

Enhanced Deep Feature Representation for Person Search

Jinfu Yang^{1,2}, Meijie Wang^{1,2}(✉), Mingai Li^{1,2}, and Jingling Zhang^{1,2}

¹ Faculty of Information Technology, Beijing University of Technology,
Beijing, China

wangmeijie1992@163.com

² Beijing Key Laboratory of Computational Intelligence and Intelligent System,
Beijing, China

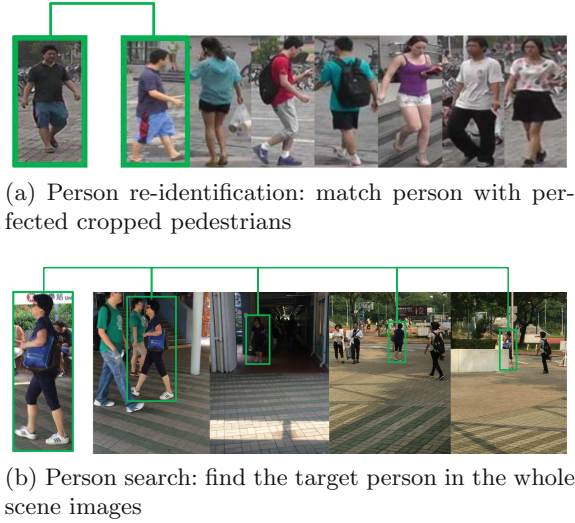
Abstract. Person re-identification (re-id) has attracted widespread attention due to its application and research significance. However, since the person re-id puts the cropped images as input, it is far from the real-world scenarios just like person search which aims at matching a target person from a gallery of the whole scene images. Person search is more difficult but more practical and meaningful. In this paper, we propose a new person search network with an enhanced feature representation. Our network mainly consists two parts, a pedestrian proposal net and an identification net. In the identification net, we utilize hand-crafted features and Convolutional Neural Network (CNN) features to get more discriminative and compact features. Experiments on a large-scale benchmark dataset demonstrate our network gets better performance than others counterparts.

Keywords: Person search · Hand-crafted features · CNN features
Feature representation

1 Introduction

Person re-identification (re-id) aims at matching the target person from different non-overlapping camera views which has been studied extensively in recent years [3, 13, 14, 17, 21, 26]. It is widely used in video surveillance, robot and image retrieval. The most challenging problem is how to correctly match the target person from a gallery images under the complex variations of human poses, significant changes in view angle, lighting, occlusion, resolution, background clutter and etc.

There are plentiful existing datasets and methods on the person re-id, but person re-id has a large difference with the real-world application. In most benchmarks [5, 10, 11, 13, 15], there are only pedestrian images which made by manually cropped, as shown in Fig. 1(a), however, finding the target person in a gallery of whole scene images is the practical goal (Fig. 1(b)). That is, most of the person re-identification methods suppose that the pedestrian detectors can perfectly



(a) Person re-identification: match person with perfected cropped pedestrians

(b) Person search: find the target person in the whole scene images

Fig. 1. Comparison of person re-identification with person search. According to real-world applications, the person search is more different but more practical and meaningful.

crop the pedestrian images, but it is unavailable. Existing pedestrian detectors would inevitably produce detection errors, which significantly harm the final performance of person search.

The conventional approaches of person search break down it into two separate tasks—pedestrian detection and person re-id. In 2014, Xu *et al.* [23] first proposed a unified framework for person search. They jointly modeled the commonness of people and uniqueness of the queried person. But a searching strategy based on sliding window brings high computational cost and blocks the scalability. And only using the hand-crafted feature maps limits the performance. In 2017, Xiao *et al.* [22] made a big step towards to the real-world application. They use a single CNN for person search. An Online Instance Matching (OIM) loss function is proposed to effectively train the network and be scalable to a mass of identities. However, the features extracted in [22] do not contain enough color features and texture features to distinguish different people. As shown in Fig. 2, the first line of Fig. 2 is the ground truth, the blue bounding box (the left one) is the target person and the green bounding boxes are the matched images with the target person. The second line is the predicted results, the green one is the correct result, and the orange ones are the false results. As we known, concatenating hand-crafted appearance features, e.g. RGB, HSV colorspace and LBP descriptor designed to overcome appearance variations in different views for person re-identification is more distinctive and reliable [20].

