Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-free Approach

Lingxiao He*[1,2], Jian Liang*[1,2], Haiqing Li[1,2], and Zhenan Sun[1,2,3]
[1] CRIPAC & NLPR, CASIA [2] University of Chinese Academy of Sciences, Beijing, P.R. China
[3] Center for Excellence in Brain Science and Intelligence Technology, CAS
{lingxiao.he, jian.liang, hqli, znsun}@nlpr.ia.ac.cn

Abstract

Partial person re-identification (re-id) is a challenging problem, where only some partial observations (images) of persons are available for matching. However, few studies have offered a flexible solution of how to identify an arbitrary patch of a person image. In this paper, we propose a fast and accurate matching method to address this problem. The proposed method leverages Fully Convolutional Network (FCN) to generate certain-sized spatial feature maps such that pixel-level features are consistent. To match a pair of person images of different sizes, hence, a novel method called Deep Spatial feature Reconstruction (DSR) is further developed to avoid explicit alignment. Specifically, DSR exploits the reconstructing error from popular dictionary learning models to calculate the similarity between different spatial feature maps. In that way, we expect that the proposed FCN can decrease the similarity of coupled images from different persons and increase that of coupled images from the same person. Experimental results on two partial person datasets demonstrate the efficiency and effectiveness of the proposed method in comparison with several state-of-the-art partial person re-id approaches. Additionally, it achieves competitive results on a benchmark person dataset Market1501 with the Rank-1 accuracy being 83.58%.

1. Introduction

Person re-identification (re-id) has witnessed great progress in recent years, existing approaches always assume that each image covers a full glance of one person. However, the assumption of person re-id on full and frontal images is easily violated in real-world applications, and we merely have access to some partial observations of each person (dubbed partial person images) for retrieval. For instance, as shown in Fig. 1, partial person images often occur when a person is occluded by moving obstacles (e.g., cars, other persons), and static obstacles (trees, barriers). Hence partial person re-id has attracted significant research attention with increasing requirement of identification from CCTV cameras and video surveillance. However, few studies have focused on how to identify an arbitrary patch of a person image, making partial person re-id a challenging problem without solutions from current approaches. From this perspective, studying the partial person re-id problem is necessary and crucial for both academic research and practical retrieval applications.

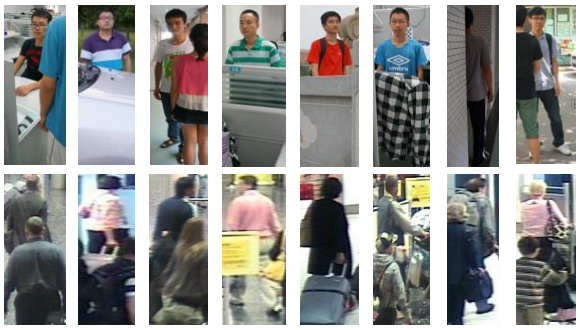

Figure 1. Examples of partial person images.

A majority of existing person re-id approaches fail to identify a person when severely partial person observations are provided. Concretely, to match an arbitrary patch of a person, some researchers resort to re-scale an arbitrary person patch to a fixed-size image. However, the performance would significantly degrade due to the undesired deformation (see Fig. 2(a)). Sliding Window Matching (SWM) [32] indeed introduces a possible solution for partial re-id by setting up a sliding window of the same size as the probe image and utilizing it to search for the most similar region within each gallery image (see Fig. 2(b)). However, SWM would not work well when the size of the probe person is bigger than the size of the gallery person. Some person re-id approaches further consider a part-based model which offers an alternative solution of partial person re-id in Fig. 2(c). Nevertheless, their computational costs are ex-

* Authors contributed equally.

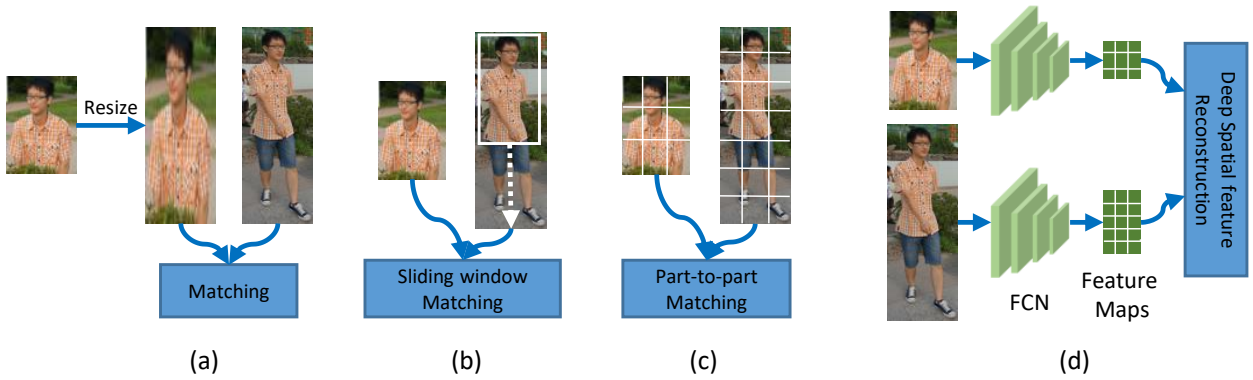

Figure 2. (a) The probe person image and gallery person image are resized to fixed-size (Resizing model). (b) Sliding window matching. (c) Part-based model. (d) The proposed Deep Spatial feature Reconstruction.

tensive and they require strict person alignment beforehand. Apart from these limitations, part-based models and SWM repeatedly extract sub-region features without sharing computation, which results in the unsatisfied computation efficiency.

