Neural Person Search Machines

Hao Liu [1] Jiashi Feng[2] Zequn Jie[3] Karlekar Jayashree[4]
Bo Zhao [5] Meibin Qi[1] Jianguo Jiang[1] Shuicheng Yan[6,2]
[1]Hefei University of Technology [2]National University of Singapore [3]Tencent AI Lab
[4] Panasonic R&D Center Singapore [5]Southwest Jiaotong University [6]360 AI Institute
{hfut.haoliu, zequn.nus}@gmail.com, elefjia@nus.edu.sg, karlekar.jayashree@sg.panasonic.com
zhaobo@my.swjtu.edu.cn, {qimeibin, jgjiang}@hfut.edu.cn, yanshuicheng@360.cn

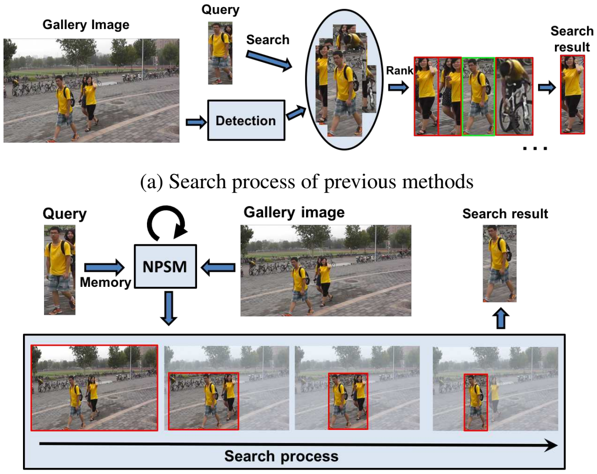
Abstract

We investigate the problem of person search in the wild in this work. Instead of comparing the query against all candidate regions generated in a query-blind manner, we propose to recursively shrink the search area from the whole image till achieving precise localization of the target person, by fully exploiting information from the query and contextual cues in every recursive search step. We develop the Neural Person Search Machines (NPSM) to implement such recursive localization for person search. Benefiting from its neural search mechanism, NPSM is able to selectively shrink its focus from a loose region to a tighter one containing the target automatically. In this process, NPSM employs an internal primitive memory component to memorize the query representation which modulates the attention and augments its robustness to other distracting regions. Evaluations on two benchmark datasets, CUHK-SYSU Person Search dataset and PRW dataset, have demonstrated that our method can outperform current state-of-the-arts in both mAP and top-1 evaluation protocols.

1. Introduction

Person search [33, 41] aims to localize a specific person matching the provided query in gallery images or video frames. It is a new and challenging task that requires to address person detection and re-identification simultaneously. It has many important applications in video surveillance and security, such as cross-camera visual tracking [22] and person verification [37]. But it is difficult in real-world scenarios due to various distracting factors including large appearance variance across multiple cameras, low resolution, cluttered background, unfavorable camera setting, *etc*.

To date, only a few methods have been proposed to address this task. In the pioneer work [33], Xiao *et al.* adopted the end-to-end person search model based on the proposed Online Instance Matching (OIM) loss function to jointly train person detection and re-identification networks. The recent work [41] also follows a similar pipeline. Generally, all of the previous person search methods are based on such a simple two-stage search strategy: first to detect all candidate persons within an image and then to perform exhaustive comparison between all possible pairs of the query and the candidates to output a search result ranked at the first place within the searched images. This pipeline has some drawbacks. Firstly, if the target person has distracting factors around, *e.g.*, another person with similar appearance, the search accuracy would be affected by the distracting



(a) Search process of previous methods

(b) Search process of the NPSM

Figure 1. Demonstration of person search process for one gallery image in previous methods and our proposed method. (a) The search process of previous methods. The query person is one-by-one compared with the detection results for one gallery image; then the search result ranked at the first place is obtained. The red boxes indicate the wrong matched results while the green box represents the truly matched person. (b) The search process of the NPSM. When a target person is searched within a whole scene, the search scope on which attention is focused is recursively shrinking with guidance from memory of the query person's appearance, which is marked in red boxes.

factors. Secondly, extra error, such as inaccurate detection, would be introduced by the two isolated frameworks, *i.e.* person detection and re-identification. See Figure 1 (a) for demonstration. The red boxes indicate the wrong matched results while the green box represents the truly matched person.

For person search, it is commonly assumed that within an image, the target person only appears at a single location. Such instance-level exclusive cues imply that instead of examining all possible persons, a more effective strategy is to only search within the regions possibly *containing the target person* in a coarse-to-fine manner. This is similar to human neural system for processing complex visual information [2,23]. More concretely, after seeing and remembering the appearance of a target person, one usually shrinks his search area from a large scope to a small one and performs matching with his memory in details within the small scope with more effort. Such a coarse-to-fine search process is intuitively useful for existing person search solutions.

