International Journal of Advanced Robotic Systems
May-June 2019: 1–14

DOI: 10.1177/1729881419858162
journals.sagepub.com/home/arx
SAGE

A cascaded multitask network with deformable spatial transform on *person search*

Yuan Hong[1,2], Hua Yang[1,2], Liangqi Li[1,2], Lin Chen[1,2] and Chuang Liu[1,2]

Abstract

This article introduces a cascaded multitask framework to improve the performance of person search by fully utilizing the combination of pedestrian detection and person re-identification tasks. Inspired by Faster R-CNN, a Pre-extracting Net is used in the front part of the framework to produce the low-level feature maps of a query or gallery. Then, a well-designed Pedestrian Proposal Network called Deformable Pedestrian Space Transformer is introduced with affine transformation combined by parameterized sampler as well as deformable pooling dealing with the challenge of spatial variance of person re-identification. At last, a Feature Sharing Net, which consists of a convolution net and a fully connected layer, is applied to produce output for both detection and re-identification. Moreover, we compare several loss functions including a specially designed Online Instance Matching loss and triplet loss, which supervise the training process. Experiments on three data sets including CUHK-SYSU, PRW and SJTU318 are implemented and the results show that our work outperforms existing frameworks.

Keywords

Person re-identification, pedestrian detection, person search, deformable pooling, spatial transformation

Date received: 29 January 2019; accepted: 22 May 2019

Topic: Vision Systems
Topic Editor: Antonio Fernandez-Caballero
Associate Editor: Kaifu Yang

Introduction

Video surveillance[1] is an important part of social security, whose effectiveness depends on whether the specific person can be found in the recording. As the complexity of video surveillance networks grows, traditional manual video monitoring method has been infeasible.[2] Apparently, it's important to find a way to obtain information from videos quickly and accurately. Thus, person search under multi-camera video surveillance network is a very challenging and practical issue. It is of great significance in real-world applications like security surveillance,[3] crowd flow monitoring[4] and human behaviour analysis.[5]

Nowadays, traditional person search in the field of computer vision can be divided into two independent tasks, pedestrian detection and person re-identification (Re-ID),[6] and achieve good accuracy in each of the two problems respectively. However, while the Re-ID works based on

[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Key Labs of Digital Media Processing and Communication, Shanghai Jiao Tong University, Shanghai, China

Corresponding author:
Hua Yang, Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University; Shanghai Key Labs of Digital Media Processing and Communication, Shanghai Jiao Tong University, Shanghai 200240, China.
Email: hyang@sjtu.edu.cn

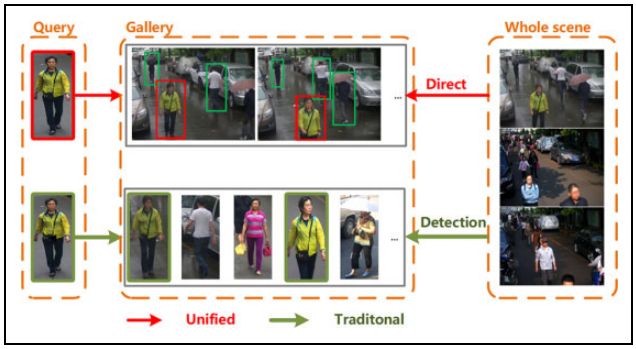

Figure 1. The difference between traditional independent person search task: searching in the sequence of cropped pedestrian images; and unified person search task: searching for the location of a specified pedestrian directly from the video frame containing several pedestrians as well as the background.

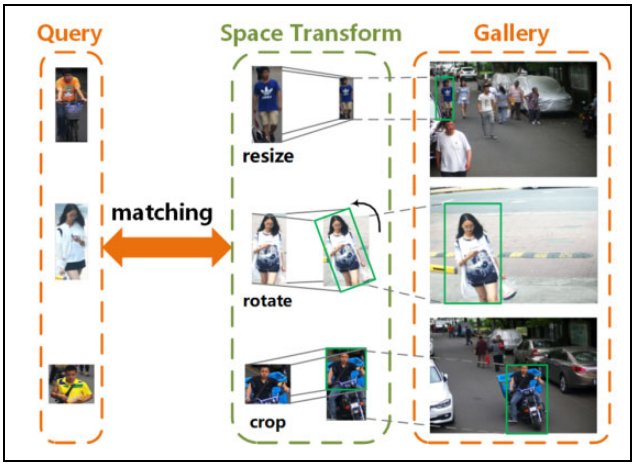

Figure 2. The spatial transformation in our DSIPN equips the network with the ability to deal with spatial variance by the way of cropping, resizing and rotating. DSIPN: Deformable Spatial Invariant Person Search Network.

detection results, it is trained with manually well-calibrated pedestrian bounding boxes,[7,8] which reduces system reliability.[9] On the one hand, just as shown in Figure 1, the person search task in a real-world application is always searching for the location of a specified pedestrian directly from the video frame containing several pedestrians as well as the background, while traditional one searching in the sequence of cropped pedestrian images. On the other hand, a network of joint pedestrian detection and person Re-ID can make full use of the correlation between two tasks and improve the performance of features.[10] And there is still much room for improvement in the end-to-end pedestrian search system.

Hence, an integrated network is needed, which can not only avoid the defects above but also improve the performance of feature extraction by making full use of the correlation between two tasks.

Accordingly, in this article, we propose a unified framework named Deformable Spatial Invariant Person Search Network (DSIPN), a network integrating pedestrian detection and person Re-ID, to solve the problem of person search.

First, the design of Deformable Pedestrian Space Transformer (DPST) offers a notable performance gain of person search. On the one hand, as shown in Figure 2, the spatial transformer equips framework with the ability to deal with spatial issues by cropping, resizing and rotating images. On the other hand, deformable pooling augments the spatial sampling locations in the modules with additional learnable offsets. What's more, inspired by Faster R-CNN,[11] which saw heavy use in object detection area, the region proposal net (RPN)-based Pedestrian Proposal Net with the spatial transformer is proposed. All these modules form the DPST to detect pedestrians and generate more robust feature after eliminating spatial variance.

Next, pair-wise or triplet distance loss functions are generally used to supervise the training process on person Re-ID research. However, when the data set is of giant scale, the sampling and computing will be complex and difficult to perform. Online learning is preferred in this case. So, we modified Online Instance Matching (OIM) loss[12] which can take into account unlabelled individuals to improve the performance of matching. Meanwhile, in order to make the most use of unlabelled samples, we also designed a strategy to form triplet loss for comparison.

To sum up, our work provides three main contributions: First, we proposed an integrated network named DSIPN, which takes panoramic images as input and outputs features for both pedestrian detection and person Re-ID. At the same time, a DPST is combined to correct the spatial variance of feature maps extracted from person samples and get more robust features with deformable pooling; second, we have improved the OIM loss function for the detection of person search tasks, which improves the accuracy and efficiency of the model training process; third, the triplet loss function for the person search problem is also proposed, which utilizes the unlabelled individual samples and further improves the accuracy of person search. Finally, notable performance gains are obtained compared to state-of-art on two public data sets CUHK-SYSU,[13] PRW[14] as well as a private data set SJTU318[15] collected by us.

This article is an extended version of our preliminary work.[15] Compared to our previous work, the framework and training supervision are improved, while more experiments are conducted, obtaining notable performance gains.

