Person re-identification in the real scene based on the deep learning

doi: https://doi.org/10.1007/s10015-021-00689-9

Noname manuscript No.
(will be inserted by the editor)

Person Re-identification in the Real-World Application Based on Deep Learning

Miaomiao Zhu · Shengrong Gong · Zhenjiang Qian · Seiichi Serikawa · Lifeng Zhang

Received: date / Accepted: date

Abstract Person re-identification aims at automatically retrieving a person of interest across multiple non-overlapping camera views. Because of increasing demand for real-world applications in intelligent video surveillance, person re-identification has become an important computer vision task and achieved high performance in recent years. However, the traditional research mainly focuses on matching cropped pedestrian images between queries and candidates on commonly used datasets and divided into two steps: pedestrian detection and person re-identification, there is still a big gap with practical applications. Under the premise of model optimization, based on the existing object detection and person re-identification, this paper achieves a one-step search of the specific pedestrians in the whole images or video sequences in the actual application scenario. The experimental results show that our method is effective in commonly used datasets and has achieved good results in real-world applications, such as finding criminals, cross-camera person tracking, and activity analysis.

Keywords Convolutional Neural Networks · Deep learning · Pedestrian detection · Person Re-identification · Real-world Application

M. Zhu · S. Serikawa · L. Zhang
Kyushu Institute of Technology, Kitakyushu, Fukuoka, Japan 804-8550
E-mail: zhu.miaomiao234@mail.kyutech.jp

S. Gong · Z. Qian
Changshu Institute of Technology, Changshu, Jiangsu, China 215500

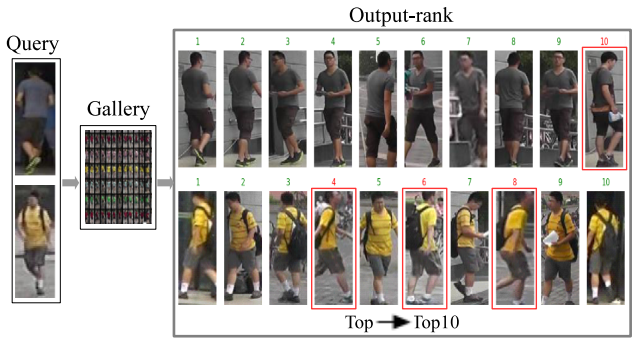
1 Introduction

Person re-identification(ReID), as an instance-level recognition problem, aims at automatically matching a target person with a gallery of pedestrian images and video sequences by multiple non-overlapping cameras, which is considered sub problem of image retrieval. Due to the increasing demand for real-world applications in intelligent video surveillance and public safety, ReID has become an important task in the field of computer vision, and has drawn a lot of attention from both academia and industry in recent years.

Depending on the improvements of deep learning and the release of many large-scale datasets, many ReID models have been proposed and have achieved high performance in the past years. For example, in 2020, FastReID [1] has achieved the best performance on Market1501 96.3% (90.3%), DukeMTMC-reID 92.4% (83.2%) and MSMT17 85.1% (65.4%) datasets at rank-1/mAP accuracy, respectively.

However, compared with face recognition, under different cameras, ReID is challenging due to the significant differences and changes of viewpoint, resolution, illumination, obstruction, pose of person, *etc.* The traditional ReID research is mainly carried out through experimental verification and evaluation on commonly used datasets, which are independent of detection and only focus on identification. In other words, the query process of ReID is divided into two steps: pedestrian detection and person re-identification.

As shown in Fig. 1(a), some sample ReID results are obtained by the method of OSNet [2] on Markket-1501. The images in the first column are the query images used to match with manually trimmed pedestrians in the second column existing dataset(gallery). The gallery set contains a large number of pedestrian



(a) Traditional ReID research, match with manually cropped pedestrians in the existing datasets.

(b) ReID in the real-world application, search a person from the whole scene images or videos.

Fig. 1 Comparison between traditional ReID research and ReID in the real-world application

images that are manually cropped in advance, while the retrieved images are sorted according to the similarity scores from left to right in the third column. The green font means correct match, the red font and the images in red rectangles indicate the false match. That is, the traditional ReID research mainly focuses on the matching between the query images and the candidate cropped pedestrian images.

