Recognizing Partial Biometric Patterns

Lingxiao He, Student Member, IEEE, Zhenan Sun, Member, IEEE, Yuhao Zhu and Yunbo Wang

Abstract—Biometric recognition on partial captured targets is challenging, where only several partial observations of objects are available for matching. In this area, deep learning based methods are widely applied to match these partial captured objects caused by occlusions, variations of postures or just partial out of view in person re-identification and partial face recognition. However, most current methods are not able to identify an individual in case that some parts of the object are not obtainable, while the rest are specialized to certain constrained scenarios. To this end, we propose a robust general framework for arbitrary biometric matching scenarios without the limitations of alignment as well as the size of inputs. We introduce a feature post-processing step to handle the feature maps from FCN and a dictionary learning based Spatial Feature Reconstruction (SFR) to match different sized feature maps in this work. Moreover, the batch hard triplet loss function is applied to optimize the model. The applicability and effectiveness of the proposed method are demonstrated by the results from experiments on three person re-identification datasets (Market1501, CUHK03, DukeMTMC-reID), two partial person datasets (Partial REID and Partial iLIDS) and two partial face datasets (CASIA-NIR-Distance and Partial LFW), on which state-of-the-art performance is ensured in comparison with several state-of-the-art approaches. The code is released online and can be found on the website: https://github.com/lingxiao-he/Partial-Person-ReID.

Index Terms—Partial Biometric Recognition, Spatial feature Reconstruction, Person Re-identification, Face Recognition

1 INTRODUCTION

Biometric recognition, especially face recognition and person re-identification (re-id), has attracted significant research attention as the demand of identification using images captured by CCTV cameras and video surveillance systems growing rapidly. In these scenarios, the random poses and perspectives of the target object, unwilling occlusions caused by other objects (e.g., hair, sunglasses even other individuals for a person or eyelids/eyelashes for irises) and only partly captured images of target objects would degrade the performance of surveillance systems. With the great progress has been made on biometric identification in recent years due to the development of deep learning, many approaches are proposed from global researches. And we consider these approaches can be divided into two generations.

The first generation approaches generally assume that each image covers full glance of one object. However, the assumption of biometric matching on full and frontal images does not always hold in real-world scenarios, where we merely have access to a few parts of images for identification. For instance shown in Fig. 1, a face is easily occluded by accessories such as sunglasses, scarfs and a person on the street can easily be occluded by moving obstacles (e.g., cars, other persons) and static ones (e.g., trees, barriers), resulting in partial observations of the target object. Besides, the frequently presented arbitrary posture of an object in video surveillance introduces additional difficulties to real-world biometric identification problems. Moreover, an object may be positioned partially outside cameras view, resulting in an arbitrary-size image. These emerging problems would reduce the performance of the first generation methods.

The drawbacks of first generation approaches makes researchers to design a framework to address partial biometric identification problems, where the second generation approaches advent. To match an arbitrary patch of an image, some researchers resort to re-scale an arbitrary patch of the image to a fixed-size image. However, the performance would be significantly degraded due to the undesired deformation. Part-based models [4], [5], [9], [17],
Multi-scale spatial feature reconstruction distance $r$ is utilized to generate spatial feature maps of a certain size. And then global averaging pooling and pyramid pooling is utilized to produce global feature and multi-scale spatial feature. For a probe and a gallery, we first extract their spatial feature pyramid pooling is utilized to produce global feature and multi-scale spatial feature. Secondly, we calculate the reconstruction coefficients $W$, and then obtain the reconstruction spatial feature $X'$ by $Y'W$. Finally, we fuse the global-to-global matching distance $d$ and spatial feature reconstruction (SFR) distance $r$. As shown in the bottom of figure, we use the triplet loss to optimize $\theta$ by using the distance metric: $s = d + r$.

Figure 2. The proposed framework. Fully Convolutional Network (FCN) is utilized to generate spatial feature maps of a certain size. And then a feature post-processing unit consist of global averaging pooling and pyramid pooling is utilized to produce global feature and multi-scale spatial feature. For a probe and a gallery, we first extract their spatial feature $X$ and $Y$ and global feature. Secondly, we calculate the reconstruction coefficients $W$, and then obtain the reconstruction spatial feature $X'$ by $Y'W$. Finally, we fuse the global-to-global matching distance $d$ and spatial feature reconstruction (SFR) distance $r$. As shown in the bottom of figure, we use the triplet loss to optimize $\theta$ by using the distance metric: $s = d + r$.

The remainder of this paper is organized as follows: In Sec. 2, we review the related work about the existing person re-id and partial person re-id algorithms. Sec. 3 introduces the technical details of spatial feature reconstruction and batch hard triplet SFR learning. Sec. 4 shows the experimental results and analyzes the performance in accuracy. Sec. 5 discuss the advantages and disadvantages of the proposed approach. Finally, we conclude our work in Sec. 6.