Related work

Pedestrian detection is an essential and significant task in a surveillance system, aiming at detecting and locating pedestrian from a complex background. Former methods such as the Integrate Channel Features detector,[16] which

extends the Viola–Jones framework,[17] relies on hand-crafted features with linear classifiers and has been improved in several ways, including ACF,[18] LDCF,[19] SCF.[20] Driven by the success of R-CNN[21] in general object detection, several deep learning based frameworks have been proposed. First, it is accomplished by combining hand-crafted features and boosted classifiers. Hosang et al.[22] use the SCF[20] pedestrian detector to propose regions and an R-CNN for classification; TA-CNN[23] employs the ACF detector[18] to generate proposals and trains an R-CNN-style network, exploits pedestrian and scene attribute labels to jointly optimize pedestrian detection with semantic tasks; the DeepParts method[24] applies the LDCF detector[19] to handle occlusion with an extensive part pool. Deep convolutional features are used to make an improvement as well. CompACT[25] proposes a complexity-aware boosting algorithm on top of hybrid hand-crafted and deep convolutional features. CCF detector[26] uses no region proposals. Zhang et al.[27] have no pyramid and are much faster and more accurate than Yang et al.[26] Early R-CNN[28] proposes proposals first and then applies classification and regression while Faster R-CNN[11] is proposed to greatly reduce the computational complexity and improves the detection accuracy. Subsequent researches find that the RPN network is quite effective in extracting pedestrian proposals, while the followed detection network is not performing well. Therefore, a series of improvements such as SA-FastRCNN,[29] HyperLearner[30] and so on are produced, basing on Faster R-CNN's framework. Another well-known network is the YOLO,[31] which considers detection as a regression problem and returns the targets' multiple locations and categories directly in the image. It is fast but incapable to deal with small and overlapping pedestrian targets.

Person Re-ID aims at matching the query person among numerous gallery samples from video sequences or static images collected from varies scenes.[32,33] It is widely used in video surveillance to perform crime prevention, cross-camera person tracking and person activity analysis, which makes it worth researching yet still challenging.[13,34–39] Existing works generally focus on three aspects: some[11,40–45] solve the problem with hand-crafted discriminative features; some[13,35,46–49] learn high-level features based on deep learning method; some[13,46–48] do innovation in the structure to gain performance. Two novel layers are designed by Ahmed et al.[46] to obtain relationships between two input person pair features; some[36,37,39,41,50–56] learn distance metrics for Re-ID; Koestinger et al.[50] proposes KISSME learning from equivalence constraints. Zhang et al.[54] learn a discriminative null space. Meanwhile, some researches addressed on abnormal images: Li et al.[51] learn a shared subspace across different scales dealing with the low-resolution person Re-ID problem. Zheng et al.[57] focuses on partially occluded images. Traditional deep learning methods for Re-ID mainly employs pair-wise or triplet distance loss functions[13,46,55,58] to supervise the training process. Li et al.[13] and Ahmed at al.[46] input a pair of cropped pedestrian images and employ a binary verification loss function. Ding et al.[46,58] exploit triplet samples to minimize the feature distance between the same person and maximize the distance between different people. But it can be considered complex if the data set is of a greater scale. Another approach is regarding Re-ID problem as a multi-classification problem and learning to classify identities with the Softmax loss function,[35] which effectively compares all samples at the same time. Also, as the number of classes increases, training the large Softmax classifier matrix would become much slower or even cannot converge.

Multitask learning is explored to improve the performance of some deep learning frameworks for Re-ID,[59] salient object detection[60] and deep cropping.[61] Also in our work, *person search* can be regarded as an integrated task combined two cascaded steps: pedestrian detection and person Re-ID. Given a query person, person search aims to match and locate all the same person that have appeared among a series of whole scene images sequence. A pedestrian detection system usually ignores the identification information of pedestrian samples in popular data sets like Caltech[62] and ETH[63] and only classifies the detected boxes as either positive or negative ones. Thus, simply combining them can't get perfect search results once the detection results are not good enough. There are only a few researchers devoting to handle person search task. Xu et al.[64] jointly models the commonness of people and the uniqueness of the queried person, using a sliding window searching strategy, which leads to low efficiency. Xiao et al.[12] and Zheng et al.[14] adopt two-stage strategies by fusing person Re-ID and detection into an integral pipeline and searching the interaction between the two tasks as well as overall performance. Xiao et al.[12] develops an end-to-end person search framework to jointly handle both aspects in a single CNN with OIM loss to train the network effectively. In some researches,[13,65–68] the Re-ID gallery only contains manually cropped pedestrian bounding boxes. Zheng et al.[14] contribute a novel large-scale data set PRW for person search and propose ID-discriminative Embedding (IDE) and Confidence Weighted Similarity (CWS) to improve the performance. Neural person search machines (NPSM)[10] coins an LSTM-based attention model, regarding person search as a detection-free process of gradually removing interference or irrelevant target persons for the query person.

Proposed method

In this section, we introduce the unified DSIPN framework which produces features for pedestrian detection and person Re-ID jointly. To share information and propagate gradients better, the DenseNet56 architecture is utilized. DenseNet[69] contains densely connected layers even between the first layer and the last layer. Such structure allows each layer in the network to make use of the input

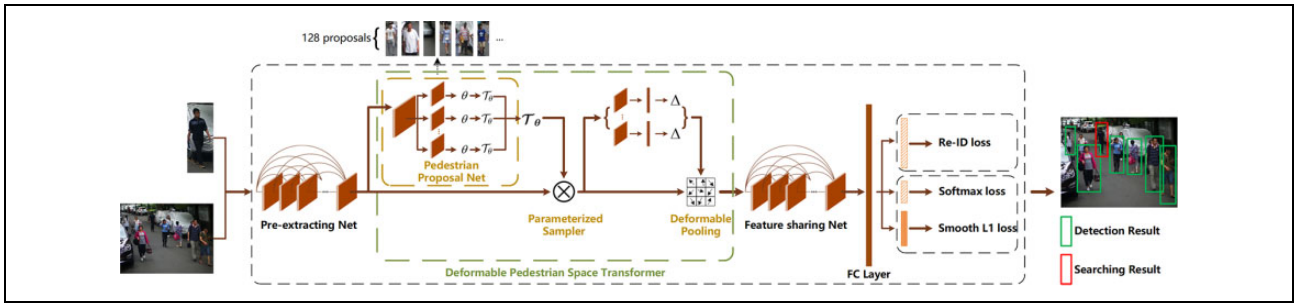

Figure 3. The structure of DSIPN. DSIPN is a cascaded structure based on DenseNet for processing pedestrian detection and person Re-ID jointly. It consists of three parts: Pre-extracting Net, DPST and Feature Sharing Net. Low-level features extracted by Pre-extracting Net are fed into DPST to generate pedestrian proposals and apply deformable spatial transformations. Feature Sharing Net extracts further down-sampled features to output results for both detection and Re-ID. DSIPN: Deformable Spatial Invariant Person Search Network; Re-ID: re-identification; DPST: Deformable Pedestrian Space Transformer.

and propagates the gradients from loss function directly to the initial layer to avoid gradient vanish. Then a DPST is incorporated into our model to improve the spatial invariance and extract more robust feature maps. In addition, an improved OIM loss is applied to supervise the training, as well as a designed strategy to form triplet samples for comparison.

Model structure

The unified person search framework consists of four main steps: image feature generation, pedestrian proposal generation, region feature extraction (pooling) and finally recognition(classification) or matching by metric learning. In our work, just as shown in Figure 3, the DSIPN model consists of three main parts. The first part is the Pre-extracting Net which extracts the low-level semantic features from the input panoramic image. The second part is the DPST which generates both pedestrian proposals and proposal features based on the low-level semantic information, performs deformable spatial transformation and scales the feature map to a fixed size. Finally, we have a Feature Sharing Net which further extracts high-level semantic features for both pedestrian detection and person Re-ID.

The Pre-extracting Net starts with a 7×7 convolution layer (stride $= 2$), batch normalization, rectified linear unit and a 2×2 max-pooling layer (stride $= 2$). Three dense blocks with 6, 12 and 24 dense layers are added behind, respectively. We set the growth layer as 32. In order to ensure that the input image would be pooled for four times (by 2×2 max pooling layer), the initial maximum pooling layer at the forefront of the pre-extracting is removed. The resolution of the output feature map with 512 channels is 1/16 of the original input image.

The DPST part starts with Pedestrian Proposal Net to generate pedestrian proposals. The following modules are a spatial transformer and a deformable pooling to implement deformable spatial transformation to the generated proposals. The structure of DPST will be further illustrated in the next subsection.

The Feature Sharing Net is composed of the final dense block of DenseNet containing 16 dense layers and a growth rate of 32, followed by a global average pooling layer to sample the feature map into a 1024-dimensional vector and three fully connected layers to map the vector to 2D, 8D, 256D respectively for classification (pedestrian or background), pedestrian position coordinate information and person Re-ID.