However, as shown in Fig. 1(b), in the real-world scenarios, the goal of ReID is to find a target person in a gallery of images or videos which come from multiple non-overlapping cameras. Compared with traditional ReID research, its purpose is to search a person from the whole scene images or videos instead of matching them with manually cropped pedestrians in the existing dataset, which is meant for video surveillance and public safety applications.

This paper contributes a complete process for a practical ReID system to achieve a one-step specific pedestrian searching from the whole images or video sequences obtained by cameras distributed in different locations. The experimental results show that our method is not only effective in commonly used datasets and achieves good results in practical application scenarios.

2 RELATED WORK

2.1 Pedestrian detection

Pedestrian detection, as an example of object detection technology, aims to determine whether there are pedestrians in images or videos captured by the camera. As shown in Fig. 1, pedestrian detection can be regarded as a key step of ReID. As described above, in the traditional ReID research, the main task of many works published on top conferences is to match with manually cropped pedestrians in the existing dataset, which are usually the pedestrian images obtained from videos through manual annotation or detection algorithms in advance. Such as Deformable Part Model (DPM) [3] and Faster R-CNN [4] are the commonly used off-the-shelf pedestrian detectors to Market-1501 and MSMT17 ReID datasets, respectively.

In the real-world application, pedestrians in the images or videos should be detected and screened out first, and then the ReID is used to identify the specific person. In recent years, object detection algorithms based on deep learning have been proposed successively and have achieved significant practical applications, such as image classification, image segmentation, *etc.* Current state-of-the-art object detection algorithms with deep learning can divide into two major categories: One is two-stage detectors, such as R-CNN, a pioneering two-stage object detector proposed by Girshick *et al.* [5] in 2014, and its variants Fast R-CNN [6] and Faster R-CNN, *etc.* The other is one-stage detectors, such as Redmon *et al.* [7] developing a real-time detector called YOLO (You Only Look Once) in 2016, and its variants YOLOv3 [8], and so on.

2.2 Person re-identification

As a follow-up process of pedestrian detection, ReID is a pedestrian classification problem, that is, given a pedestrian to be detected, find all the pedestrian images in different camera scenes. In 2014, Li *et al.* [9] introduced CNN to ReID for the first time, using the FPNN network (Filter paring neural network) to train and get True or False by comparing two patches to determine if a pair of input images belong to the same ID.

In general, the ReID method based on deep learning usually includes three steps. The first step is to train the classification network on the training set. The pedestrian ID number in the training set is the number of network categories. Then, after the network converges, the output of full connected layers is used as the feature representation. The last step is to calculate Euclidean distance and judge their similarity of all image features.

Table 1 Statistics of some commonly used datasets for ReID

Dataset	Years	Identities	Images	Eval.
VIPeR	2007	632	1,264	CMC
CUHK03	2014	1,467	13,164	CMC
Market-1501	2015	1,501	32,668	C&M
DukeMTMC	2017	1,812	36,411	C&M
MSMT17	2018	4,101	126,441	C&M
SYSU-30k	2020	30,000	29606,918	C&M

Therefore, the research of ReID based on deep learning is mainly divided into two parts: representation learning and metric learning.

At present, a number of datasets for image-based and video-based ReID have been released. VIPeR [10], the first dataset used for ReID, was published in 2007. CUHK03 [9], the first dataset large enough for training deep learning models was released in 2014. Besides, the commonly used datasets include Market-1501 [11], DukeMTMC-reID [12] and MSMT17 [13], *etc.* In 2020, Wang *et al.* [14] released a large-scale ReID dataset SYSU-30k. This dataset contains 30,000 identities and a total of 29606,918 images, which is equivalent to 30 times that of ImageNet. The basic description of some commonly used datasets for ReID is shown in Table 1, images include training, test set (gallery) and query, C&M means both CMC and mAP are evaluated.

The statistics of paper submissions on ReID published in top-tier computer vision conferences (CVPR, ICCV/ECCV) show that ReID has attracted more and more attention and has become a topical research in computer vision in recent years. For example, in 2020, there were 34 and 18 papers published on CVPR and ECCV, respectively. Although numerous ReID datasets and algorithms have been proposed, the performance on these benchmarks has improved substantially. However, the existing ReID benchmarks and algorithms mainly focuses on matching cropped pedestrian images between queries and candidates, and there is still a significant gap between the problem setting itself and real-world applications.