Inspired by above observations, we propose a new and more effective person search strategy and develop the Neural Person Search Machines (NPSM). Compared to the search process in previous methods, our NPSM (Figure 1 (b)) takes the query person as memory to recursively guide the model to shrink the search region and judge whether the current region contains the target person or not. This process would include more contextual cues beneficial for person matching. In Figure 1 (b), the red box in each image from left to right corresponds to a region that can be focused on, and the arrow indicates a search process which can be considered as the continuous shrinkage of the focus region. Additionally, those irrelevant regions can be ignored after every shrinkage of a subregion from a big region, which can reduce the interference of other unimportant regions.

To model the above person search process, we need to solve the following two non-trivial problems: 1) integrating information of the query person into the search process as memory to exclude interference from impossible candidates; 2) judging which subregion should be focused on in the bigger region at each recursive step in the coarse-to-fine search process under the guidance of memory.

For localizing the target person in a sequence correctly and fully exploiting the context information, we propose a neural search architecture to selectively concentrate on an effective subregion of the input region, and meanwhile ignore other perceived information from distracting subregions in a recursive way. Take the third subregion of the search process in Figure 1 (b) for example, the proposed NPSM would highlight the truly matched person at the left side of the region and ignore the similar person at the right side. Considering the specific ability of Long Short-Term Memory (LSTM) [10] to partially allow or deny information to flow into or out of its memory component, we build our Neural Search Networks (NSN) upon Convolutional LSTM (Conv-LSTM) [34] units which are capable of preserving spatial information from the spatio-temporal sequences.

Different from the vanilla Conv-LSTM, we augment our NSN by equipping it with external primitive memory that contains appearance information of the query and helps identify the candidate regions at the coarse level and discards irrelevant regions. The external primitive memory thus enables the query to be involved in the representation learning for person search as well as the recursive search process with region shrinkage.

To sum up, in this work we go beyond the standard LSTM based models and propose a new person search approach called Neural Person Search Machines (NPSM) based on the Conv-LSTM [34], which contains the context information of each person and employs the external memory about the query person to guide the model to attend to the right region. Our approach is able to achieve better performance compared with other methods, as validated by experimental results.

We make the following contributions to person search:
1) We redefine the person search process as a detection free procedure of recursively focusing on the right regions.
2) We coin a novel method more robust to distracting factors benefiting from contextual information.
3) We propose a new neural search model that can integrate the query person information into primitive memory to guide the model to recursively focus on the effective regions.

2. Related Work

Person search can be regarded as the combination of person re-identification and person detection. Most of existing works of person re-identification focus on designing hand-crafted discriminative features [5, 8, 17], learning deep learning based high-level features [1,14,18,19,31,32] and learning distance metrics [12, 15, 20, 27, 38]. Recent deep learning based person re-identification methods [1, 14, 18, 19] re-design the structure of the deep network to improve performance. For instance, [1] designed two novel layers to capture relationships between two views of a person pair. Among distance metric learning methods, [12] proposed KISSME (KISS MEtric) to learn a distance metric from equivalence constraints. Additionally, [38] proposed to solve the person re-id problem by learning a discriminative null space of the training samples while [15] proposed a method learning a shared subspace across different scales to address the low resolution person re-identification problem.

For person detection, Deformable Part Model (DPM) [6], Aggregated Channel Features (ACF) [4] and Locally Decorrelated Channel Features (LDCF) [21] are three representative methods relying on hand-crafted features and linear classifiers to detect pedestrians. Recently, several deep learning-based frameworks have been proposed. In [29], DeepParts was proposed to handle occlusion with an extensive part pool. Besides, [3] proposed the CompACT boosting algorithm learning complexity-aware detector cascades for person detection. In our knowledge, two previous works [33, 41] address person search by fusing person re-identification and detection into an integral pipeline to consider whether any complementarity exists between the two tasks. [33] developed an end-to-end person search framework to jointly handle both aspects with the help of Online Instance Matching (OIM) loss while [41] proposed ID-discriminative Embedding (IDE) and Confidence Weighted Similarity (CWS) to improve the person search performance. However, these two works simply focus on how the interplay of pedestrian detection and person re-identification affects the overall performance, and they still isolate the person search into two individual components (detection and re-identification), which would introduce extra error, *e.g.* inaccurate detection. In this paper, we regard person search as a detection-free process of gradually removing interference or irrelevant target persons for the query person.

Recently, LSTM based attention methods have shown good performance in image description [35], action recognition [16, 25] and person re-identification [18]. In [35], Xu *et al.* showed how the learned attention can be exploited to give more interpretability into the model generation process, while [16,25] adopted attention methods to recognize important elements in video frames based on the action that is being performed. Moreover, [18] formulated an attention model as a triplet recurrent neural network which dynamically generates comparative attention location maps for person re-identification. Analogously, our proposed NPSM also has such a locally emphasizing property, but NPSM is a query-aware model while the above attention-based methods all adopt a blind attention mechanism.

3. Proposed Neural Person Search Machines

In this section, we present the architecture details of the proposed Neural Person Search Machines (NPSM), and explain how it works with the primitive memory modeling to facilitate person search.

3.1. Architecture Overview

The overall architecture is shown in Figure 2. It consists of two components, *i.e.* Primitive Memory and Neural Search Networks. We propose to solve person search by recursively shrinking the search area from the whole image to the precise bounding box of the person of interest. And each region in the shrinking search process would contain the contextual information of the final search result. Besides recursively utilizing the contextual cues, NPSM provides extra robustness to interference from other distracting subregions for the model in the search process.

