Person Search via A Mask-Guided Two-Stream CNN Model

Di Chen[1], Shanshan Zhang[1]⋆, Wanli Ouyang[2], Jian Yang[1]⋆, and Ying Tai[3]

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information
of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding
for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] The University of Sydney, SenseTime Computer Vision Research Group,
Sydney, New South Wales, Australia
[3] Youtu Lab, Tencent
{dichen, shanshan.zhang, csjyang}@njust.edu.cn,
wanli.ouyang@sydney.edu.au, yingtai@tencent.com

Abstract. In this work, we tackle the problem of person search, which is a challenging task consisted of pedestrian detection and person re-identification (re-ID). Instead of sharing representations in a single joint model, we find that separating detector and re-ID feature extraction yields better performance. In order to extract more representative features for each identity, we propose a simple yet effective re-ID method, which models foreground person and original image patches individually, and obtains enriched representations from two separate CNN streams. On the standard person search benchmark datasets, we achieve mAP of 83.0% and 32.6% respectively for CUHK-SYSU and PRW, surpassing the state of the art by a large margin (more than 5pp).

Keywords: Person search; Pedestrian detection; Person re-identification; Foreground

1 Introduction

The task of person search is first introduced by [1], which unifies the pedestrian detection and person re-identification in a coherent system. A typical person re-identification method aims to find matchings between the query probe and the cropped person image patches from the gallery, thus requiring perfect person detection results, which are hard to obtain in practice. In contrast, person search, which searches the queried person over the whole image instead of comparing with manually cropped person image locally, is closer to the real world applications. However, considering the tasks of detection and re-ID together brings domain-specific difficulties: large appearance variance across cameras, low resolution, occlusion, *etc*. In addition, sharing features between detection and re-ID also accumulates errors from each of them, *e.g.* false alarms, misalignments and inexpressive person descriptors, which further jeopardizes the final person search performance.

⋆ Corresponding authors.

Following [1], a few other works [2,3,4,5] have also been proposed for person search. Most of them [3,4,5] focus on an end-to-end solution based on Faster R-CNN [6]. Specifically, an axillary fully-connected (FC) layer is added upon the top convolutional layer of Faster R-CNN to extract discriminative features for re-identification. During training, they optimize a joint loss which is composed of the Faster R-CNN losses and a person categorization loss. However, we argue that it is not appropriate to share representations between the detection and re-ID tasks, as their goals contradict with each other. For the detection task, all people are treated as one class, and the goal is to distinguish them from the background, thus the representations focus on the commonness of different people, *e.g.* the body shape; while for the re-ID task, different people are deemed as different categories, and the goal is to maximize the differences in between, thus the representations focus on the characteristics of each identity, *e.g.* clothing, hairstyle, *etc*. In other words, the detection and re-ID tasks aim to model the inter-class and intra-class difference for people respectively. Therefore, it makes more sense to separate the two tasks rather than solving them jointly.

In the community of person re-ID, it is widely believed that discriminative information lies in foreground, while background is one of the detrimental factors and ought to be neglected or removed during feature extraction [7,8]. An intuitive idea would be to extract features on the foreground person patch only while ignoring the background area. However, simply abandoning all background information may harm the re-ID performance from two aspects. Firstly, the feature extraction procedure may gather errors from imperfect or noisy segmentation masks, *i.e.* identification information loss caused by fractional body shape. Secondly, background information sometimes acts as useful context, *e.g.* attendant suitcases, handbags or companions. Casting out all background area would neglect some informative cues for the re-ID problem. Therefore, we argue that it is more suitable to consider a compromised strategy of paying extra attention on the foreground person while also using the background as a complementary cue.

Inspired by the above discussions, we propose a new approach for person search. It consists of two stages: pedestrian detection and person re-identification. We solve them separately, without sharing any representations. Furthermore, we propose a Two-stream CNN to model foreground person and original image independently, which aims to extract more informative features for each identity and still consider the complementarity of the background. The whole framework is demonstrated in Fig. 1, and we will talk about more details in Sec. 3.

In summary, our contributions are three-folds:

- To the best of our knowledge, this paper is the first work showing that for the person search problem, better performance can be achieved by solving the pedestrian detection and person re-identification tasks separately rather than jointly.
- We propose a Mask-Guided Two-Stream CNN Model (MGTS) for person re-id, which explicitly makes use of one stream from the foreground as the emphasized information and enriches the representation by incorporating another separate stream from the original image.
- Our proposed method achieves mAP of 83.0% and 32.6% on CUHK-SYSU [3] and PRW [2] benchmarks respectively, which improves over the state-of-the-arts by a large margin (more than 5pp).
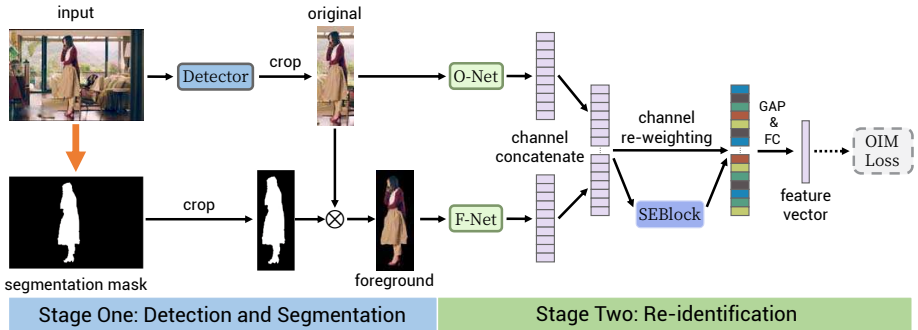
Fig. 1. The proposed framework for person search. It is composed of two stages: 1) Detection and segmentation. We use an adapted Faster R-CNN [9] as our pedestrian detector; the segmentation mask is generated by a MS COCO pre-trained FCIS model [10] without any fine-tuning; 2) Re-identification. The feature extractor is supervised by an Online Instance Matching (OIM) loss [3]. Please note that the detector and re-ID feature extractor are trained independently.

2 Related Work

We first review existing works on person search, which is a recently proposed topic. Since our person search method is composed of two stages: pedestrian detection and person re-identification, we also review some recent works in both fields.

Person Search. Person search has drawn much research interest since the publication of two large scale datasets: CUHK-SYSU [3] and PRW [2]. Zheng *et al*. [2] conduct a detailed survey on various *separated models* and propose to solve the person search problem in two separate models, one for detection and another for re-ID. Other works propose to solve the problem in an *end-to-end fashion* by employing the Faster R-CNN detector [6] for pedestrian detection and share the base network between detection and re-identification [3]. In [4], feature discriminative power is increased by introducing center loss [11] during training. Liu *et al*. [5] improve the localization policy of Faster R-CNN by recursively shrink the search area from the whole image till achieving precise location of the target person. In this paper, we first make a systematic comparison between *separated models* and *joint models*, and show that a separated solution improves both the detection and re-identification results.

Pedestrian Detection. Pedestrian detection is canonical object detection, especially when hand-crafted features are widely used. The classic HOG descriptor [12] is based on local image differences, and successfully represents the special head-shoulder shape of pedestrians. A deformable part model (DPM) [13] is proposed to handle deformations and still uses HOG as basic features. More recently, the integral channel feature (ICF) detectors [14,15,16] become popular, as they achieve remarkable improvements while running fast. In recent years, convnets are also employed in pedestrian detection and further push forward the progress [17,18,19,20]. Some works use the R-CNN architecture, which relies on ICF for proposal generation [17,18]. Aiming for an end-to-end procedure, Faster R-CNN[6] is adopted and it achieves top results by applying proper adaptations [9,21]. Therefore, we use the adapted Faster R-CNN detector in this paper.

Person Re-ID. Early person re-identification methods focus on manually designing discriminative features [22,23,24,25,26], using salient regions [27], and learning distance metrics [28,29,30,31,32]. For instance, Zhao *et al.* [25] propose to densely combine color histogram and SIFT features as the final multi-dimensional descriptor vector. Kostinger *et al.* [28] present KISS method to learn a distance metric from equivalence constraints. CNN-based models have attracted extensive attentions since the successful applications by two pioneer works [33,34]. Most of those CNN models can be categorized into two groups. The first group uses the siamese model with image pairs [33,34,35,36,37,38] or triplets [39,40] as inputs. The main idea of these works is to minimize the feature distance between the same person and maximize the distance between different people. The second group of works formulate the re-identification task as a classification problem [41,42,2]. The main drawback of classification models is that they require more training data. Xiao *et al.* [41] propose to combine multiple datasets for training and improve feature learning by domain guided dropout. Zheng *et al.* [42,2] point out that classification models are able to reach higher accuracy than siamese model, even without careful sample choosing.

Recently, attention mechanism [43,37,44,45,46] has been adopted to learn better features for person re-ID. For instance, HydraPlus-Net [43] aggregates multiple feature layers within the spatial attentive regions extracted from multiple layers in the network. PDC model [46] enriches the person representation with pose-normalized images and re-weights the features by channel-wise attention. In this paper, we also formulate person re-identification as a classification problem, and we propose to emphasize foreground information in the aggregated representation by adding an axillary stream with spatial attention (instance mask) and channel-wise re-weighting (SEBlock), which is similar to HydraPlus-Net and PDC model. However, our work differs from them in that the attention mechanism in our work is introduced with a different motivation, which is to consider the foreground-background relationship instead of local-global or part-whole relationship. In addition, the architecture of our model is more clear and concise, along with more practical training strategy without multi-staged fine-tuning.

3   Method

As shown in Fig. 1, our proposed person search method consists of two stages: pedestrian detection and re-identification. In this section, we first give an overview of our framework, and then describe more details for both stages individually.

