End-to-End Detection and Re-identification Integrated Net for Person Search

Zhenwei He[0000−0002−6122−9277] and Lei Zhang[0000−0002−5305−8543]

School of Microelectronics and Communication Engineering, Chongqing University
{hzw,leizhang}@cqu.edu.cn

Abstract. This paper proposes a pedestrian detection and re-identification (re-id) integrated net (I-Net) in an end-to-end learning framework. The I-Net is used in real-world video surveillance scenarios, where the target person needs to be searched in the whole scene videos, and the annotations of pedestrian bounding boxes are unavailable. Comparing to the successful OIM method [31] for joint detection and re-id, we have three distinct contributions. First, we implement a Siamese architecture instead of one stream for an end-to-end training strategy. Second, a novel on-line pairing loss (OLP) with a feature dictionary restricts the positive pairs. Third, hard example priority softmax loss (HEP) with little computation cast is proposed to deal with the online hard example mining. We show our results on CUHK-SYSU and PRW datasets. Our method narrows the gap between detection and re-identification, and achieves a superior performance.

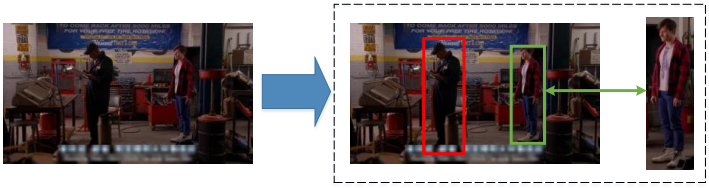
Keywords: Person Search · Deep Learning.

1 Introduction

Real-world video surveillance tasks such as criminals search [29], multi-camera tracking [26] need to search the target person from different scenes. In other words, the algorithms for real-world person search tasks are asked to find the target person from a whole image. Therefore, this problem is generally issued by two separate steps: person detection from single image and person re-identification (re-id). These two problems are challenging due to the influences of poses, viewpoints, lighting, occlusion, resolution, background *etc.* Therefore, they have been paid too much attention in recent research [3], [2], [27], [21].

Although numerous endeavor on person detection and re-identification has been made, most of them handle these two problems independently. The traditional methods divide the person search task into two sub-problems. First, a detector is implemented to predict the bounding boxes of the persons from the images. Second, the detected persons are cropped based on the bounding boxes for the further re-identification task. Actually, most advanced re-identification method are modeled on the manually cropped pedestrian images [3], [15], [14], and the manually cropped pedestrian samples are much better than the specially trained detector because of inevitable false detection. Additionally, person

(a) Traditional method



(b) Our method

Fig. 1. Comparison of the traditional method and our method. In the Figure 1.a, Traditional method need detect all persons in the picture and cropped them for the further re-id model. Our new method(I-Net) in Figure 1.b can find the target person from a whole image directly.

search task should be a joint work of detection and re-id, which requires the close cooperation of both two parts. Therefore, in this paper, we propose a new model to jointly train these two parts in a unified deep framework end-to-end, which aims to search the target person from whole image scene.

Specifically, to reduce the gap between traditional algorithms and practical applications, we propose an Integrated Net (I-Net) to simultaneously learn the detector and re-identifier by an end-to-end manner. Figure. 1 shows the difference between traditional scheme and the proposed I-Net in application scenario. Different from the traditional re-id method, which got a separated two steps(detection and re-identification) for the person search task, our I-Net can predict the location (bounding box) of the target person from a whole image. The joint learning of the detector and re-identifier in I-Net for person search brings lots of benefits: On one hand, the co-learning of the detector and re-identifier helps the re-id part to handle the misalignments of bounding boxes and the false positives provided by the detector, such that the re-id part can be more robust than independent training. On the other hand, the detection and re-id reuse the same features which can accelerate the search speed.

Triplet loss [24] is a loss function for the verification problems, it's widely used for the re-id tasks [6], [19], [7]. Triplet loss sometimes encounters the stagnate problem during the training phase because the condition of the loss function can be easily satisfied. Further, in order to remit the stagnate problem and achieve

the purpose of co-learning of person detection and re-id, an on-line pairing loss (OLP loss) and a hard example priority softmax loss (HEP loss) are proposed in I-Net. By storing the features of different persons in a dynamic dictionary, a positive pair and lots of negative pairs can be captured for every iterations. Such that, the positive pairs is restricted by more negative pairs than the Triplet loss [24], which is helpful to remit the stagnate problem. HEP Loss is an auxiliary loss based on softmax, which considers the hard example priority strategy. This loss function drops some easily distinguished samples to reduces computation cast and achieves a better results than traditional softmax. Besides, comparing to the OIM loss [31] which generates positive and negative pairs with out of date features, our model based on the Siamese architecture generates real-time positive pairs during the training phase. We achieved an advanced performance on the person search task with our special designed loss function for joint training. The contribution and novelty of this paper are summarized in three folds:

- We propose a Siamese structure based Integrated Net (I-Net) for an end-to-end learning strategy. Based on the special net structure, our I-Net can generate a better real-time positive pairs than OIM [31], which leads to a better performance.
- On-line pairing (OLP) loss with an online feature dictionary is proposed to strictly restrict the positive pair. Since that, it's more helpful to remit the stagnate problem in the training phase.
- A hard example priority strategy is implemented in Hard example priority (HEP) loss which focus on the online hard example mining. Benefit from the results of OLP, we achieve a better performance than traditional softmax but with less computation cast.