In this paper, we propose a novel and fast partial person re-id framework that matches a pair of person images of different sizes (see Fig. 2(d)). In the proposed framework, Fully Convolutional Network (FCN) is utilized to generate correspondingly-size spatial feature maps, which can be considered a pixel-level feature matrix. Motivated by the remarkable successes achieved by dictionary learning in face recognition [12, 22, 27], we develop an end-to-end model named Deep Spatial feature Reconstruction (DSR), which expects that each pixel in the probe spatial maps can be sparsely reconstructed on the basis of an entire gallery spatial maps. In this manner, the model is independent on the image size, which naturally jumps the time-consuming alignment step. Specifically, we design an objective function for FCN which encourages that the reconstruction error of the spatial feature maps extracted from the same persons is smaller while spatial feature maps from different identities cannot well reconstruct each other (i.e., larger reconstruction error). Generally, the major contributions of our work are summarized as four-fold:

- We propose a freestyle approach named Deep Spatial feature Reconstruction (DSR) for partial person re-id, which is alignment-free and flexible to arbitrary-sized person images.
- We first integrate sparse reconstruction learning and deep learning in a unified framework, and train an end-to-end deep model through minimizing the reconstruction error for coupled person images from the same identity and maximizing that of different identities.
- Besides, we further replace the pixel-level reconstruction with a block-level one, and develop a multi-scale (different block sizes) fusion model to enhance the performance.
- Experimental results demonstrate that the proposed approach achieves impressive results in both accuracy and efficiency on Partial-REID [32] and Partial-iLIDs [31] databases.

The remainder of this paper is organized as follows. In Sec. 2, we review the related work about FCN, Sparse Representation Classification (SRC), and existing partial person re-id algorithms. Sec. 3 introduces the technical details of deep spatial feature reconstruction. Sec. 4 shows the experimental results and analyzes the performance in computational efficiency and accuracy. Finally, we conclude our work in Sec. 5.

2. Related Work

Since the proposed model is a deep feature learning method for partial person re-id based on Fully Convolutional Network and Sparse Representation Classification, we briefly review some related works in this section.

Fully Convolutional Network. FCN only contains convolutional layers and pooling layers, which have been applied into spatially dense tasks including semantic segmentation [1, 2, 6, 16, 19] and object detection [5, 14, 17, 18]. Shelhamer *et al.* [14] introduced a FCN that is trained end-to-end, pixels-to-pixels on semantic segmentation, which outperformed the state of the art model without additional machinery. Liu *et al.* [11] proposed single shot multi-box detector (SSD) based on FCN that can detect objects quickly and accurately. Besides, FCN also has been exploited in visual recognition. He *et al.* [7] introduced a spatial pyramid pooling (SPP) layer imposed on FCN to produce fixed-length representation from arbitrary-size inputs.

Sparse Representation Classification. Wright *et al.* [22] introduced a well-known method, SRC for face recognition,

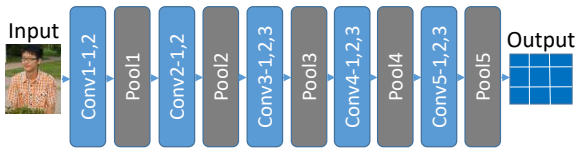

Figure 3. Fully convolutional network.

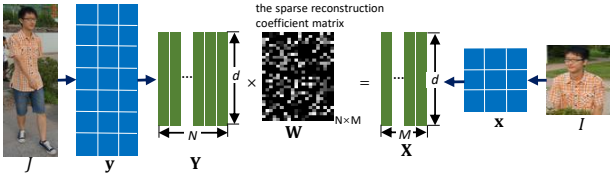

Figure 4. Deep Spatial feature Reconstruction.

achieving a robust performance under occlusions and illumination variations. Further studies [4, 27, 24, 23] based on SRC about face recognition have also been conducted. SRC has been also applied to signal classification [8], visual tracking [15], and visual classification [26], etc.

Partial Person Re-identification. Partial person re-id has become an emerging problem in video surveillance. However, few methods consider how to match an arbitrary patch of a person image. To address this problem, many methods [3, 6] warp an arbitrary patch of an image to a fixed-size image, and then extract the fixed-length feature vector for matching. However, such processing way would result in unwanted deformation. Besides, part-based models offer a kind of solution to partial person re-id. Patch-to-patch matching strategy is employed to handle occlusion and cases where the target is partially out of camera's view. Zheng *et al.* [32] proposed a local patch-level matching model called Ambiguity-sensitive Matching Classifier (AMC) that was based on dictionary learning with explicit patch ambiguity modeling, and introduced a global part-based matching model called Sliding Window Matching (SWM) that can provide complementary spatial layout information. But, the computation cost of AMC+SWM is rather extensive because it runs feature extractor many times without sharing computation. Furthermore, similar occlusion problem also occurs in partial face recognition, Liao *et al.* [12] proposed an alignment-free approach called multiple keypoints descriptor SRC (MKD-SRC), where multiple affine invariant keypoints are extracted for facial features representation and sparse representation based on classification (SRC) [22] is then used for face classification. However, the performance of keypoint-based methods is not quite satisfying with hand-crafted local descriptors. To this end, we propose a fast and accurate method, Deep Spatial feature Reconstruction (DSR), to address partial person images.