3.1   Overview

A panoramic image is first fed into a pedestrian detector, which outputs several bounding boxes along with their confidence scores. We remove the bounding boxes whose confidence scores are lower than a given threshold. Only the remaining ones are used by the re-ID network.

A post-processing is implemented on the detected persons before they are sent to the re-ID stage. Specifically, we expand each RoI (Region of Interest) with a ratio of $\gamma(\gamma > 1)$ to include more context and crop out the person from the whole image. In

order to separate the foreground person from background, we apply an off-the-shelf instance segmentation method FCIS [10] on the whole image, and then designate the person to the right mask via majority vote. After that, for each person, we obtain a pair of images, one containing only the foreground person, and the other containing both the foreground and the background. An illustration is shown in Fig. 2.

Next up, in the re-ID model, the pair images go through two different paths, namely *F-Net* and *O-Net*, for individual modeling. Feature maps from the two paths are then concatenated and re-weighted by an SEBlock [47]. After channel re-weighting, we pool the two dimensional feature maps into feature vectors using Global Average Pooling (GAP). Finally, the feature vectors are projected to an L2-normalized d-dimensional subspace as the final identity descriptor.

The pedestrian detector and re-ID model are trained independently. In order to avoid the mistakes resulting from the detector, we use the ground truth annotations instead of detections to train the re-ID model.

3.2 Pedestrian Detection

We use a Faster R-CNN [6] detector for pedestrian detection. The Faster R-CNN architecture is composed of a base network for feature extraction, a region proposal network (RPN) for proposal generation and a classification network for final predictions.

In this paper, we use VGG16 [48] as our base network. The top convolutional layer 'conv5_3' produces 512 channels of feature maps, where the image resolution is reduced by a factor of 16. According to [9], up-sampling the input image is a reasonable way for compensation.

RPN is built upon 'conv5_3' to predict pedestrian candidate boxes. We follow the anchor settings in [6] and set uniform scales ranging from the smallest and biggest persons we want to detect. The RPN produces a large number of proposals, so we apply a humble Non-Maximum Suppression (NMS) with an Intersection over Union (IoU) threshold of 0.7 to remove repeating ones and also cut off low-scoring ones by a given threshold.

The remaining proposals are then sent to the classification network, where an RoI pooling layer ($512 \times 7 \times 7$) is used to generate the same length of features for each proposal. The final detection confidence and corresponding bounding box regression parameters are regressed by fully connected layers. After bounding box regression, another NMS with IoU threshold of 0.45 is applied and low-scoring detections are cut off.

The base net, RPN and classification network are jointly trained using Stochastic Gradient Descent (SGD).

3.3 Person Re-ID via A Mask-Guided Two-Stream CNN Model

After RoIs for each person are obtained (either from a detector or ground truth), we aim to extract discriminative features. First of all, we expand each RoI by a ratio of γ to include more context. Then, we propose a two-stream structure to extract features for foreground person and whole image individually. The features from both streams are

Algorithm 1 Foreground Separation

Input: RoI $\mathbf{b} \in \mathbb{N}^4$, expand ratio $\gamma \in \mathbb{R}$, image $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$, instance mask $\mathbf{M} \in \mathbb{N}^{h \times w}$
Output: Masked image for an instance $\mathbf{I}'_{\mathbf{M}} \in \mathbb{R}^{h' \times w' \times 3}$

1: $\mathbf{b}^\gamma \leftarrow$ expand $\mathbf{b}$ by γ
2: Crop out image patch $\mathbf{I}'$ from image $\mathbf{I}$ according to $\mathbf{b}^\gamma$
3: Crop out mask patch $\mathbf{M}'$ from mask $\mathbf{M}$ according to $\mathbf{b}^\gamma$
4: Find the dominant instance k inside $\mathbf{M}'$ by majority vote
5: Binarize $\mathbf{M}'$ by $\mathbf{M}'_b \leftarrow (\mathbf{M}' == k)$
6: Cast mask onto image by element-wise production:
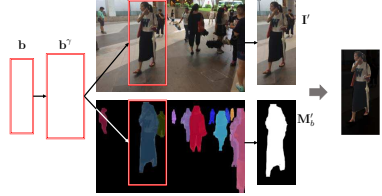$\mathbf{I}'_{\mathbf{M}} \leftarrow (\mathbf{M}'_b \odot \mathbf{I}')$

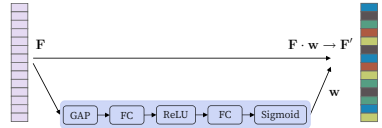**Fig. 2.** Illustration of foreground separation

Fig. 3. Demonstration of an SEBlock [47]

concatenated as enriched representations for the RoIs and a re-weighting operation is applied to highlight more informative features while suppressing less useful ones.

Foreground Separation. The key step is to separate foreground and background for each RoI. We first apply an instance segmentation method FCIS [10] on the whole image to obtain segmentation masks for persons. After that, we associate each RoI with its corresponding mask by majority vote. Those pixels inside and outside the mask boundary are considered as foreground and background respectively. We describe the detailed separation procedure in Algorithm 1 and show an example in Fig. 2.