2 Related Work

In recent years, lots of researches on person detection and re-identification were conducted independently. In this section, we will introduce some closely related works for a better presentation of our paper.

Person re-identification. Most re-id work focus on two aspects: feature representation [1], [38], [5] and similarity metric [17], [23], [16]. All of these methods achieved great success in the past few years, Chen *et al.* [5], jointly optimized the two tasks simultaneously for re-id, Cheng *et al.* [6] presented a novel multi-channel parts-based model with triplet loss. Additionally, in order to remit the influence of stagnate problem of triplet loss, Chen *et al.* [4] enlarged the three-steam network to quadruplet network such that one more negative pair can be obtained to strength the condition. For our special designed OLP loss function, we generate amounts of negative samples to remit the stagnate problem during the training phase and a better results for the person search task is achieved.

Pedestrian detection. Traditional pedestrian detection methods are based on hand-crafted features and Adaboost classifiers, such as ACF [8], LDCF [20], Checkerboards [34] and Integral Channels Features (ICF) [9]. These methods

dominated the detection field for years due to their effectiveness. Recently, convolutional neural network (CNN) based deep learning methods have achieved significant progress in pedestrian detection. Tian *et al.* [28] jointly optimized the pedestrian detection with semantic tasks, including pedestrian attributes and scene attributes. Song *et al.* [27] combined multiple deep networks with one fully-connected layer to improve the detection accuracy. In [33], CNN features extracted by RPN [22] are fed into the random forest for pedestrian detection. In our work, we implemented our detection module based on the Faster-RCNN [22], which is used to generate proposals for our further re-identification part.

End-to-end person search. Recently, some works were proposed to address the person search task. ID-discriminative Embedding (IDE) and Confidence Weighted Similarity (CWS) were proposed by Zheng *et al.* [37]. While NPSM [18] based on LSTM was used to reduce the region containing the target automatically. Xiao *et al.* [31] jointly trained both two tasks with OIM loss, which updated features of a labeled identity every hundreds of iterations. As a result, the out-of-date features stored by OIM can't properly compute the loss function. Benefiting from the Siamese structure, our I-Net can generate a real-time positive pairs during the training phase and lead to a better performance.

3 The Proposed I-Net

We propose a new I-Net framework that jointly handles the pedestrian detection and person re-id into an end-to-end Siamese network. The architecture is shown in Figure. 2. Given a pair of images with the persons with same identity, two pedestrian proposal networks (PPN) with shared parameters are learnt to predict the proposals of pedestrians from the two images. The feature maps pooled by region of interest (ROI) pooling layer are then fed into the fully-connected (fc) layers to extract 256-D L2-normalized features for the re-id task. After that, these features are stored in an on-line dynamic feature dictionary, in order to generate one positive pair and lots of negative pairs for OLP loss and HEP loss.

3.1 Deep Model Structure

The basic model of I-Net is the VGG16 architecture [25]. which has 5 stacks of convolutional part, including 2, 2, 3, 3, 3 convolutional layers for each stack. 4 max-pooling layer are followed on the first 4 stacks. On the top of $conv5\_3$ layer, we generate feature maps with 512 channels which are used to predict pedestrian proposals. A $512 \times 3 \times 3$ convolutional layer is first added to get the features for pedestrian proposals. Similar to faster RCNN [22], we then associate 9 anchors at each feature map, and a softmax classifier (cls.) is used to predict whether the anchor is a pedestrian or not. A SmoothL1Loss (reg.) is used for bounding box regression. Finally, 128 proposals for each image after the non-maximum suppression (NMS) are obtained. In fact, the two branches of the PPN are shared same parameters during the training phase.