3. The Proposed Approach

3.1. Fully Convolutional Network

Deep Convolutional Neural Networks (CNNs), as feature extractors in visual recognition task, require a fixed-size input image. However, it is impossible to meet the requirement since partial person images have arbitrary sizes/scales. In fact, the requirement comes from fully-connected layers that demand fixed-length vectors as inputs. Convolutional layers operate in a sliding-window manner and generate correspondingly-size spatial outputs. To handle an arbitrary patch of a person image, we discard all fully-connected layers to implement Fully Convolutional Network that only convolution and pooling layers remain. Therefore, FCN still retains spatial coordinate information, which is able to extract spatial feature maps from arbitrary-size inputs. The proposed FCN is shown in Fig. 3, it contains 13 convolution layers and 5 pooling layers, and the last pooling layer produces identity feature maps.

3.2. Deep Spatial Feature Reconstruction

In this section, we will introduce how to measure the similarity between a pair of person images of different sizes. Assume that we are given a pair of person images, one is an arbitrary patch of person image I (partial person), and the other is holistic person image J. Correspondingly-size spatial feature maps $\mathbf{x} = conv(I, \theta)$ and $\mathbf{y} = conv(J, \theta)$ are then extracted by FCN, where θ denotes the parameters in FCN. $\mathbf{x}$ denotes a vectorized $w \times h \times d$ tensor, where w, h and d denote the height, the width and the number of channel of $\mathbf{x}$, respectively. As shown in Fig. 3, we divide $\mathbf{x}$ into N blocks $\mathbf{x}_n$, $n = 1, \ldots, N$, where $N = w \times h$, and the size of each block is $1 \times 1 \times d$. Denote by

$$\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \in \mathbb{R}^{d \times N} \tag{1}$$

the block set, where $\mathbf{x}_n \in \mathbb{R}^{d \times 1}$. Likewise, $\mathbf{y}$ is divided into M blocks as

$$\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_M\} \in \mathbb{R}^{d \times M}, \tag{2}$$

then $\mathbf{x}_n$ can be represented by linear combination of $\mathbf{Y}$. That is to say, we attempt to search similar blocks to reconstruct $\mathbf{x}_n$. Therefore, we wish to solve for the sparse coefficients $\mathbf{w}_n$ of $\mathbf{x}_n$ with respect to $\mathbf{Y}$, where $\mathbf{w}_n \in \mathbb{R}^{M \times 1}$. Since few blocks of $\mathbf{Y}$ are expected for reconstructing $\mathbf{w}_n$, we constrain $\mathbf{w}_n$ using ℓ_1-norm. Then, the sparse representation formulation is defined as

$$\min_{\mathbf{w}_n} ||\mathbf{x}_n - \mathbf{Y}\mathbf{w}_n||_2^2 + \beta||\mathbf{w}_n||_1, \tag{3}$$

where β ($\beta = 0.4$ is fixed in our experiment) controls the sparsity of coding vector $\mathbf{w}_n$. $||\mathbf{x}_n - \mathbf{Y}\mathbf{w}_n||_2$ is used to

Algorithm 1 Spatial Feature Reconstruction.

Input: A probe person image I of an arbitrary-size; a gallery person image J.
Output: Similarity score d.

1: Extract probe feature maps $\mathbf{x}$ and gallery feature maps $\mathbf{y}$.
2: Divide $\mathbf{x}$ and $\mathbf{y}$ into multiple blocks: $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{x}_M\}$.
3: Solve equation (3) to obtain sparse reconstruction coefficient matrix $\mathbf{W} = \{\mathbf{w}_1, \cdots, \mathbf{w}_N\}$.
4: Solve equation (4) to obtain similarity score.

measure the similarity between $\mathbf{x}_n$ and $\mathbf{Y}$. For N blocks in $\mathbf{X}$, the matching distance can be defined as

$$d = \frac{1}{N}||\mathbf{X} - \mathbf{Y}\mathbf{W}||_F^2, \tag{4}$$

where $\mathbf{W} = \{\mathbf{w}_1, \cdots, \mathbf{w}_N\} \in \mathbb{R}^{M \times N}$ is the sparse reconstruction coefficient matrix. The whole matching procedure is exactly our proposed Deep Spatial feature Matching (DSR). As such, DSR can be used to classify a probe partial person, which does not need additional person alignment. The flowchart of our DSR approach is shown in Fig. 4 and the overall DSR approach is outlined in Algorithm 1.