In this paper, we propose a novel person search network which could effectively fuse the hand-crafted feature maps and deep feature maps. Hand-crafted features are the Ensemble of Local Features, named ELF16 and deep features

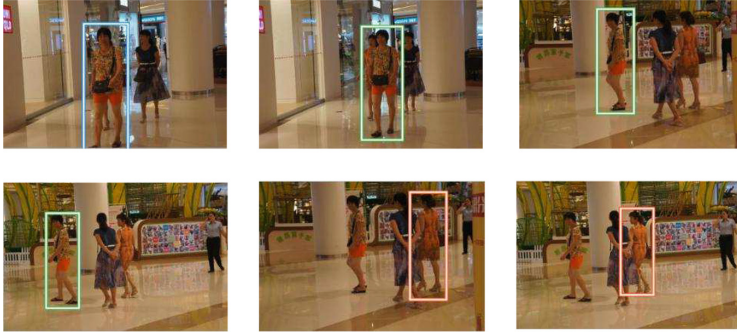


Fig. 2. Some example results of person search. It is clear that the features [22] do not contain enough color features and texture features to distinguish between different people. (Color figure online)

are CNN features. The proposed network consists of two parts: a pedestrian proposal net and an identification net. The identification net can extract more complementary features for person search. Experimental results on a large-scale benchmark dataset demonstrate the effectiveness of our enhanced deep feature representation.

2 Method

Our network architecture is shown in Fig. 3. A given whole scene image is fed into a basic CNN stage as input. And the raw pixels are transformed to CNN features through the basic CNN stage. Next these features are fed into a pedestrian proposal net to predict bounding boxes of candidate pedestrians. Then they are fed into a RoI-Pooling layer [6], followed by an identification net to extract L2-normalized features. The identification net includes two parts. One is the rest

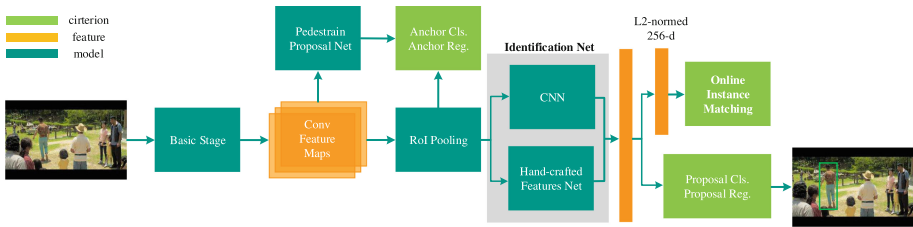


Fig. 3. The architecture of person search network. An input image is fed into the basic stage to extract feature maps. Then the pedestrian proposal net is used to generate bounding boxes of candidate person, followed by an identification net that contains a CNN and a hand-crafted feature net. Finally, the extracted features are projected to a L2-normalized 256-d subspace, and Online Instance Matching loss function is employed to train the network.

part of ResNet-50, the other is the hand-crafted features net. We combine these two parts into a full-fledged feature map via Buffer Layer and Fusion Layer. At training stage, in order to supervise the identification net, the OIM loss function is employed to train the pedestrian proposal net via a multi-task manner. At testing stage, we compute the distances between the target person with each of the gallery person. In the next sections, we introduce the details of the CNN model structure and the hand-crafted features as well as the details of the fusion process.

2.1 Model Structure

Our model is built upon the ResNet-50 [9]. The first layer of the ResNet-50 has a 7×7 convolution filter, named conv1. The following layers consist of four blocks. Each of layers has 3, 4, 6, 3 residual units, named conv2_x to conv5_x. We regard the layers from conv1 to conv4_3 as the basic stage. The input image produces 1024 feature channels through the basic stage, and the resolutions of the feature maps is 1/16 of the original image.

In order to detect pedestrian bounding boxes, we build a pedestrian proposal network upon these features. In order to make the feature maps more suitable for pedestrian detection task, we use a $3 \times 3 \times 512$ filter to transform the feature maps. Following [18], at each feature map location we adopt 9 anchors, and use a softmax classifier and a linear regression. The softmax classifier is to predict whether each anchor is a pedestrian or not and the linear regression is to adjust their bounding boxes. In order to reduce redundancy, we utilize a non-maximum suppression (NMS) on these bounding boxes and then reserve the top 128 bounding boxes as our pedestrian proposals.

