Spatial Invariant Person Search Network

Liangqi Li, Hua Yang(✉), and Lin Chen

Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai, China
{Lewis_lee,hyang,SJChenLin}@sjtu.edu.cn

Abstract. A cascaded framework is proposed to jointly integrate the associated pedestrian detection and person re-identification in this work. The first part of the framework is a Pre-extracting Net which acts as a feature extractor to produce low-level feature maps. Then a PST (Pedestrian Space Transformer), including a Pedestrian Proposal Net to generate person candidate bounding boxes, is introduced as the second part with affine transformation and down-sampling models to help avoid the spatial variance challenges related to resolutions, viewpoints and occlusions of person re-identification. After further extracting by a convolutional net and a fully connected layer, the resulting features can be used to produce outputs for both detection and re-identification. Meanwhile, we design a directionally constrained loss function to supervise the training process. Experiments on the CUHK-SYSU dataset and the PRW dataset show that our method remarkably enhances the performance of person search.

Keywords: Person re-identification · Person search
Spatial transformation

1 Introduction

Pedestrian detection and person re-identification (Re-ID) are of great significance in real-world applications like security surveillance, crowd flow monitoring and human behavior analysis [1]. To date, they are usually regarded as two isolate problems in computer vision research. A pedestrian detection system usually ignores the identification information of pedestrian samples in the popular datasets like Caltech [2] and ETH [3], and only classifies the detected boxes as either positive or negative ones. On the other hand, Re-ID aims at matching the query person among a lot of gallery samples from video sequences or static images collected from a variety cameras. Most Re-ID benchmarks [4,5] built on the datasets, such as CUHK03 [4] and Market1501 [6], which have manually cropped bounding boxes of individual persons.

However, there are no pre-cropped individual images in real-world situations. Even though utilizing pedestrian detectors, apart from filtering false alarms like backgrounds bounding boxes, it is still tedious to assign an ID to each sample. Generally, person Re-ID is established on the basis of pedestrian detection and

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 122–133, 2018.
https://doi.org/10.1007/978-3-030-03335-4_11

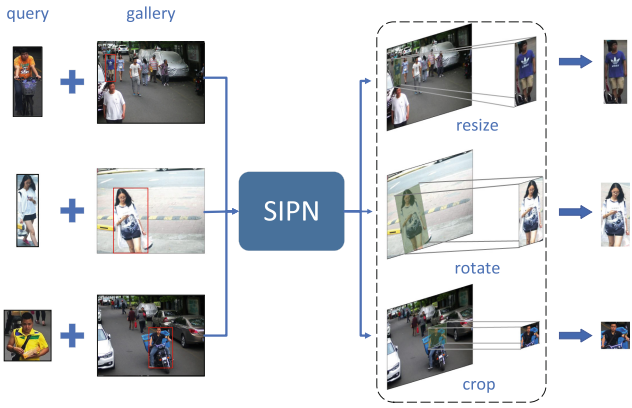

Fig. 1. Spatial transformations in our SIPN. It is difficult to straightly compare the input queries and galleries because of the giant spatial variance between them. But after implementing spatial transformations like resizing, rotating and cropping, the objective samples would be easier to identify.

the results from pedestrian detection can influence the accuracy of person Re-ID. From our observation, the two tasks can be integrated into a unified framework to improve convenience and performances, especially on the person Re-ID problem. Such cascaded task is called person search in this work.

There are only a few researchers devoting to handle this task. Pioneer works [7] and [8] just adopted simple two-stage strategies to jointly address the person search problem, and NPSM [9] coined an LSTM-based attention model to straightly search person from the image. However, all these approaches ignored the new challenges in person search task. Different from a traditional person Re-ID problem, pedestrians appear at a range of scales and orientations in person search scenes, which is much more in line with a real-world situation. Moreover, there are much more spatial variance challenges raised by multifarious resolutions, viewpoints and occlusions in person search. A lot of methods use spatial transformations like cropping, resizing and rotating for data augmentation. That is to say, vanilla Convolutional Neural Network (CNN) based models lack capabilities to cope with such spatial variance. We coin a new model named Spatial Invariant Person search Network (SIPN) to handle such challenges. As shown in Fig. 1, our SIPN can implement spatial transformations such as cropping, resizing, and rotating to make the detected samples spatially invariant. This means that, with SIPN, features extracted for identification will be much robuster. In contrast, traditional CNN, which is used as a workhorse in computer vision, can only guarantee the translation invariance of input samples.

