End-to-End Deep Learning for Person Search

Tong Xiao[1*] Shuang Li[1*] Bochao Wang[2] Liang Lin[2] Xiaogang Wang[1]
[1]The Chinese University of Hong Kong
[2]Sun Yat-Sen University
{xiaotong,sli,xgwang}@ee.cuhk.edu.hk, wangboch@mail2.sysu.edu.cn, linliang@ieee.org

Abstract

Existing person re-identification (re-id) benchmarks and algorithms mainly focus on matching cropped pedestrian images between queries and candidates. However, it is different from real-world scenarios where the annotations of pedestrian bounding boxes are unavailable and the target person needs to be found from whole images. To close the gap, we investigate how to localize and match query persons from the scene images without relying on the annotations of candidate boxes. Instead of breaking it down into two separate tasks—pedestrian detection and person re-id, we propose an end-to-end deep learning framework to jointly handle both tasks. A random sampling softmax loss is proposed to effectively train the model under the supervision of sparse and unbalanced labels. On the other hand, existing benchmarks are small in scale and the samples are collected from a few fixed camera views with low scene diversities. To address this issue, we collect a large-scale and scene-diversified person search dataset, which contains 18,184 images, 8,432 persons, and 99,809 annotated bounding boxes. We evaluate our approach and other baselines on the proposed dataset, and study the influence of various factors. Experiments show that our method achieves the best result.

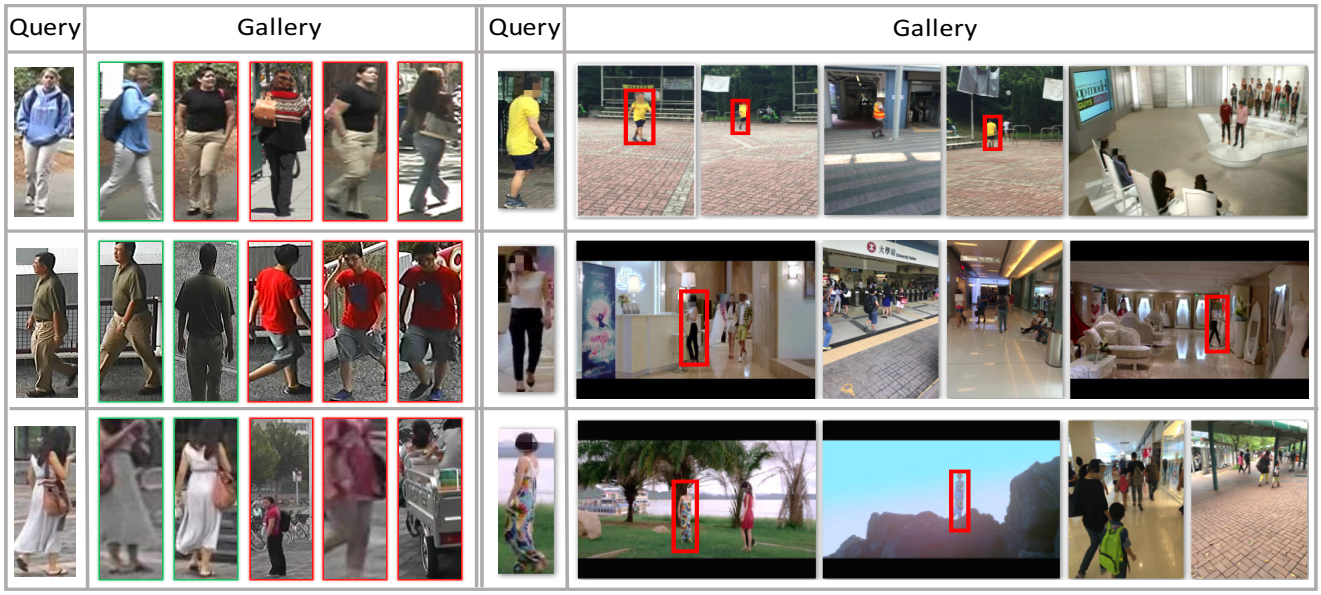
1. Introduction

Person re-identification (re-id) targets on matching pedestrian images across camera views. It is a fast growing research area [41, 18, 20] and has many important applications in video surveillance and multimedia, such as pedestrian retrieval [22], cross-camera visual tracking [34], and activity analysis [36]. This problem is particularly challenging because of complex variations of viewpoints, poses, lighting, occlusions, resolutions, background clutter and camera settings.