The proposed NPSM is trained end-to-end to learn to make a decision on the subregion attention at each recursive step and finally localize the person of interest. The Neural Search Network enables the model to automatically focus on relevant regions, and the Primitive Memory that models the representation of the query person continuously provides extra cues for every search step and facilitates more precise localization of persons. After performing the recursive region shrinkage, the model reaches a search result with the biggest confidence as the final search result with an gallery image. Note that, different from previous works [33, 41], our method is detection-free and includes no Non-Maximum Suppression (NMS), as it is a search process performing simultaneous region shrinking and person identification. When the searching is finished, there will be only one bounding box person search result left.

3.2. Person Search with NPSM

As aforementioned, we redefine the person search process as the recursive region shrinking process. It is equivalent to recursively focusing on a subregion containing the person of interest from a bigger region. Here we describe the details of our proposed NPSM and explain how to perform the recursive region shrinking to search the target person for each gallery image.

3.2.1 Neural Search Networks

Learning to search for a person from a big region to a specific person region within the gallery image can be deemed as a sequence modeling problem. Specifically, the shrinking regions constitute a sequence. Thus a natural choice for the model candidates is the Recurrent Neural Network (RNN) or LSTM based RNN. However, the vanilla LSTM [10] only models sequence information through fully connected layers and requires vectorizing 2D feature maps. This would result in the loss of spatial information, harming person localization performance. In order to preserve the spatial structure of the regions over the shrinking process shown in Figure 2, we design a new network called Neural Search Network (NSN) based on Convolutional LSTM (Conv-LSTM) [34] for each recursive step. Conv-LSTM replaces the fully connected multiplicative operations in an LSTM unit with convolutional operations. Different from it, the NSN has an additional memory component recording the query.

Conv-LSTM can be used for building attention networks that can learn to pay attention to critical regions within feature maps. Thus, Conv-LSTM based NSN is also equipped with attention mechanism to learn to gradually shrink the region and selectively memorize the contextual information contained in the searched bigger region at each recursive

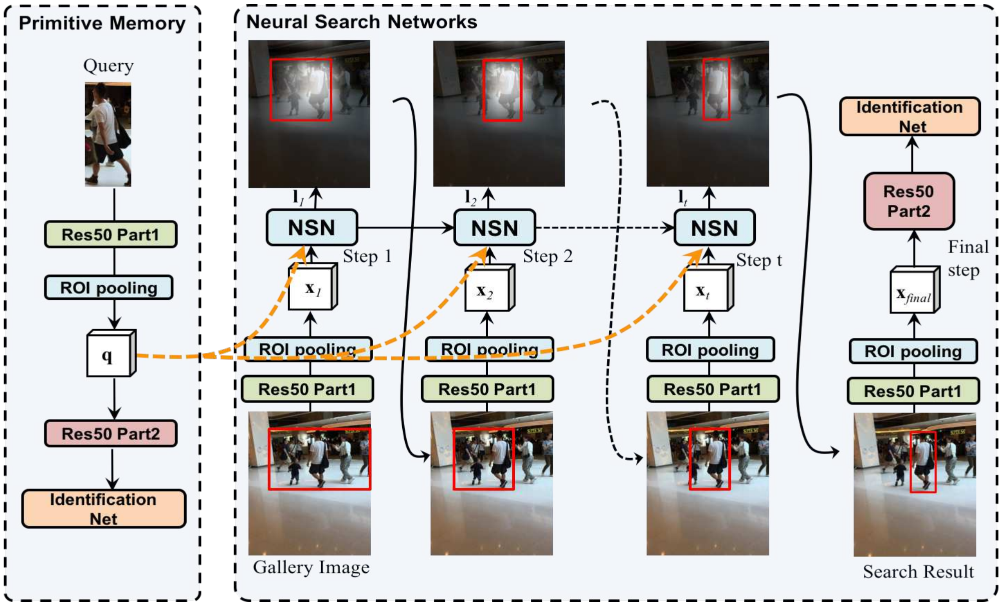


Figure 2. Architecture of our proposed Neural Person Search Machines (NPSM). It consists of two components, *i.e.* Primitive Memory and Neural Search Networks. It works by recursively shrinking the search area from the whole image to the precise bounding box of the person of interest under the guidance (orange dotted lines) of Primitive Memory. And each region in the shrinking search process would contain the contextual information of the final search result. Red boxes denote the shrinking regions highlighted at different recursive steps in our NPSM. "Res50 Part1" corresponds to the *conv1* to *conv4_3* of ResNet-50 while "Res50 Part2" represents the *conv4_4* to *conv5_3* of ResNet-50. Best viewed in color.

step. However, our neural search model has a unique feature that distinguishes it from a plain attention model: in addition to gallery images, a query illustrating the search target is also input to the search network. Traditional attention networks cannot well model such extra cues. In this work, we propose to model such query information into the primitive memory in order to facilitate person search.