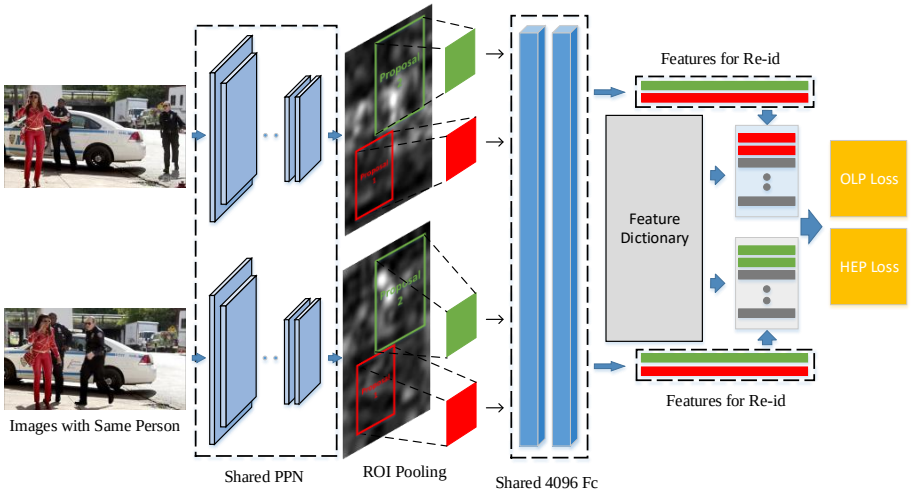


Fig. 2. The net structure of our I-Net. A pair of images including the person with same identity are fed into the two PPNs of the I-Net respectively. Two PPNs shared same parameters generate proposals for pedestrians. These proposals are then fed into the further fc layers to get the corresponding features. The input features and the feature dictionary are then used to form positive and negative pairs, which participate the computation of OLP loss and HEP loss for joint training.

A ROI pooling layer [11] is integrated into I-Net to pool the generated proposals from both two PPNs. The pooled features from both two branches are then fed into the two fc layers of 4096 neurons. In order to remove the false positives of proposals, a two class softmax layer (person vs. non-person) is trained to classify them. Then a 256-D L2-normalized features generated by an extra fc layer which are then fed into the OLP loss and HEP loss for guiding the whole training phase. With the SoomthL1loss which is used to correct the bounding boxes of proposals, the proposed I-Net can be jointly trained for simultaneous person detection and re-identification in an end-to-end architecture. The whole network structure is shown in the fig.2

3.2 On-line Pairing Loss (OLP)

128 proposals per image are learned by PPNs are then fed into the re-identification part. For person re-id, the proposal features can be divided into 3 types, including background (B), persons with identity information (p-w-id) and persons without identity information (p-w/o-id). The division depends on the IOU between the proposals and the ground truth. As shown in Figure. 3, the B, p-w-id and p-w/o-id are represented by red, green, and yellow bounding box, respectively. In the OLP loss, an on-line feature dictionary is designed where the features of p-w-id, p-w/o-id, and a part of B are stored with corresponding labels. Note that, the

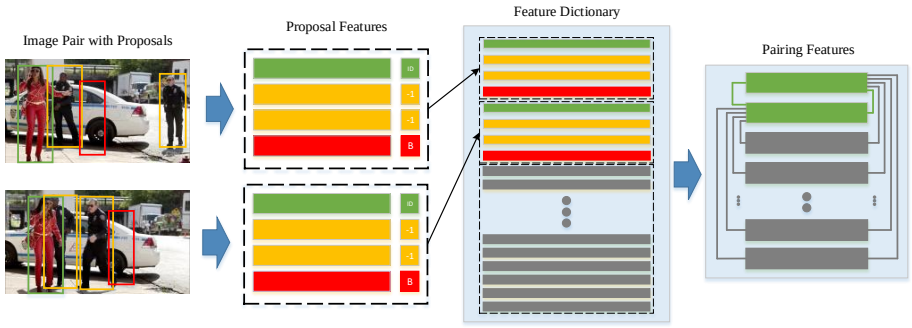


Fig. 3. The procedure of OLP. The features of proposals including B (red), p-w-id (green), and p-w/o-id (yellow and labeled as -1) are extracted. These features are stored into the feature dictionary. OLP loss uses the features extracted by I-Net to compose positive pairs(collected with green line) while the features in the feature dictionary is used to construct negative pairs(collected with gray lines).

number of the stored features depends on the number of proposals per PPN. Specifically, we use 40 times number of proposals per PPN in our experiment. Notably, once the number of features in the dictionary reaches the maximum number, the out-of-date feature will be replaced.

In order to minimize the discrepancy of the features from the same id, while maximizing the discrepancy of different, we use the person proposals from the two stream of I-Net and the stored feature dictionary to establish positive and negative pairs. Suppose that the proposal group for loss computation is $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{n}_1, \mathbf{n}_2..., \mathbf{n}_k)$, where $(\mathbf{p}_1, \mathbf{p}_2)$ stands for proposals from the same identity person generated by I-Net in forward propagation and $(\mathbf{n}_1, \mathbf{n}_2..., \mathbf{n}_k)$ are the features stored in the dictionary. For each proposal group, we tend to formulate two symmetrical subgroups by taking $\mathbf{p}_1$ and $\mathbf{p}_2$ as anchor, alternatively. For example, when $\mathbf{p}_1$ is regarded as anchor, the $(\mathbf{p}_1, \mathbf{p}_2)$ is the positive pair. $(\mathbf{p}_1, \mathbf{n}_1)$, $(\mathbf{p}_1, \mathbf{n}_2)$, ... , $(\mathbf{p}_1, \mathbf{n}_k)$ are negative pairs. Alternatively, when $\mathbf{p}_2$ is regarded as anchor, $(\mathbf{p}_2, \mathbf{p}_1)$ is the positive pair. $(\mathbf{p}_2, \mathbf{n}_1)$, $(\mathbf{p}_2, \mathbf{n}_2)$, ... , $(\mathbf{p}_2, \mathbf{n}_k)$ are negative pairs. Considering the large amount of negative samples, the OLP loss is established based on softmax function. Suppose we get m subgroups in one iteration, and $\mathbf{x}_A^i$, $\mathbf{x}_p^i$, $(\mathbf{x}_{n_1}^i, \mathbf{x}_{n_2}^i ..., \mathbf{x}_{n_k}^i)$ stand for the anchor, positive and negative features of i^{th} subgroup, respectively. The OLP loss function is represented as follows.