3.3. Fine-tuning on Pre-trained FCN with DSR

We train the FCN with a particular identification signal that classifies each person images (320×120 in our experiment) into different identities. Concretely, the identification is achieved by the last pooling layer connected with an entropy-loss (see Fig. 5(a)). To further increase the discriminative ability of deep features extracted by FCN, fine-tuning with DSR is adopted to update the convolutional layers, the framework is shown in Fig. 5(b).

The DSR signal encourages the feature maps of the same identity to be similar while feature maps of the different identities stay away. The DSR can be regarded as verification signal, the loss function is thus defined as

$$\mathcal{L}_{veri} = \min_{\theta}(\theta, \mathbf{W}) \frac{\alpha}{N} \min_{\mathbf{W}} ||\mathbf{X} - \mathbf{Y}\mathbf{W}||_F^2 + \beta||\mathbf{W}||_1 \tag{5}$$

where $\alpha = 1$ means that the two features are from the same identity and $\alpha = -1$ for different identities.

We employ an alternating optimization algorithm to optimize $\mathbf{W}$ and θ in the objective $\mathcal{L}_{veri}$.

Step 1: fix θ, optimize $\mathbf{W}$. The aim of this step is to solve sparse reconstruction coefficient matrix $\mathbf{W}$. For solving optimal $\mathbf{W}$, we solve $\mathbf{w}_1, \ldots, \mathbf{w}_N$ respectively, hence, equation (3) is further rewritten as

$$\min_{\mathbf{w}_n} \frac{1}{2}\mathbf{w}_n^T\mathbf{Y}^T\mathbf{Y}\mathbf{w}_n - \mathbf{x}_n^T\mathbf{Y}\mathbf{w}_n + \beta||\mathbf{w}_n||_1. \tag{6}$$

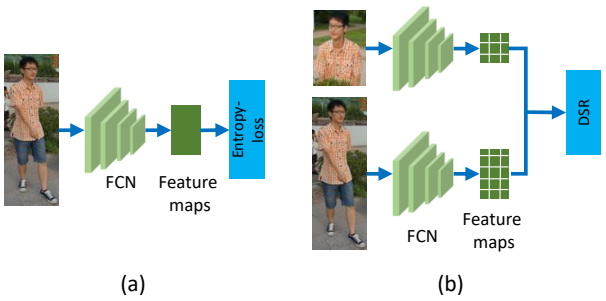

Figure 5. (a) Train FCN with identification signal (entropy-loss). (b) Fine-tune on pre-trained FCN using DSR.

Algorithm 2 Feature Learning with DSR.

Input: Training data I and J. The parameter of indicator value α and sparsity strength β. Pre-trained FCN parameter θ.
Output: FCN parameter θ.

1: Extract multiple blocks $\mathbf{X}$ and $\mathbf{Y}$.
2: $t + 1 \leftarrow t$
3: Compute the reconstruction error by $\mathcal{L}_{veri}(\mathbf{W}, \theta)$.
4: Update the sparse reconstruction coefficient matrix $\mathbf{W}$ using Equation (6).
5: Update the gradients of $\mathcal{L}_{veri}(\mathbf{W}, \theta)$ with respect to $\mathbf{X}$ and $\mathbf{Y}$.
6: Update the parameters θ by $\theta^{t+1} = \theta^t - \alpha(\frac{\partial \mathcal{L}_{veri}}{\partial \mathbf{X}}\frac{\partial \mathbf{X}}{\partial \theta^t} + \frac{\partial \mathcal{L}_{veri}}{\partial \mathbf{Y}}\frac{\partial \mathbf{Y}}{\partial \theta^t})$
7: **end while**

We utilize the feature-sign search algorithm adopted in [9] to solve an optimal $\mathbf{w}_n$.

Step 2: fix $\mathbf{w}_c$, optimize θ. To update the parameters in FCN, we then calculate the gradients of $\mathcal{L}_{veri}(\theta)$ with respect to $\mathbf{X}$ and $\mathbf{Y}$

$$\begin{cases} \frac{\partial \mathcal{L}_{veri}(\theta)}{\partial \mathbf{X}} = 2\alpha(\mathbf{X} - \mathbf{Y}\mathbf{W}) \\ \frac{\partial \mathcal{L}_{veri}(\theta)}{\partial \mathbf{Y}} = -2\alpha(\mathbf{X} - \mathbf{Y}\mathbf{W})\mathbf{W}^T. \end{cases} \tag{7}$$

Clearly, FCN supervised by DSR is trainable and can be optimized by standard Stochastic Gradient Descent (SGD). In Algorithm 2, we summarize the algorithm details of feature learning with DSR.

We directly embed the proposed DSR into FCN to train an end-to-end deep network, which can improve the overall performance. It is noteworthy that person images in each training pair share the same scale.