At the end of the model, a Softmax classifier is used to deal with 2D vector, which outputs the classification of pedestrian detection. The 8D vector is fine-tuned by linear regression to generate the corresponding refined localization coordinates. As for the 256-dimensional vectors, they will be L2-normalized first and then to compare with corresponding feature vectors of the target person for person Re-ID. Here we apply triplet loss and OIM loss for supervision separately for comparison.

Deformable Pedestrian Space Transformer

As Figure 1 shows, the spatial variance exists between pedestrian samples, which is caused by viewpoints, occlusions, and resolution, etc.

Though general CNN defines an exceptionally powerful class of models, it can only guarantee the translation invariance of the input samples. Also, handcrafted design of invariant algorithms cannot meet the demand of dealing overly complex transformations, known or unknown, let alone large unknown ones.

So, in our work, we introduce a new learnable module: The DPST, which brings our framework the ability to actively apply deformable spatial transformations on feature maps, conditional on the feature map itself, without any extra training supervision or modification to the optimization process.

The first step is applying a Pedestrian Proposal Net to generate pedestrian proposals. Inspired by Faster R-CNN, k kinds of scales as well as aspect ratios of anchors (k^2 kinds

in total) are present in the DPST to predict the pedestrian position. We first apply a 3×3 convolution layer (stride = 1) and get a feature map of 512 channel, for every position in which k^2 anchors were predicted. Then the feature map is feed into three kinds of 1×1 convolutional layers (stride = 1) to generate feature maps with channel dimensions of $2k^2$, $4k^2$ and $6k^2$, respectively, which have the same size as the output feature map of the pre-extracting network. These three kinds of features generated correspond to binary categories, location information and spatial transformation coefficients, respectively.

After these steps, we have predicted k^2 candidate boxes at each position of the feature map. Given the input image of size $H \times W$, the feature map size of the pre-extracting network output is approximately $\frac{1}{16}H \times \frac{1}{16}W$, and the number of candidate frames generated is approximately $\frac{1}{16}H \times \frac{1}{16}W \times 9$. Assuming H is 560 and W is 1000, the amount of anchors is $36 \times 63 \times 9 \approx 20k$. However, for an image, it is unnecessary to use all these anchors. Therefore, according to the predicted category scores and non-maximum suppression method, 128 candidate anchors are selected as the final output proposal boxes. Coordinate offsets and transformation parameters are selected correspondingly.

Spatial transformer. Then the DPST will implement spatial transformation to these proposals with transformation parameters generated before. Let the input pedestrian feature map be $U \in \mathbb{R}^{H \times W \times C}$, where W,H and C denotes the width, height and channel of the input feature map. And the output feature map is $V \in \mathbb{R}^{H' \times W' \times C}$. The spatial transformation is accomplished by a parameterized sampling grid $G^t = \{G_i^t\} = \{(x_i^t, y_i^t)\}$ representing the target location set of the output feature map. As an element of a genetic feature map, each output pixel coordinate (x_i^t, y_i^t) is computed by applying a sampling kernel centred at a corresponding position $G_i^s = (x_i^s, y_i^s)$ in the input source feature map. So the transformation $\mathcal{T}_\theta$ used in this article is a 2D affine transformation described as

$$G_i^s = \begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta G_i^t = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (1)$$

Here we use normalized coordinates, such that $-1 \le x_i, y_i \le 1$.

Parameterized sampler. After getting the transformation parameter, the parameterized sampler is applied with a set of sampling points G_i^s to sample the origin feature into the transformed feature.

Since the location of $G_i^s = (x_i^s, y_i^s)$ may not be integers, in order to perform such a spatial transformation in our network which allows applying cropping, translation, rotation and scaling operations to the input feature maps, a sampler must take the set of sampling points $G_i^s = (x_i^s, y_i^s)$ then output the sampled transformed feature map V. This can be written as

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \phi_x) k(y_i^s - m; \phi_y) \quad \forall i \in [1, ..., H'W'], \forall c \in [1, ..., C] \quad (2)$$

$k()$ defines the sampling kernel applying image interpolation where ϕ_x and ϕ_y are the parameters. U_{nm}^c is the value at location (n, m) in channel c of the input feature, while U_i^c is the output value at location (x_i^t, y_i^t) in the target feature map. In this article, we choose the bilinear sampling kernel, giving

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

For this kind of sampling, in order to allow back propagation of the loss through the optimization, we define the gradients with respect to U_{nm}^c and x_i^s, y_i^s

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

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

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

Obviously, for transformation parameters θ, $\frac{\partial x_i^s}{\partial \theta}$ and $\frac{\partial y_i^s}{\partial \theta}$ can also be delivered. So, the sampling process is differentiable and able for gradients to flow back to the transformation parameters θ and localization network.

For each proposal, the sampler outputs a transferred feature map for the proposal as well as the whole scene image.

Deformable pooling. In order to convert the input regions of arbitrary size into features with a fixed size, Region of interest pooling (also known as RoI pooling) is widely used in regular region proposal-based object detection methods.[11,21,28,70]

In spatial invariant person search network (SIPN),[15] by setting the sampled feature V to a fixed size, the parameterized sampler plays a similar role with RoI pooling. However, the Space Transformer just applied a single kind of

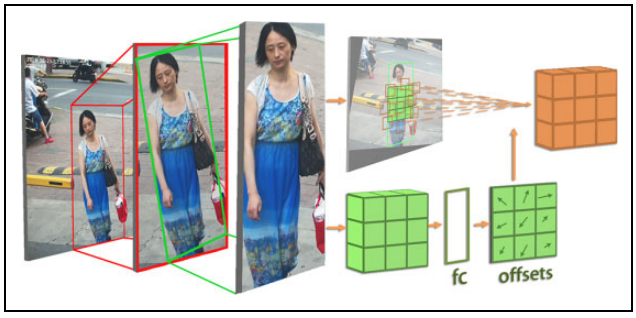

Figure 4. The details of deformable space transform: For every proposal generated by Pedestrian proposal Net, first we apply spatial transform to the origin proposal features extracted from the whole image feature, then a general pooling layer with an FC layer to get the offsets of every grid for deformable pooling. Finally, we apply deformable pooling at the transformed whole scene image to extract the final size-fixed features of the corresponding proposal by regarding it as a RoI.

offset for all sampling points in the feature map, which is less likely to take advantages of the spatial information. What's more, after pedestrian proposal's generation, the background information is not well used either.

Therefore, on the basis of regular pooling, a deformable pooling containing two stages is designed. At the first stage, we apply general pooling to the transformed feature of a proposal. Then, in the second stage, the pedestrian proposal is regarded as an RoI. During the pooling, for each bin, an offset is learned, then added to the bin centre.

For general RoI pooling, given the input transformed feature map V and a RoI b, the features W is generated as

$$W = RegionFeat(V, b) \tag{7}$$

Specifically, in the current RoI pooling practice, RoI b is divided into $k \times k$ bins. After pooling, the output is a $k \times k$ feature map W. For $(i,j)_{th}$ bin $(0 \leq i,j < k)$, we have

$$W(i,j) = \frac{1}{N_{ij}} \sum_{p \in (i,j)_{th} bin} V(p) \tag{8}$$

where N_{ij} is the number of pixels in the bin.

Similarly as in equation (8), offsets $\Delta_{p_{ij}}$, where $0 \leq i,j < k$, are added to the spatial bins positions then

$$W(i,j) = \frac{1}{N_{ij}} \sum_{p \in (i,j)_{th} bin} V(p + \Delta_{p_{ij}}) \tag{9}$$

For the same reason that $\Delta_{p_{ij}}$ may not be an integer, it is implemented via bilinear interpolation as in equation (3).

Figure 4 illustrates how the offsets are obtained. For every proposal in an image, space transformation is applied to it to generate the transformed feature. Then a general pooling is applied to the transformed feature of the proposal, followed by a fully connected layer (FC) which generates the normalized offsets $\Delta_{\hat{p}_{ij}}$.