Although numerous re-id datasets and algorithms have been proposed in recent years and the performance on these benchmarks have been improved substantially, there is still a big gap with practical applications. In most benchmarks [18, 38, 9, 39, 22, 17], a query person is matched with manually cropped pedestrians in the gallery (as shown in Fig. 1(a)) instead of searching for the target person from whole images. Under the protocols of these benchmarks, the developed re-id algorithms assumed perfect pedestrian detection. However, the annotations of pedestrian bounding boxes are unavailable in real-world scenarios. Existing pedestrian detectors will unavoidably produce false alarms, misdetections, and misalignments, which could harm the re-id result. Under such circumstances, current re-id algorithms cannot be applied directly to real surveillance systems. Thus, it is urgent and important to solve this problem by proposing new benchmarks and effective approaches for person search, *i.e.*, searching for target persons in whole images, which is more difficult than the challenges posed by existing re-id benchmarks. Some targets could be missed in the first detection step. Therefore, the challenges from a large number of false alarms and misdetections cannot be fully studied based on these benchmarks.

To close the gap between traditional re-id research and practical applications, we investigate how to localize and match query persons from the scene image without relying on the candidate annotations. Different from conventional approaches that break down this person search problem into two separate tasks—pedestrian detection and person re-identification, we propose an end-to-end deep learning framework to jointly handle the challenges from both the aspects. Joint optimization brings multiple benefits, for example, the learned detector could allow some false alarms that can be easily handled by re-id, while focusing on other hard samples. On the other hand, the detector and re-id parts can better fit each other to reduce the influence of detection misalignments. We share a fully convolutional neural network to extract features for detecting pedestrians and producing discriminative re-id features. A random sampling softmax loss is proposed to effectively train the model under the supervision of sparse and unbalanced labels. By uni-

*Tong Xiao and Shuang Li are co-first authors with equal contributions.



(a) Existing Datasets (b) Our Dataset

Figure 1. Samples of pedestrians in existing benchmarks and our dataset. Query in each row of (a) is collected from VIPeR [9], CUHK03 [18], and Market-1501 [38], respectively. Their gallery sets consist of manually cropped pedestrians with green border representing correct matching while red border is wrong. Samples in (b) are from our dataset. For each cropped query person, some gallery images contain the target person (labeled by red bounding boxes) while the rest serve as distractors

fying the whole process into a single deep neural network, the searching performance and speed get substantially improved compared with traditional approaches.

To the best of our knowledge, existing benchmarks are small in scale and the samples are collected from a few fixed camera views, which have low scene diversities. To justify our proposed pipeline and also serve the community, we collect a large-scale and scene-diversified person search dataset. It includes $18,184$ images and $8,432$ query people. They are collected from hundreds of scenes from street and movie snapshots. We manually annotate all the people inside each image, resulting in $99,809$ bounding boxes, and match the same person across different images. Some example are shown in Fig. 1(b).

The contributions of this paper are summarized in three folds. (1) We propose an end-to-end deep learning framework to search for the query persons from whole images in the gallery, which is much closer to real applications. Instead of simply combining the pedestrian detector and person re-id, we jointly optimize these two objectives in a unitary process. Several training strategies are proposed to effectively train the model and the experimental results show that our framework outperforms other baselines. (2) We collect a large-scale benchmark dataset for person search, covering hundreds of scenes from street and movie snapshots. Rich annotations and protocols are also provided to facilitate the experiments on training and testing the models. The dataset and codes will be released to the public. (3) A full set of benchmark is established on our dataset, which consists of the performance of our method, as well as other pedestrian detection and person re-id baselines. We also study the influence of various factors empirically, including detection recall, gallery size, occlusion and resolution.

2. Related Work

Person Re-identification and Pedestrian Detection. Existing works of person re-identification focus on manually designing discriminative features [35, 11, 37], automatically learning features with convolutional neural networks (CNN) [18, 1], learning feature transforms across camera views [25, 24, 30], and learning distance metrics [40, 10, 26, 23, 21]. Li *et.al.* [18] and Ahmed *et.al.* [1] designed specific CNN models for person re-id. Both the networks utilized a pair of cropped pedestrian images as input and employed a verification loss function to train the parameters. Ding *et.al.* [5] utilized triplet samples for training CNNs to minimize the feature distance between the same person, while maximize the distance between different people. Several recent works addressed on improving person re-id performance on abnormal pedestrian images. Li *et.al.* [19] proposed a multi-scale metric learning method for re-id under low-resolution images, while Zheng *et.al.* [42] proposed a local-global matching framework for paritially occluded pedestrian images.

	Ours	Market-1501 [38]	CUHK03 [18]	VIPeR [9]	i-LIDS [39]	GRID [22]	CUHK01 [17]
# identities	**8,432**	1,501	1,360	632	119	250	971
# bboxes	**99,809**	32,643	13,164	1,264	476	500	1,942

Table 1. Statistics of our person search dataset and existing re-id datasets. Note that in our dataset, the bounding boxes are only used for training and evaluation. Person search methods will be tested with the whole images as input instead of the bounding boxes

For pedestrian detection, DPM [7] and ACF [6] were the most commonly used pedestrian detectors. They relied on hand-crafted features and linear classifiers to detect pedestrians. Recent years, CNN-based pedestrian detectors have also been developed. Various factors, including CNN model structures, training data, and different training strategies are studied empirically in [14]. Tian *et.al.* [33] utilized pedestrian and scene attribute labels to train CNN pedestrian detectors in a multi-task manner. Cai *et.al.* [4] proposed a complexity-aware boosting algorithm for learning CNN detector cascades.