3.4. Multi-scale Block Representation

Invariance to varying probe scale is a challenging problem for an arbitrary patch of a person image. Unlike holistic person image, we can directly resize the person image to a fixed size. With regard to a partial person image, it is dif-

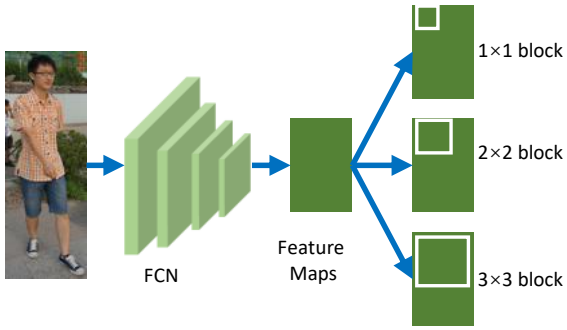

Figure 6. Multi-scale block representation.

ficult to determine its scale explicitly. Therefore, the scales between a partial person and a holistic person are easily mismatching, which results in the degraded performance. In Sec. 3.2, we use single-scale block (1×1 block), it is not very robust to scale variations. To alleviate the influence of scale mismatching, multi-scale block representation is also proposed in DSR (see Fig. 6). In our experiment, we adopt 3 different scale blocks: 1×1, 2×2 and 3×3 and extract these blocks in a sliding-window manner (stride is 1 block).

In order to keep the dimensions consistent, 2×2 and 3×3 blocks are resized to 1×1 block by average pooling. The resulting blocks are all pooled in the block set. The main purpose of multi-scale block representation is to improve the robustness of scale variation. Experiment results show that such processing stages can effectively improve the performance of partial person re-id.

Unlike some detection-based models that perform a multi-scale operation in image-level, the designed multi-scale block representation here is operated in feature-level. Therefore, the computation cost of feature extraction is extensive inevitably. It is also apparent that multi-scale block representation in the feature-level is quite efficient by sharing computation.

4. Experiments

In this section we mainly focus on five aspects below, 1). explore the influence of deformable person images; 2). the benefits of multi-scale block representation; 3). comparisons with other partial person re-id approaches; 4). computational time of various partial person re-id approaches; 5). effectiveness of fine-tuning with DSR.

4.1. Experiment Settings

Network Architecture. The designed Fully Convolutional Network (FCN) is shown in Fig. 3. The Market1501 dataset [21] is used to train the FCN with a 1,500-way softmax to obtain a pre-trained model and the size of network input is 320 × 120. 3,000 positive pairs of person images and

Dataset	Individual	Image	Gallery	Probe
Partial REID [32]	60	600	5	5
Partial-iLIDS [31]	119	476	3	1

Figure 7. Examples of partial persons in Partial REID (a) and P-iLIDS Dataset (b) Datasets.

3,000 negative pairs of person images are used to fine-tune on pre-trained FCN with DSR. For each pair, one is a holistic person image and the other one is an arbitrary patch of a person image.

Datasets. Partial REID dataset is a specially designed partial person Dataset that includes 600 images from 60 people, with 5 full-body images and 5 partial images per person. These images are collected at a university campus from different viewpoints, background and different types of severe occlusion. The examples of partial persons in the Partial REID dataset are shown in Fig. 7(a). The region in the red bounding box is the partial person image. The probe set consists of all partial images per person, and the holistic person images are used as the gallery set. Partial-iLIDS is a simulated partial person dataset based on iLIDS [31]. The iLIDS contains a total of 476 images of 119 people captured by multiple non-overlapping cameras. Some images in the dataset contain people occluded by other individuals and luggage. Fig. 7(b) shows some examples of individual images from the iLIDS dataset. For the occluded individuals, the partial observation is generated by cropping the non-occluded region of one image of each person to construct the probe set. The non-occluded images of each person are selected as a gallery set. There are $p = 60$ and $p = 119$ individuals in each test set for the Partial REID and Partial-iLIDS datasets respectively. One and five partial person images of each person are used as a probe set for the Partial REID and Partial-iLIDS datasets, respectively.

Evaluation Protocol. In order to show the performance of the proposed approach, we provide the average Cumulative Match Characteristic (CMC) curves for close-set experiment and Receiver Operating Characteristic (ROC) curves for verification experiment to evaluate our algorithm.

Benchmark Algorithms. Some existing partial person re-identification methods are used for comparison, including part-based matching method Ambiguity-sensitive Matching (AMC) [32], global-to-local matching method Sliding Win-

Table 1. Influence of person image deformation (rank-1 accuracy).

Method	Partial REID		Partial-iLIDS	
	$N = 1$	$N = 3$	$N = 1$	$N = 3$
Resizing model	19.33	26.00	21.85	28.57
DSR	**39.33**	**49.33**	**51.06**	**54.58**

dow Matching (SWM) [32], AMC+SWM [32] and Resizing model (see Fig. 2(a)). For AMC, features are extracted from a 64×64 support area, and these support areas are densely sampled with an overlap of half of the height/width of the supporting area in both horizontal and vertical directions. Each region is represented by the fine-tuning FCN, creating a 2,048-dimensional feature vector (the output size is $2 \times 2 \times 512$ in the designed FCN).

Settings. Single-shot and multi-shot experiments are conducted respectively. Single-shot experiment means that single ($N = 1$) person image is used as gallery image for each individual. Multi-shot experiment means that multiple ($N > 1$) person images are used as gallery images for each individual.