2 Literature Review

As our approach is expected to settle multiple biometric identification problems yet current existing approaches are specialized to one of person re-identification, partial person re-identification or partial face re-identification, we would love to review some of them and give comparisons in Sec. 4 to show our approach holds state-of-the-art on these problems without any specializing and pre-alignment.
2.1 Person Re-identification

Part-based models [4], [5], [17], [40], [47], [39] are widely applied to person re-identification since they could achieve significant performance. Zhao et al. [47] proposed a novel Spindle Net based on human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion. Li et al. [17] design a Multi-Scale Context-Aware Network (MSCAN) to learn powerful features over full body and body parts, which can well capture the local context knowledge by stacking multi-scale convolutions in each layer. Moreover, instead of using predefined rigid parts, they proposed to learn and localize deformable pedestrian parts using Spatial Transformer Networks (STN) with part-based representation. Besides, Sun et al. [39] proposed a network named Part-based Convolutional Baseline (PCB) that outputs a convolutional descriptor consisting of several part-level features. PCB is able to lay emphasis on the content consistency within each part. However, these methods require the presence of certain person components and pre-alignment.

Mask-guided models [16], [28], [36] provide a solution for person re-identification. Mask as external cue helps to remove the background clutters in pixel-level and contain body shape information. Song et al. [36] introduced the binary segmentation masks to construct synthetic RGB-Mask pairs as inputs, then they design a mask-guided contrastive attention model (MGCAM) to learn features separately from the body and background regions. Kalayeh et al. [16] proposed a person re-identification model that integrated human semantic parsing in person re-identification. Similar to [36], Qi et al. [28] combined source images with person masks as the inputs to remove the appearance variations (illumination, pose, occlusion, etc.). Although mask-guided approaches can achieve satisfying performance, they extremely rely on accurate pedestrian segmentation model, otherwise, it would result in poor performance.

Pose-guided models [37], [38], [21], [29] utilize skeleton as a external cue in person re-identification to reduce the part misalignment problem. Each part can be well located using person landmarks. Su et al. [37] proposed a Pose-driven Deep Convolutional (PDC) model to learn improved feature extractors and matching models from end-to-end, PDC can explicitly leverages the human part cues to alleviate the pose variations. Suh et al. [38] proposed a two-stream network that consisted appearance map extraction stream and body part map extraction stream. And then a part-aligned feature map is obtained by a bilinear mapping of the corresponding local appearance and body part descriptors. Except for the person alignment, some works [21], [29] proposed pose-transferrable models that combined pose estimation and Generative Adversarial Networks (GAN) to augment training samples. The same as the mask-guided models, pose estimation may fail to work due to the loss of person component and severe occlusions.

Attention-based models [19], [34], [18], [44], [54] take advantages of attention mechanism to extract more discriminative feature. In fact, attention mechanism is a feature selection approach. Li et al. [19] formulated a novel Harmonious Attention CNN (HA-CNN) model for joint learning of soft pixel attention and hard regional attention along with simultaneous optimisation of feature representations, dedicated to optimise person re-id in uncontrolled (misaligned) images. Si et al. [34] proposed a dual attention mechanism, in which both intra-sequence and inter-sequence attention strategies are used for feature refinement and feature-pair alignment, respectively. Besides, attentive spatial-temporal networks [18], [44], [54] are widely used in video-based person re-identification task.

2.2 Partial Person Re-identification

Partial person re-id has become an emerging problem in video surveillance. To address this problem, many methods [7], [8] warp an arbitrary patch of an image to a fixed-size image, and then extract fixed-length feature vectors for matching. However, such method would result in undesired deformation. Part-based models are considered as a solution to partial person re-id. Patch-to-patch matching strategy is employed to handle occlusions and cases where the target is partially out of the camera’s view. Zheng et al. [51] proposed a local patch-level matching model called Ambiguity-sensitive Matching Classifier (AMC) 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. However, the computation cost of AMC+SWM is rather expensive as features are calculated repeatedly without further acceleration.