$$L_{OLP} = -\frac{1}{m}\sum_{i=1}^{m} log\frac{e^{d(\mathbf{x}_A^i, \mathbf{x}_p^i)}}{e^{d(\mathbf{x}_A^i, \mathbf{x}_p^i)} + \sum_{j=1}^{k} e^{d(\mathbf{x}_A^i, \mathbf{x}_{n_j}^i)}} \tag{1}$$

where the function $d(\cdot)$ stands for the cosine distance of two features. Because these features are L2-normalized, the cosine distance can be directly computed by the inner product. In gradient computation, we only calculate the deviation from the anchor feature.

Then, the deviation of the OLP loss function with respect to $\mathbf{x}_A^i$ for the i^{th} subgroup can be calculated as:

$$\frac{\partial L_{OLP}}{\partial \mathbf{x}_A^i} = (q^i - 1)\mathbf{x}_P^i + \sum_{l=1}^{k} (\hat{q}_l^i \mathbf{x}_{n_l}^i) \tag{2}$$

where q^i and $\hat{q}_l^i$ are expressed in equation (3) and (4).

$$q^i = \frac{e^{d(\mathbf{x}_A^i, \mathbf{x}_p^i)}}{e^{d(\mathbf{x}_A^i, \mathbf{x}_p^i)} + \sum_{j=1}^{k} e^{d(\mathbf{x}_A^i, \mathbf{x}_{n_j}^i)}} \tag{3}$$

$$\hat{q}_l^i = \frac{e^{d(\mathbf{x}_A^i, \mathbf{x}_{n_l}^i)}}{e^{d(\mathbf{x}_A^i, \mathbf{x}_p^i)} + \sum_{j=1}^{k} e^{d(\mathbf{x}_A^i, \mathbf{x}_{n_j}^i)}}, l = 1, ..., k \tag{4}$$

Following the standard BP optimization in CNN, stochastic gradient descent (SGD) is adopted in the training phase of I-Net.

From the OLP loss (1), we can see that a large number of negative features can be processed at one time by utilizing the cosine distance guided softmax function. In terms of the softmax character, the cosine distance between anchor and positive samples tends to be maximized which improves the performance.

3.3 Hard Example Priority Softmax loss (HEP)

The OLP loss function aims to restrict the cosine distance of positive pairs to be larger than negative pairs. However, the identity information is not fully exploited. Therefore, we propose a HEP softmax loss function for person identity classification. Different from the traditional softmax loss, we propose a hard example priority strategy, which focuses on the classes of hard negative pairs.

Suppose that there are C identities in the dataset. The HEP loss function aims to classify all proposals (except p-w/o-id) into $C + 1$ classes (C classes plus B). In order to calculate the HEP loss, we annotate the proposals based on IOU between the proposal and the ground truth. To this end, when the cosine distances among the positive pair and negative pairs are computed by the OLP loss as described in Section 3.2, we can find out the top 20 maximum distances of the negative pairs, and record their corresponding labels as priority classes. In order to keep the fixed total number of the priority classes, we also randomly choose an uncertain number of classes from the remaining classes, such that totally $M (M << C + 1)$ classes are selected to compute the final loss. Finally, the HEP softmax loss function is represented as:

$$L_{HEP} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{M} \mathbf{1}(label = j) log \frac{e^{x_j^i}}{\sum_{k=1}^{M} e^{x_k^i}} \tag{5}$$

where x_j^i stands for the i-th proposal's score from the classifier and j stands for the j-th class. Suppose that $\mathbf{L}$ stands for the pool of priority classes. The

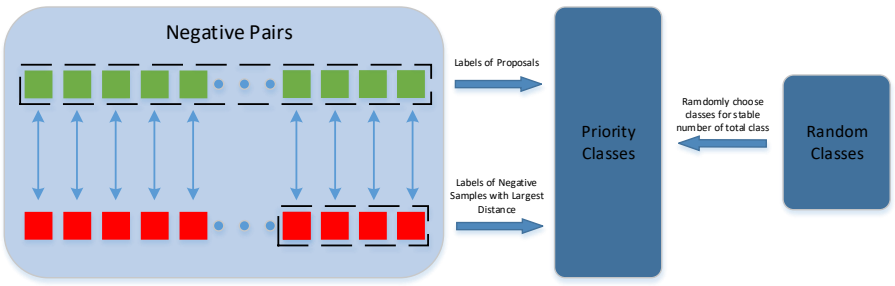


Fig. 4. The protocol for choosing the priority class for the HEP loss function, First, the labels of PPN generated proposals during the forward propagation were selected as priority class. Second, labels of negative samples from negative pairs which got the largest consine distances were selected. At last, same random classes were chosen to make sure the total number of priority class fixed.

protocol for choosing the M classes with hard example priority strategy is summarized as below, which can be recognized in Fig. 4.