4.2. Influence of Person Image Deformation

Fig. 2(a) shows the details of the resizing model, person images in the gallery and probe set are all re-sized to 320×120. FCN is used as feature extractor and 15,360-dimension feature vector is produced for each person image. In the single-shot experiment, we use Euclidean distance to measure the similarity of a pair of person images in the Resizing model. In the multi-shot experiment, we return the average similarity between a probe person image and multiple person images of an individual. For DSR, we only adopt single-scale block representation (1×1 block) in this experiment. Table. 1 shows the experimental results on Partial REID and Partial-iLIDS datasets. Regardless of single-shot experiments or multi-shot experiments, the gap between resizing model and DSR is very large. Such experimental results convincingly show that person image deformation would produce a significant influence on recognition performance. For example, an upper part of the person image is re-sized to fixed-size, which results in the entire image to be stretched vertically.

4.3. Multi-scale Block Representation Benefits

To evaluate the performance of the proposed DSR with regard to the multi-scale block representation, we pool different-size blocks into the gallery and probe block set. 3 different fusion ways are adopted: 1×1 blocks, 1×1 blocks combined with 2×2 and 1×1 blocks, 2×2 blocks combined with 3×3 blocks. Results are shown in Fig. 8. DSR achieve the best performance when gallery and probe block set contain 1×1, 2×2 and 3×3 blocks. Experimental results suggest that multi-scale block representation is effective. The single-scale block contains more local information, while the multi-scale block is able to provide complementary information to make DSR more robust to scale variation.

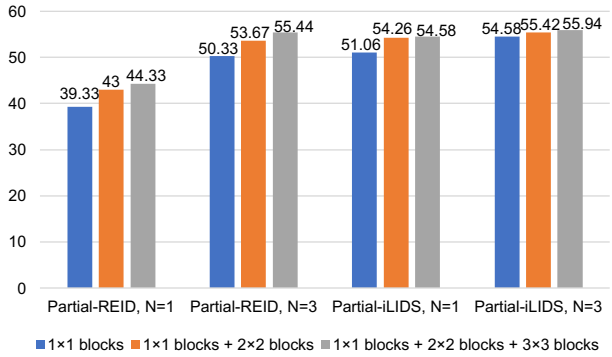

Figure 8. Rank-1 accuracy of DSR with single-scale block representation and multi-scale block representation.

Table 2. Performance comparison on single-shot experiment.

Method	Partial REID		Partial-iLIDS	
	$r = 1$	$r = 3$	$r = 1$	$r = 3$
Resizing model	19.33	32.67	21.85	36.97
SWM [32]	24.33	45.00	33.61	47.06
AMC [32]	33.33	46.00	46.78	64.75
AMC+SWM [32]	36.00	51.00	49.58	63.34
DSR (single-scale)	39.33	55.67	51.06	61.66
DSR (multi-scale)	**43.00**	**60.33**	**54.58**	**64.50**

4.4. Comparison to the State-of-the-Art

We compare the proposed DSR to the state-of-the-art methods, including AMC, SWM, AMC+SWM and Resizing model, on the Partial REID and Partial-iLIDS datasets. There are $p = 60$ and $p = 119$ individuals in each of the test sets for the Partial REID and Partial-iLIDS datasets respectively. For DSR, we report the results using single-scale block representation and multi-scale bloc representation. For AMC+SWM, the weights of AMC and SWM are 0.7 and 0.3, respectively. Both the single-shot setting and the multi-shot setting are conducted in this experiment.

Single-shot experiments. Table 2 shows the single-shot experimental results. We find the results on Partial REID and Partial-iLIDS are similar. The proposed method DSR outperforms AMC, SWM, AMC+SWM and Resizing model. DSR takes full advantage of FCN that operate in a sliding-window manner and outputs feature maps without deformation. AMC is a local-to-local matching method that achieves comparable performance because background patches can be automatically excluded due to their low visual similarity. Thus, it is somewhat robust to occlusion. However, it is difficult to select satisfactory support area size and stride making it not robust to scale variation. SWM is a local-to-global matching method, which requires that

Table 3. Performance comparison on multi-shot experiment.

Method	Partial REID		Partial-iLIDS	
	$r=1$	$r=3$	$r=1$	$r=3$
Resizing model	26.00	37.00	28.57	43.67
SWM [32]	34.33	47.67	35.33	49.67
AMC [32]	42.33	55.67	44.67	56.33
AMC+SWM [32]	44.67	56.33	52.67	63.33
DSR (single-scale)	49.33	65.67	54.67	64.33
DSR (multi-scale)	**53.67**	**72.33**	**55.46**	**68.07**

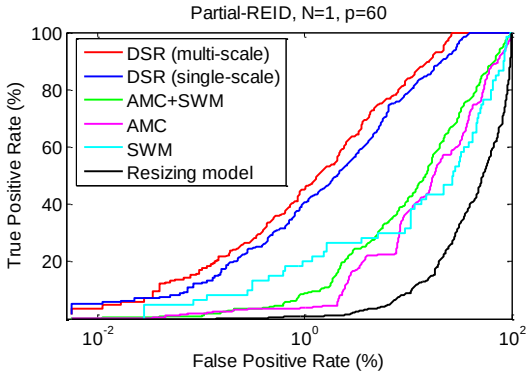

Figure 9. ROC curves of various partial person re-id approaches on Partial REID Dataset.