Two-Stream Modeling. After foreground separation, image pairs of each person are fed into the MGTS model. Specifically, foreground images go through F-Net and original images go through O-Net. Both F-Net and O-Net share the same architecture, but their network parameters are not shared. The corresponding feature maps, denoted as $\mathbf{F_F}, \mathbf{F_O} \in \mathbb{R}^{c \times h'' \times w''}$, are produced individually. Here c denotes the number of channels, h'', w'' are the height and width of $\mathbf{F_F}$ and $\mathbf{F_O}$. The feature maps are then concatenated along the channel axis as $\mathbf{F} \in \mathbb{R}^{2c \times h'' \times w''}$.

Feature Re-Weighting. We further re-weight all the feature maps with an SEBlock [47], which models the interdependencies between channels of convolutional features. The architecture of an SEBlock is illustrated in Fig. 3. It is defined as a transformation from $\mathbf{F}$ to $\mathbf{F}'$:

$$\mathbf{F}' = \mathbf{F} \cdot \mathbf{w}, \quad \text{where} \tag{1}$$
$$\mathbf{w} = [w_1, w_2, \ldots, w_{2c}], \quad w_i \in [0, 1]$$
$$= \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 f_{GAP}(\mathbf{F}))),$$

σ and δ refer to the Sigmoid activation and the ReLU [49] function respectively. $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrix of two FC layers. f_{GAP} is the operation of GAP and

Fig. 4. Instance segmentation results on CUHK-SYSU (the first row) and PRW (the second row) generated by the open sourced version of FCIS [10]

· denotes channel-wise multiplication. SEBlock learns to selectively emphasis informative features and suppress less useful ones by re-weighting channel features using the weighting vector $\mathbf{w}$. In this way, foreground and background information are fully explored and re-calibrated, and hence help to optimize the final feature descriptor for person re-identification.

The re-weighted feature maps $\mathbf{F}'$ are then pooled to feature vectors $\mathbf{f} \in \mathbb{R}^{2c}$ by GAP, and further projected to a L2-normalized d-dimensional subspace by an FC layer:

$$\mathbf{x}^T = \frac{\mathbf{W}^T \mathbf{f}^T}{\|\mathbf{W}^T \mathbf{f}^T\|_2}, \mathbf{W} \in \mathbb{R}^{2c \times d}. \tag{2}$$

The whole MGTS model is trained with ground truth RoIs and supervised by an Online Instance Matching loss (OIM) [3]. The OIM objective is to maximize the expected log-likelihood:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}}[\log p_t], \quad \text{where} \tag{3}$$

$$p_t = \frac{\exp(\mathbf{v}_t^T \mathbf{x}/\tau)}{\sum_{j=1}^{L} \exp(\mathbf{v}_j^T \mathbf{x}/\tau) + \sum_{k=1}^{Q} \exp(\mathbf{u}_k^T \mathbf{x}/\tau)},$$

p_t denotes the probability of $\mathbf{x}$ belonging to class t. τ is a temperature factor similar to the one in Softmax function. $\mathbf{v}_t$ is the class central feature vector of the t-th class. It is stored in a lookup table with size L and incrementally updated during training with a momentum of η:

$$\mathbf{v}_t \leftarrow \eta \mathbf{v}_t + (1 - \eta)\mathbf{x}, \tag{4}$$

where $\mathbf{u}_k$ is a feature vector for an unlabeled person. A circular queue of size Q is used to store $\mathbf{u}_k$ vectors. It pops out old features and pushes in new features during training.

4   Experiments

In this section, we first introduce the datasets and evaluation protocols we use in our experiments, followed by some implementation details. After that, we show experimental

results with comparison to state-of-the-art methods, followed by an ablation study to verify the design of our approach.

4.1   Datasets

CUHK-SYSU. CUHK-SYSU [3] is a large-scale person search database consisted of street/urban scene images captured by a hand-held camera or selected from movie snapshots. It contains $18,184$ images and $96,143$ pedestrian bounding boxes. There are a total of $8,432$ labeled identities, and the rest of the pedestrians are served as negative samples for identification. We adopt the standard train/test split provided by the dataset, where the training set includes $11,206$ images and $5,532$ identities, while the testing set contains $2,900$ probe persons and $6,978$ gallery images. Moreover, each probe person corresponds to several gallery subsets with different sizes, which are defined in the dataset.

PRW. The PRW dataset [2] is extracted from video frames captured with six cameras in a university campus. There are $11,816$ frames annotated with $34,304$ bounding boxes. Among all the pedestrians, 932 identities are tagged and the rest of them are marked as unknown persons similar to CUHK-SYSU. The training set includes $5,134$ images with 482 different persons. The testing set contains $2,057$ probe persons and $6,112$ gallery images. Different from CUHK-SYSU, the whole gallery set serves as the search space for each probe person.