3 PROPOSEED APPROACH

3.1 Overall framework

Like any advanced algorithm, the ultimate goal of ReID research needs contribute to practical application. In 2014, Xu *et al.* [15] took the first step in this work. They introduced the problem of person search to the community, and proposed a sliding window searching strategy based on pedestrian detection and person matching scores. However, the performance is limited by the handcrafted features, and the sliding window framework is not scalable. Subsequently, in 2017, Xiao *et al.* [16,17] proposed a new deep learning framework for person search. Instead of breaking ReID down into two separate tasks-pedestrian detection and person re-identification, an Online Instance Matching(OIM) loss function is proposed to train the network to close the gap between traditional ReID method and real application scenarios. However, in the implementation process, it need to design a pedestrian proposal net, identification net and a large-scale and scene-diversified person search dataset, respectively, to train the network in a multitasking way, which also has certain limitations in practical applications.

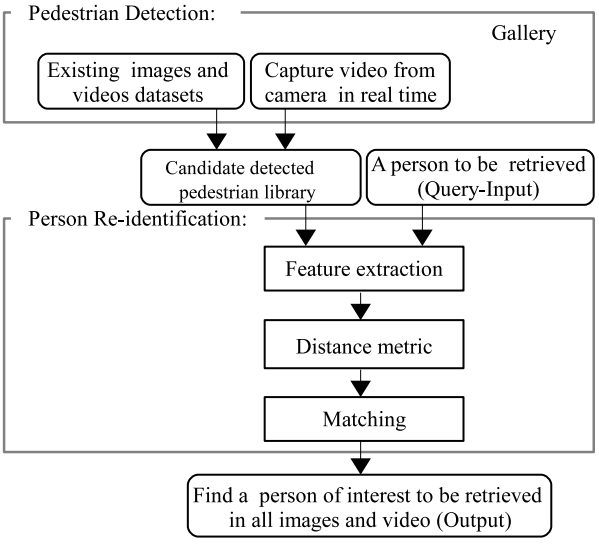


Fig. 2 Complete process for a practical ReID system, including Pedestrian detection and ReID

Similar to what the author described in previous works [18], the purpose of this paper is to achieve a one-step search for specific pedestrians in the scene images and videos captured by different cameras, and the complete process in a practical ReID system is shown in Fig. 2. The system includes two parts, pedestrian detection, and ReID. Unlike the traditional ReID method, the proposed system aims to combine pedestrian detection and ReID to perform one-step pedestrian detection and search, the biggest advantage of which is effectively and directly using the existing pedestrian detection and ReID models.

3.2 Model selection and optimization

First of all, as described in section 2.1, there are many high-performance object detection algorithms for the pedestrian detection model. After comparison and anal-

Table 2 Comparison with the state-of-the-art methods on imaged-based ReID datasets in recent years, i.e., the Market-1501 and DukeMTMC dataset

Methods	Years	Market-1501		DukeMTMC	
		Rank1	mAP	Rank1	mAP
PCB [19]	2018	92.3	77.4	81.8	66.1
BNNeck [20]	2019	94.5	85.9	86.4	76.4
OSNet [2]	2019	94.8	84.9	88.6	73.5
FastReID [1]	2020	96.3	90.3	92.4	83.2