We now elaborate on the new Neural Search Networks (NSN) of our NPSM, tailored for the person search task. In the NSN component, the query person information, denoted as $\mathbf{q}$, is integrated into the computation within gates and cell states in a way to bias the updating of internal states towards emphasizing information relevant to the query while forgetting irrelevant information. Here the query feauture $\mathbf{q}$ is extracted from the query image through the pre-trained "Res50 part1" (*conv1* to *conv4_3* of ResNet-50 [9]) which is the same as the one extracting features from gallery images. The cell and gates in the NSN are defined as

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{w}_{xi} * \mathbf{x}_t + \mathbf{w}_{hi} * \mathbf{h}_{t-1} + \mathbf{w}_{qi} * \mathbf{q} + b_i\right)\\
\mathbf{f}_t &= \sigma\left(\mathbf{w}_{xf} * \mathbf{x}_t + \mathbf{w}_{hf} * \mathbf{h}_{t-1} + \mathbf{w}_{qf} * \mathbf{q} + b_f\right)\\
\mathbf{o}_t &= \sigma\left(\mathbf{w}_{xo} * \mathbf{x}_t + \mathbf{w}_{ho} * \mathbf{h}_{t-1} + \mathbf{w}_{qo} * \mathbf{q} + b_o\right) \qquad (1)\\
\mathbf{g}_t &= \tanh\left(\mathbf{w}_{xc} * \mathbf{x}_t + \mathbf{w}_{hc} * \mathbf{h}_{t-1} + \mathbf{w}_{qc} * \mathbf{q} + b_c\right)\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t\\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right),
\end{aligned}$$

where $*$ represents the convolutional operation and $\odot$ is the Hadamard product, $\mathbf{w}_{x\sim}$, $\mathbf{w}_{h\sim}$ are two-dimensional convolutional kernels and $\mathbf{x}_t$ which is the feature map of the region highlighted by the previous time-step denotes the input at time step t. The input gate, forget gate, output gate, hidden state and memory cell are denoted as $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$, $\mathbf{h}_t$, $\mathbf{c}_t$ respectively, which are all three-dimensional tensors retaining spatial dimensions. With their control, the contextual information can be selectively memorized. Note the query person information $\mathbf{q}$ is independent of the time step t, therefore serving as the global primitive memory that guides the person search procedure continuously. The effect of such memory information over the states is modeled through the parameter $\mathbf{w}_{q\sim}$.

3.2.2 Region Shrinkage with Primitive Memory

As stated above, the goal of NPSM is to effectively shrink regions containing the target person based on the neural search mechanism, guided by the primitive memory. That is, the NPSM will decide which local region should be focused on at each recursive step in the search process as shown in Figure 2. Through this way, more context information would be included from a large region and the number of irrelevant person candidates with the target person would be recursively reduced in the search process. In this subsection, we introduce how the subregion of each recursive time-step is generated and shrunk from the bigger region of the previous time-step.

Here we define the region covered by the highlighted proposal bounding boxes induced by current attention map as follows:

$$\mathbf{R} = (\min(\boldsymbol{\theta}_{x1}), \min(\boldsymbol{\theta}_{y1}), \max(\boldsymbol{\theta}_{x2}), \max(\boldsymbol{\theta}_{y2})),$$

where $\boldsymbol{\theta}_{x1}, \boldsymbol{\theta}_{y1}, \boldsymbol{\theta}_{x2}, \boldsymbol{\theta}_{y2}$ are the top left and lower right corner coordinates of all the highlighted bounding boxes from a predefined collection, generated by an unsupervised object proposal model (*e.g.*, Edgeboxes [42]). Then we separate the region $\mathbf{R}$ into several candidate subregions according to the relationship of each contained bounding box in the region $\mathbf{R}$. In this paper, we choose the Euclidean distance as the metric of the relationship defined as

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum\nolimits_{i=1}^{2} (a_i - b_i)^2}, \tag{2}$$

where $\mathbf{a}$ and $\mathbf{b}$ are the centre coordinates of two proposal bounding boxes *A* and *B* respectively. Specifically, $\mathbf{a} = (a_1, a_2)$, $\mathbf{b} = (b_1, b_2)$, $a_1 = a_{x1} + 0.5\,(a_{x2} - a_{x1})$, $a_2 = a_{y1} + 0.5\,(a_{y2} - a_{y1})$, $b_1 = b_{x1} + 0.5(b_{x2} - b_{x1})$, $b_2 = b_{y1} + 0.5(b_{y2} - b_{y1})$. (a_{x1}, a_{y1}) and (a_{x2}, a_{y2}) are the top left and lower right coordinates of bounding box *A* while (b_{x1}, b_{y1}) and (b_{x2}, b_{y2}) are the top left and lower right coordinates of bounding box *B*. Then the proposal bounding boxes can be grouped into C clusters according to their relationships. The corresponding subregions covered by proposals are $\mathbf{R}^{sub}_{(C)}$. We denote the parent region to generate subregions $\mathbf{R}^{sub}_{(C)}$ as $\mathbf{R}^{par}$.

At each recursive step t, the proposed NSN outputs an attention map which predicts the scores (reflecting confidence on containing the target person given the primitive memory information) of shrinking to region $\mathbf{R}^{sub}_{t,(C)}$ after NSN taking input the parent region $\mathbf{R}^{par}_{t-1}$ at the previous step $t-1$.