Finally, the offsets $\Delta_{\hat{p}_{ij}}$ are transformed to $\Delta_{p_{ij}}$ by element-wise product with the pedestrian's width and height, as $\Delta_{p_{ij}} = \gamma \Delta_{\hat{p}_{ij}} \circ (w,h)$. The FC layer is learned by back propagation, the gradient with respect to the offset $\Delta_{p_{ij}}$ can be computed by

$$\frac{\delta y(i,j)}{\delta \Delta_{p_{ij}}} = \frac{1}{N_{ij}} \sum_{p \in (i,j)_{th} bin} \frac{\delta V(p + \Delta_{p_{ij}})}{\delta p_{ij}} \tag{10}$$

here $V(p + \Delta_{p_{ij}})$ is computed via bilinear interpolation as equation (3). And the gradient w.r.t., the normalized offsets $\Delta_{\hat{p}_{ij}}$ can be easily obtained via computing derivatives in $\Delta_{p_{ij}} = \gamma \Delta_{\hat{p}_{ij}} \circ (w,h)$.

As stated above, in DSIPN, we use such a pedestrian transformation network to prevent spatial variance of detected proposals and a deformable pooling to extract more robust feature for Re-ID in person search.

Loss function

The training of DSIPN can be divided into two main parts: the training of pedestrian detection and person Re-ID.

For pedestrian detection, the result outputs by two fully connected layers: classification layer and regression layer, whose output dimension is 2 and 4, respectively. Therefore, we use Softmax loss $\mathcal{L}_{cls}$ and smooth L1 loss $\mathcal{L}_{loc}$ together to supervise detection learning

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{loc} \tag{11}$$

where λ is the hyper-parameter of the network that balances the two loss functions.

For person Re-ID, the loss functions can be roughly divided into classification-based loss functions and comparison-based loss functions. However, in person search task, pedestrians are divided into two categories: one is *Labelled Identity* who occurs more than once and can be used as the matching target; the other is *Unlabelled Identity* who appears only once in the entire data set. Since it is not possible to treat *Unlabelled Identity* as a category, simply adopting a classification loss function is unachievable.

To deal with the problem, we propose an improved OIM loss, which contains directional constraints and is capable of taking advantages of unlabelled identities features.

Taking OIM for example, when using classification-based RE-ID loss function for pedestrian search tasks, the training progress is regarded as a multi-classification task. However, in the test progress, since there is no pedestrian intersection between the test set and the training set, the trained model cannot be used directly in the test set. Hence, the final fully connected layer should be removed and pedestrian identification is performed by measuring the distance or the similarity between them. Thus, training and testing use different judging methods. So, our work also considers a unified approach, which uses the Re-ID loss

function based on comparison to supervise the pedestrian search network.

The details of the two algorithms are introduced in the following subsections. In summary, the overall loss function used by DSIPN is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{loc} + \eta\mathcal{L}_{reid} \tag{12}$$

where η is the hyper-parameter of the equilibrium Re-ID loss function which is set according to the Re-ID loss function we used.

Improved OIM in person search

OIM loss is first presented by Xiao et al.[12] By making full use of non-labelled individual samples in the data set as negative samples, and basing on the overall samples instead of the samples in the current small patch, the loss function becomes much easier to converge.

For all labelled identities, OIM created a *Look-Up Table (LUT)* to store the normalized features of each pedestrian identity. Assume we have L labelled identities in the training set, and the final output feature map of DSIPN has the dimension of D, then we will have LUT as $\mathbf{V} \in \mathbb{R}^{D\times L}$. Each column of the matrix $\mathbf{v_t}$ represents a feature vector of a labelled identity. Given an output feature $\mathbf{x} \in \mathbb{R}^D$, the similarity between the labelled identity and all pedestrians in the data set can be computed as $\mathbf{V}^T\mathbf{x}$. If $\mathbf{x}$ represents a sample of labelled identity t, the t_{th} column will be updated as

$$\mathbf{v_t} \leftarrow \gamma\mathbf{v_t} + (1-\gamma)\mathbf{x} \tag{13}$$

where $\gamma \in [0, 1]$. Then $\mathbf{v_t}$ will be L2-normalized.

For unlabelled identities, OIM creates a Circular Queue which stores the normalized features to represent them. As the length of the queue is specified to Q, it can be defined as $\mathbf{U} \in \mathbb{R}^{D\times Q}$. During propagation, we insert samples of unlabelled identity $\mathbf{x}$ into the tail of the queue while popping the out-of-date feature out to update $\mathbf{Q}$. Meanwhile, the similarity between the identity vector $\mathbf{x}$ and the samples in the Circular Queue $\mathbf{U}$ can be expressed as $\mathbf{U}^T\mathbf{x}$. Then we supervise the training by minimizing distance between same identities $\mathbf{V}^T\mathbf{x}$ while maximizing $\mathbf{U}^T\mathbf{x}$.

The DPST will produce 128 proposals during the process. Although some of the proposals have relatively high intersection-over-union (IoU) with ground truth (GT), which seems suitable for updating LUT, the noise caused by background or missing information still exists. Their overlapping with the real box is misleading. Taking such multiple-target candidate boxes as positive samples may results in identity feature vector's disorder and lack of representation.

Therefore, in our work, more restrictive restrictions are added: just as shown in Figure 5, we only update LUT with ground truth feature $\mathbf{x_t}$, $t \in \mathbb{G}$. Here $t \in \mathbb{G}$ denotes the

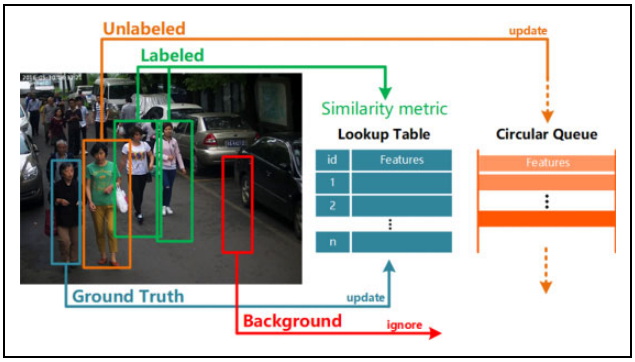

Figure 5. The strategy of improved OIM loss function. There are four kinds of bounding boxes in a whole scene image: GT is used to update the Look-up Table by equation (13); proposals regarded as a sample of labelled identity is used to compute the similarity and loss for back propagation; proposals regarded as a sample of unlabelled identity is going to be inserted at the tail of the Circular Queue while popping the head feature out; proposals regarded as a background is ignored in the loss function. OIM: Online Instance Matching; GT: ground truth.

ground truth bounding boxes of the identity t. The update process of LUT is

$$\mathbf{v_t} \leftarrow \gamma\mathbf{v_t} + (1-\gamma)\mathbf{x}\,, \text{if } t \in \mathbb{G} \tag{14}$$

In this way, LUT will be more robust and calculation cost will be reduced while updating. As Xiao et al.[12] declare the probability of $\mathbf{x}$ being identified as a sample of identity i is defined by a Softmax function

$$p_i = \frac{\exp(\mathbf{v_i^T x})}{\sum_{j=1}^{L}\exp(\mathbf{v_j^T x}) + \sum_{k=1}^{Q}\exp(\mathbf{u_k^T x})} \tag{15}$$

while the probability of being a sample of the i_{th} unlabelled individual in the Circular Queue is

$$q_i = \frac{\exp(\mathbf{u_i^T x})}{\sum_{j=1}^{L}\exp(\mathbf{v_j^T x}) + \sum_{k=1}^{Q}\exp(\mathbf{u_k^T x})} \tag{16}$$

Finally, we maximize the log-likelihood, the improved OIM loss function is expressed as

$$\mathcal{L}_{reid_{oim}} = E_{\mathbf{x}}[\log(p_t)] \tag{17}$$

The overall loss function is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{loc} + \eta\mathcal{L}_{reid_{oim}} \tag{18}$$

Pedestrian triplet loss in person search

The previous section mentions that person search can be supervised with the improved OIM loss function. However, through experiments, we find that changing the size of the Circular Queue or even removing it has little effect on the result of person search, which means that the design

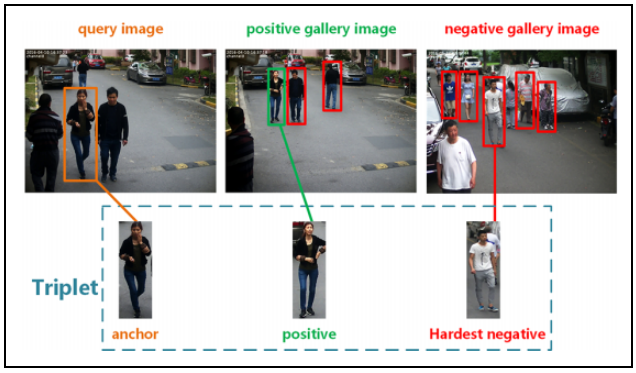

Figure 6. The strategy of triplet loss: (1) A whole scene image is selected first to be a query image and a query with its GT location from the image is chosen subsequently to be an anchor pedestrian. (2) Images which contain samples of the same identity as the query make up positive gallery with positive samples to be traversed. (3) The rest images make up the negative gallery, from which a hardest negative sample is finally chosen. GT: ground truth.

of the Circular Queue does not make full use of unlabelled individuals, and it necessitates a number of redundant learnable parameters. To make better use of unlabelled identities in person search data set, we propose a triplet loss that regards all unlabelled identities as negatives.