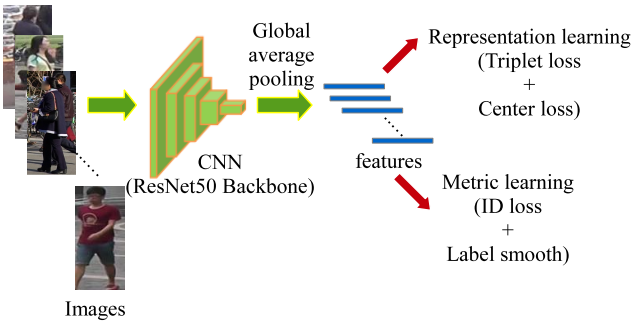


Fig. 3 The pipeline of our used strong ReID baseline

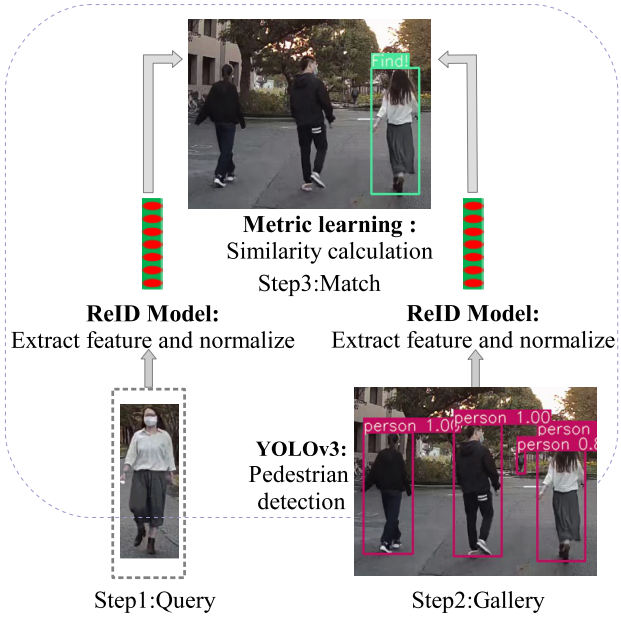


Fig. 4 The implementation of proposed a real-world application ReID system

ysis, YOLOv3, as one of the most popular real-time object detection algorithms, has a high detection effect and speed. For example, it is 1000 times faster than R-CNN and 100 times faster than Fast R-CNN. Therefore, we use the COCO dataset, containing more pedestrians and VOC dataset for joint training based on the YOLOv3, and implement it based on Keras in this paper. The related works have been done in previous papers, and now the weight file after training is used directly.

For the ReID model, the numerous performances have also been improved substantially in recent years. Some state-of-the-art algorithms published in CVPR, ICCV, and ECCV top conferences during 2018-2020 are listed in Table 2. Compared with many works, BNNeck [20] is found to have the best performance acquired by global features in all ReID works. At the same time, none of the existing ReID models addresses Omni-scale feature learning. OSNet [2] can well capture the local discriminative features, and it is also the key to overcoming the challenges of the ReID work. Therefore, in the part of ReID model, it is mainly implemented by combining BNNeck and OSNet. Followed those two open-source as our standard baseline, the strong ReID pipeline baseline used in this paper is shown in Fig. 3.

As shown in Fig. 3, similar to most of the existing ReID models, we adopt the ResNet-50 [21] as our basic CNN model. During the training stage, we initialize the ResNet-50 with pre-trained parameters on ImageNet and change the dimension of the fully connected layer to N, which denotes the number of identities in the training dataset. Simultaneously, to enhance the generalization ability of the ReID model, we use the training sets of three datasets, Market-1501, DukeMTMC-reID, and MSMT17, to form a new dataset called Market-7414 dataset. It contains a total of 7,414 identities. Simultaneously, GAN method is used to expand the dataset [22]. All the naming of the new dataset is consistent with Market-1501 dataset.

Finally, after optimization, the ReID model used in this paper achieves the best performance (95.7%) on Market1501 at rank-1 accuracy. Here, we will not go into the details of the model optimization. The key point of the method designed in this paper is that the latest model with good performance can be used at any time.

3.3 Combination of pedestrian detection and Person re-identification

After selecting and optimizing the pedestrian detection and ReID models respectively, we start to combine the two models to build a one-step person search system. Finally, the implementation of our proposed real-world application ReID system is shown in Fig. 4.

As shown in Fig. 4, the query process is divided into three steps. Firstly, use the ReID model to extract features for the query set. Secondly, use the pedestrian detection algorithm YOLOv3 to detect pedestrians in the images or videos sequence in the gallery set. Then,

Fig. 5 Experiment on commonly used image-based ReID datasets (A part of the output results)

use the ReID model to extract features for all pedestrians detected in the image or video. Finally, calculate the similarity of the distance between the query and gallery set, match the results, make the candidate pedestrians images with a smaller distance (or higher similarity) as search results, and then output them in the terminal and specified folder.

In the code implementation part, load the pedestrian detection and ReID models directly that have been trained in advance and remove the training part of the two source codes. Only the inference code has retained the inference to determine whether the person is the one to be retrieved. The ultimate goal is to achieve a one-step person search from the whole scene images or videos.