More specifically, at each time step (corresponding to shrinking to one region), NSN takes input the query person feature $\mathbf{q}$ and the region feature $\mathbf{x}_t$ extracted from pretrained "Res50 part1" which denotes the *conv1* to *conv4_3* of ResNet-50 [9]. Here, we add a Region of Interest (ROI) pooling layer following "Res50 part1" to make sure the regions of different sizes can have feature maps of the same size $K \times K \times D$. Compared with the standard LSTM based model relying on multi-layer perceptron, NSN uses convolutional layers to integrate the region representation with primitive memory and produce attention maps. Specifically, at each time step t, an attention score map of size $K \times K$ for $K \times K$ locations is computed:

$$\mathbf{z}_t = \mathbf{w}_z * \tanh\left(\mathbf{w}_{qa} * \mathbf{q} + \mathbf{w}_{ha} * \mathbf{h}_t + b_a\right) \tag{3}$$

$$\mathbf{l}^{i,j}_t = \frac{\exp(\mathbf{z}^{ij}_t)}{\sum_i \sum_j \exp(\mathbf{z}^{ij}_t)}. \tag{4}$$

The score for location (i, j) is denoted as $\mathbf{l}^{i,j}_t$.

Then, in the process of region shrinkage, the NSN computes the average scores of different subregions from the parent region. NSN highlights the subregion with the maximum score as the region to be searched in the next step. This computation would be performed many times until the search path reaches the final proposal. The average score of the subregion is computed as follows:

$$\mathbf{S}_t = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{l}^{i,j}_t, \tag{5}$$

where m and n are the height and the width of the subregion respectively. $\mathbf{l}^{i,j}_t$ corresponds to the score map defined in Eqn. (4) generated on the parent region. In other words, our model does not stick to the single region. If some regions not highlighted before receive higher attention at certain search step, our model would jump to that region with higher intra-region confidence scores. In this way, the accumulative error in the shrinkage process can be alleviated. Note that our NSN serves as a region shrinkage method. In other words, our NSN only outputs the most similar proposal with the query person in each gallery image. Therefore, the features of the query person image and the final search result are extracted from the "Identification Net" (orange boxes in Figure 2) of the trained model when the searching is finished. Here, the "Identification Net" takes input the output of "Res50 Part2" (pink boxes in Figure 2) representing the *conv4_4* to *conv5_3* of ResNet-50. And it consists of a global average pooling layer and a 256-dimension Fully Connected layer. Then the *cosine* similarity between the features of the query person and the final person search result is computed for evaluation.

3.3. Training Strategy

Here we detail the training of the proposed model. Firstly, we use the architecture proposed in OIM [33] to *pre-train* the Fully Convolutional Networks (FCN) including both "Res50 part1" and "Res50 part2'. Then for the region at each recursive time-step, the feature is extracted from the ROI pooling layer after the pre-trained "Res50 part1". After that, all the features are fed to the NSN and we add a convolutional layer of size $1 \times 1 \times 2$ after output of each time step to calculate the "region shrinkage loss". Here we adopt segmentation alike softmax loss as the "region shrinkage loss". The supervision label of each time step is defined as

$$\mathbf{U}_t = \begin{cases} \mathbf{1}, \text{if } G \in R_t \\ \mathbf{0}, \text{otherwise}, \end{cases} \tag{6}$$

where G is the ground truth bounding box of the target person while R_t is the region box reached at the tth time step. This training strategy enables the proposed network to produce proper attention maps that fall into the region containing the target person as tight as possible. In other words, our NPSM is expected to predict the probability of the target person appearing at each location in a gallery image.

Besides, to make the learned feature more discriminative, we add an identification loss following the "Identification Net", which takes input the output feature $\mathbf{u}$ of "Identification Net" and is defined as

$$P(z=c|\mathbf{u}) = \frac{\exp(S_c\mathbf{u})}{\sum_k \exp(S_k\mathbf{u})}, \tag{7}$$

$$L_{iden} = -\log(P(z=c|\mathbf{u})). \tag{8}$$

where there are a total of *I* identities, *z* is the identity of the person, and *S* is the softmax weight matrix while S_c and S_k represent the *c*th and *k*th column of it, respectively.

4. Experiments

4.1. Datasets and Evaluation Protocol

4.1.1 Datasets

CUHK-SYSU: CUHK-SYSU [33] is a large-scale person search dataset with diverse scenes, containing 18,184 images, 8,432 different persons, and 96,143 annotated pedestrians bounding boxes. Each selected query person appears in at least two images captured from different viewpoints. The images present large variations in viewpoint, lighting, resolution, occlusion and background, intensively reflecting the real application scenarios and scene diversity. We use the official training/test split provided by the dataset. The training set contains 11,206 images and 5,532 query persons. Within the testing set, the query set contains 2,900 persons and the gallery contains 6,978 images in total.