The triplet loss function is first proposed by Schroff et al. in FaceNet[71] and has been widely used in image retrieval tasks. Chen et al.[72] also explored triplet loss to a deep quadruplet network for person Re-ID. It is formulated as following

$$\mathcal{L}_{triplet} = \sum_i [D(x_i^a, x_i^p) - D(x_i^a, x_i^n) + m]_+ \tag{19}$$

here x_i^a, x_i^p, x_i^n represents feature of an anchor image of a specific person, a positive image of the same person and a negative image of any other person, respectively. $[x]_+$ represents $\max(0)$. In FaceNet,[71] $D(x,y) = ||x-y||_2^2$. And in some Re-ID works,[10,14] cosine similarity is used to measure the similarity between features of target person and candidate person, contrary to the distance measurement like Euclidean distance, which is used in our experiment

$$D(x,y) = \sqrt{\sum_j (x_j - y_j)^2} \tag{20}$$

However, different from normal face recognition or person Re-ID, we are not sure how many pedestrians there in a whole scene image so the pedestrians cannot be directly utilized for triplet or divided into batches of triplets. To solve this problem, we developed a strategy that utilizes the whole scene image to form a triplet. Specifically, as shown in Figure 6, we select a whole scene image first to be an anchor image and choose one pedestrian with its GT location from the image to be an anchor pedestrian called query. Then for the positive gallery, we select images which contain samples of the same identity as the query. The rest images make up the negative gallery. For each query image, we traversed each image from the positive gallery as a positive sample, along with a randomly chosen image from the negative gallery as a negative sample. Finally, we collect the output pedestrian proposals, divide them into positive samples and negative samples according to their real labels to form triplets. The triplet loss function of the current network parameters is defined as

$$\mathcal{L}_{trp}(\omega) = \sum_{\substack{p,n \\ i_a = i_p \neq i_n}} [m + D_{a,p} - D_{a,n}]_+ \tag{21}$$

where the parameter ω that minimizes $\mathcal{L}_{trp}(\omega)$ is to be estimated during optimization. i_a, i_p, i_n represent the identity label (person ID) of the anchor samples, positive samples and negative samples, respectively. $D_{a,p}$ and $D_{a,n}$ are corresponding distances. m is a manually designed margin between them which means $D_{a,n}$ is supposed to be at least m bigger than $D_{a,p}$. However, simply throwing all samples may cause explosive calculation. For an anchor, assuming there are N_p positive samples and N_n negative samples, then we will have $N_p N_n$ triplets. Therefore, we apply a strategy of hard mining with samples to find the 'hardest' samples whose loss would be computed to optimize the net work. 'Hardest' positive sample p^* is the positive sample with the largest distance among all positive samples from the anchor pedestrian, while 'hardest' negative sample n^* is the closest negative sample. The final loss function is defined as

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{loc} + \eta \mathcal{L}_{reid_{trp}} \tag{22}$$

where

$$\mathcal{L}_{reid_{trp}} = [m + D_{a,p^*} - D_{a,n^*}]_+ \tag{23}$$

Experiments

To demonstrate the effectiveness of our approach and study the impact of various factors on person search performance, we conduct several comprehensive experiments on three person search data sets.

In this section, we first introduce the data sets we used. Then evaluation metrics and training settings are shown in the next two subsections. In the following subsection, we reveal the performance of our DSIPN and compare our work with previous separate and joint works for person search. At last, we discuss the influence of various factors, including DPST, improved OIM, triplet loss and different backbone network structures.

Data sets

Currently, the problem of person search is not widely considered. The images in traditional person Re-ID data sets

are cropped in advance, makes them unable to be utilized in person search. Our work is tested on two public person search data sets: CUHK-SYSU,[13] PRW[14] and a private data set SJTU318[15] collected by ourselves.

CUHK-SYSU is composed of two parts: pedestrian pictures in urban area taken by hand-held cameras and screenshots of movies. It contains 18,184 images with 96,143 pedestrians bounding boxes, 8432 labelled identities in total. We take 5532 identities as the training set. The left is used for testing. In CUHK-SYSU, pedestrian samples with most parts blocked by obstacles or poor postures (sitting, kneeling, etc.) are not labelled. Additionally, the same pedestrian with a large change in appearance (such as different clothes or decorations) is labelled as different identities. Samples with an image of height less than 50 pixels are not labelled as well, due to identification difficulty. In conclusion, this data set is quite suitable for person search.

PRW (Person Re-identification in the Wild) data set is built from 10 h of surveillance video recorded in Tsinghua University. Five 1080 × 1920 HD and one 576 × 720 SD cameras are used. It contains 11,816 images with 43,110 pedestrians bounding boxes, 933 labelled identities in total. We take 483 identities as the training set. The left 450 identities are used for testing. Multiple-source cameras and diverse filming angles bring challenges when applying. There are fewer pedestrians in this data set, but more samples for each individual. However, if different cameras are taken as another data dimension, searching target range can be expanded to 2057. The person recorded in the current camera will be searched in images taken by different cameras. This cross-camera searching is closer to the actual application scenario, while much more difficult.

SJTU318 is another large-scale person search data set collected by ourselves. It is transferred from raw uptown surveillance videos. There are twelve 1200 × 1600 HD cameras in total. As shown in Figure 7, these twelve cameras distribute on the gates and roads of the uptowns. We not only sampled daytime surveillance videos but also involved some nightly scene. Our data set consists of 14,610 whole scene images including 63,755 pedestrian bounding boxes, among which 13,067 pedestrians are annotated with 621 IDs. We picked 244 identities for training and 202 identities for testing. There are samples of the same identity wearing different clothes or hairstyle. So, it is a rather challenging data set due to the lower resolution of the person in the image, changes of light, scenes and pedestrian appearances. In general, the data set is much closer to real-life scenarios.

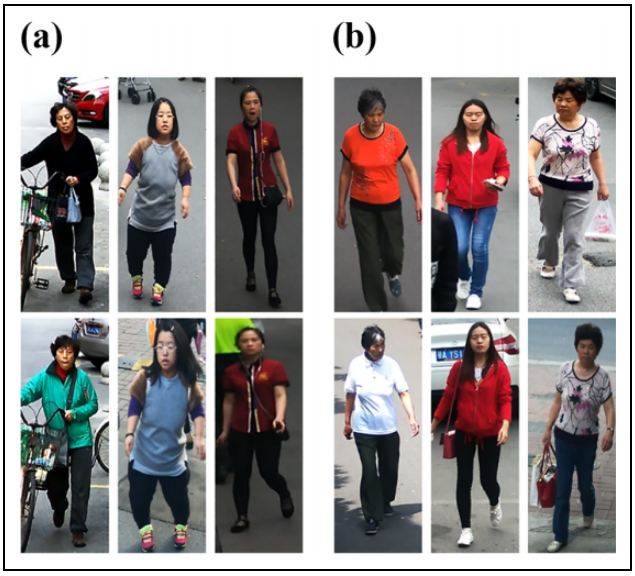

Figure 7. Several samples from our private data set SJTU318; (a) Outdoor cameras are distributed on the gates and roads of the uptowns, which brings changes of light and scene. (b) Challenging pedestrian samples in the data set, some pairs of pedestrians have different appearances.