Meanwhile, SIPN acts as a cascaded framework to implement an end-to-end pipeline. It is a CNN-based model that not only produces pedestrian candidate bounding boxes, but also goes further to extract features of these boxes to identify persons. We follow Faster R-CNN which saw heavy use in object

detection area to design a network like RPN (Region Proposal Network) to generate pedestrian proposals. As stated above, the performance of person Re-ID is usually influenced by some spatial variances. We therefore combine the pedestrian proposal net with a spatial transformer to form a PST (Pedestrian Space Transformer) in our SIPN to detect pedestrians and also do away with the spatial variance problem.

To share information and propagate gradients better, the DenseNet [10] architecture is utilized. DenseNet contains densely connected layers even between the first layer and the last layer. Such structure allows any layer in the network to make use of input and may propagate the gradients from loss function directly to the initial layer to avoid gradient vanish. With DenseNet, the model can extract deeper features for detection and classification.

In general person Re-ID research, some methods [11,12] use pair-wise or triplet distance loss functions to supervise the training process. But it can be considerably complex if the dataset is of a greater scale. Xiao *et al.* [7] proposed an Online Instance Matching (OIM) loss function with which we do not need to optimize each pair of instances. The OIM loss merges instances features of the same identity offline, which makes the training of the model easier, faster and better. On the contrary, original OIM loss merges all instances which may obtain disturbing features. This work modifies OIM to only merge ground truth features thereby improving the performance of person Re-ID.

To sum up, our work provides two main contributions. First, we design a PST in our SIPN to produce pedestrian candidate bounding boxes and prevent spatial variance of person samples; Second, the improved OIM loss is proposed to learn features more effectively and conducts to robuster performance for the person search task.

2 Proposed Method

In this paper, a unified SIPN framework is proposed to process pedestrian detection and person Re-ID jointly. We adopt DenseNet [10] as a feature extractor to extract shared features for both detection and Re-ID. The PST is incorporated into our model to improve the spatial invariance of the feature maps. Finally, an improved OIM loss is applied in the model to supervise the training.

2.1 Model Structure

The cascaded SIPN model consists of three main parts. As shown in Fig. 2, the first part is the Pre-extracting Net which extracts features from a whole scene image. The second part is the PST which generates the parameters to spatially transform the feature maps and down-samples them to a fixed size. The third part of the model is a Feature Sharing Net, which further extracts features to be used for pedestrian detection as well as person Re-ID. Here are the details of our model whose structure is based on DenseNet.

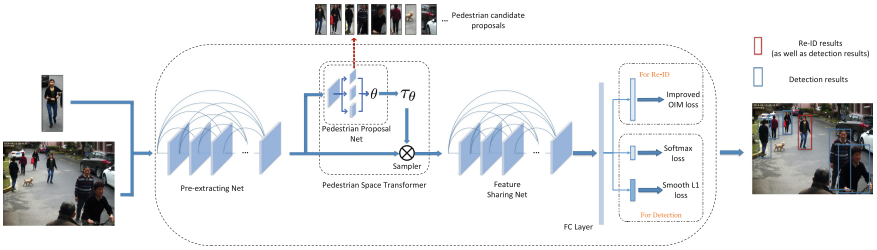

Fig. 2. Our cascaded DenseNet-based structure SIPN for processing pedestrian detection and person Re-ID jointly. Pre-extracting Net extracts low-level features which are then fed to our PST to produce pedestrian proposals and apply spatial transformations. Feature Sharing Net extracts further down-sampled features to output results for both detection and Re-ID.

There is a 7×7 convolutional layer in the front of Pre-extracting Net, followed by 3 dense blocks with 6, 12, and 24 dense layers respectively. The growth rate set in this paper is 32. We removed the initial pooling layer to make sure that input images would be pooled 4 times (by 2×2 max-pooling layer) through the Pre-extracting Net. The output will have a 1/16 resolution of the original image with 512 channels.

The Next part is our PST. It consists of a Pedestrian Proposal Net to generate pedestrian proposals and a spatial Transformer to implement transformation to these proposals. The structure of PST will be further illustrated in the next subsection.

The Features Sharing Net, a dense block with 16 dense layers and a growth rate of 32, is then built on the top of these feature maps to extract shared features. These feature maps are then pooled to 1024-dimensional vectors by an average pooling layer. And we raise 3 fully connected layers to map the vectors to 256D, 2D and 8D respectively.