4 EXPERIMENTS

4.1 Experiments on commonly used datasets

First of all, we evaluate our proposed method on the commonly used benchmark datasets, including three image-based ReID datasets, *i.e.*, Market-1501, DukeMTMC-reID, MSMT17, and one video-based dataset, MARS, respectively. Some sample retrieval results are shown in Fig. 5.

When experimenting on Market-1501, we directly select an image in the query set (without changes the file name) as the query image randomly. We then use the testing set (also without changes the file name) as a candidate pedestrian library (gallery) for one-step person retrieval. But when experimenting on MSMT17 dataset, we can randomly select one image from any camera as the query set in the testing set and use the rest of the images to form the gallery set. Before querying, the file name of images in the query set needs to be modified and strictly consistent with the Market-1501 dataset, while the gallery set does not need to be modified.

Finally, many test results show that the average accuracy of a single query on the commonly used ReID datasets surpasses 90%, and we can conclude that our proposed method can be further applied to find specific pedestrians in practical application scenarios.

4.2 Experiments in real-world scenery

To further verify the effectiveness of our proposed method in actual application scenarios, three cameras are placed in different locations to obtain video clips containing different poses and viewpoints of multiple pedestrians, including the basis pre-processing such as cropping the videos as required, converting videos into images (at intervals 5 frames) and so on. Finally, our gallery set contains 1,318 images and 3 videos in total, and the file name is consistent with Market-1501 dataset.

In our setting, the query task of ReID is divided into two classes, image-based and video-based. Specifically, the first step is to give an original scene image or video containing the pedestrian to be retrieved. The second step is to crop a specific person from the original scene images or videos (from any image in the gallery set). The third step is to save the cut pedestrian image in the query set. Here, the image's file name to be retrieved in the query set must be strictly consistent with the Market-1501 dataset. Then, use the ReID model to extract the query set's features, including the normalization of the query features. The last step is to output the results, including pedestrian detection, features extraction, and normalization for each original image and video sequence in the gallery set, and perform distance similarity with the query set one by one. Finally, some gallery images contain the target person labeled by bounding boxes (random use of one color). The search results are output through two channels in real-time. One is the file output in the specified folder, and the other is the terminal information. Some sample retrieval results are shown in Fig. 6 and Fig. 7, respectively.

We all know that the interclass (instance/identity) variations are typically significant due to the changes in camera viewing conditions. Different cameras may

Fig. 6 Experiment in real-world application scenario, the search results come from gallery multiple non-overlapping cameras

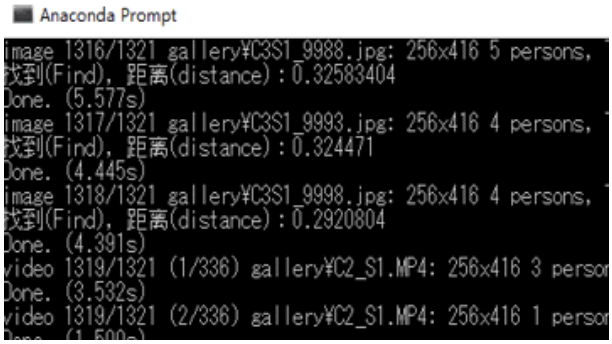


Fig. 7 Terminal display results of one-step person search

present different styles. The second row of Fig. 6 shows the retrieved pedestrian's original input images under different camera viewpoints, background clutter, human poses, resolution, *etc.* To improve the efficiency of the retrieval, the image can be retrieved one by one, while the video can be retrieved every few frames. We can conclude that our proposed method has achieved good results in practical application scenarios through many samples and experiments. However, we can also see that occlusion and resolution are the two most important factors affecting the retrieval results' accuracy.

Simultaneously, as shown in Fig. 7, the images in the gallery set are retrieved one by one, and the video that has been cut in advance is also retrieved frame by frame. Moreover, we can obtain the retrieval results information of all the files in the gallery set in real-time, including the image or video file name, the distance value of the detected pedestrians, a total of several pedestrians were detected, and the time spent.

As far as we know, ReID has broad application prospects and needs in the field of intelligent video surveillance and public safety in real-world scenarios. The method proposed in this paper improves the availability of multi-camera monitoring, such as public security criminal investigation, finding criminals, and cross-camera person tracking, *etc.*

5 CONCLUSION

In this paper, compared with the traditional research that divides the problem of ReID into two steps: pedestrian detection and ReID, a novel and complete practical ReID system is designed to achieve a one-step search of the specific pedestrians in images or video sequences in actual application scenarios. For images and video sequences obtained by cameras distributed in different locations, given a person-of-interest to be queried, our method's goal is to search a person from the whole scene images or videos directly instead of matching them with manually cropped pedestrians in the existing dataset. Finally, the experimental results show that our method is effective in commonly used datasets and has achieved good results in practical application scenarios.