PRW: The PRW dataset [41] is extracted from one 10-hour video captured on a university campus. The dataset includes 11,816 video frames of scenes captured by 6 cameras. In total 11,816 frames are manually annotated, giving 43,110 pedestrian bounding boxes. Among them, 34,304 pedestrians are annotated with 932 IDs. It provides 5,134 frames of 482 different persons for training. The provided testing set contains 2,057 query persons and a gallery of 6,112 images.

4.1.2 Evaluation Protocol

We adopt the mean Averaged Precision (mAP) and the top-1 matching rate as performance metrics, which are also used in OIM [33] and [41]. Using the mAP metric, person search performance is evaluated in a similar way as detection, reflecting the accuracy of detecting the query person from the gallery images. The top-1 matching rate treats person search as a ranking and localization problem. A matching is counted if a bounding box among the top-1 predicted boxes overlaps with the ground truth larger than the threshold 0.5.

4.2. Implementation Details

In this paper, the Fully Convolutional Networks (FCN) including both "Res50 part1" and "Res50 part2" are *pre-trained* by using the architecture proposed in OIM [33]. For the input region at each time-step, we apply an ROI pooling layer on the $conv4\_3$ convolutional features of it to normalize all the features to the same size of 14 × 14 × 1024. For query person images, we also extract their 14 × 14 × 1024 convolutional features in the same way. These features are then fed into the NPSM architecture. In particular, within NSN, the convolutional kernels for input-to-input states and state-to-state transitions are fixed as 3 × 3 with 1024 channels. At each recursive search step, we set the number C of subregions covered by clustered proposals to 3. We implement our network using Theano [28] and Caffe [11] deep learning framework. The training of the NPSM converges in roughly 50 hours for CUHK-SYSU dataset and 40 hours for PRW dataset on on a machine with 64GB memory, NVIDIA GeForce GTX TITAN X GPU and Intel i7-4790K Processor. The initial learning rate is 0.001 and decays at the rate of 0.9 for the weight updates of RMSProp [30]. Additionally, we manually augment the data by performing random 2D translation. The speed of our method is close to real-time. For one gallery image, our model takes round 1s to output a final searched result. However, overhead of ranking over gallery is dominating. For the CUHK-SYSU with gallery size of 100, calculating cosine similarity between search result from all the gallery images and query for ranking takes round 20s. For the PRW with 6,112 gallery images, ranking over gallery takes round 15 minutes.

4.3. Ablation Study

In this subsection, we perform several analytic experiments on CUHK-SYSU benchmark to investigate the contribution of each component in our proposed NPSM architecture. We analyze attention prediction, contextual cue and primitive memory of query person. In total we have three variants by training the model based on different combinations of the above factors. And the gallery size is set to 100. The details and corresponding results are shown in Table 1.

As aforementioned, we employ the framework in OIM [33] which involves none of three factors, as the baseline. Based on this framework, the results of OIM [33] are obtained. In the method named "NPSM w/o C", we remove the "contextual cue and primitive memory integration" part (corresponding to Eqn. (1)) of the NSN in our proposed NPSM. Instead, at each recursive step, we replace the "contextual que and primitive memory integration" part with a 3 × 3 × 1024 convolutional layer followed by the concatenation of the FCN ("Res50 part1") feature map of the query person (primitive memory) and the current step region ($\mathbf{q}$ and $\mathbf{x}_t$). Moreover, for each recursive step, we only keep the shrinking region generation method and the attention score prediction model (Eqn. (4) and (5)) to predict the attention score map. This setting makes our NPSM a simple version without contextual cues involved but still with the attention prediction ability. Furthermore, in the method named "NPSM w/o A&C", we further remove the attention prediction model and only generate the shrinking region as the input of each recursive step and add a 1024-dimension fully connected (FC) layer and a 2-dimension

Table 1. Results of ablation study on CUHK-SYSU dataset with 100 gallery size setting. Legend: **A**: Attention prediction model, **C**: Contextual cue, **P**: Primitive memory of query person. “w/o A&C’ and “w/o C” are short for “without Attention prediction model and Contextual cue” and “without Contextual cue but with Attention prediction model” respectively.

	A	**C**	**P**	**mAP(%)**	**top-1(%)**
OIM (Baseline)	×	×	×	75.5	78.7
NPSM w/o A&C	×	×	✓	56.5	62.5
NPSM w/o C	✓	×	✓	72.5	76.3
NPSM	✓	✓	✓	**77.9**	**81.2**

FC layer after the output (concatenation of the FCN feature map of query person (primitive memory) and the current region) of each recursive step. And the 2-dimension FC layer aims at predicting the score of each highlighted subregion from the parent region. From comparison between the results of “OIM” and “NPSM w/o A&C”, we can see that simply using primitive memory of query without contextual cues involved to search for a target person in the recursive way can not achieve satisfactory results. From the result of “NPSM w/o C” , we find that the sightly higher performance is achieved than the baseline due to usage of the attention model which can introduce more spatial location information than the “NPSM w/o A&C”. However, both “NPSM w/o A&C” and “NPSM w/o C” lack a contextual cue memory mechanism. In other words, the above methods are unable to memorize the context information provided in a larger region through previous recursive steps. From the result of “NPSM” overtaking the baseline method “OIM” by 2.4% and 2.5% for mAP and top-1 evaluation protocol, we find that the neural search mechanism induced by our proposed NPSM is beneficial for person search performance, and memory of query person can also effectively guide the neural search model to find the right person.