4.2   Evaluation Protocol

Pedestrian Detection. Average Precision (AP) and recall are used to measure the performance of pedestrian detection. A detection bounding box is considered as a true positive if and only if its overlap ratio with any ground truth annotation is above 0.5.

Person Search. We adopt the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC top-K) as performance metrics for re-identification and person search. The mAP metric reflects the accuracy and matching rate of searching a probe person from gallery images. CMC top-K is widely used for person re-identification task, where a matching is counted if there is at least one of the top-K predicted bounding boxes overlapping with the ground truth with an IoU larger than or equal to a threshold. The threshold is set to 0.5 throughout the paper.

4.3   Implementation Details

We implement our system with Pytorch. The VGG-based pedestrian detector is initialized with an ImageNet-pretrained model. It is trained using SGD with a batch size of 1. Input images are resized to have at least 900 pixels on the short side and at most $1,500$ pixels on the long side. The initial learning rate is 0.001, decayed by a factor of 0.1 at 60K and 80K iterations respectively and kept unchanged until the model converges at 100K iterations. The first two convolutional blocks ('conv1' and 'conv2') are frozen during training, while other layers are updated.

The RoI expand ratio γ is set to 1.3. Both F-Net and O-Net of our MGTS model are based on ResNet50 [50] and truncated at the last convolutional layer ('conv5_3'). The

input image patches are re-scaled to an arbitrary size of 256×128 and the batch size is set to 128. The model is trained with an initial learning rate of 0.001, decayed to 0.0001 after 11 epochs and kept identical until early cutting after epoch 15. The temperature scalar τ, circular queue size Q and momentum η in OIM loss are set to $1/30$, 5000 and 0.5 respectively. The feature dimension d is set to 128 through out the paper if not specified.

As of foreground extraction, we use the off-the-shelf instance segmentation method FCIS trained on COCO trainval35k [51] without any fine-tuning[1]. Sample results of instance masks from CUHK-SYSU and PRW are shown in Fig. 4, where we can see FCIS generalizes well to both datasets.

4.4   Comparison with State-of-the-Art Methods

In this subsection, we report our person search results on the CUHK-SYSU and PRW datasets, with a comparison to several state-of-the-art methods, including OIM [3], IAN [4], NPSM [5] and IDE [2]. Other than the above joint methods, we also compare with some methods which also solve the person search problem in two steps of pedestrian detection and person re-identification, similar to our method. These methods use different pedestrian detectors (DPM [13], ACF [52], CCF [53], LDCF [54]), person descriptors (BoW [55], LOMO [26], DSIFT [25]) and distance metrics (KISSME [28], XQDA [26]).

Results on CUHK-SYSU. Table 1 shows the person search results on CUHK-SYSU with a gallery size of 100. We follow the notations defined in [3], where "CNN" denotes the Faster R-CNN detector based on ResNet50 and "IDNet" represents a re-identification net defined in [3]. Our VGG-based detector is marked as "CNN$_v$". ID-NetOIM is a re-identification net trained with ground truth RoIs and supervised by an OIM loss. Compared with OIM, CNN$_v$ + IDNetOIM slightly improves the performance by solving detection and re-identification tasks in two independent models. By further employing our proposed MGTS model, we achieve 83.0% mAP. Our final model outperforms the state-of-the-art method by more than 5 pp w.r.t. mAP, and 2.5 pp w.r.t. CMC top-1.

Moreover, we evaluate the proposed method (CNN$_v$ + MGTS) under different gallery sizes along with other competitive methods. Figure 5 shows how the mAP changes with a varying gallery size of [50, 100, 500, 1000, 2000, 4000]. We can see that all methods suffer from a performance degeneration as the gallery size increases. However, our method outperforms others under different settings, which indicates the robustness of our method. Besides, we notice that the gap between our method and others become even larger as gallery size increases.

We also show some qualitative results of our method and the competitive baseline OIM in Fig. 7. As can be seen in the figure, our method performs better on hard cases where gallery persons wear similar clothes with the probe person, possibly with the help of context information in the expanded RoI, *e.g.* accompanied person (Fig. 7(a), 7(d)), handrail (Fig. 7(d)), baby carriage (Fig. 7(e)) *etc*. It is also more robust on hard cases where gallery entries share both similar foreground and background to the probe