1. The label indexes of generated proposals from the input image pair in the forward propagation are first stored in the priority classes pool $\mathbf{L}$ to ensure the ground truth class.
2. For each subgroup (described in Section.3.1), the label indexes of negative samples from the top 20 negative pairs with the maximum distances are stored in the priority classes pool $\mathbf{L}$. This step makes sure the hard classes are priority considered.
3. If the size of pool $\mathbf{L}$ is still smaller than M (a preset value), then we randomly generate the classes indexes without repetition and store them in the chosen priority pool $\mathbf{L}$.

This strategy ensures that the classes of hard samples are priority selected. That is, if a person with identity is hard to distinguish from others, then the proposals of the corresponding identity must participate the HEP softmax loss computation in Eq.(5). Additionally, the loss function has little computation cast because the cosine distances are pre-computed by OLP.

4 Experiments

To evaluate the effectiveness of our approach, we conduct a number of experiments on the CUHK-SYSU dataset [30] and PRW dataset [37]. We first compare our results to state-of-the-art methods on both two datasets in section 4.2 and section 4.3. Then the discussion of our model is presented in section 4.3.

4.1 Experimental Setup

Implementation Details. Our I-Net is implemented on Caffe [13] and py-faster-rcnn [22] platform. VGG16 [25] is the basic network of I-Net and the trained model in [30] was used for network initialization. The first two stacks of convolutional layers are frozen while training our net. The two streams of the I-Net shared the same parameters for both initialization and training. The pedestrian proposal network (PPN) at each branch generates 128 proposals. We randomly choose 5 background proposals per image to store in the feature dictionary. In I-Net, all loss functions are imposed the same loss weight. The learning rate is initialized to 0.001, and drops to 0.0001 in 50k iterations. Totally, 60k iterations are set to insure convergence.

CUHK-SYSU Dataset. The CUHK-SYSU dataset [30] is a large person search dataset containing 18184 images from the hand-held cameras and movie snapshots, which has a large variations in viewpoint, lighting, resolution, *etc.* It annotated 8432 different persons and 96143 bounding boxes. Each labeled person has at least two images from different viewpoints. The dataset provided 11206 images with 5532 identities as the training set, and 6978 images with 2900 identities for test. Our experiments follow the protocols provided by the dataset.

PRW Dataset. The PRW dataset [37] is extracted from 10-hour video captured by 6 cameras, which 5 are 1080×1920 HD and 1 is 576×720 SD. Totally 11816 frames are manually annotated, which contains 43110 pedestrian bounding boxes, among which 34304 pedestrians are annotated with 932 IDs. The PRW dataset provides 5134 frames with 482 labeled identities as training set while 6112 frames with 450 labeled identities are used as testing set. The PRW dataset ask the model to search the target person from the whole testing set, which is challenging.

4.2 Experiments on CUHK-SYSU

In this section, we perform several experiments on the CUHK-SYSU datasets to investigate the effectiveness of our I-Net. For baseline comparisons, we select three pedestrian detection methods and five person re-id approaches, which then results in 15 baselines. The three detection methods, CCF [32], Faster-RCNN [22] with resnet50 [12] and ACF [8], are used for detecting pedestrians. Besides, we also use the detection ground truth of the test set as the perfect detector. For the re-identification problem, we evaluate several famous re-id feature representation methods including DenseSIFT-ColorHist (DSIFT) [35], Bag of Words (BoW) [36], Local Maximal Occurrence (LOMO) [17] and ID-Net(The re-identification part of OIM [31]) to extract features for re-identification, while Euclidean, Cosine similarity, KISSME [23], and XQDA [17] are used as metric method for them. We combine these detection methods and re-identification methods as the examples of traditional person search model.

Further, we implement the OIM loss model [31], the End-to-End model (initialized model) [30] and NPSM [18] as the competitor, which addressed the

Table 1. Comparisons between our framework and other methods on CUHK-SYSU

Detector	Re-id Method	mAP(%)	Top-1(%)
ACF	DSIFT [35]+Euclidean	21.7	25.9
	DISFT [35]+KISSME [23]	32.3	38.1
	BOW [36]+KISSME [23]	42.4	48.4
	LOMO [17]+XQDA [17]	55.5	63.1
	IDNet [31]	56.5	63.0
CCF	DSIFT [35]+Euclidean	11.3	11.7
	DISFT [35]+KISSME [23]	13.4	13.9
	BOW [36]+KISSME [23]	26.9	29.3
	LOMO [17]+XQDA [17]	41.2	46.4
	IDNet [31]	50.9	57.1
CNN	DSIFT [35]+Euclidean	34.5	39.4
	DISFT [35]+KISSME [23]	47.8	53.6
	BOW [36]+KISSME [23]	56.9	62.3
	LOMO [17]+XQDA [17]	68.9	74.1
	IDNet [31]	68.6	74.8
GT	DSIFT [35]+Euclidean	41.1	45.9
	DISFT [35]+KISSME [23]	56.2	61.9
	BOW [36]+KISSME [23]	62.5	67.2
	LOMO [17]+XQDA [17]	72.4	76.7
	IDNet [31]	73.1	78.3
End-to-End(Initialized model) [30]		55.7	62.7
OIM [31]		75.5	78.7
NPSM [18]		77.9	81.2
I-Net(ours)		**79.5**	**81.5**