In the cause of finding the target person in these proposals, we build an identification net upon these proposals to further extract the feature maps that contain CNN feature maps and hand-crafted feature maps, and compare with the features of the target person. First, we use an RoI-Pooling layer [6] to extract a $14 \times 14 \times 1024$ feature map from the basic stage for each proposal. Then the extracted feature maps are fed into the identification net and a global average pooling layer to generate a 2048D feature map. Since there are some malposition, distortion and misalignments in the pedestrian proposals, a softmax classifier and a linear regression are applied to reject non-persons and adjust the location. Finally, the feature maps are mapped into a 256D subspace through a L2-normalization and the cosine similarities are computed to inference whether it is the same person. In the training phase, we use the OIM loss function to train the whole network. The whole network is jointly trained in a multi-task learning manner, rather than using the alternative optimizations in [18].

2.2 The Detail of the Identification Net Structure

The identification net contains two parts as shown in Fig. 4. The first part is the rest of the ResNet-50, i.e., conv4_4 to conv5_3, following as [22]. The convolution in pedestrian proposal network allows part displacement and visual changes to

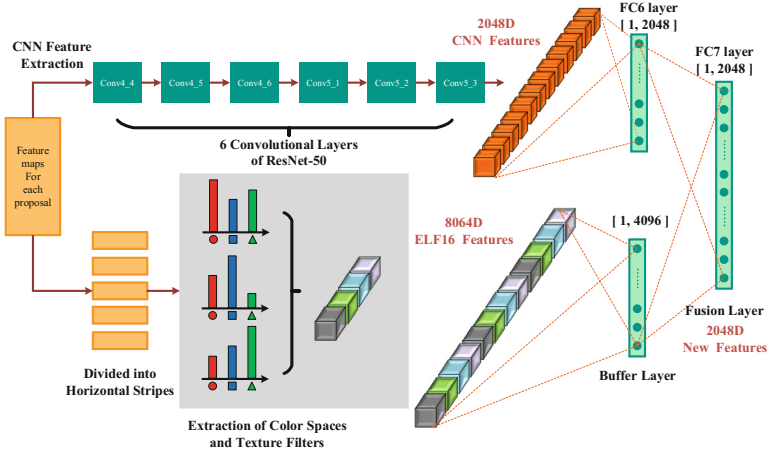


Fig. 4. ELF16 feature maps and CNN feature maps of the identification net. (Color figure online)

be alleviated in this part. At the same time, convolution kernels provide different descriptions for pedestrian images. In addition, pooling and LRN layers provide nonlinear expression of corresponding description, which significantly reduces the overfitting problem. So we build the identification net upon pedestrian proposal network and RoI-Pooling layer to extract more discriminatory. The second part extracts additional hand-crafted features from the same feature map, followed by a Buffer Layer which is a 4096D-output fully connection layer. A fusion layer is also used to combine feature maps of the two parts into a full-fledged feature map.

In the lower part of Fig. 4, ELF [8] is extracted. In a general way, the input image of the pedestrian is partitioned into several horizontal stripes. Each horizontal stripe extracts different color and texture information. Specifically, we partition the feature maps of each pedestrian proposal into 16 horizontal stripes. And the color feature maps and the texture feature maps are extracted respectively from each horizontal stripe. The color features contain RGB, HSV, LAB, XYZ, YCbCr and NTSC and the texture features contain Gabor, Schmid and LBP. Each channel of the above features extracts a L1-normalized 16D histogram. All histograms are concatenate together to form a single feature vector, named ELF16.

In order to effectively fuse CNN feature maps with ELF16, we use a buffer layer and a fusion layer. In our network, the ELF16 could affect the parameters of the whole network. Generally speaking, the fusion feature maps should be more discriminative than both CNN feature maps and ELF16. It is important for the buffer layer in our network, since it bridges the gap between the ELF16 and CNN feature maps with enormous difference, moreover, it can ensure the convergence.

If the input of the fusion layer is

$$x = [ELF16, CNN_Features] \quad (1)$$

Then the output of the fusion layer is computed by:

$$Z_{Fusion}(x) = h(W_{Fusion}^T x + b_{Fusion}) \quad (2)$$

The activation function is denoted as $h(\cdot)$. And we adopt the ReLU and dropout layers, the dropout ratio is set as 0.5. On the basis of the back propagation(BP) algorithm, the parameters of l^{th} layer after a new iteration are as following:

$$\begin{aligned} W_{new}^{(l)} &= W^{(l)} - \alpha \left[\left(\frac{1}{m} \Delta W^{(l)} \right) + \lambda W^{(l)} \right] \\ b_{new}^{(l)} &= b^{(l)} - \alpha \left[\left(\frac{1}{m} \Delta b^{(l)} \right) \right] \end{aligned} \quad (3)$$

where parameters α , m and λ are set as [1].