Evaluation metrics

Our work divides the above three data sets into training sets and test sets, ensuring no same identities shared. During testing, a target pedestrian picture or panoramic picture with target pedestrian location information is given, called a query. Our target is to match all the present identities of query in the gallery of whole scene images as well as detect every other pedestrian. Considering the difficulty of the experiment, we set multiple gallery sizes to evaluate the performance of person searches for each data set: 100 for CUHK-SYSU and SJTU318, 1000 for PRW. The following results are reported using the protocol with such gallery sizes if not specified.

Similar to the person Re-ID task, person search is a branch of retrieval task. So, we use the Cumulative Matching Characteristics (CMC top-K) and mean Average Precision (mAP) to evaluate the performance of our framework. CMC top-K calculates the probability that the results can do correct prediction which overlaps the GT with IoU ≥ 0.5 within the top-K predicted bounding boxes. An average precision (AP) is computed for each target person image based on the precision–recall curve. mAP is the average of APs and focuses on the entire output list, which means that with higher mAP the last one of the positive samples is more likely to be placed in the front of the list.

Training settings

The experiments are all implemented on PyTorch with Python3.6, the operating system is Ubuntu16.04 with 1080Ti. We use the Stochastic Gradient Descent to train DSIPN; random horizontal flipping is used to apply data augmentation; batch size is set to 1. We initialize the learning rate to 0.0001 then reduce it to $\frac{1}{10}$ of the original after the whole data set being iterated for 3 times and 6 times; the entire data set is trained and iterated 10 times. VGG, ResNet and DenseNet are used as backbone network for comparison.

Table 1. Searching performance of DSIPN with improved OIM loss on CUHK-SYSU data set with CMC top-1 and mAP comparing with previous related work.

	CCF	ACF	FRCN	GT
CMC top-1 (%)				
DSIFT + Euclidean[12]	11.7	25.9	39.4	45.9
DSIFT + KISSME[12]	13.9	38.1	53.6	61.9
BoW + Cosine[12]	29.3	48.4	62.3	67.2
LOMO + XQDA[12]	46.4	63.1	74.1	76.7
IDNet[12]	57.1	63.0	74.8	78.3
OIM (baseline)[12]	–	–	78.7	80.5
Yang et al.[9]	–	–	80.6	–
NPSM[10]	–	–	81.2	–
SIPN[15]	–	–	84.0	84.2
Ours	–	–	**84.1**	84.3
mAP (%)				
DSIFT + Euclidean[12]	11.3	21.7	34.5	41.1
DSIFT + KISSME[12]	13.4	32.3	47.8	56.2
BoW + Cosine[12]	26.9	42.4	56.9	62.5
LOMO + XQDA[12]	41.2	55.5	68.9	72.4
IDNet[12]	50.9	56.5	68.6	73.1
OIM (baseline)[12]	–	–	75.5	77.9
Yang et al.[15]	–	–	77.8	–
NPSM[10]	–	–	77.9	–
SIPN[15]	–	–	84.2	84.5
Ours	–	–	**84.5**	84.6

DSIPN: Deformable Spatial Invariant Person Search Network; OIM: Online Instance Matching; GT: ground truth; CMC: Cumulative Matching Characteristics; map: mean Average Precision; BoW: Bag of Word; LOMO: Local Maximal Occurrence; DSIFT: Dense Scale-invariant feature transform. The best results are marked in bold.

Table 2. Searching performance of DSIPN with improved OIM loss on the PRW data set with CMC top-1 and mAP comparing with previous related work.

Method	mAP (%)	Top-1 (%)
DPM-Alex + LOMO + XQDA[10]	13.0	34.1
DPM-Alex + IDE_{det}[10]	20.3	47.4
DPM-Alex + IDE_{det} + CWS[10]	20.5	48.3
ACF-Alex + LOMO + XQDA[10]	10.3	30.6
ACF-Alex + IDE_{det}[10]	17.5	43.6
ACF-Alex + IDE_{det} + CWS[10]	17.8	45.2
LDCF + LOMO + XQDA[10]	11.0	31.1
LDCF + IDE_{det}[10]	18.3	44.6
LDCF + IDE_{det} + CWS[10]	18.3	45.5
OIM[12]	21.3	49.9
NPSM[10]	24.2	53.1
SIPN[15]	28.2	57.8
Ours	**34.3**	**68.4**

DSIPN: Deformable Spatial Invariant Person Search Network; OIM: Online Instance Matching; CMC: Cumulative Matching Characteristics; map: mean Average Precision; LOMO: Local Maximal Occurrence; CWS: Confidence Weighted Similarity; IDE: ID-discriminative Embedding; DPM: Deformable Parts Model. The best results are marked in bold.

Table 3. Detection performance of DSIPN on CUHK-SYSU, PRW and SJTU 318 data sets with recall and AP comparing with previous related work.

	CUHK-SYSU		PRW		SJTU318	
	Recall (%)	AP (%)	Recall (%)	AP (%)	Recall (%)	AP (%)
OIM (baseline)[12]	79.49	74.93	90.20	84.26	73.97	60.03
SIPN[15]	78.45	75.14	89.91	85.60	72.46	59.98
Ours	**79.54**	**75.35**	**89.64**	**85.72**	**73.91**	**61.60**

DSIPN: Deformable Spatial Invariant Person Search Network; OIM: Online Instance Matching; AP: average precision. The best results are marked in bold.

When experimenting with the performance of improved OIM, we set different sizes of Circular Queue: 5000 for CUHK-SYSU, 500 for PRW and 250 for SJTU318.

Performance of DSIPN

In order to illustrate the robustness of DSIPN, we test it against some existing works.

When dealing with person search problem, some choose to use a two-phase model, which detects pedestrian first, Re-ID later. ACF, CCF and Faster CNN (FRCN) are used in the first phase. DenseSIFT-ColorHist, Bag of Words (BoW), and Local Maximal Occurrence (LOMO) are used in the second phase. Distance metrics involved are Euclidean, Cosine Similarity, KISSME and XQDA. Integrated models similar to our work are proposed as well. Such as one based on Faster R-CNN and ResNet-50 proposed by Xiao et al.,[12] one adding artificial design features proposed by Yang et al.,[9] and another attention model based on LSTM proposed by Liu et al.[10]

Tables 1 and 2 show the performance and comparison of DSIPN with improved OIM loss on CUHK-SYSU and PRW with CMC top-1 and mAP, respectively. Especially, in Table 1, we also use the GT bounding boxes as the results of a perfect detector in CUHK-SYSU data set. The results show that the performance of our cascaded framework is extremely close to the results even with the perfect detector, which further proved that the performance of detection has little effect.

Meanwhile, as shown in Table 3, we measure the detection ability by recall (%) and AP (%) on the three data sets too. DSIPN outperforms previous frameworks. It also has a good detection performance as well, which demonstrates that by spatially transforming and deformable pooling feature maps of pedestrians, the performance of person search can be improved.

Comparison and impact of factors

Impact of DPST. The DSPT from our work achieves the same function as RPN network in Faster R-CNN with a sampler capable of deformable spatial transforming. At the same time, the pedestrian feature map is spatially transformed to prevent or correct the spatial variation caused

Table 4. The contribution of DPST: Comparison between OIM baseline and DPST (with OIM loss function and Resnet50 for backbone) on PRW and CUHK-SYSU data sets.

		mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
PRW	OIM (baseline)[12]	21.3	49.9	72.9	81.5
	DSIPN	**33.4**	**64.6**	**83.7**	**89.6**
CUHK-SYSU	OIM (baseline)[12]	75.5	78.7	82.1	85.8
	DSIPN	**83.7**	**80.0**	**84.4**	**88.3**

DPST: Deformable Pedestrian Space Transformer; DSIPN: Deformable Spatial Invariant Person Search Network; OIM: Online Instance Matching; mAP: mean Average Precision.The best results are marked in bold.