---

[1] https://github.com/msracver/FCIS

Table. 1. Comparison of results on CUHK-SYSU with gallery size of 100

Method	mAP(%)	top-1(%)
CNN + DSIFT + Euclidean	34.5	39.4
CNN + DSIFT + KISSME	47.8	53.6
CNN + BoW + Cosine	56.9	62.3
CNN + LOMO + XQDA	68.9	74.1
CNN + IDNet	68.6	74.8
OIM [3]	75.5	78.7
IAN [4]	76.3	80.1
NPSM [5]	77.9	81.2
Ours(CNN$_v$ + IDNetOIM)	75.8	79.5
Ours(CNN$_v$ + MGTS)	**83.0**	**83.7**
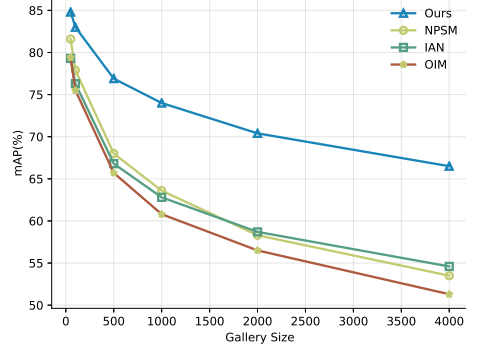
Fig. 5. Performance comparison on CUHK-SYSU with varying gallery sizes

person (Fig.7(b), 7(c)), where more subtle differences like hairstyle and gender shall be excavated from the emphasized foreground person. Fig. 7(f) shows a failure case where both OIM and MGTS suffer from bad illumination condition, which is rather challenging and needs more efforts in future works.

Results on PRW. In Table 2 we report the evaluation results on the PRW dataset. A number of combinations of detection methods and re-identification models are explored in [2]. Among them, DPM [13] + AlexNet [56]-based R-CNN with ID-discriminative Embedding (IDE$_{det}$) and Confidence Weighted Similarity(CWS) achieves the best performance. In contrast, joint methods including OIM, IAN and NPSM, all achieve better results. But it is unclear whether the improvement comes from a joint solution or the usage of deeper networks (ResNet50/ResNet101) and a better performed detector (Faster R-CNN).

For fair comparison, we also employ ResNet50 and Faster R-CNN in our framework, and achieve significant improvements compared to joint methods. Specifically, we outperform the state of the art by 8.4 pp and 10.2 pp w.r.t. mAP and top-1 accuracy. These results again demonstrate the effectiveness of our proposed method.

4.5   Ablation Study

From the experimental results in Sec. 4.4, we obtain significant improvement to our baseline method OIM [3] on both standard benchmarks. The major differences between our method and OIM are as follows: (1) We solve the pedestrian detection and person re-identification tasks separately rather than jointly, i.e. we do not share features between them. (2) In the re-identification net, we model the foreground and original image in two parallel streams so as to obtain enriched representations.

In order to understand the impact of the above two changes, we conduct analytical experiments on CUHK-SYSU at a gallery size of 100, and provide discussions in the following.

Integration vs. Separation. We investigate the impact of sharing features between detection and re-identification task on both performance.

Table. 2. Comparison of results on PRW

Method	mAP(%)	top-1(%)
DPM-Alex + LOMO + XQDA	13.0	34.1
DPM-Alex + IDE_{det}	20.3	47.4
DPM-Alex + IDE_{det} + CWS	20.5	48.3
ACF-Alex + LOMO + XQDA	10.3	30.6
ACF-Alex + IDE_{det}	17.5	43.6
ACF-Alex + IDE_{det} + CWS	17.8	45.2
LDCF + LOMO + XQDA	11.0	31.1
LDCF + IDE_{det}	18.3	44.6
LDCF + IDE_{det} + CWS	18.3	45.5
OIM [3]	21.3	49.9
IAN [4]	23.0	61.9
NPSM [5]	24.2	53.1
Ours(CNN_v + IDNetOIM)	28.2	66.7
Ours(CNN_v + MGTS)	**32.6**	**72.1**

Table. 3. Integration/Separation study on CUHK-SYSU with gallery size of 100. (a): Detector performance comparison between a jointly trained model and a vanilla detector; OIM-ours is our re-implementation of OIM. (b): Re-ID performance comparison for an integrated person search model and a naïve re-ID model

(a)

Method	Joint	AP(%)	Recall(%)
OIM-ours	✓	69.5	75.6
CNN [3]	✗	78.0	75.7

(b)

Method	Joint	mAP(%)	top-1(%)
GT + OIM [3]	✓	77.9	80.5
GT + IDNetOIM	✗	78.5	81.7

In Table 3(a), we compare the detection performance of a jointly trained model and a vanilla detector. We can see that the jointly trained detector under-performs the vanilla one by 8.5 pp w.r.t. AP while reaching the same recall.

Similarly, we make a comparison on re-ID performance between a jointly trained model and a vanilla re-ID net in Table 3(b). The person search performance of a jointly trained OIM method is 0.6 pp and 1.2 pp lower in mAP and top-1 accuracy than a vanilla re-ID net (IDNetOIM).

From the above comparisons, we conclude that joint training harms both the detection and re-ID performance, thus it is a better solution to solve them separately.

Visual Component Study. In this part, we study the contributions of foreground and original image information to a re-ID system. To exclude the influence of detectors, all the following models are trained and tested using ground truth RoIs on CUHK-SYSU with a gallery size of 100. They are based on ResNet50 and supervised by an OIM loss.

Four variants to the input RoI patch and their combinations are considered: (1) Original RoI (**O**); (2) Masked foreground (**F**); (3) Masked background (**B**); (4) Expanded RoI (**E**) with a ratio of γ.