Datasets and Benchmarks. Many person re-identification datasets have emerged in recent years, including VIPeR [9], ETHZ [29], i-LIDS [39], PRID2011 [13], RGB-D [2], GRID [22], CUHK01 [17], CUHK02 [16], and Multi-Camera Surveillance Database [3]. However, these datasets provided only manually cropped pedestrian images, and thus are not suitable for our person search problem. CUHK03 [18] provided pedestrian bounding boxes generated by the Deformable Part-based Model (DPM) [7] aside from manually cropped images. But the detection false alarms were manually removed. It can only be used to evaluate the influence of detection misalignments. Market-1501 [38] consisted of more than $32,000$ DPM detected bounding boxes, including both the true pedestrian detections and some false alarms. However, they did not provide the original whole images, thus if a person was missed by the detector, he/she was also excluded from the query and gallery sets. The influence of misdetections was not clear. On the other hand, since it fixed DPM as the detector, we cannot evaluate how different detectors would affect the performance of the person search algorithms. It is also impossible to develop end-to-end learning methods for person search based on these benchmarks.

From another perspective, the numbers of camera views in the above datasets are small (< 10). It is uncertain about the generalization capability of an algorithm given a pair of new camera views without extra training samples. In our dataset, the camera views are not fixed. It contains various viewpoints and background scenes to test the robustness of the algorithms. We summarize the statistics of our dataset and some other existing ones in Table 1.

3. A New Benchmark for Person Search

To close the gap between traditional re-id methods and real application scenarios, we contribute a large-scale benchmark dataset for comprehensive evaluation of person search from whole images. The dataset can be divided into two parts according to the image sources: street snaps and movies. For street snaps, $12,490$ images and $6,057$ query persons are collected with hand-held cameras across hundreds of scenes. Note that the total number of pedestrians contained in the $12,490$ images is much larger than $6,057$. Each selected query person appears in at least two images captured from different viewpoints. We have made efforts on including variations of viewpoints, lighting, resolutions, occlusions, and background as much as possible during image collection, in order to intensively reflect the real application scenarios and increase scene diversity.

We choose movies and TV dramas as another source for collecting images, because they provide more diversified scenes and more challenging viewpoints. Some examples are shown in Fig. 1(b). $5,694$ images and $2,375$ query persons are selected from movies and TV dramas. We exclude query persons who appeared with half bodies or abnormal poses such as sitting or squatting. Query persons who change clothes and decorations in different videos frames are not selected in our dataset, since person re-id and person search recognize identities based on body shape and clothes. It is also ensured that the selected video frames are from different scenes or had large variation on viewpoints.

In order to evaluate the influence of detectors on the re-identification performance, the bounding boxes of all the pedestrians whose heights larger than 50 pixels are manually annotated. The same persons appeared in different images are associated.

3.1. Evaluation Protocols

The dataset contains $18,184$ images and $8,432$ selected query persons in total. Each query person appears in at least two images. Each image may contain more than one query person and many background people. The data is partitioned into a training set and a test set. The training set contains $11,206$ images and $5,532$ query persons. The test set contains $6,978$ images and $2,900$ query persons. The training and test sets have no overlap on images or query persons.

We design several evaluation protocols for testing. For

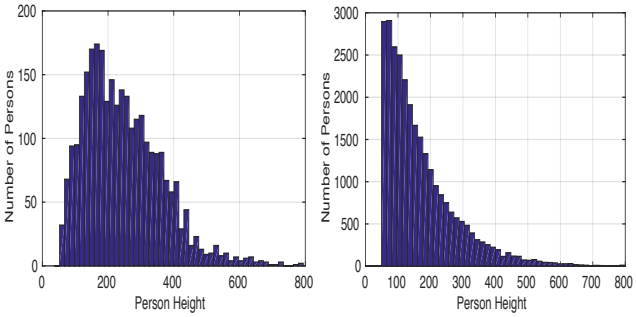


Figure 2. Distributions of the heights of test persons (left) and background persons (right) in test images. We keep the pedestrians higher than 50 pixels. The heights of persons spread in a wide range. The selected query persons appear larger than background persons in general

(a) (b) (c)