Table 5. Comparison of searching performance of DSIPN between multiple backbone network structures on PRW and CUHK-SYSU data sets.

		mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
PRW	VGG16	21.8	33.3	63.8	75.3
	Res34	32.9	79.1	82.5	85.2
	Res50	34.1	63.6	84.3	89.9
	Dense121	**34.31**	**68.35**	**86.92**	**91.69**
CUHK-SYSU	VGG16	44.4	42.4	64.5	72.0
	Res34	58.7	59.7	76.3	81.5
	Res50	83.9	80.9	85.1	88.9
	Dense121	**84.49**	**84.10**	**88.34**	**90.76**

DSIPN: Deformable Spatial Invariant Person Search Network; mAP: mean Average Precision. The best results are marked in bold.

by different resolutions or viewing angles. And with deformable pooling, more robust feature vectors which facilitate the determination of pedestrian identity would be extracted. The baseline model[12] proposed by Xiao et al. is applied with OIM loss and ResNet-50. In Table 4, we compare the baseline model[12] without DPST with our model. Here backbone network ResNet-50 and OIM loss were applied in our model in order to demonstrate the compact of DPST. The comparison accomplished on PRW and CUHK-SYSU data sets shows that DPST makes a notable improvement on the performance of the network.

Comparison of different backbone networks. DenseNet-121 is used as backbone network in our work. To analyse the influence of different network structures on pedestrian searching performance, we also applied VGG16, ResNet-34 and ResNet-50 as backbone. VGGNet is very appealing because of its uniform architecture. ResNet (Residual Neural Network) can easily enjoy accuracy gains from greatly increased depth, producing high-quality results. And in DenseNet, all layers have direct access to every feature map from all preceding layers, while exhibiting no optimization difficulties. We compare the performance of these architectures to extract deeper features for person search. Table 5 shows the comparison between multiple backbone network structures on PRW data set, where DPST is combined in the backbone. The results show that DenseNet outperforms the others.

Table 6. Comparisons of searching performance of DSIPN between using OIM loss and our improved OIM loss function for supervision on PRW, CUHK-SYSU and SJTU318 data sets.

		mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
PRW	OIM loss[12]	33.6	55.7	75.2	81.9
	Improved OIM	**34.3**	**68.4**	**87.0**	**91.7**
CUHK-SYSU	OIM loss[12]	84.2	83.9	87.5	90.6
	Improved OIM	**84.5**	**84.1**	**88.3**	**90.8**
SJTU318	OIM loss[12]	31.2	66.0	72.6	74.5
	Improved OIM	**32.3**	**67.8**	**76.7**	**78.7**

DSIPN: Deformable Spatial Invariant Person Search Network; OIM: Online Instance Matching; mAP: mean Average Precision.The best results are marked in bold.

Table 7. Comparisons of performance of DSIPN between using OIM loss and triplet loss for supervision on PRW, CUHK-SYSU and SJTU318 data sets.

		mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
PRW	OIM loss[12]	29.6	55.7	75.2	81.9
	Triplet loss	**34.5**	**68.8**	**87.9**	**93.1**
CUHK-SYSU	OIM loss[12]	84.2	**83.9**	87.5	**90.6**
	Triplet loss	**84.3**	83.8	**87.6**	89.6
SJTU318	OIM loss[12]	31.2	66.0	72.6	74.5
	Triplet loss	**34.6**	**69.1**	**77.2**	**79.2**

DSIPN: Deformable Spatial Invariant Person Search Network; OIM: Online Instance Matching; mAP: mean Average Precision.The best results are marked in bold.

Loss function comparison. Two loss functions are conducted in our work: an improved OIM loss function and triplet loss function. Tables 6 and 7 show the experiment results of the two loss functions with OIM loss function, both of which contribute to the person search task. Moreover, during the training process, with improved OIM loss, the optimization is easier to converge. Triplet loss takes more time, but it has a slightly higher precision on the PRW and SJTU318, and the performance on CUHK-SYSU is similar to that of OIM loss.

Conclusion

In this article, we propose a DenseNet-based cascaded network structure DSIPN to solve the problem of person search. A DPST is introduced in DSIPN, in which pedestrian proposals are generated and pedestrian feature maps are spatially transformed. With the deformable pooling after it, more robust and spatially invariant features are extracted in the subsequent network. We also compare two

loss functions: improved OIM loss, which reduces the amount of computation while considering unlabelled samples; triplet loss, which makes better use of unlabelled samples in the data set. All in all, DSIPN is able to solve the person search problem end-to-end, and simultaneously output the results of pedestrian detection and person Re-ID. Its performance is also improved compared to state-of-art works.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by National Natural Science Foundation of China (NSFC, grant nos 61771303 and 61671289), Science and Technology Commission of Shanghai Municipality (STCSM, grant nos 17DZ1205602 and 18DZ1200102) and SJTU-Yitu/Thinkforce Joint laboratory for visual computing and application. Director Fund of PSRPC.

ORCID iD

Yuan Hong https://orcid.org/0000-0001-7333-4499

References

1. Wang X. Intelligent multi-camera video surveillance: a review. *Pattern Recogn Lett* 2013; 34(1): 3–19.
2. Loy CC, Xiang T, and Gong S. Multi-camera activity correlation analysis. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, Florida, USA, 20–25 June 2009, pp. 1988–1995.
3. Gong S, Cristani M, Loy CC, et al. The re-identification challenge. In: Gong S, Cristani M, and Yan S, et al. (eds) *Person re-identification*. London: Advances in Computer Vision and Pattern Recognition, Springer, 2014, pp. 1–20.
4. Zhang J, Zheng Y, Qi D, et al. Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artif Int* 2018; 259: 147–166.
5. Gong S and Xiang T. *Visual analysis of behaviour – from pixels to semantics*. London: Springer, 2011.
6. Gong S, Cristani M, Yan S, et al. (eds) *Person re-identification*. London: Advances in Computer Vision and Pattern Recognition, Springer, 2014.
7. Chen S, Guo C, and Lai J. Deep ranking for person re-identification via joint representation learning. *IEEE Trans Image Proc* 2016; 25(5): 2353–2367.
8. Bak S, and Brémond F. Re-identification by covariance descriptors. In. Gong S, Cristani M, Yan S, et al. (eds) *Person re-identification*. London: Advances in Computer Vision and Pattern Recognition, Springer, 2014, pp. 71–91.
9. Yang J, Wang M, Li M, et al. Enhanced deep feature representation for person search. In: *Proceedings, Part III Computer Vision – Second CCF Chinese Conference, CCCV 2017*, Tianjin, China, 11–14 October 2017, pp. 315–327.
10. Liu H, Feng J, Jie Z, et al. Neural person search machines. In: *IEEE International Conference on Computer Vision, ICCV 2017*, Venice, Italy, 22–29 October 2017, pp. 493–501.
11. Ren S, He K, Girshick RB, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, Montreal, Quebec, Canada, 7–12 December 2015, pp. 91–99.
12. Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, 21–26 July 2017, pp. 3376–3385.
13. Li W, Zhao R, Xiao T, et al. DeepReID: deep filter pairing neural network for person re-identification. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, Columbus, OH, USA, 23–28 June 2014, pp. 152–159.
14. Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, 21–26 July 2017, pp. 3346–3355.
15. Li L, Yang H, and Chen L. Spatial invariant person search network. In: *Proceedings, Part II Pattern Recognition and Computer Vision – First Chinese Conference, PRCV 2018*, Guangzhou, China, 23–26 November 2018, pp. 122–133.
16. Dollár P, Tu Z, Perona P, et al. Integral channel features. In: *Proceedings British Machine Vision Conference, BMVC 2009*, London, UK, 7–10 September 2009, pp. 1–11.
17. Viola PA and Jones MJ. Robust real-time face detection. *Int J Comput Vision* 2004; 57(2): 137–154.
18. Dollár P, Appel R, Belongie SJ, et al. Fast feature pyramids for object detection. *IEEE Trans Pattern Anal Mach Intell* 2014; 36(8): 1532–1545.
19. Nam W, Dollár P, and Han JH. Local decorrelation for improved pedestrian detection. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, Montreal, Quebec, Canada, 8–13 December 2014, pp. 424–432.
20. Benenson R, Omran M, Hosang JH, et al. Ten years of pedestrian detection, what have we learned? In: *Proceedings, Part II Computer Vision – ECCV 2014 Workshops*, Zurich, Switzerland, 6–7, 12 September 2014, pp. 613–627.
21. Girshick RB, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, Columbus, OH, USA, 23–28 June 2014, pp. 580–587.
22. Hosang JH, Omran M, Benenson R, et al. Taking a deeper look at pedestrians. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, 7–12 June 2015, pp. 4073–4082.
23. Tian Y, Luo P, Wang X, et al. Pedestrian detection aided by deep learning semantic tasks. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, 7–12 June 2015, pp. 5079–5087.

24. Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 1904–1912.
25. Cai Z, Saberian MJ, and Vasconcelos N. Learning complexity-aware cascades for deep pedestrian detection. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 3361–3369.
26. Yang B, Yan J, Lei Z, et al. Convolutional channel features. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 82–90.
27. Zhang L, Lin L, Liang X, et al. Is faster R-CNN doing well for pedestrian detection? In: *Proceedings, Part II Computer Vision – ECCV 2016 – 14th European Conference*, Amsterdam, The Netherlands, 11–14 October 2016, pp. 443–457.
28. Girshick RB. Fast R-CNN. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 1440–1448.
29. Li J, Liang X, Shen S, et al. Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans Multimedia* 2018; 20(4): 985–996.
30. Mao J, Xiao T, Jiang Y, et al. What can help pedestrian detection? In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, 21–26 July 2017, pp. 6034–6043.
31. Redmon J, Divvala SK, Girshick RB, et al. You only look once: unified, real-time object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, 27–30 June 2016, pp. 779–788.
32. Gheissari N, Sebastian TB, and Hartley RI. Person reidentification using spatiotemporal appearance. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, New York, NY, USA, 17–22 June 2006, pp. 1528–1535.
33. Zajdel W, Zivkovic Z, and Kröse BJA. Keeping track of humans: Have I seen this person before? In: *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, *ICRA 2005*, Barcelona, Spain, 18–22 April 2005, pp. 2081–2086.
34. Chu X, Ouyang W, Li H, et al. Structured feature learning for pose estimation. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, 27–30 June 2016, pp. 4715–4723.
35. Xiao T, Li H, Ouyang W, et al. Learning deep feature representations with domain guided dropout for person re-identification. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, 27–30 June 2016, pp. 1249–1258.
36. Liao S and Li SZ. Efficient PSD constrained asymmetric metric learning for person re-identification. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 3685–3693.
37. Paisitkriangkrai S, Shen C, and van den Hengel A. Learning to rank in person re-identification with metric ensembles. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, 7–12 June 2015, pp. 1846–1855.
38. Zheng L, Shen L, Tian L, et al. Scalable person re-identification: a benchmark. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 1116–1124.
39. Zheng W, Gong S, and Xiang T. Person re-identification by probabilistic relative distance comparison. In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, Colorado Springs, CO, USA, 20–25 June 2011, pp. 649–656.
40. Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features. In: *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010*, San Francisco, CA, USA, 13–18 June 2010, pp. 2360–2367.
41. Gray D and Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: *Proceedings, Part I Computer Vision – ECCV 2008, 10th European Conference on Computer Vision*, Marseille, France, 12–18 October 2008, pp. 262–275.
42. Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, 7–12 June 2015, pp. 2197–2206.
43. Hamdoun O, Moutarde F, Stanciulescu B, et al. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*, Stanford, CA, USA, 7–11 September 2008, pp. 1–6.
44. Wang X, Doretto G, Sebastian T, et al. Shape and appearance context modeling. In: *IEEE 11th International Conference on Computer Vision, ICCV 2007*, Rio de Janeiro, Brazil, 14–20 October 2007, pp. 1–8.
45. Zhao R, Ouyang W, and Wang X. Unsupervised salience learning for person re-identification. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 23–28 June 2013, pp. 3586–3593.
46. Ahmed E, Jones MJ, and Marks TK. An improved deep learning architecture for person re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, 7–12 June 2015, pp. 3908–3916.
47. Liu H, Feng J, Qi M, et al. End-to-end comparative attention networks for person re-identification. *IEEE Trans Image Proc* 2017; 26(7): 3492–3506.
48. Liu H, Jie Z, Karlekar J, et al. Video-based person re-identification with accumulative motion context. *IEEE Trans Circuit Syst Video Techn* 2018; 28(10): 2788–2802.
49. Wu L, Shen C, and van den Hengel A. PersonNet: person re-identification with deep convolutional neural networks. *CoRR* 2016; abs/1601.07255.
50. Köstinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints. In: *2012 IEEE*

Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012, pp. 2288–2295.
51. Li X, Zheng W, Wang X, et al. Multi-scale learning for low-resolution person re-identification. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 3765–3773.
52. Liu H, Qi M, and Jiang J. Kernelized relaxed margin components analysis for person re-identification. *IEEE Signal Proc Lett* 2015; 22(7): 910–914.
53. Tao D, Guo Y, Song M, et al. Person re-identification by dual-regularized KISS metric learning. *IEEE Trans Image Proc* 2016; 25(6): 2726–2738.
54. Zhang L, Xiang T, and Gong S. Learning a discriminative null space for person re-identification. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, 27–30 June 2016, pp. 1239–1248.
55. Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, 27–30 June 2016, pp. 1335–1344.
56. Prosser BJ, Zheng W, Gong S, et al. Person re-identification by support vector ranking. In: *Proceedings British Machine Vision Conference, BMVC 2010*, Aberystwyth, UK, 31 August–3 September 2010, pp. 1–11.
57. Zheng W, Li X, Xiang T, et al. Partial person re-identification. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, 7–13 December 2015, pp. 4678–4686.
58. Ding S, Lin L, Wang G, et al. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recogn* 2015; 48(10): 2993–3003.
59. Chen W, Chen X, Zhang J, et al. A multi-task deep network for person re-identification. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 4–9 February 2017, pp. 3988–3994.
60. Wang W, Shen J, Dong X, et al. Salient object detection driven by fixation prediction. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, 18–22 June 2018, pp. 1711–1720.
61. Wang W and Shen J. Deep cropping via attention box prediction and aesthetics assessment. In: *IEEE International Conference on Computer Vision, ICCV 2017*, Venice, Italy, 22–29 October 2017, pp. 2205–2213.
62. Dollár P, Wojek C, Schiele B, et al. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans Pattern Anal Mach Intell* 2012; 34(4): 743–761.
63. Ess A, Mueller T, Grabner H, et al. Segmentation-based urban traffic scene understanding. In: *Proceedings British Machine Vision Conference, BMVC 2009*, London, UK, 7–10 September 2009, pp. 1–11.
64. Xu Y, Ma B, Huang R, et al. Person search in a scene by jointly modeling people commonness and person uniqueness. In: *Proceedings of the ACM International Conference on Multimedia, MM '14*, Orlando, FL, USA, 03–07 November 2014, pp. 937–940.
65. Gray D, Brennan S, and Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. In: *Proceeding IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, Rio de Janeiro, Brazil, 14 October 2007, vol. 3, pp. 1–7, Citeseer.
66. Hirzer M, Beleznai C, Roth PM, et al. Person re-identification by descriptive and discriminative classification. In: *Proceedings Image Analysis – 17th Scandinavian Conference, SCIA 2011*, Ystad, Sweden, May 2011, pp. 91–102.
67. Li W and Wang X. Locally aligned feature transforms across views. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 23–28 June 2013, pp. 3594–3601.
68. Zheng W, Gong S, and Xiang T. Associating groups of people. In: *Proceedings British Machine Vision Conference, BMVC 2009*, London, UK, 7–10 September 2009, pp. 1–11.
69. Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, 21–26 July 2017, pp. 2261–2269.
70. Dai J, Li Y, He K, et al. R-FCN: object detection via region-based fully convolutional networks. In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, Barcelona, Spain, 5–10 December 2016, pp. 379–387.
71. Schroff F, Kalenichenko D, and Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, 7–12 June 2015, pp. 815–823.
72. Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, 21–26 July 2017, pp. 1320–1329.