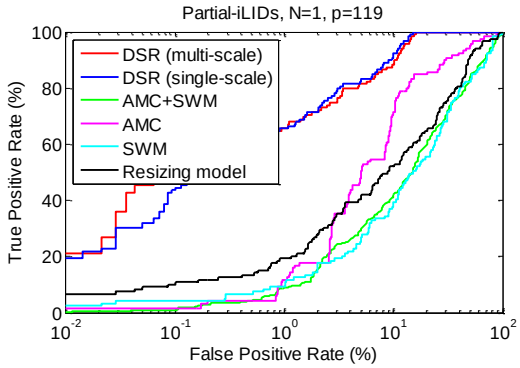

Figure 10. ROC curves of various partial person re-id approaches on Partial-iLIDS Dataset.

the probe size is smaller than the gallery size. Search manner in SWM would ignore some detailed information about a person image. AMC+SWM perform as well as DSR because local features in AMC combined with global features in SWM makes it robust to occlusion and view/pose various. Similar results are also observed from the ROC curves shown in Fig. 9 and Fig. 10. Obviously, DSR shows small intra-distance and large inter-distance.

As shown in Fig. 11, we illustrate how to find the most similar person image to an input probe image. Four blocks are respectively reconstructed by all blocks from gallery feature maps, then the reconstruction errors are averaged to find the minimum one. Looking carefully the reconstruction coefficients, the feature blocks from the probe could be well

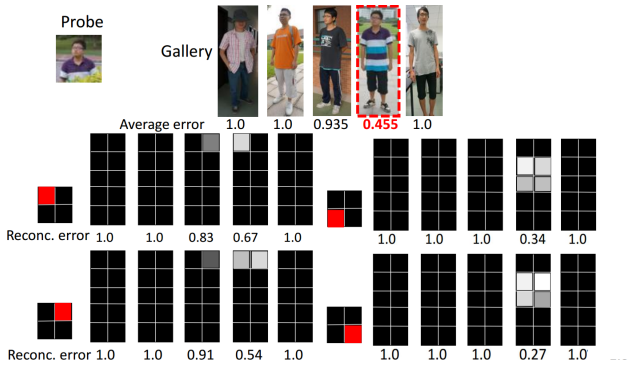

Figure 11. Examples of searching similar blocks.

reconstructed by similar feature blocks from the gallery person. Even though the gallery size and the viewpoint or position of person (unalignment) changes, we can also use DSR to find similar gallery blocks to reconstruct probe blocks, and finally return the minimum reconstruction error.

Multi-shot experiments. DSR approach is evaluated under the multi-shot setting (N=3) on Partial REID and Partial-iLIDS datasets. The results are shown in Table 3. Similar results are obtained in the single-shot experiment. Specifically, the results show that multi-shot setup helps to improve the performance of DSR since DSR increases from 39.33% to 49.33% on Partial REID dataset and from 51.06% to 54.67% on Partial-iLIDS dataset.

4.5. Computational Efficiency

Our implementation is based on the publicly available code of *MatConvnet* [20]. All experiments in this paper are trained and tested on PC with 16GB RAM, i7-4770 CPU @ 3.40GHz. Single-shot and multi-shot experiments on Partial REID dataset are conducted to test the computational time of identifying a probe person image. For DSR, we use single-scale block representation (1×1 block) and multi-scale block representation (1×1 and 2×2 blocks). Table 4 shows the computational time of various partial person re-id approaches, which suggests that the propose DSR outperforms other approaches in computation efficiency. DSR with single-scale block representation and multi-scale block representation respectively take 0.269s and 0.278s to identify a person image. For AMC, it costs more computational time than DSR because it repeatedly runs FCN for each sub-region without sharing computation. For SWM, it sets up a sliding window of the same as the probe person image to search for similar sub-region within each gallery image. Generally, many sub-regions would generate by the sliding window, which increases extensive computational time of feature extraction. Besides, when given a new probe person image, it requires regenerating sub-region by the sliding window of the same as the probe image. DSR performs better than the Resizing model, the computational cost of feature extraction would increase after resizing.

Table 4. Computational time comparison on Partial REID dataset.

Method	Computational time (s)	
	$N = 1$	$N = 3$
Resizing model	0.326	0.371
AMC [32]	0.972	1.213
SWM [32]	81.519	237.144
DSR (single-scale)	**0.269**	**0.265**
DSR (multi-scale)	0.278	0.285

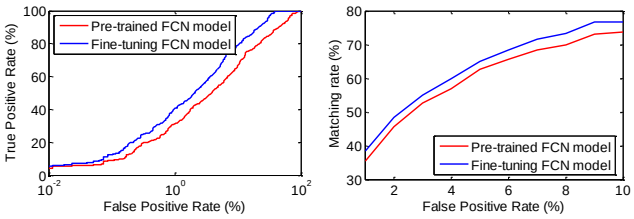

Figure 12. ROC curves and CMC curves on Partial REID dataset using pre-trained FCN model and fine-tuning FCN model (N=1).