At the end of the model, a Softmax classifier is used to output the classification (pedestrian or not) and a linear regressor is utilized to generate the corresponding refined localization coordinates. As for the 256 dimensional vectors, they will firstly be L2-normalized and then compared with corresponding feature vectors of target person for inference.

2.2 Pedestrian Space Transformer

In person search scenes, the performance will inevitably be influenced by the spatial variance of the samples. There are a number of reasons that can lead to increased spatial variance such as viewpoints, lights, occlusions, resolutions and so on. In an attempt to prevent spatial variances, we propose a PST to apply spatial transformations on the feature maps produced by the Pre-extracting Net.

At first, it is necessary to localize pedestrians by using the Pedestrian Proposal Net. Inspired by Faster R-CNN [13], we predict nine manually designed anchors with different scales and aspect ratios at each position of the feature map

extracted by a 3×3 convolutional layer. Then we feed the feature into three kinds of 1×1 convolutional layers (as shown in Fig. 2) to predict the scores, coordinate offsets and transformation parameters of the anchors. The three convolutional layers are called score layer, coordinate layer and parameter layer with two, four and six filters respectively, which means for each anchor, there will be two scores (pedestrian and background), four coordinates and six transformation parameters. However, it is unnecessary to use all these anchors (taking 51×35 as an example size of the feature map, there will be $51 \times 35 \times 9 \approx 160k$ anchors in total) predicted from the feature map as proposals because most of them are redundant. So, we sort the anchors by their scores and implement Non-Maximum Suppression (NMS) to pick out 128 anchors as the final proposals. Coordinate offsets and transformation parameters are selected correspondingly.

Then the PST will implement spatial transformations to these proposals with transformation parameters generate before. In order to perform spatial transformation for the input feature map $U \in \mathbb{R}^{H \times W \times C}$ with width W, height H and C channels, we compute output pixels at a particular location defined by the proposals. Specifically, the output pixels are defined to lie on a regular grid $G = \{G_i\}$ of pixels $G_i = (x_i^t, y_i^t)$, forming an output feature map $V \in \mathbb{R}^{H' \times W' \times C}$, where H' and W' are the height and width of the grid. The transformation τ_θ used in this paper is a 2D affine transformation described as

$$\begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} = \tau_\theta G_i = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

where (x_i^t, y_i^t) are the target coordinates of the regular grid in the output feature map, and (x_i^s, y_i^s) are the source coordinates in the input feature map.

Such a transformation allows to apply cropping, translation, rotation, scaling and skew operations to the input feature maps, and needs only 6 parameters θ to be produced by the Pedestrian Proposal Net. To perform a spatial transformation on the input feature map, a sampler must take the set of sampling points $\tau_\theta(G)$. Meanwhile, the pedestrian proposals with a range of scales also need a sampler to resize them to an identical size. In [13], this function is finished by the ROI-pooling layer. While in our SIPN, we use a robuster sampler which can also do spatial transformation along with the input feature map U to produce the sampled output feature map V. In detail, each (x_i^s, y_i^s) coordinate in $\tau_\theta(G)$ defines the spatial location in the input where a sampling kernel is applied to get the value at a particular pixel in the output V. Just like [14], that is defined as

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - m; \Phi_y) \tag{2}$$

where $i \in [1...H'W']$, $c \in [1...C]$, Φ_x and Φ_y are the parameters of a generic sampling kernel $k()$ which defines the image interpolation, U_{nm}^c is the value at location (n, m) in channel c of the input, and V_i^c is the output value for pixel i at location (x_i^t, y_i^t) in channel c. Taking bilinear sampling kernel as an example, we can reduce Eq. 2 to

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \tag{3}$$

and the partial derivatives are

$$\frac{\partial V_i^c}{\partial U_{nm}^c} = \sum_n^H \sum_m^W \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \tag{4}$$

$$\frac{\partial V_i^c}{\partial x_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |y_i^s - n|) \begin{cases} 0 \text{ if } |m - x_i^s| \geq 1 \\ 1 \text{ if } m \geq x_i^s \\ -1 \text{ if } m < x_i^s \end{cases} \tag{5}$$

$$\frac{\partial V_i^c}{\partial y_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - n|) \begin{cases} 0 \text{ if } |m - y_i^s| \geq 1 \\ 1 \text{ if } m \geq y_i^s \\ -1 \text{ if } m < y_i^s \end{cases} \tag{6}$$

This gives us a differentiable sampling mechanism that allows gradients to flow back.