In future work, the research will continue to achieve end-to-end ReID, including improving query speed and accuracy. Meanwhile, it is found that the current ReID model will not work when the clothes change. The research is mainly concentrated on short-term scenarios. In order to meet the need for real-life, it urgently needs to consider the more complex variability. The problem of clothes changes in real-world scenarios and performing real-time pedestrian detection and query on mobile devices will be the main research directions and content of our future work.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61972059, Grant 61702055, and Grant 61773272, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191474 and Grant BK20191475, in part by the Natural Science Foundation of Jiangsu Province in China under grant No. BK20191475, in part by the fifth phase of "333 Project" scientific research funding project of Jiangsu Province in China under grant No. BRA2020306, and in part by the Qing Lan Project of Jiangsu Province in China under grant No. 2019.

References

1. He L., Liao X., Liu W., *et al.*: FastReID: A Pytorch Toolbox for Real-world Person Re-identification. arXiv: 2006.02631v4 (2020)
2. Zhou K., Yang Y., Cavallaro A., *et al.*: Omni-Scale Feature Learning for Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3702-3712 (2019)
3. Felzenszwalb P. F., Girshick R. B., McAllester D., *et al.*: Object Detection with Discriminatively Trained Part Based Models. In IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1627-1645 (2010)
4. Ren S., He K., Girshick R., *et al.*: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1137-1149 (2017)
5. Girshick R., Donahue J., Darrell T.: Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.580-587 (2014)
6. Girshick R.: Fast R-CNN. In The IEEE International Conference on Computer Vision (ICCV), pp.1440-1448 (2015)
7. Redmon J., Divvala S., Girshick R.: You Only Look Once: Unified, RealTime Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.779-788 (2016)
8. Redmon J., Farhadi A.: YOLOv3: An Incremental Improvement. arXiv: 1804.02767 (2018)
9. Li W., Zhao R., Xiao T., *et al.*: DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.152-159 (2014)
10. Gray D., Brennan S., Tao H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS) (2007)
11. Zheng L., Shen L., Tian L., *et al.*: Scalable Person Re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.1116-1124 (2015)
12. Zheng Z., Zheng L., Yang Y.: Unlabeled Samples Generated by GAN Improve the Person re-identification Baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.3754-3762 (2017)
13. Wei L., Zhang S., Gao W., *et al.*: Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.79-88 (2018)
14. Wang G., Wang G., Zhang X., *et al.*: Weakly Supervised Person Re-ID: Differentiable Graphical Learning and A New Benchmark. In IEEE Transactions on Neural Networks and Learning Systems (TNNLS) (2020)
15. Xu Y., Ma B., Huang R., *et al.*: Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness. In Proceedings of the 22nd ACM international conference on Multimedia, pp.937-940 (2014)
16. Xiao T., Li S., Wang B., *et al.*: End-to-End Deep Learning for Person Search. arXiv: 1604.01850v1 (2016)
17. Xiao T., Li S., Wang B., *et al.*: Joint Detection and Identification Feature Learning for Person Search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3415-3424 (2017)
18. Zhu M., Gong S., Qian Z., *et al.*: Person Re-identification on Mobile Devices Based on Deep Learning. In The 8th IIAE International Conference on Industrial Application Engineering 2020 (ICIAE), pp.253-260 (2020)
19. Sun Y., Zheng L., Yang Y., *et al.*: Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In Proceedings of the European Conference on Computer Vision (ECCV), pp.480-496 (2018)
20. Luo H., Gu Y., Liao X., *et al.*: Bag of Tricks and A Strong Baseline for Deep Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4321-4329 (2019)
21. He K., Zhang X., Ren S., *et al.*: Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770-778 (2016)
22. Zhu M., Gong S., Qian Z., *et al.*: A Brief Review on Cycle Generative Adversarial Networks. In The 7th IIAE International Conference on Intelligent Systems and Image Processing (ICISIP), pp.235-242 (2019)