4.6. Contribution of Fine-tuning with DSR

In section 3.3, DSR is used to fine-tune on the pre-trained FCN to learn more discriminative spatial features. To verify the effectiveness of fine-tuning FCN with DSR, we conduct the single-shot experiment on Partial REID dataset. We compare the pre-trained FCN (FCN training only with softmax loss is regarded as a pre-trained model) to the fine-tuning FCN with DSR (fine-tuning model). Fig. 12 shows ROC curves and CMC curves of these two models. Experimental results show that the fine-tuning FCN model performs better than the pre-trained model, which indicates that fine-tuning with DSR can learn more discriminative spatial deep features. Pre-trained model with softmax loss training can only represent the probability of each class that a person image belongs to. For the fine-tuning model, DSR can effectively reduce the intra-variations between a pair of person images of the same individual.

4.7. Evaluation on Holistic Person Image

To verify the effectiveness of DSR on holistic person re-identification, we carry out additional holistic person re-id experiments on Market1501 dataset [30]. Market1501 is one of the largest benchmark dataset that contains 1,501 individuals which are captured by six surveillance cameras in campus. Each individual is captured by two disjoint cameras. Totally it consists of 13,164 person images and each individual has about 4.8 images at each **viewpoint**. We follow the standard test protocol, i.e., 751 individuals are used for training and 750 individuals are used for testing. The ResNet50 pre-trained on ImageNets is used as the baseline model. For DSR, feature maps extracted from *res5c* are used as identity feature. We respectively adopt single-scale representation (1×1) and multi-scale representation (1×1, 2×2 and 3×3) in feature representation term. Experimental results in Table 5 suggest that DSR achieves the best performance. We draw three conclusions: 1) DSR is very effective compared to Euclidean distance because DSR can automatically search similar feature blocks for best matching; 2) Multi-scale presentation can achieve better results because it avoids the influence of scale variations; 3) Training model with DSR effectively learns more discriminative deep spatial features, which encourages the feature maps of the same identity to be similar while feature maps of the different identities are pushed far apart.

Table 5. Experimental results on Market1501 with single query.

Method	$r = 1$	mAP
BOW [30]	34.38	14.10
MSCAN [10]	80.31	57.53
Spindle [28]	76.90	-
Re-ranking [33]	77.11	63.63
CADL [13]	80.85	55.58
CAMEL [25]	54.50	26.30
DNSL+OL-MANS [34]	60.67	-
DLPAR [29]	81.00	-
Resnet50-pool5 +Euclidean distance (baseline model)	77.40	55.64
Resnet50-res5c (single-scale)+DSR	82.72	61.25
Resnet50-res5c (multi-scale)+DSR	**83.58**	**64.25**

5. Conclusion

We have proposed a novel approach called Deep Spatial feature Reconstruction (DSR) to address partial person re-identification. To get rid of the fixed input size, the proposed spatial feature reconstruction method provides a feasibility scheme where each channel in the probe spatial feature map is linearly reconstructed by those channels of a gallery spatial image map, it also avoids the trivial alignment-free matching. Furthermore, we embed DSR into FCN to learn more discriminative features, such that the reconstruction error for a person image pair from the same person is minimized and that of image pair from different persons is maximized. Experimental results on the Partial REID and Partial-iLIDS datasets validate the effectiveness and efficiency of DSR, and the advantages over various partial person re-id approaches are significant. Additionally, the proposed method is also competitive in the holistic person dataset, Market1501.

Acknowledgments This work is supported by the Beijing Municipal Science and Technology Commission (Grant No. Z161100000216144) and National Natural Science Foundation of China (Grant No. 61427811). Special thanks to Dangwei Li who supports our experiments.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 2

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 2

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, 2014. 3

[4] S. Gao, I. W.-H. Tsang, and L.-T. Chia. Kernel sparse representation for image classification and face recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 2010. 3

[5] R. Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[7] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1904–1916, 2015. 2

[8] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in neural information processing systems (NIPS)*, 2007. 3

[9] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems (NIPS)*, 2007. 4

[10] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[11] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[12] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(5):1193–1205, 2013. 2, 3

[13] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*. Springer, 2016. 2

[15] X. Mei and H. Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(11):2259–2272, 2011. 3

[16] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*, 2015. 2

[19] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(4):640–651, 2017. 2

[20] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *ACM international conference on Multimedia (ACM MM)*. ACM, 2015. 7

[21] J. Wang and S. Li. Query-driven iterated neighborhood graph search for large scale indexing. In *ACM international Conference on Multimedia (ACM MM)*, 2012. 5

[22] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(2):210–227, 2009. 2, 3

[23] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang. A two-phase test sample sparse representation method for use with face recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 21(9):1255–1262, 2011. 3

[24] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. *European Conference on Computer Vision (ECCV)*, 2010. 3

[25] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[26] X.-T. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing (TIP)*, 21(10):4349–4360, 2012. 3

[27] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 2, 3

[28] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[29] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. 2017. 8

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 8

[31] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 649–656, 2011. 2, 5

[32] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3, 5, 6, 7, 8

[33] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. 2017. 8

[34] J. Zhou, P. Yu, W. Tang, and Y. Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 8