4.4. Comparison with State-of-the-art Methods

We compare NPSM with several state-of-the-arts, including end-to-end person search ones proposed by Xiao *et al.* [33] and Zheng *et al.* [41] and some other methods combining commonly used pedestrian detectors (DPM [6], ACF [4], CCF [36], LDCF [21] and their respective R-CNN [7]) with hand-crafted features (BoW [40], LOMO [17], DenseSIFT-ColorHist (DSIFT) [39]) and distance metrics (KISSME [12], XQDA [17]).

4.4.1 Results on CUHK-SYSU

We report the person search performance on CUHK-SYSU with 100 gallery size setting in Table 2, where “CNN” represents the detector part (Faster-RCNN [24] with ResNet-50) and “IDNet” denotes the re-identification part in the framework of OIM [33]. Compared with CNN+IDNet, the OIM achieves performance improvement by introducing joint optimization of the detection and identification parts, but still follows the isolated “detection+re-identification” two-stage strategy in the person search process. Comparatively, our proposed NPSM is a detection-free method and solves localization and re-identification of the query person simultaneously by introducing the query-aware region shrinkage mechanism which can include more contextual information beneficial for search accuracy. It can be verified by results shown in Table 2. NPSM beats all compared methods consistently for both the mAP and top-1 metrics.

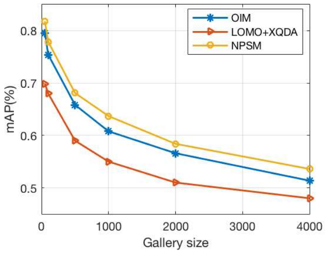


Figure 3. Test mAPs of different approaches under different gallery sizes.

Table 2. Comparison of NPSM’s performance on CUHK-SYSU with 100 gallery size setting with the state-of-the-arts.

Method	**mAP(%)**	**top-1(%)**
ACF [4]+DSIFT [39]+Euclidean	21.7	25.9
ACF+DSIFT+KISSME [12]	32.3	38.1
ACF+BoW [40]+Cosine	42.4	48.4
ACF+LOMO+XQDA [17]	55.5	63.1
ACF+IDNet [33]	56.5	63.0
CCF [36]+DSIFT+Euclidean	11.3	11.7
CCF+DSIFT+KISSME	13.4	13.9
CCF+BoW+Cosine	26.9	29.3
CCF+LOMO+XQDA	41.2	46.4
CCF+IDNet	50.9	57.1
CNN [24]+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 [33](Baseline)	75.5	78.7
NPSM	**77.9**	**81.2**

Moreover, Figure 3 shows the mAP of the compared methods with different gallery sizes, including [50, 100, 500, 1000, 2000, 4000]. One can see that the mAP drops gradually as the gallery size increases, but our method can still outperform all other methods under different gallery size settings. In particular, NPSM improves average performance per gallery size over OIM [33] by around 2%.

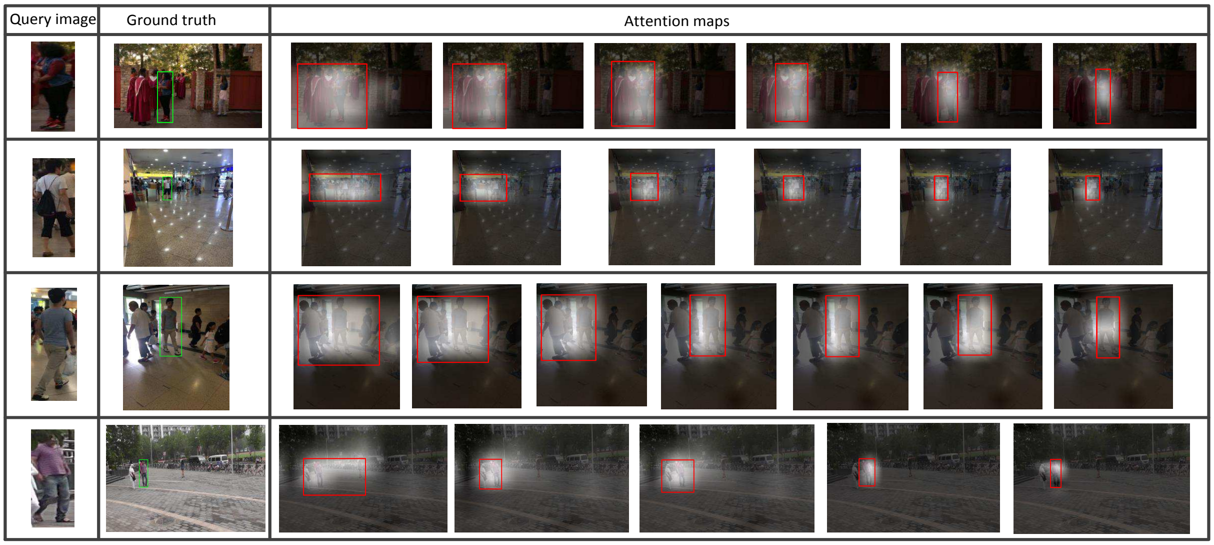


Figure 4. Attention maps learned by our NPSM model for different testing person samples in CUHK-SYSU and PRW dataset. The first three rows are from CUHK-SYSU, while the bottom row is from PRW. Green boxes represent the ground truth boxes while red boxes are the region bounding boxes highlighted by our NPSM model.