2.3 How Do ELF16 and CNN Feature Maps Make More Complementary?

We effectively fuse CNN feature maps which do not contain enough color features and texture features to distinguish different people and ELF16 to get better performance. If the parameters of the whole network are affected by the ELF16, i.e., the gradient of the network parameters are adjusted according to ELF16, then CNN feature maps and ELF16 could be more complementary with each other, as the buffer layer and fusion layer are used to make our features more discriminative in different images.

Denote ELF16 as \tilde{x} and conv5_3 feature map as x , and the weight connecting the j^{th} node in n^{th} node in $(n+1)^{th}$ layer as W_{ij}^n . Denote $Z_j^n = \sum_j W_{ji}^{n-1} a_i^{n-1}$ where $a_i^{n-1} = h(Z_i^{n-1})$. Then

$$\delta_i^n = \frac{\partial J}{\partial Z_i^n} \quad (4)$$

Note that $Z_j^6 = \sum_j W^{n-1} x_i^{n-1}$. According to back propagation, \tilde{x} influence $\frac{\partial J}{\partial W_{ij}^5}$. According to this, ELF16 \tilde{x} could affect the parameters of CNN feature maps to make more complementary. And

$$\frac{\partial J}{\partial W_{ij}^5} = x_j \delta_j^6 \quad (5)$$

where

$$\delta_i^6 = \left(\sum_j W_{ji}^6 \delta_j^7 \right) h'(Z_i^6) \delta_j^7 = \left(\sum_k W_{kj}^7 \delta_k^8 \right) h'(Z_j^7) \quad (6)$$

At the same time, \tilde{x} could affect δ_j^7 by,

$$Z_k^7 = \sum_j W_{kj}^6 a_j^6 + \sum_j \tilde{W}_{kj}^6 \tilde{a}_j^6 \quad (7)$$

where

$$\tilde{a}_j^6 = h \left(\sum \tilde{W}_{ji}^5 \tilde{x}_i \right) \quad (8)$$

3 Experiments

3.1 Dataset

We use a large-scale person search dataset [22] to evaluate our approach. The dataset comes from hundreds of scenes of street in an urban city and movie snapshots with the variations of lighting, viewpoints, and background conditions. The dataset contains 18,184 images, 8,432 identities, and 96,143 pedestrian bounding boxes. The statistics of the dataset are shown in Table 1. However, the images which the person who appear with half bodies or abnormal poses are not be annotated.

Table 1. Statistics of the dataset with respect to data sources and training/test splits.

Source/split	# image	# pedestrians	# identities
StreetSnap	12,490	75,845	6,057
Movie & TV	5,694	20,298	2,375
Training	11,206	55,272	5,532
Test	6,978	40,871	2,900
Overall	18,184	96,143	8,432

3.2 Evaluation Protocols and Metrics

The dataset is split into a training set and a test set, and there are no overlapped images or labeled identifies between the two sets. Further, the test identity instances are split into queries and galleries. We randomly choose one of these identity instances as the query for each of the 2900 test identifies. At the same time, all the images that contain the other instances and some images which do not contain this person with randomly sample constitute the homologous gallery set. Different queries correspond different galleries and jointly covered all the 6,978 test images.

We use cumulative matching characteristic (CMC top-K) and mean averaged precision (mAP) to evaluate the proposed method. The CMC top-K is inspired from the person re-id task, when one can match with anyone of the top-K predicted bounding boxes which overlaps with intersection-over-union (IoU) greater or equal to 0.5 of the ground truths. The mAP is inherited from the object detection task to judge the validity of predicted bounding boxes.

3.3 Experiment Settings

We use a pre-trained caffemodel that implemented the person search network in [22]. We regard the first convolution layer and the batch normalization layers as constant affine transformations in the basic stage. All the loss weights are the same. There are two scene images in each mini-batch. The initial learning rate is set as 0.00001, decreased by one-tenth of the initiative after 40K iterations. And the network took 50K iterations to converge.

We compare our method with traditional methods which regard the person search as two separate task, that is, pedestrian detection task and person re-id task. In our experiments, there are 3 pedestrian detection and 5 person re-id methods, besides the person search network [22]. For pedestrian detection, we employ the off-the-shelf detectors, including CCF [24], ACF [4] and Faster R-CNN [18] with ResNet-50.