Table. 4. Visual component study. Legend: **O**: Original image; **F**: Masked image with foreground people only; **B**: Masked image with background only; **E**: Expand RoI with a ratio of γ

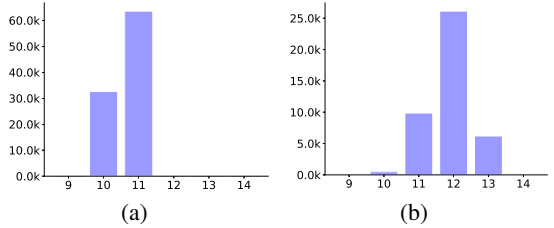| O	F	B	E	mAP(%)	top-1(%)
✓				78.5	81.7
	✓			75.3	78.7
		✓		34.2	35.9
✓			✓	77.7	81.1
		✓	✓	38.7	40.0
✓	✓		✓	89.1	90.0

(a)   (b)

Fig. 6. SEBlock weights statistics for CUHK-SYSU (a) and PRW (b). x-axis denotes $N^{20}(F)$: the number of foreground-related weights among the top 20 largest ones. y-axis denotes the number of training instances

Comparison results are shown in Table 4, from which we make the following observations:

1. Background is an important cue for re-ID. The mAP drops by 2.8 pp when only the foreground is used, while discarding all background. More interestingly, using only background information yields an mAP of 34.2%, which can be further pushed to 38.7% if RoI is expanded.
2. Modeling foreground and original image in two streams improves the results significantly. The two-stream model **O+F+E** reaches 89.1% mAP, surpassing the one-stream model **O+E** by 11.4 pp.

4.6   Model Inspection

To further understand the respective impact of the two streams, we provide an analysis on the SEBlock weights of foreground vs. original image representations. The analysis is implemented based on the trained models in Sec.4.4, to which we feed all training samples in CUHK-SYSU ($96, 143$ proposals) and PRW ($42, 871$ proposals) respectively. For sample i, we compute three metrics: (1) average weight of F-Net channels, $Avg_i(F)$; (2) average weight of O-Net channels, $Avg_i(O)$; (3) number of channels that among top 20 of the whole network while come from F-Net, $N^{20}(F)$, the histograms of $N^{20}(F)$ for all training samples from two datasets are shown in Fig. 6.

Based on analyzing the above statistics, we have the following findings:

1. The inequation $Avg_i(F) > Avg_i(O)$ holds for all samples. It demonstrates that in general the foreground patch contributes more than the original patch to the final feature vector, as it involves more informative cues for each identity.
2. From Fig. 6, the majority of the top 20 channels come from the foreground stream for most samples. This observation indicates that the most informative cues are from the foreground patch.
3. Although the majority of the top 20 channels are represented by the foreground patch, we still observe that quite a few top channels are from the original patch. This is a good evidence showing the context information contained in the original image patch is helpful for the re-identification task.

Table. 5. Re-ID performance of MGTS on CUHK-SYSU with different γ

Value of γ	1.0	1.1	1.2	1.3	1.4	1.5
mAP (%)	85.6	85.4	88.9	**89.1**	87.8	87.1
top-1 (%)	86.6	86.2	89.8	**90.0**	88.2	87.8

Table. 6. Comparison results on CUHK-SYSU with gallery size of 100. The three models are all trained and tested with ground truth bounding boxes. Overall runtime includes pedestrian detection, mask generation (if used) and person feature extraction

Mask	**mAP** (%)	**top-1** (%)	**Overall Runtime** (s)
-	78.5	81.7	0.65
FCIS	**89.1**	**90.0**	1.27
Bounding Box	85.1	86.0	0.69

Moreover, the impact of the amount of context information is inspected by changing the RoI expansion value γ. We conduct a set of experiments on CUHK-SYSU with a gallery size of 100 and list the results in Table 5, from which we can draw the intuitive conclusion that **1)** the γ is relatively stable when $\gamma \in [1.2\ 1.5]$; and **2)** a proper amount of context information is better than no context, while too much background could be harmful.

4.7   Runtime Analysis and Acceleration

We implement our runtime analysis on a Tesla K80 GPU. For testing a 1500×900 input image, our proposed approach takes around 1.3 seconds on average, including 626 ms for pedestrian detection, 579 ms for segmentation mask generation and another 64 ms for person re-identification.

We notice half of the computational time is used to generate the segmentation mask. In order to accelerate, we propose an alternative option to use the tight ground truth bounding boxes as 'weak' masks instead of the 'accurate' FCIS masks. The results are presented in Tab. 6, from which we can see that using 'accurate' FCIS masks yields better performance than using bounding box masks. However, using bounding boxes as weak masks still achieves promising results, which outperforms the single stream model without using any masks by a large margin ($\sim$ 7pp mAP) at comparable time cost. Therefore, our proposed method can be accelerated by a factor of $\sim$2x with an acceptable performance drop, while still surpassing the state-of-the-art results.