As stated above, SIPN uses such PST to prevent spatial variance of detected proposals and extract robuster features for Re-ID in person search. It should be noted that there are a batch of transformation parameters matrices τ_θ (128 in our work) for an input image.

2.3 Improved OIM

Common loss functions in person Re-ID, such as verification loss or identification loss, are not suitable for the person search problem. Moreover, in person search datasets, unlabeled identities that have no target to match to are redundant for Re-ID and can not be taken into consideration.

To deal with these problems, we propose an improved OIM loss that contains directional constraints and takes advantages of the unlabeled identities features. OIM loss presented by Xiao *et al.* [7] makes a Look-Up Table (LUT) for labeled identities. Suppose there are L labeled identities in the dataset, and we have the LUT $V \in \mathbb{R}^{D \times L}$, where D denotes the dimension of final normalized features. Given an output normalized feature vector $x \in \mathbb{R}^D$, we can compute the similarity with all L identities by $V^T x$. During the back propagation, if the label ID of x is t, then we update the t-th column of the LUT by $v_t \leftarrow \gamma v_t + (1 - \gamma)x$, where $\gamma \in [0, 1]$, and then L2-normalize v_t. On the other hand, a circular queue is built to store features of those unlabeled identities. It is possible to customize the length of the circular queue, for example Q. Denoting the circular queue by $U \in \mathbb{R}^{D \times Q}$, it is simple to compute similarities between x and those unlabeled identities by $U^T x$. If x belongs to any unlabeled identities, we may push x into U and pop the out-of-date feature to update the circular queue. In a word, LUT is used to store representative features for the labeled identities to guide the matching process.

However, even after NMS, there are still lots of bounding boxes that are regarded as labeled identities in an image. Although these bounding boxes have different overlaps with ground-truth ones, OIM uses all their features to update the LUT. However, bounding boxes with background areas or lacking important information may corrupt the features to be merged and their features are naturally unrepresentative for those identities. Therefore, by denoting ground-truth bounding boxes as $\mathbb{G}$, we only merge features of the bounding boxes by

$$v_t \leftarrow \gamma v_t + (1-\gamma)x, \quad \text{if } x \in \mathbb{G} \tag{7}$$

to make the LUT robust and further reduce calculations for Re-ID in this paper.

Following [7], the probability of x being recognized as the identity with class-id i is defined by a Softmax function

$$p_i = \frac{\exp(v_i^T x)}{\sum_{j=1}^{L} \exp(v_j^T x) + \sum_{k=1}^{Q} \exp(u_k^T x)}, \tag{8}$$

and the probability of being recognized as the i-th unlabeled identity in the circular queue is

$$q_i = \frac{\exp(u_i^T x)}{\sum_{j=1}^{L} \exp(v_j^T x) + \sum_{k=1}^{Q} \exp(u_k^T x)}. \tag{9}$$

The improved OIM objective is to maximize the expected log-likelihood

$$\mathcal{L} = \mathrm{E}_x[\log(p_t)]. \tag{10}$$

3 Experiments

To demonstrate the effectiveness of our method, several comprehensive experiments were conducted on two public person search datasets.

3.1 Datasets

CUHK-SYSU. CUHK-SYSU was collected by hand-held cameras and movie snapshots. It contains 18,184 images and 96,143 pedestrians bounding boxes in total. Among all the pedestrians bounding boxes, there are 8,432 labeled identities and 5,532 of them are used as training split.

Person Re-identification in the Wild. Person Re-identification in the Wild (PRW) [8] dataset was transferred from raw videos collected in Tsinghua University. A total of 6 cameras were used, among which five are 1080×1920 HD and one is 576×720 SD. PRW consists of 11,816 whole scene images including 43,110 pedestrian bounding boxes, among which 34,304 pedestrians are annotated with an ID. For the training set, we pick up 483 identities.

3.2 Evaluation Metrics

We split the two datasets both into a training subset and a test subset following the principles that there are no similar identities between two subsets. During the inference stage, given an input target person image, the aim is to find the same person from the gallery. The gallery must include all whole scene images that contain the pedestrian samples of the target person. To make the problem more challenging, we can customize the size of the gallery ranging from 50 to 4,000 for CUHK-SYSU and 1,000 to 4,000 for PRW.