4.4.2 Results on PRW

On PRW dataset, we conduct experiments to compare NPSM with some state-of-the-art methods combining different detectors (respective R-CNN [7] detectors of DPM [6], CCF [36],ACF [4], LDCF [21]) and recognizers (LOMO, XQDA [17], IDE$_{det}$, CWS [41]). Among them, AlexNet [13] is exploited as the base network for the R-CNN detector. Although VGGNet [26] and ResNet [9] have more parameters and are deeper than AlexNet, according to [41], AlexNet can achieve better performance than the other two for DPM and ACF incorporating different recognizers. The results are shown in Table 3. Because the OIM method is the baseline of our NPSM, we implement the source code provided in OIM [33] to obtain the baseline result on the PRW dataset. Compared with the result, our NPSM outperforms it by 2.9% and 3.2% for mAP and top-1 accuracy separately. Besides, compared with all other state-of-the-arts considering five bounding boxes per gallery image, our method achieves better performance by only keeping one bounding box for testing per gallery image.

In Figure 4, we visualize some attention maps produced by our NPSM for testing samples from CUHK-SYSU and PRW datasets, which are all ranked top 1 in search results. The first three rows are from CUHK-SYSU, while the bottom row is from PRW. We observe that our NPSM can effectively shrink the search region to the correct person region guided by primitive memory of the query person.

5. Conclusions

In this work, we introduced a novel neural person search machine solving person search through recursively localizing effective regions, with guidance from the memory of the query person. Extensive experiments on two public benchmarks demonstrated its superiority over state-of-the-arts in most cases and the benefit to recognition accuracy in person search.

Table 3. Comparison of NPSM's performance on PRW with state-of-the-arts.

Method	**mAP(%)**	**top-1(%)**
DPM-Alex+LOMO+XQDA [17]	13.0	34.1
DPM-Alex+IDE$_{det}$ [41]	20.3	47.4
DPM-Alex+IDE$_{det}$+CWS [41]	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(Baseline)	21.3	49.9
NPSM	**24.2**	**53.1**

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61371155, Grant 61174170, and Grant 61632007. The work of Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112 and IDS R-263-000-C67-646.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE CVPR*, pages 3908–3916, 2015.

[2] J. R. Anderson. *Cognitive psychology and its implications*. WH Freeman/Times Books/Henry Holt & Co, 1990.

[3] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *IEEE CVPR*, pages 3361–3369, 2015.

[4] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE TPAMI*, 36(8):1532–1545, 2014.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE CVPR*, pages 2360–2367. IEEE, 2010.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE TPAMI*, 38(1):142–158, 2016.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE CVPR*, pages 2288–2295, 2012.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[14] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE CVPR*, pages 152–159, 2014.

[15] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *IEEE ICCV*, pages 3765–3773, 2015.

[16] Z. Li, E. Gavves, M. Jain, and C. G. Snoek. Videolstm convolves, attends and flows for action recognition. *arXiv preprint arXiv:1607.01794*, 2016.

[17] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE CVPR*, pages 2197–2206, 2015.

[18] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE TIP*, 26(7):3492–3506, 2017.

[19] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *arXiv preprint arXiv:1701.00193*, 2017.

[20] H. Liu, M. Qi, and J. Jiang. Kernelized relaxed margin components analysis for person re-identification. *IEEE Signal Processing Letters*, 22(7):910–914, 2015.

[21] W. Nam, P. Dollar, and J. H. Han. Local decorrelation for improved pedestrian detection. *NIPS*, 1:424–432, 2014.

[22] J. Niño-Castañeda, A. Frías-Velázquez, N. B. Bo, M. Slembrouck, J. Guan, G. Debard, B. Vanrumste, T. Tuytelaars, and W. Philips. Scalable semi-automatic annotation for multi-camera person tracking. *IEEE TIP*, 25(5):2259–2274, 2016.

[23] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience the Official Journal of the Society for Neuroscience*, 13(11):4700–19, 1993.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[25] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang. Person re-identification by dual-regularized kiss metric learning. *IEEE TIP*, 25(6):2726–2738, 2016.

[28] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[29] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *IEEE ICCV*, pages 1904–1912, 2015.

[30] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.

[31] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016.

[32] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016.

[33] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. *arXiv:1604.01850*, 2017.

[34] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, pages 802–810, 2015.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *IEEE ICCV*, pages 2048–2057, 2015.

[36] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *IEEE ICCV*, pages 82–90, 2015.

[37] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *AAAI*, 2016.

[38] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. *arXiv preprint arXiv:1603.02139*, 2016.

[39] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE CVPR*, pages 3586–3593, 2013.

[40] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE ICCV*, pages 1116–1124, 2015.

[41] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. *arXiv preprint arXiv:1604.02531*, 2016.

[42] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. Springer, 2014.