Figure 3. Examples of query persons from different subsets: (a) high resolution without occlusion, (b) occlusion, (c) low-resolution

each test person, its query image is cropped from one of the test images containing the person. Its gallery set includes the other test images containing the query person, as well as some randomly selected test images not containing the query person. Different queries have different gallery sets. Moreover, our test images are captured from many scenes and viewpoints. Therefore, it is very difficult for the person search algorithms to overfit to particular scenes, camera views, or gallery sets. In order to evaluate the influence of gallery size, we have designed different evaluation protocols by setting the gallery size to 50, 100, 500, $1,000$, $2,000$, and $4,000$. Each image contains 5.3 background persons on average. If the gallery size is set to 100, a query person has to be distinguished from around $5,300$ background persons and thousands of non-pedestrian bounding boxes, which is challenging. Fig. 2 shows the distributions of bounding box heights of test query persons and background persons in test images. We can see that the heights of persons spread in a large range, and the selected query persons are larger than background persons in general. Note that we only annotate pedestrians higher than 50 pixels. In our experiments, pedestrian detectors do not evaluate bounding boxes whose heights are smaller than 50.

Low-resolution Subset. In order to evaluate the influence of resolution on person search, we construct a subset which contains query persons of low resolutions. Fig. 2(a) shows the distribution of the heights of bounding boxes for query persons and we take 10% of the query persons with the smallest heights into this subset. This subset contains 290 cropped low-resolution pedestrian regions as queries. Again, each query has a different gallery set. Some examples of selected low-resolution query persons are shown in Fig. 3(c) and high resolution samples are shown in Fig. 3(a) for comparison. There is a separate protocol to evaluate person search with low-resolution queries on this subset.

Occlusion Subset. Occlusion is another factor affecting person search. We identify 187 occluded query persons from the test set and add them to this subset. The occlusion of each person is larger than 40%. Some examples are shown in Fig. 3(b).

3.2. Evaluation Metrics

Person search is different from re-id which only requires matching with cropped images in the gallery. Therefore, two evaluation metrics are used for person search. The first metric is mean Averaged Precision (mAP). It treats person search as a detection problem, *i.e.* detecting the query person from all the images in the gallery. For each query person, there are only a few ground truth bounding boxes among millions of candidate windows in the gallery. A candidate window is considered as positive if its overlap with the ground truth is larger than 0.5. The mAP is a commonly used metric for object detection. For each query person, an AP is calculated from its Precision-Recall curve, and the mAP is calculated by averaging APs over all the queries.

The second metric is the top-k matching rate on bounding boxes. A matching is counted if a bounding box among the top-k predicted boxes overlaps with the ground truth larger than the threshold. It has the constraint that each ground truth can at most nominate one bounding box which has the largest matching score as true. This metric is different from mAP, since it treats person search as a ranking and localization problem. The metric is motivated by the classification/localization task of ImageNet [28].

4. Method

To address the problem of person search, we propose an end-to-end deep learning framework which jointly handles the pedestrian detection and the person re-identification. As shown in Fig. 4, the framework consists of three parts. First, we utilize a fully convolutional neural network (FCN) to extract feature maps from an input image of arbitrary size. Then, a pedestrian proposal network is built on the top of the feature maps to predict pedestrian bounding boxes. At

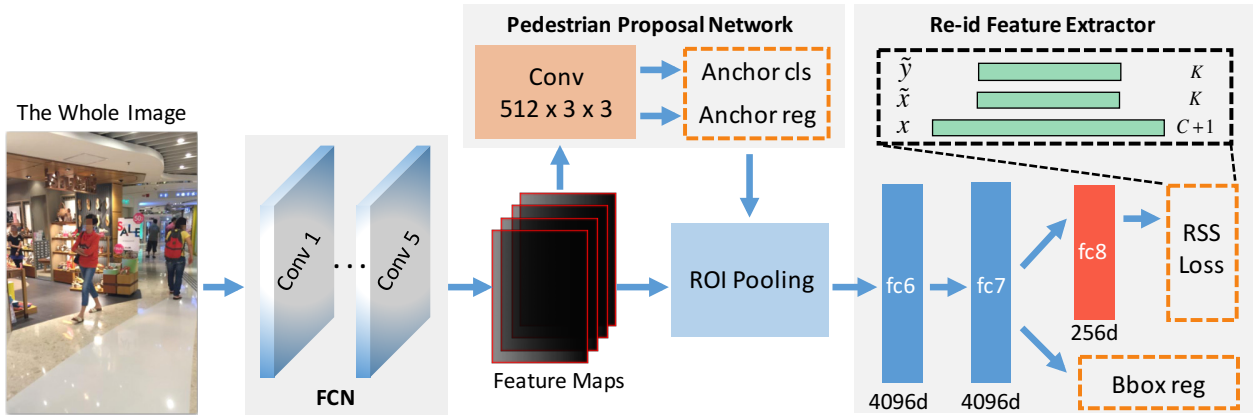


Figure 4. Overview of our framework. Given a whole image, we first utilize a fully convolutional network to extract feature maps. Then we deploy a convolution layer with 512 3×3 filters on the top of the feature maps, followed by sibling anchor classification (denoted by Anchor cls) and regression layers (denoted by Anchor reg) to predict pedestrian ROIs. These ROIs are then used to pool the feature vector for each candidate box on the convolutional feature maps. Three fully connected layers are utilized to produce the final feature vector (fc8) for computing distances. The boxes with dashed orange borders represent the loss layers

last, for each confident candidate box, we use the ROI pooling technique [12] to pool a fixed-length feature vector inside its region on the convolutional feature maps, followed by several fully connected layers to produce the final feature vector for re-id. By sharing the FCN for pedestrian detection and re-id feature extraction, we could accelerate the person search process. In the following, we will describe the network structure, as well as the training and test procedures.