Following the common person Re-ID researches, the cumulative matching characteristics (CMC top-K) metric is used in our person search problem. CMC counts a matching as there are at least one of the top-K predicted bounding boxes overlaps with the ground-truths with intersection-over-union (IoU) lager or equal to 0.5. We also adopt mean averaged precision (mAP) as used on ILSVRC object detection criterion [15] as the other metric. An averaged precision (AP) is computed for each target person image based on the precision-recall curve. Therefore, mAP is the average APs across all target person images.

3.3 Training Settings

Our experiments are carried on PyTorch [16]. Using batch size 1 like Faster R-CNN [13], we apply stochastic gradient descent (SGD) to train models for 6 epochs. The initial learning rates are set to 0.0001 for both datasets and are divided by 10 after 2 and 4 epochs gradually. As a comparison, we adopt both VGG, ResNet [17] and DenseNet [10] as our CNN-based architectures for experiments. We use Softmax loss and Smooth L1 loss to supervise the detection process during the training stage. On the other hand, an improved OIM loss is utilized to supervise the Re-ID process, for which the sizes of circular queue are set to 5000 and 500 for CUHK-SYSU and PRW respectively.

3.4 Results

We make comparisons with some priori works including separate pedestrian detection & person Re-ID methods and three unified approaches. In the separate methods, there are three main detectors, ACF, CCF and Faster R-CNN [13] (FRCN). We also use several popular Re-ID feature representations like DenseSIFT-ColorHist (DSIFT), Bag of Words (BoW) and Local Maximal Occurrence (LOMO) along with some distance metrics such as Euclidean, Cosine similarity, KISSME, and XQDA. As for the unified approaches, Xiao *et al.* [7] proposed an end-to-end method based on Faster R-CNN and ResNet-50. Yang *et al.* [18] added hand-crafted features. Liu *et al.* [9] used an LSTM-based attention model. Tables 1 and 2 shows the results conducted on CUHK-SYSU dataset with a gallery size of 100. We also carry experiments on the more challenging dataset - PRW with gallery size 1000. Table 3 shows the results.

From Tables 1, 2 and 3 we can see that our model outperforms separate pedestrian detection & person Re-ID methods as well as the three unified frameworks.

Table 1. CMC top-1 results comparisons between our method and priori works on CUHK-SYSU dataset.

CMC top-1 (%)	CCF	ACF	FRCN
DSIFT+Euclidean [7]	11.7	25.9	39.4
DSIFT+KISSME [7]	13.9	38.1	53.6
BoW+Cosine [7]	29.3	48.4	62.3
LOMO+XQDA [7]	46.4	63.1	74.1
OIM (Baseline) [7]	–	–	78.7
Yang *et al.* [18]	–	–	80.6
NPSM [9]	–	–	81.2
Ours	–	–	**86.0**

Table 2. mAP results comparisons between our method and priori works on CUHK-SYSU dataset.

mAP (%)	CCF	ACF	FRCN
DSIFT+Euclidean [7]	11.3	21.7	34.5
DSIFT+KISSME [7]	13.4	32.3	47.8
BoW+Cosine [7]	26.9	42.4	56.9
LOMO+XQDA [7]	41.2	55.5	68.9
OIM (Baseline) [7]	–	–	75.5
Yang *et al.* [18]	–	–	77.8
NPSM [9]	–	–	77.9
Ours	–	–	**85.3**

It demonstrates that by spatially transforming feature maps of pedestrians, the performance of person search can be improved.

3.5 Pedestrian Detection Results

We also acquire pedestrian detection results on both CUHK-SYSU and PRW datasets as shown in Table 4. It can be observed that our method does not harm the performance of pedestrian detection in person search tasks.

4 Discussion

4.1 Influence of PST

Ren *et al.* [13] used RPN to produce proposals that may contain objects. Sizes of the proposals are various but the network needs a fixed input size so that ROI pooling layer was proposed to pool features maps into identical sizes. In

Table 3. Comparison between baseline and our method on PRW dataset.