same person search problem. All our experiments are following the protocol of the CUHK-SYSU dataset, and we test the models on the gallery size of 100.

In experiment, the top-1 accuracy and the mAP (mean average precision) are used for evaluating the person re-identification performance. Specifically, the re-identification results are shown in Table 1, from which we can see that the proposed I-Net achieves a top-1 accuracy of 81.5% and mAP of 79.5%, which beats all compared methods. From Table 1, we can observe that the state-of-the-art end-to-end person search methods(OIM, NPSM, I-Net) always outperform the traditional person search methods(detection+person re-id). It is noteworthy that even the perfect detector(ground truth) with re-id methods had a inferior results compared to the end-to-end method, which demonstrates that it is important and necessary to integrate detection and re-id together for joint modeling. Benefited from siamese architecture, our model generated a real-time positive and negative pairs which leads to 4% higher in mAP than our baseline OIM. Additionally, our method also outperforms NPSM by 1.6% in mAP with lower computation cast.

4.3 Experiments on PRW dataset

Table 2. Results of PRW datasets

Methods	mAP(%)	Top-1(%)
DPM [10]+BOW [36]	9.7	31.1
DPM [10]+IDE$_{det}$ [37]	18.8	45.9
DPM-Alex+LOMO+XQDA [17]	13.0	34.1
DPM-Alex+IDE$_{det}$ [37]	20.3	47.4
DPM-Alex+IDE$_{det}$ + CWS [37]	20.5	48.3
ACF [8]+LOMO+XQDA [17]	10.5	30.9
ACF [8]+IDE$_{det}$ [37]	17.5	43.8
ACF-Alex+LOMO+XQDA [17]	10.3	30.6
ACF-Alex+IDE$_{det}$ [37]	17.5	43.6
ACF-Alex+IDE$_{det}$ + CWS [37]	17.8	45.2
LDCF [20]+BOW [36]	9.1	29.8
LDCF [20]+LOMO+XQDA [17]	11.0	31.1
LDCF [20]+IDE$_{det}$ [37]	18.3	44.6
LDCF [20]+IDE$_{det}$ + CWS [37]	18.3	45.5
OIM(baseline) [31]	23.0	46.7
I-Net(ours)	**25.6**	**48.7**

We also conduct some experiments on PRW [37] dataset by using I-Net, OIM (baseline) [31] and some state-of-the-art detection and re-id methods. For the detection, the DPM [10], ACF [8] and their related RCNN methods and LDCF [20] are implemented, and the LOMO [17]+XQDA [17], bag of words vector [36], IDE$_{det}$, CWS [37] are used for re-identification. At last, we get 14 kinds of methods by combining these detector and re-identification methods. For the RCNN, AlexNet is implemented as the base network according to [37]. On the other hand, OIM was also implemented on the dataset as an end-to-end person search method. The results are shown in Table 2, from which we can see that our method achieves the best results. It's noteworthy that our method also outperforms OIM by 2.6% in mAP and 2.0% in top-1.

4.4 Model Discussion

In this section, we discuss the effectiveness of Joint Loss and the influence of feature dictionary size. All results in this section are tested on CUHK-SYSU dataset.

Analysis of Joint Loss. Our HEP loss can be recognized as a variation of softmax which treats the task as a classification problem. While the traditional softmax loss function have to compute all classes of the dataset every iteration, our HEP loss function gets the priority classes based on the cosine distances of negative pairs. This strategy makes our loss function achieves better performance

and reduces the computation cast. Otherwise, as an auxiliary loss function of OLP, it can promote the effectiveness of the OLP loss function. In order to take an insight of the joint training between OLP loss, softmax and HEP loss, three different cases: OLP only, OLP with softmax loss, and OLP with HEP are discussed, respectively.The OIM which is our baseline is also shown.

Table 3. Comparisons among different loss types.