4.1. Model Structure

We adopt the commonly used VGG16 [31] model for our convolutional layers (conv1 to conv5) in FCN and the first two fully connected layers (fc6 and fc7) in re-id feature extraction. For the pedestrian proposal network, we add a $512 \times 3 \times 3$ convolutional layer on the top of the feature maps, and follow [27] to associate 9 anchors at each feature map location. The module will then predict the score of person / non-person and the bounding box regression result for each anchor.

After the ROI pooling, we summarize the features for each pedestrian proposal region by using two fully connected layers (fc6 and fc7). However, the dimension of fc7 is 4096, which slows down the feature distance computation for re-id. Therefore, we append another fully connected layer (fc8) after fc7, which reduces the dimension to 256 and serves as the final feature vector for re-id. Moreover, another bounding box regression layer is employed on the top of fc7 to further improve the localization ability for better box alignments.

4.2. Training Phase

At the training phase, several loss functions are deployed to train the network to detect pedestrians and produce discriminative features for re-id. We use the smoothed-L1 loss [8] for the two bounding box regression layers. For the pedestrian proposal module, a softmax loss is employed to classify pedestrian / non-pedestrian. While for the re-id feature extraction part, we also add a softmax loss layer on the top of fc8, but it aims at classifying the person identities. It is shown in several previous works that such kind of classification task could greatly benefit the feature learning [32]. The feature representation in fc8 has to be highly discriminative, since it is required to distinguish a large number of identities in the training set. Specific to our framework, we further include a background class along with the identities to suppress false alarms generated by the pedestrian proposal net. The overall loss is the sum of the previous four loss functions, and the gradients w.r.t. the network parameters are computed through backpropagation.[1]

A key challenge to train the proposed framework lies in the identity softmax classification loss. First, our training set has 5532 identities, thus the softmax target is very sparse. Second, due to computation cost on large images, each minibatch consists of only two input scene images, which often contain no more than ten different training identities. Thus the minibatch label distribution mismatches dramatically with the dataset label distribution and lacks diversity. When combined together, these two issues make it extremely difficult for the net to receive proper gradients. In practice, we found that the training loss would not decrease if we directly finetune the network from the ImageNet pretrained VGG16 model.

To overcome this problem, we first need to set a good initial point for the softmax classifier. Specifically, we crop the

[1] In practice we did not backpropagate the losses of the re-id parts to the pedestrian proposal module through the ROI pooling layer.

ground truth bounding boxes for each training person and randomly sample the same number of background boxes. Then we shuffle the boxes, resize them to 224×224, and finetune the VGG16 model to classify the boxes with batch size of 256. This finetuning process works properly because of the rich label diversity inside each minibatch, and the resulting model is used as the initial point for training the whole framework.

Next, we propose a random sampling softmax (RSS) loss layer to replace the original one. For the original softmax loss, the gradients could favor only a few number of classes that appear in a minibatch, while severely suppress the other classes. The RSS loss layer solves this problem by randomly selecting a subset of softmax neurons for each input sample to compute the loss and gradients. Detailed formulation is given below.

Suppose the target classes are from 1 to $C+1$, where class $C+1$ is the background and the others are the identities. Denote each data sample by $\{x, t\}$, where $x \in \mathbb{R}^{C+1}$ is the classifier scores (input of the softmax) and t is a 1-of-$(C+1)$ binary vector representing the label. Then the original softmax loss can be written as

$$l = -\sum_{i=1}^{C+1} t_i \log y_i, \quad \text{where } y_i = \frac{e^{x_i}}{\sum_{j=1}^{C+1} e^{x_j}}. \tag{1}$$

The RSS loss will randomly select $K (K \ll C+1)$ dimensions from x and t to compute the loss and gradients. Suppose the selected indices are $i_1, i_2, \ldots, i_K$, the sampled classifier scores and label vector can be denoted by $\tilde{x} = (x_{i_1}, x_{i_2}, \ldots, x_{i_K})^T$ and $\tilde{t} = (t_{i_1}, t_{i_2}, \ldots, t_{i_K})^T$. Then the RSS loss function is defined as

$$\tilde{l} = -\sum_{i=1}^{K} \tilde{t}_i \log \tilde{y}_i, \quad \text{where } \tilde{y}_i = \frac{e^{\tilde{x}_i}}{\sum_{j=1}^{K} e^{\tilde{x}_j}}. \tag{2}$$

Suppose the label class is c, we define the following rules to generate the indices:

1. Set i_1 to c;
2. If $c = C+1$, sample $i_2, \ldots, i_K$ from $\{1, \ldots, C\}$ uniformly;
3. If $c \neq C+1$, set i_2 to $C+1$, and sample $i_3, \ldots, i_K$ from $\{1, \ldots, C\} \setminus c$ uniformly.