Method	mAP (%)	top-1 (%)
DPM-Alex+LOMO+XQDA [9]	13.0	34.1
DPM-Alex+IDE$_{\text{det}}$ [9]	20.3	47.4
DPM-Alex+IDE$_{\text{det}}$+CWS [9]	20.5	48.3
ACF-Alex+LOMO+XQDA [9]	10.3	30.6
ACF-Alex+IDE$_{\text{det}}$ [9]	17.5	43.6
ACF-Alex+IDE$_{\text{det}}$+CWS [9]	17.8	45.2
LDCF+LOMO+XQDA [9]	11.0	31.1
LDCF+IDE$_{\text{det}}$ [9]	18.3	44.6
LDCF+IDE$_{\text{det}}$+CWS [9]	18.3	45.5
OIM (baseline)	21.3	49.9
NPSM [9]	24.2	53.1
Ours	**39.5**	**59.2**

Table 4. Detection results comparison between baseline and our method on both datasets.

	CUHK-SYSU		PRW	
	Recall (%)	AP (%)	Recall (%)	AP (%)
Baseline	79.49	74.93	90.20	84.26
Ours	78.45	75.14	89.91	85.60

Table 5. Comparison between baseline and our model with PST on PRW dataset.

	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)
Baseline	21.3	49.9	72.9	81.5
Ours	**33.9**	**51.9**	**73.7**	**82.3**

our SIPN, the PST has the similar function to resize feature maps. Our PST is also able to prevent the spatial variance problem caused by different resolutions, viewpoints and occlusions. Table 5 shows that the performance of our PST is better than the model proposed by Xiao *et al.* [7]. When comparing the influence of PST, we maintain the same network architecture, ResNet-50, as Xiao *et al.*

4.2 Comparison Between Different Network Architectures

In this paper, we adopt DenseNet [10] as our CNN-based architecture to extract high-level features and compare the results between multiple structures such as VGG, ResNet and DenseNet as shown in Table 6. It can be observed that DenseNet, which connects layers densely and utilizes multiple-scale feature maps, outperforms a little bit.

Table 6. Comparisons between multiple network structures on PRW dataset. Res34, Res50 and Dense121 refer to ResNet-34, ResNet-50 and DenseNet-121 respectively.

	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)
VGG16	20.8	35.3	59.8	70.3
Res34	31.1	49.1	71.5	79.2
Res50	33.9	51.9	73.7	82.3
Dense121	**38.7**	**58.0**	**76.8**	**82.9**

4.3 Results on the Improved Loss Function

Xiao *et al.* [7] used the OIM loss to reduce calculations by merging features of the same identity. However, those features belonging to proposals produced by the model may be disturbing because they may have some redundant information such as backgrounds or lack some essential information. In this work, we suggest to just merge features of ground-truth bounding boxes. Obviously, as shown in Table 7, our improved loss function has better performance since we only use ground-truth features to update the LUT, which makes it much robuster.

Table 7. Comparisons between using OIM loss and our improved one on PRW.

	mAP (%)	top-1 (%)	top-5 (%)	top-10 (%)
OIM loss	38.7	58.0	76.8	82.9
Ours	**39.5**	**59.2**	**77.6**	**83.4**

5 Conclusion

In this paper we propose a cascaded DenseNet-based framework SPIN to process pedestrian detection and person re-identification jointly. The PST is used to generate pedestrian proposals and avoid spatial variance challenges of person search task. After comparing several different network architectures, we adopt DenseNet as our base model to extract much richer features. At last, an improved OIM loss function with directional constrains is utilized to further improve the performance of our model.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (NSFC, Grant No. 61771303 and 61671289), Science and Technology Com- mission of Shanghai Municipality (STCSM, Grant Nos. 17DZ1205602,18DZ1200102), and SJTU-Yitu/Thinkforce Joint laboratory for visual computing and application.

References

1. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1259–1267, June 2016
2. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
3. Ess, A., Müller, T., Grabner, H., Van Gool, L.J.: Segmentation-based urban traffic scene understanding. In: BMVC, vol. 1, p. 2. Citeseer (2009)
4. 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)
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. 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)
7. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3376–3385. IEEE (2017)
8. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q., et al.: Person re-identification in the wild. In: CVPR, vol. 1, p. 2 (2017)
9. Liu, H., et al.: Neural person search machines. In: ICCV, pp. 493–501 (2017)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, no. 2, p. 3 (2017)
11. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
12. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel CNN with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)
13. 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)
14. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
15. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**, 211–252 (2015)
16. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)
17. 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)
18. Yang, J., Wang, M., Li, M., Zhang, J.: Enhanced deep feature representation for person search. In: Yang, J. (ed.) CCCV 2017. CCIS, vol. 773, pp. 315–327. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-7305-2_28