our pedestrian detection method with the state-of-the-art Faster R-CNN method, following the same evaluation protocol as proposed in [3]. We plotted the miss detection rate against False Positives Per Image (FPPI) with a fixed 0.5 IoU threshold in Fig. 4. Our method always achieves lower miss rate than Faster R-CNN. Faster R-CNN gets its lowest miss rate of 20.25% at 21.72% FPPI, while our method achieves 12.1% at 9.89% FPPI. The log-average miss rate (MR) is also commonly used to summarise the detector performance, computed by averaging miss rate at nine FPPI rate points, evenly spaced in log-space in the range 10^{-2} to 10^{0}. Our method obtains 24.86% MR, while Faster R-CNN exhibits a much higher one of 37.24%.

According to [3], when miss detection rates cover the full range of FPPI, log-average miss rate is similar to the performance at 10^{-1} FPPI. As we do not have all the data needed for averaging, in our experiment, the miss rate at 10^{-1} FPPI is a more appropriate single metric to measure the detector performance. This rate is 28.40% and 12.1% respectively for Faster R-CNN and our method.

We compare the runtime of Faster R-CNN [12], OpenPose [1] and our method on a laptop with a Intel i7-7700HQ CPU and NVIDIA GTX-1070 GPU. The software was implemented with Python 3 and Tensorflow 1.4. The image resolution is 1920×1080 for Faster R-CNN and our method, while it is 368×654

for OpenPose. As shown in Table 2, Faster R-CNN processes an image in circa 0.95s, while OpenPose takes about 10 times longer at much lower image resolution. Compared to them, the runtime of our method is 2.293s on average, which fluctuates drastically with the varying number of crowd patches and pedestrians.

6 Conclusion

In this paper, we argued that the problem of pedestrian detection and ReID in crowd scenes has been somewhat neglected. We described a new ReID dataset, collected at the Jiangnan University, with many instances of crowd scenes designed to fill this gap. The main purpose of the paper was to address the problem of ground truth annotation of the new dataset and to develop the key steps of the annotation process. In particular, we proposed a pedestrian detection method robust to crowd occlusion, and an automatic data association technique to link individual frame detections to extended tracks.

Our proposed pedestrian detection method combines the advantages of a fast global pedestrian detection method, and a computationally demanding, but accurate, bottom-up part-based human detection procedure. We first apply the Faster R-CNN as a global model to generate pedestrian proposals. In sparse proposal regions the subjects are located using NMS. In the dense proposal regions, proposals are merged to create crowd patches. An advanced, part-based model human pose estimator is subsequently engaged to detect individual pedestrians in the crowd patches. The bounding boxes derived inside and outside a crowd from both stages constitute our final detection result. Our experimental results on a representative subset of crowded scenes from DukeMTMC dataset demonstrate that our method is superior to Faster R-CNN, and strikes a balance in efficiency between the applied global and part-based methods.

Acknowledgements. This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1, EPSRC/dstl/MURI project EP/R018456/1, the National Natural Science Foundation of China (61373055, 61672265, 61602390, 61532009, 61571313), Chinese Ministry of Education (Z2015101), Science and Technology Department of Sichuan Province (2017RZ0009 and 2017FZ0029), Education Department of Sichuan Province (15ZB0130), the Open Research Fund from Province Key Laboratory of Xihua University (szjj2015-056) and the NVIDIA GPU Grant Program.

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Real time multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
2. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE TPAMI **36**(8), 1532–1545 (2014)
3. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE TPAMI **34**(4), 743–761 (2012)

4. Gou, M., Karanam, S., Liu, W., Camps, O., Radke, R.: Dukemtmc4reid: a large-scale multi-camera person re-identification dataset. In: CVPR Workshops (2017)
5. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: International Workshop on Performance Evaluation for Tracking and Surveillance (2007)
6. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21227-7_9
7. Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. arXiv preprint arXiv:1705.02950 (2017)
8. Li, W., Zhao, R., Xiao, T., Wang, X.: Deep filter paring neural network for person re-identification. In: CVPR (2014)
9. Loy, C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. In: CVPR (2009)
10. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: NIPS, pp. 424–432 (2014)
11. Ouyang, W., Zeng, X., Wang, X.: Partial occlusion handling in pedestrian detection with a deep model. IEEE TCSVT **26**(11), 2123–2137 (2016)
12. 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)
13. 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
14. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: ICCV, pp. 1904–1912 (2015)
15. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: detecting pedestrians in a crowd. arXiv preprint arXiv:1711.07752 (2017)
16. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
17. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 443–457. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_28
18. Zhang, S., Benenson, R., Schiele, B.: Citypersons: a diverse dataset for pedestrian detection. arXiv preprint arXiv:1702.05693 (2017)
19. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 868–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_52
20. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV (2015)
21. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
22. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV (2017)