These rules make sure that the background class will always be selected. This helps the net to better distinguish between identities and false alarms predicted by the detector net. K is a hyperparameter that controls how many classes will be suppressed. We find that choosing $K = 100$ works well in practice.

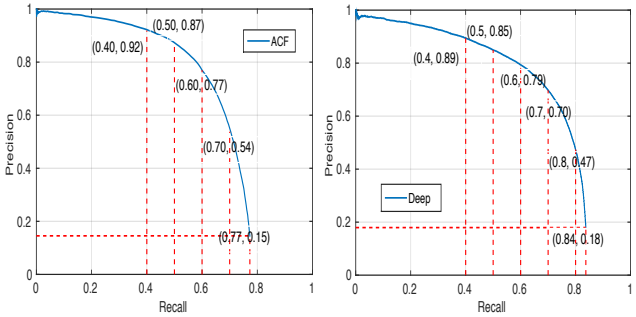


Figure 5. The precision-recall curves of the ACF and the Deep pedestrian detectors on our test set. Some (recall, precision) data points are annotated on the curves

4.3. Test Phase

During the test phase, for each gallery image, we can get the features (fc8) of all the candidate pedestrians by performing the network forward computation once. While for the query image, we replace the pedestrian proposals with the given bounding box, and then do the forward computation to get its feature vector (fc8). After that, we compute the pairwise Euclidean distances between the query features and those of the gallery candidates, which are utilized later for re-id evaluation. Notice that we decouple the feature computations for the query and gallery images, which is different from previous deep learning re-id approaches [18, 1]. Thus the gallery features can be reused for other queries, which further accelerates the search process. Be reminded that the purpose of the net in Fig. 4 is to train the feature extractor for re-id. Although C identities are used in training, the feature extractor can be applied to new identities outside the training data. It has been proved to be effective in face verification applications [32].

5. Experiments

To investigate the influence of various factors on person search and demonstrate the effectiveness of our proposed approach, we conduct several groups of experiments on the new benchmark. In this section, we first detail the baseline methods and experiment settings in Section 5.1. Then we evaluate the baselines with separate detection and re-id in Section 5.2. Section 5.3 shows the effectiveness of our approach, and some ablation studies are conducted in Section 5.4. At last, we present the influence of various factors, including detection recall, gallery size, person occlusion, and image resolution.

5.1. Baseline Methods and Experiment Settings

A straightforward solution to searching a person inside an image is to break down the problem into two subtasks—pedestrian detection and person re-identification, and choose algorithms separately for each task. Therefore,

	Euclidean		KISSME [15]		BoW [38]		IDNet		Ours	
	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1
ACF [6]	21.7	25.9	32.3	38.1	42.4	48.4	47.3	53.3		
Deep	28.0	33.0	39.1	44.9	48.3	54.1	52.6	59.0	**55.7**	**62.7**
GT	41.1	45.9	56.2	61.9	62.5	67.2	66.5	71.1		

Table 2. Comparisons between our approach and four person re-identification methods under two pedestrian detectors (ACF and Deep), as well as the ground truth boxes (GT). The DenseSift+ColorHist [37] features are used for Euclidean and KISSME

we combine several existing pedestrian detectors and person re-id methods as our baselines.

For the pedestrian detection task, ACF [6] is a widely used pedestrian detector. We train it on our dataset by using the publicly available tools provided by the authors. On the other hand, the FCN and pedestrian proposal network in our framework can also be treated as a detector. We thus discard the re-id parts and train only the remaining layers. The resulting model serves as another baseline pedestrian detector (Deep). The precision-recall curves of these two detectors on our test set are shown in Fig. 5. At last, we utilize the ground truth bounding boxes as a perfect detector (GT) to compare with other detection results.

For the person re-identification task, we use the BoW [38] feature with cosine distance, as well as the DenseSift+ColorHist [37] feature (reduced to 256 dimension by using PCA) with Euclidean and KISSME [15] distance metric. The KISSME metric is learned on our training set. Moreover, by discarding the pedestrian proposal network in our framework and training the remaining net to classify identities from cropped pedestrian images, we get another baseline re-id method (IDNet). Note that its model capacity is approximately the same as our framework.