2.3 Partial Face Recognition

Many approaches [14], [20], [41] proposed for solving partial face recognition are keypoint-based. Hu et al. [14] proposed an approach based on SIFT descriptor [22] representation that does not require alignment, and the similarities between a probe patch and each face image in the gallery are computed by the instance-to-class (I2C) distance with the sparse constraint. Liao et al. [20] proposed an alignment-free approach called multiple key points descriptor SRC (MKD-SRC), where multiple affine invariant key points were extracted for facial features representation and sparse representation based on classification (SRC) [42] was used for classification. Weng et al. [41] proposed a Robust Point Set Matching (RPSM) method based on SIFT descriptor, SURF descriptor [2] and LBP [1] histogram for partial face matching. Their approach first aligned the partial faces and then computed the similarity of the partial face and a gallery face image. However, the computational cost of each algorithms is expensive and the required alignment step limits its practical applications. Besides, region-based models [3], [9], [23], [24], [25], [32], [33] also offered a solution for partial face recognition. They only required face sub-regions as input, such as eye [32], nose [32], half (left or right portion) of the face [9], or the periorocular region [26]. He et al. [11] proposed a Dynamic Feature Matching (DFM) model and achieves the highest performance (94.96%)for partial face recognition on CASIA-NIR-Distance database [12]. However, these methods require the presence of certain facial components and pre-alignment. To this end, we propose an alignment-free partial re-identification algorithm.
that achieves better performance with higher computation efficiency.

3 The Proposed Approach

We will give a clear explanation of the proposed approach in this section from network definition to loss construction. The code is available on https://github.com/lingxiao-he/Partial-Person-ReID.

3.1 Architecture of Deep Network

For a quick view, the feature matching process is shown in Fig. 2. In the proposed network, a Fully Convolution Network (FCN) is adopted to extract spatial features, which are post-processed by a unit consist of two feature extraction branches are implemented: global features are extracted by global average pooling layer (GAP) and multi-scale spatial features are extracted by pyramid pooling layer. Then, multi-scale spatial features are fed to SFR, a dictionary learning based reconstruction mechanism supporting matches on arbitrary sized inputs, in feature matching step. Finally, the matching score equals to the weighted sum of results from global matching and SFR matching.

3.1.1 FCN Encoder

Models pre-trained on ImageNet [6] such as VGG [35] and ResNet [10] can be viewed as a stack of multi-stage convolution layers and a sequence of fully-connected layers. Here we make use of those convolution layers (FCN) in ResNet as our feature encoder. The parameters of the encoder will be fine-tuned in the training process.

3.1.2 Feature Representation

This part introduces the two branches in feature representation step. Basically, global averaging pooling (GAP) produces one scalar representing the feature of whole picture and pyramid pooling gives a batch of features calculated on different receptive fields, which leads better performance in matching objects in arbitrary size and posture.

Global Feature. Global feature is wildly exploited in modern person re-id algorithms. Basicly, Global Averaging Pooling (GAP) realized by a single averaging layer takes the feature maps from FCN as input and outputs one scalar value each image as its global feature. As tested in existing re-id methods that global feature holds relative valid information for matching, we make it in consideration as one of our reference.

Pyramid Feature. Invariance to varying person scale is a challenging problem for an arbitrary-size person image. It is difficult to align arbitrary-size person image to pre-defined scale. Therefore, the scales between two person images are easily mismatched, resulting in the degraded performance. To this end, we propose pyramid pooling layer to extract multi-scale spatial features to alleviate the influence of scale mismatching.

As shown in Fig. 3, pyramid pooling (PP) layer consists of multiple average pooling layers of different kernel sizes so that it has different receptive fields. For a 256 × 128 input person image, we implement 4 pooling layers of different sizes: 1 × 1, 2 × 2, 3 × 3, and 4 × 4 in the pyramid pooling layer. The pyramid pooling layer filters the output spatial features at the stride of 1 to generate multi-scale spatial features. The output spatial features inferred by pooling layer of small kernel size generate dense spatial features, and each spatial feature represents the local feature of the small local region. The output spatial features inferred by pooling layer of large kernel size generate sparse spatial features, and each spatial feature represents the relatively large source region. Finally, we concate these output spatial features to obtain multi-scale spatial features. And the multi-scale features are defined as PP(fθ(x)).

3.1.3 Spatial Feature Reconstruction

Spatial feature reconstruction (SFR) between a pair of person images is introduced in this part. As shown in Fig. 4, for a pair of given person images: x and y with different sizes, correspondingly-size multi-scale spatial features X = PP(fθ(x)) = \{x_1, \ldots, x_N\} ∈ ℝ^{d × N} and Y = PP(fθ(y)) = \{y_1, \ldots, y_M\} ∈ ℝ^{d × M} are then extracted, where θ denotes the parameters of FCN. Then, x_n can be represented by linear combination of Y. That is to say, we attempt to search similar spatial features in Y to reconstruct x_n. Therefore, we wish to solve for the linear representation coefficients w_n of x_n with respect to Y, where w_n ∈ ℝ^M. We constrain w_n using ℓ2-norm. Then, the linear representation formulation is defined as

\[ \mathcal{L}(w_n) = \min_{w_n} \|x_n - Yw_n\|_2^2 + \beta|w_n|_2, \]

(1)

For N spatial features in X, the Eq. (1) can be rewritten as

\[ \mathcal{L}(W) = \min_W \|X - YW\|_2^2 + \beta|W|_F, \]

(2)
Algorithm 1 Spatial Feature Reconstruction (SFR).