Loss Type	mAP(%)	Top-1(%)
OLP only	73.6	76.2
OIM	75.5	78.7
OLP with softmax	78.7	80.9
OLP with HEP	79.5	81.5

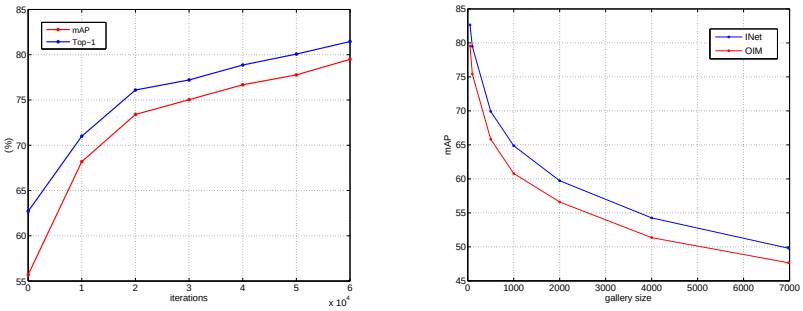
The results with four types of loss functions are shown in Table 3. We can see that a single OLP loss without joint training achieves a inferior result. By adding the softmax loss in our framework for the joint training, both of the mAP and top-1 exceed the OIM loss function by 3.2% and 2.5%, respectively. Such that, the effectiveness of joint loss is demonstrated. Further, the joint training of OLP and HEP outperforms the traditional softmax by 0.8% in mAP and 0.6% in top-1, which shows that the drop of some easily distinguished samples helped the model improve its effectiveness and reduce the computation cast.

Table 4. Influence of the number of features stored in OLP.

	Number of Features		
mAP(%)	20×128	40×128	80×128
OLP only	73.2	74.3	72.9
OLP+HEP	78.4	**79.5**	79.1
Top-1(%)	20×128	40×128	80×128
OLP only	76.0	77.5	75.6
OLP+HEP	80.7	**81.5**	81.1

Influence of Stored Features. The number of features from the proposals stored in the dictionary is a major parameter which influences our model. This parameter is set as 40 times number of proposals generated by each PPN. Such that $40 \times 128 = 5120$ features will be stored. To explore the influence of the feature dictionary, OLP only and the joint training of OLP and HEP with different number of stored features have been tested. The results are shown in Table 4.

The experiments show that 40 times proposal number achieves the best result. Both 20 times and 80 times the proposal number get a inferior results. This kind of phenomenon may expose the two problems mentioned before. A large



(a) Accuracy during the training phase (b) Accuracy changed with gallery size

Fig. 5. Experiment results. figure(a) is the trend of mAP and top-1, which was changed during the training phase. figure(b) is the mAP from different gallery size.

feature dictionary size may make the stored feature stored out of date, which may influence the quality of negative pairs. On the contrary, a small feature dictionary size means a smaller number of negative pairs, which can't fully restrict the positive pair, such that the training phase may stagnated.

Accuracy during the training phase Our I-Net is training for 60k iterations. The variation trend of during the training phase is shown in Figure. 5(a), which shows that both mAP and top-1 index are increasing during the whole training phase. Both mAP and top-1 are increasing sharply during the training phase, and the learning rate is changed at 50k iterations. The model is convergence for 60k iterations.

Gallery Size. Person search problem should be more challenging when the gallery size is growing up. Therefore, we evaluate our method on different gallery size from 50 to 6978 (full set) on the CUHK-SYSU. All test images are covered even in a small gallery size according to the dataset testing protocol. The result is shown in Figure.5(b). As the gallery size increased, the model suffer a significantly descend in both mAP and top-1. It's noteworthy that our I-Net still beats the OIM methods on all gallery sizes. More hard samples are chosen as distracter while the gallery size extends, which increases the the difficulty to find the target person. The person search task based on the large gallery size is a challenging task, which may be our future mission.

Visualization of person search results. The Fig. 6 presents some person search results of our I-Net on the CUHK-SYSU dataset. Rows 1, 2 are successful matched samples which find the target person from other images, and row 3 gets a false alarm because of the similar clothes. The row 4 is a failure case, which may caused by the dusky light condition of the query person.

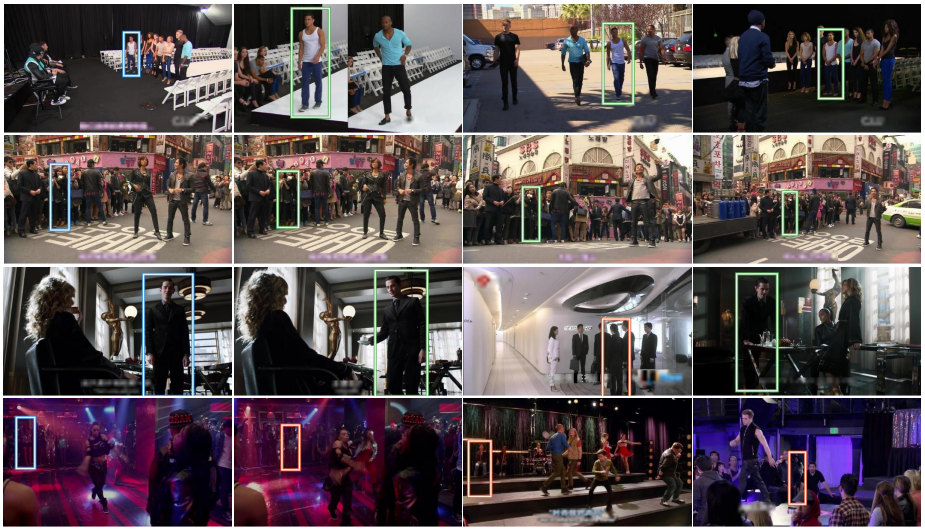


Fig. 6. Visualization of some person search results on the CUHK-SYSU dataset. Every row of the figure is a result of a query. The cyan bounding boxes from the pictures of first column are target persons. The last three columns are matched samples. The red bounding boxes are false alarms, and green bounding boxes are successful matches.