Our method is trained end-to-end that jointly handles the pedestrian detection and the person re-identification. The model structure and training strategies are listed in Section 4. The initial learning rate is set to 1e-3, and decreases to 1e-4 after 100K iterations. The training process converges at 120K iteration. In the following experiments, we will report the performance based on the test protocol where the gallery size is 100 if not specified.

5.2. Evaluation of Baselines with Separate Detection and Re-ID

The results of our approach and different baselines under two evaluation metrics are summarized in Table 2. We first conduct analysis on the combination of baseline detection and re-id methods. To understand how different pedestrian detectors would affect the person search results, we evaluate the performance of ACF and Deep detectors on our dataset by drawing their Precision-Recall curves in Fig. 5. Deep outperforms ACF in pedestrian detection and also leads to better person search performance when it is combined with different person re-id algorithms, as shown in Table 2. This is consistent with our common knowledge. Since a better pedestrian detector generates fewer false alarms, fewer misdetections, and more accurate boxes, it makes the re-id subtask much easier.

From Table 2, we also observe that the performance of using ground truth pedestrian bounding boxes (GT) are much better than using the other detectors. This verifies our conjecture that although existing re-id algorithms have achieved impressive results on various benchmarks, it does not mean that they are ready for real-world applications. Because detector remains to be one of the major factors that affect the person search performance.

On the other hand, the relative performance of different re-id algorithms are consistent across all the detectors. It implies that existing person re-id datasets could still have their research values, even though there is a gap between these datasets and real-world application scenarios. If a re-id algorithm performs better on manually cropped pedestrian bounding boxes (GT), it also has better performance on ACF and Deep. Although this phenomenon matches with our common cognition, it is not obvious without strong evidence. We are the first to provide detailed empirical analysis on the influence of detectors and re-id.

5.3. Effectiveness of Our Approach

From Table 2, we can see that our proposed framework achieves the best mAP and top-1 accuracy among all the methods, except for those utilizing ground truth bounding boxes. The only difference between our method and Deep+IDNet is that we jointly optimize the detector and re-id feature extractor in an end-to-end manner, while Deep+IDNet is trained separately on these two subtasks. The 3% performance gap between Deep+IDNet and our approach shows that with joint optimization, detector and re-id feature extractor could better fit to each other. Detector could allow some false alarms easily handled by re-id while focusing on other hard samples. Re-id could better adapt to the boxes generated by the detector and help reduce the false alarms and misalignments. Moreover, our approach is about $40\times$ faster than Deep+IDNet in practice, because sharing the underlying convolutional layers for both the pedestrian detection and the re-id feature extraction saves lots of com-

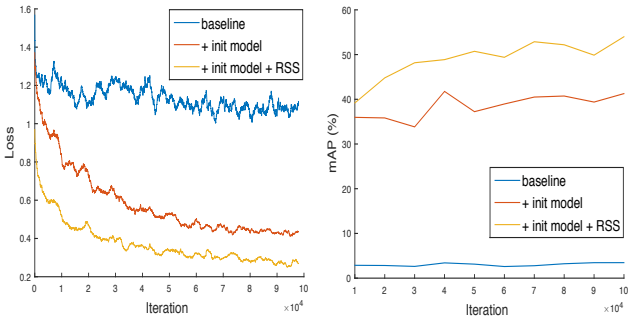
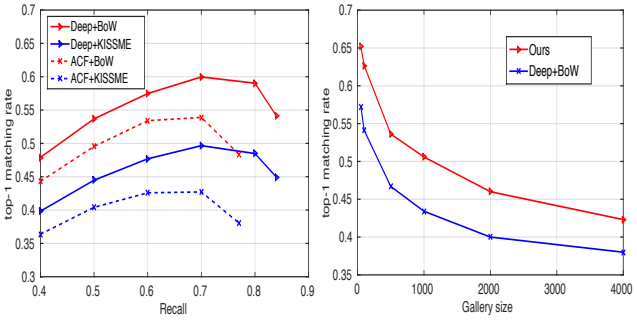


Figure 6. Training loss and test mAP of different models used in ablation experiments



(a) (b)

Figure 7. The top-k matching rates of (a) four baseline methods under different detection recalls, (b) our method and Deep+BoW with different gallery sizes

putation.

5.4. Ablation Studies of Our Training Strategies

To show the effectiveness of the training strategies proposed in Section 4.2, we conduct several ablation experiments. First, we establish a baseline by training the framework without the proposed strategies, *i.e.*, using the traditional softmax loss layer for identity classification and fine-tuning the whole net directly from the ImageNet pretrained VGG16 model. Next, we change the initial model from ImageNet pretrained VGG16 to the one stated in Section 4.2. At last, we replace the traditional softmax layer with the proposed RSS layer. We plot training loss and test mAP of each method w.r.t. the training iteration in Fig. 6.