For person re-id, we employ popular feature representations of re-id which are DenseSIFT-ColorHist (DSIFT) [25], Bag of Words (BoW) [26] and LOMO [14]. And every feature representation and a specific distance metric are jointly used. The distance metrics contain Euclidean, Cosine similarity, KISSME [10] and XQDA [14]. We use the remaining net without the pedestrian proposal network in [22], and give the cropped pedestrian images as input images to train a baseline re-id method that named IDNet.

3.4 Comparison with Detection and Re-ID

We compare our method with the above mentioned methods which include a person search framework that was proposed by Xiao *et al.* in [22] and 15 baseline combinations that separate the person search problem into pedestrian detection and person re-id tasks. The results are shown in Tables 2 and 3. It is shown that our network outperforms the others. Thanks to the joint optimization of the pedestrian detection and person re-id, we can see that both the person search method [22] and ours outperforms the others by a large margin.

Table 2. Comparison of our network with separate pedestrian detection + person re-identification method and the person search network that was proposed by Xiao *et al.* [22] in terms of CMC top-1.

CMC top-1 (%)	CCF	ACF	CNN	GT
DSIFT + Eudclidean	11.7	25.9	39.4	45.9
DSIFT + KISSME	13.9	38.1	53.6	61.9
BoW + Cosine	29.3	48.4	62.3	67.2
LOMO + XQDA	46.4	63.1	74.1	76.7
IDNet	57.1	63.0	74.8	78.3
Xiao <i>et al.</i> [1]	—	—	78.7	80.5
Ours	—	—	80.6	82.8

Table 3. Comparison of our network with separate pedestrian detection + person re-identification method and the person search network that was proposed by Xiao *et al.* [22] in terms of mAP.

mAP (%)	CCF	ACF	CNN	GT
DSIFT + Eudclidean	11.3	21.7	34.5	41.1
DSIFT + KISSME	13.4	32.3	47.8	56.2
BoW + Cosine	29.9	42.4	56.9	62.5
LOMO + XQDA	41.2	55.5	68.9	72.4
IDNet	50.9	56.5	68.6	73.1
PSNet	—	—	75.5	77.9
Ours	—	—	77.8	80.2

From Tables 2 and 3 for each person re-id method, we can see that the person search performance are significantly affected by the detectors. It may not a good idea to use an off-the-shelf detector to employ the existing person re-id methods for the real-world person search. Otherwise, the pedestrian detector will impede the recall of the better person re-id methods.

On the other hand, the performance of different person re-id methods is consistent with all the detectors. We can see that IDNet has similar performance with LOMO+XQDA when employing ACF and CNN detectors or GT. However, it is significantly better than LOMO+XQDA method when using the CCF detector. We can also see that our person search has better performance than Xiao *et al.* [22], since our network can extract more effective features to distinguish different people.

The performance of person search is directly affected by the detector in the separate network. Poor performance of detector could not get a good performance of person search task. The two parts could not affect each other, i.e. the pedestrian detection network only affect the person re-id network, and the person re-id network could affect the pedestrian detection network. However, in our work, these two parts could affect each other to make two parts more complementary.

3.5 Experiments on Effectiveness of Fusion Features

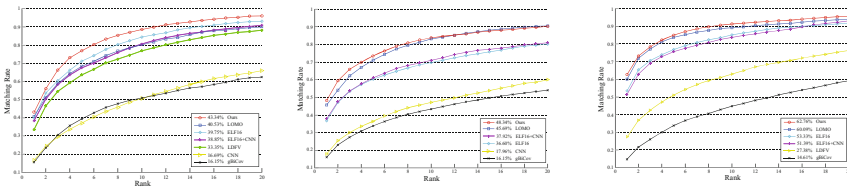
We evaluated our features on three benchmark datasets of person re-identification to demonstrate the effectiveness of our fusion features.

Datasets and Experimental Settings. We use three publicly available datasets: VIPeR [7], CUHK01 [12] and PRID450s [19]. Each dataset was captured from two disjoint camera views under significant misalignment, light change and body part distortion. There is a brief introduction for the datasets in Table 4.

Table 4. Statistics of the dataset with respect to data sources and training/test splits.