5   Conclusion

In this paper, we propose a novel deep learning based approach for person search. The task is accomplished in two steps: we first apply a Faster R-CNN detector for pedestrian detection on gallery images; and then make matchings between the probe image and output detections via a re-identification network. We obtain better results by training the detector and re-identification network separately, without sharing representations. We further enhance the re-identification accuracy by modeling the foreground and original

Fig. 7. Qualitative search results from OIM [3] (upper row in each sub-figure) and our method (bottom row). The ranking and similarity score are shown under each image patch; "✓" denotes the correct matching. OIM would mistakenly assign high similarity to a wrong person wearing similar clothes to the probe, while our model successfully ranks it down. When faced with similar background, our method could emphasize the subtle difference on foreground people and return the correct ranking. (f) is a failure case with low visibility.

image patches in two subnets to obtain enriched representations. Experimental results show that our proposed method significantly outperforms state-of-the-art methods on two standard benchmarks.

Inspired by the success of utilizing the segmented foreground patch for additional feature extraction, for future work, we will explore to optimize the segmentation masks and identification accuracy in a joint framework, so as to obtain finer masks.

Acknowledgments

References

1. Xu, Y., Ma, B., Huang, R., Lin, L.: Person search in a scene by jointly modeling people commonness and person uniqueness. In: ACM'MM. (2014) 1, 2

2. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: CVPR. (2017) 2, 3, 4, 8, 9, 10

3. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR. (2017) 2, 3, 7, 8, 9, 10, 11, 14

4. Xiao, J., Xie, Y., Tillo, T., Huang, K., Wei, Y., Feng, J.: Ian: The individual aggregation network for person search. arXiv preprint arXiv:1705.05552 (2017) 2, 3, 9, 10, 11

5. Liu, H., Feng, J., Jie, Z., Jayashree, K., Zhao, B., Qi, M., Jiang, J., Yan, S.: Neural person search machines. In: ICCV. (2017) 2, 3, 9, 10, 11

6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. TPAMI **39**(6) (2017) 2, 3, 5

7. Le, C.V., Hong, Q.N., Quang, T.T., Trung, N.D.: Superpixel-based background removal for accuracy salience person re-identification. In: ICCE-Asia. (2017) 2

8. Nguyen, T.B., Pham, V.P., Le, T.L., Le, C.V.: Background removal for improving saliency-based person re-identification. In: KSE. (2016) 2

9. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR. (2017) 3, 5

10. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: CVPR. (2017) 3, 5, 6, 7

11. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV. (2016) 3

12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 3

13. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI (2010) 3, 9, 10

14. Dollar, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC. (2009) 3

15. Zhang, S., Bauckhage, C., Cremers, A.B.: Informed haar-like features improve pedestrian detection. In: CVPR. (2014) 3

16. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: CVPR. (2015) 3

17. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR. (2016) 3

18. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: Towards reaching human performance in pedestrian detection. TPAMI (2018) 3

19. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV. (2013) 3

20. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR. (2012) 3

21. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in cnns. In: CVPR. (2018) 3

22. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: ICCV. (2007) 4

23. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV. (2008) 4

24. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. (2010) 4

25. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR. (2013) 4, 9

26. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR. (2015) 4, 9
27. Zhao, R., Oyang, W., Wang, X.: Person re-identification by saliency learning. IEEE transactions on pattern analysis and machine intelligence **39**(2) (2017) 356–370 4
28. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR. (2012) 4, 9
29. Li, X., Zheng, W.S., Wang, X., Xiang, T., Gong, S.: Multi-scale learning for low-resolution person re-identification. In: ICCV. (2015) 4
30. Liao, S., Li, S.Z.: Efficient PSD constrained asymmetric metric learning for person re-identification. In: ICCV. (2015) 4
31. Paisitkriangkrai, S., Shen, C., Van Den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: CVPR. (2015) 4
32. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR. (2016) 4
33. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: ICPR. (2014) 4
34. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: Deep filter pairing neural network for person re-identification. In: CVPR. (2014) 4
35. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR. (2015) 4
36. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: ECCV. (2016) 4
37. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing **26**(7) (2017) 4
38. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: CVPR. (2018) 4
39. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition **48**(10) (2015) 4
40. 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: CVPR. (2016) 4
41. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR. (2016) 4
42. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: MARS: A video benchmark for large-scale person re-identification. In: ECCV. (2016) 4
43. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: ICCV. (2017) 4
44. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. arXiv preprint arXiv:1802.08122 (2018) 4
45. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR. (2017) 4
46. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV. (2017) 4
47. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017) 5, 6
48. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015) 5
49. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: ICML. (2010) 6
50. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 8

51. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. (2014) 9
52. Dollar, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. TPAMI **36**(8) (2014) 9
53. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: ICCV. (2015) 9
54. Nam, W., Dollar, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: NIPS. (2014) 9
55. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV. (2015) 9
56. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 10