From Fig. 6 we can see that the training loss of the baseline model remains at a high level throughout the training process. Setting a good initial point enables the loss to decrease, but the test mAP improves very slowly and converges to about 40%. This phenomenon shows that if the softmax target is very sparse and the minibatch contains only a few label classes, the gradients would be biased on these classes at each SGD iteration. Thus the whole network cannot be updated efficiently. However, after replacing the traditional softmax layer with the proposed RSS layer, we can find that the training loss decreases much faster and converges to a better local minimum. Meanwhile the test mAP increases significantly and converges to about 55%. This big performance improvement shows that our proposed RSS loss is effective to solve the problem.

5.5. Influence of Various Factors

Detection Recall. We investigate how detection recalls would affect the person search performance by setting different thresholds on detection scores. A lower threshold reduces the misdetections (increases the recall) but results in more false alarms. We choose different recall rates for the Deep and the ACF detectors, ranging from 40% to 90%, and use the generated bounding boxes to do the person re-id with BoW and KISSME. The results of top-1 accuracy w.r.t. the recall rate of different baseline combinations are shown in Fig. 7(a). It is observed that all baselines achieve the highest top-1 accuracy when the recall rate is 70%, which indicates that both the false alarms and misdetections could affect person search. We should find a best trade-off in real applications.

	Deep+KISSME		ACF+BoW		Ours	
	mAP	top-1	mAP	top-1	mAP	top-1
Whole	39.1	44.9	42.4	48.4	55.7	62.7
Occlusion	18.2	17.7	29.1	32.1	39.7	43.3
Low-resolution	43.8	42.8	44.9	53.8	35.7	42.1

Table 3. Experimental results of our algorithm and two combinations of detectors and person re-identification on the occlusion subset, low-resolution subset, and the whole test set

Gallery Size. Person search could become more challenging when the gallery size increases. We regard the gallery size as a hyperparameter and investigate its influence through a group of experiments. We vary the gallery size from 50 to 4,000, and test our approach and Deep+BoW accordingly. The results of top-1 accuracy w.r.t. the gallery size are shown in Fig. 7(b). It can be seen that when the gallery size increases from 50 to 4,000, more distractors are included and the top-1 accuracies drop by half for both the methods.

Occlusion and Resolution. We construct two subsets by selecting abnormal samples from the whole test set as stated in Section 3.1. One consists of low-resolution query persons and the other consists of partially occluded persons. The gallery size is fixed as 100 and several methods are evaluated on these subsets. The results are shown in Table 3. It is observed that all the methods perform significantly worse on the occlusion subset than on the whole test set. On the other hand, resolution is not a major factor that affects the re-id methods with hand-crafted features. The results of Deep+KISSME and ACF+BoW on this low-resolution subset are even better than those on the whole test set. However, for our method, since the ROI pooling is not effective in

extracting features inside small pedestrian bounding boxes, the results drop down on the low-resolution subset significantly. This phenomenon indicates that our method can still be improved to better handle small pedestrians.

6. Conclusions

In this paper, we target on the problem of searching query persons from whole images. Different from conventional approaches that break down the problem into two separate tasks—pedestrian detection and person re-identification, we develop an end-to-end deep learning framework to jointly handle both aspects with the help of the proposed random sampling softmax loss. A large-scale and scene-diversified person search dataset is contributed, which contains 18,184 images, 8,432 persons, and 99,809 annotated bounding boxes. We evaluate our approach and other baselines on the dataset, and our method of joint detection and re-id achieves the best result. We also study the influence of various factors. It is observed that detectors greatly affect the person search performance of baseline method and there is still a big gap between using the ground truth bounding boxes and the automatically detected ones. This phenomenon again shows the importance of jointly considering both the detection and the re-id for the person search problem.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2, 6

[2] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *ECCV*, 2012. 3

[3] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *Digital Image Computing Techniques and Applications*, 2012. 3

[4] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015. 3

[5] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *PR*, 2015. 2

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *TPAMI*, 2014. 3, 7

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 3

[8] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 5

[9] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007. 1, 2, 3

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008. 2

[11] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ICDSC*, 2008. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 2015. 5

[13] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*. 2011. 3

[14] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015. 3

[15] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 7

[16] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 3

[17] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 1, 3

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 3, 6

[19] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015. 2

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1

[21] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015. 2

[22] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009. 1, 3

[23] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 2

[24] F. Porikli. Inter-camera color calibration by correlation model function. In *ICIP*, 2003. 2

[25] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, 2008. 2

[26] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 2

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 2015. 5

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2014. 4

[29] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *Computer Graphics and Image Processing*, 2009. 3

[30] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2014. 5

[32] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 5, 6

[33] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*, 2015. 3

[34] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 2013. 1

[35] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 2

[36] X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *TPAMI*, 2010. 1

[37] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2, 7

[38] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 3, 7

[39] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009. 1, 3

[40] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 2

[41] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *TPAMI*, 2013. 1

[42] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, 2015. 2