Dataset	# identities	# images	# images in training sets	# camera views
VIPeR	632	1,264	316	2
CUHK01	971	3,884	485	2
PRID450s	450	900	225	2

In each experiment, the training set was half of the identities via randomly selected and the rest as testing set. During training, we learn the projection matrix W using the training set. During testing, we get the final features by $x' = W^T x$ then measure the distance between two inputs. In order to make the results reliable and stable, we repeated 10 times for each experiment and then computed the average Rank- i accuracy rate. CMC was present in Fig. 5.



(a) VIPeR on Mirror KMFA (b) CUHK01 on Mirror KMFA (c) PRID450s on Mirror KMFA

Fig. 5. CMC Curves of our experiments on three datasets with Mirror KMFA metric learning algorithm.

During testing, we regard one image which was chosen from View 2 as a probe and we also regard all images which in View 1 as the gallery. However, in the CUHK01 dataset, there are two images of the same person in one view, one image of each identity was randomly chosen and regard it as the gallery.

In [2], Chen *et al.* proposed a high-performance metric learning algorithm named Mirror Kernel Marginal Analysis (KMFA) for person re-id. Our experiments employ this metric learning algorithm, and the setting as the [2].

Features. In our experiments, we use six methods to extract feature maps, which contain LDFV [15], gBiCov [16], CNN features, LOMO features [14], ELF16 features and our features. Specifically, the CNN features are the same as the upper network of the Fig. 4. And our features are the fusion features of the Fig. 4. Because of the copyrights of the LDFV features code, we only experiment the LDFV features on VIPeR dataset. In order to extract features efficiently, we resized all images to 224×224 .

In addition, we enhanced the features representation (CNN+ELF16) to prove the effectiveness of our approach. The features of CNN+ELF16 are the concatenation of CNN features to ELF16.

Experiments on the Features. The Fig. 5(a)–(c) are the results of our experiments on VIPeR, CUHK01 and PRID450s with Mirror KMFA, respectively. We can see that our features outperform other standalone features. And our features significantly outperformed than CNN+ELF16 features, which maybe according to this two reasons:

1. *The CNN features of the experiment were trained to make them complementary with the ELF16 features, which it is not optimal to simply cascade CNN features with ELF16.*
2. *We use the buffer layer and fusion layer to automatically tune the parameters of the network, and the performance of the fused features is better than others.*

4 Conclusion

In this paper, we propose an enhanced feature representation for person search. It jointly handles the pedestrian detection and person re-id tasks into a single CNN. And in person re-id part we jointly use CNN feature maps and ELF16 which include color features and texture features. Then the buffer layer and fusion layer could regularize the network to make the identification net more focus on extracting complementary features maps. Experiments on a large-scale person search dataset show the effectiveness of our person search network.

Acknowledgments. This work is partly supported by the National Natural Science Foundation of China under Grant nos. 61573351, and 81471770.

References

1. Bottou, L.: Stochastic gradient descent tricks. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) *Neural Networks: Tricks of the Trade*. LNCS, vol. 7700, pp. 421–436. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_25
2. Chen, Y.C., Zheng, W.S., Lai, J.: Mirror representation for modeling view-specific transform in person re-identification. In: *IJCAI*, pp. 3402–3408. Citeseer (2015)
3. Chu, X., Ouyang, W., Li, H., Wang, X.: Structured feature learning for pose estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4715–4723 (2016)
4. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(8), 1532–1545 (2014)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010)
6. Girshick, R.: Fast R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448 (2015)
7. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: *Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3 (2007)

8. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_21
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2288–2295. IEEE (2012)
11. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person reidentification. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1622–1634 (2013)
12. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_3
13. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
14. 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 (2015)
15. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 413–422. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_41
16. Ma, B., Su, Y., Jurie, F.: Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image Vis. Comput. **32**(6), 379–390 (2014)
17. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1855 (2015)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
19. Roth, P.M., Hirzer, M., Köstinger, M., Beleznai, C., Bischof, H.: Mahalanobis distance learning for person re-identification. In: Gong, S., Cristani, M., Yan, S., Loy, C.C. (eds.) Person Re-Identification. ACVPR, pp. 247–267. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4_12
20. Wu, S., Chen, Y.C., Li, X., Wu, A.C., You, J.J., Zheng, W.S.: An enhanced deep feature representation for person re-identification. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8. IEEE (2016)
21. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1249–1258 (2016)
22. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the CVPR (2017)
23. Xu, Y., Ma, B., Huang, R., Lin, L.: 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. ACM (2014)

24. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 82–90 (2015)
25. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3586–3593 (2013)
26. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)