5 Conclusions

In this paper, we introduce a novel end-to-end learning framework for large-scale person search. By jointly modeling pedestrian detection and re-identification, an integrated convolutional neural network (I-Net) is proposed. Specifically, a novel on-line pairing loss (OLP) and hard example priority based softmax (HEP) are proposed for supervising the joint training. For OLP loss, we propose to design a feature dictionary which is used to store a large amounts of features, such that more negative pairs can be obtained to improve the training effect. Besides that, we propose a hard example priority strategy based HEP loss, which has improved the effectiveness as well as the efficiency. By testing the I-Net on the CUHK-SYSU [30] and PRW [37] dataset, the effectiveness of our I-Net is validated.

6 Acknowledgements

This work is supported by National Natural Science Fund of China(Grant 61771079) and the Fundamental Research Funds for the Central Universities.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Computer Vision and Pattern Recognition. pp. 3908–

3916 (2015)
2. Cao, C., Wang, Y., Kato, J., Zhang, G., Mase, K.: Solving occlusion problem in pedestrian detection by constructing discriminative part layers. In: Applications of Computer Vision. pp. 91–99 (2017)
3. Chen, S.Z., Guo, C.C., Lai, J.H.: Deep ranking for person re-identification via joint representation learning. IEEE Transactions on Image Processing **25**(5), 2353–2367 (2016)
4. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification (2017)
5. Chen, W., Chen, X., Zhang, J., Huang, K.: A multi-task deep network for person re-identification (2017)
6. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Computer Vision and Pattern Recognition. pp. 1335–1344 (2016)
7. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition **48**(10), 2993–3003 (2015)
8. Dollar, P., Belongie, S., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(8), 1532 (2014)
9. Dollar, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings (2009)
10. Felzenszwalb, P.F., Girshick, R.B., Mcallester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis & Machine Intelligence **32**(9), 1627–1645 (2010)
11. Girshick, R.: Fast r-cnn. Computer Science (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778 (2016)
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia. pp. 675–678 (2014)
14. Li, W., Wang, X.: Locally aligned feature transforms across views. In: Computer Vision and Pattern Recognition. pp. 3594–3601 (2013)
15. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Asian Conference on Computer Vision. pp. 31–44 (2012)
16. Li, X., Zheng, W.S., Wang, X., Xiang, T.: Multi-scale learning for low-resolution person re-identification. In: IEEE International Conference on Computer Vision. pp. 3765–3773 (2015)
17. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Computer Vision and Pattern Recognition. pp. 2197–2206 (2015)
18. Liu, H., Feng, J., Jie, Z., Jayashree, K., Zhao, B., Qi, M., Jiang, J., Yan, S.: Neural person search machines (2017)
19. Liu, J., Zha, Z.J., Tian, Q.I., Liu, D., Yao, T., Ling, Q., Mei, T.: Multi-scale triplet cnn for person re-identification. In: ACM on Multimedia Conference. pp. 192–196 (2016)
20. Nam, W., Dollar, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: NIPS. pp. 1–9 (2014)

21. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship in pedestrian detection. In: Computer Vision and Pattern Recognition. pp. 3222–3229 (2013)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence **39**(6), 1137–1149 (2017)
23. Roth, P.M., Wohlhart, P., Hirzer, M., Kostinger, M., Bischof, H.: Large scale metric learning from equivalence constraints. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2288–2295 (2012)
24. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering pp. 815–823 (2015)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014)
26. Song, B., Kamal, A.T., Soto, C., Ding, C., Farrell, J.A., Roychowdhury, A.K.: Tracking and activity recognition through consensus in distributed camera networks. IEEE Transactions on Image Processing **19**(10), 2564–2579 (2010)
27. Song, H., Wang, W., Wang, J., Wang, R.: Collaborative deep networks for pedestrian detection. In: IEEE Third International Conference on Multimedia Big Data. pp. 146–153 (2017)
28. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks pp. 5079–5087 (2014)
29. Wang, X.: Intelligent multi-camera video surveillance: A review. Elsevier Science Inc. (2013)
30. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search (2016)
31. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
32. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: IEEE International Conference on Computer Vision. pp. 82–90 (2015)
33. Zhang, L., Lin, L., Liang, X., He, K.: Is faster r-cnn doing well for pedestrian detection? pp. 443–457 (2016)
34. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection pp. 1751–1760 (2015)
35. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3586–3593 (2013)
36. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision. pp. 1116–1124 (2015)
37. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild pp. 3346–3355 (2016)
38. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person re-identification (2016)