**Input:** A probe person image $x$ of an arbitrary-size; a gallery person image $y$.

**Output:** Similarity score $D_s$.

1. Extract probe multi-scale spatial feature $X$ and gallery multi-scale spatial feature $Y$.
2. Solve equation (2) to obtain reconstruction coefficient matrix $W$.
3. Solve equation (12) to obtain reconstruction score.

where $W = \{w_1, \ldots, w_N\} \in \mathbb{R}^{M \times N}$, and $\beta$ controls the smoothness of coding vector $W$.

We use the least square algorithm to solve $W$, so $W = (Y^T Y + \beta \cdot I)^{-1} Y^T X$. Let $M = X - YW$, then the spatial feature reconstruction between $X$ and $Y$ can be defined as

$$D_s(X, Y) = tr(\sqrt{M^T M})/N \quad (3)$$

where $D_s(\cdot; \cdot)$ is Spatial Feature Reconstruction between a pair of person images.

### 3.2 Loss Function

Though pairwise loss with $\ell_1$ regularization in our previous work in [13], we replace it in this paper by proposed batch hard triple loss with $\ell_2$ regularization, which is found performs better than earlier implementation.

#### 3.2.1 Batch Hard Triplet Loss

The goal of triplet embedding learning is to learn a function $f_\theta(x)$. Here, we want to ensure that an image $x^a_i$ (anchor) of a specific person is closer to all other images $x^p_i$ (positive) of the same person than it is to any image $x^n_i$ (negative) of any other person. Thus, we want $D(x^a_i, x^p_i) + m < D(x^a_i, x^n_i)$, where $D(\cdot; \cdot)$ is Euclidean measure between a pair of person images. So the Triplet Loss with $N$ samples is defined as

$$L_{tr}\theta = \frac{1}{N} \sum_{i=1}^{N} [m + D(g^a_i, g^p_i) - D(g^a_i, g^n_i)] \quad (4)$$

where $m$ is a margin that is enforced between positive and negative pairs, and $g^a_i = GAP(f_\theta(x^a_i))$, $g^p_i = GAP(f_\theta(x^p_i))$ and $g^n_i = GAP(f_\theta(x^n_i))$.

To effectively select triple samples, batch hard triplet loss modified by triplet loss is adopted: the core idea is to form batches by randomly sampling $P$ subjects, and then randomly sampling $K$ images of each subject, thus resulting in a batch of $PK$ images. Now, for each anchor sample in the batch, we can select the hardest positive and hardest negative samples within the batch when forming the triplets for computing the loss, which is called as **Batch Hard Triplet Loss**:

$$L_{BH}\theta = \frac{1}{P \cdot K} \sum_{i=1}^{P} \sum_{a=1}^{K} [m + \max_{p=1, \ldots, K} D(g^a_i, g^p_i) \quad \text{hardest positive}]$$

$$- \min_{n=1, \ldots, K} D(g^a_i, g^n_i) \quad \text{hardest negative}] \quad (5)$$

which is defined for a mini-batch $B$ and where a data point $x^a_i$ corresponds to the $j$-th image of the $i$-th person in the batch. This results in $PK$ terms contributing to the loss. Additionally, the selected triplets can be considered moderate triplets, since they are the hardest within a small subset of the data, which is exactly what is best for learning with the triplet loss.

#### 3.2.2 SFR Embedded Batch Hard Triplet

**Batch Hard Triplet Spatial Feature Reconstruction** is proposed to improve the discriminative of spatial features (see Fig. 5). It encourages the spatial features of the same identity to be similar while spatial features of the different identities stay away. Batch Hard Triplet Spatial Feature Reconstruction can be defined as

$$L(\theta) = \frac{1}{PK} \sum_{i=1}^{P} \sum_{a=1}^{K} [m + \max_{p=1, \ldots, K} D(g^a_i, g^p_i) + D_s(X^a_i, X^p_i) \quad \text{hardest positive}]$$

$$- \min_{n=1, \ldots, K} D(g^a_i, g^n_i) + D_s(X^a_i, X^n_i) \quad \text{hardest negative}] \quad (6)$$

where $D(\cdot; \cdot)$ is Euclidean distance, $D_s(\cdot; \cdot)$ is Spatial Feature Reconstruction distance.

It can be seen that, the similarity distance consists of global feature matching distance (Euclidean distance) and local feature matching distance (spatial feature reconstruction).

#### 3.2.3 Optimization

We employ an alternating optimization method to optimize $\theta$.

**step 1:** fix $\theta$, obtain $W^{ap}_{i}$ and $W^{an}_{i}$. The aim of this step is to solve linear reconstruction coefficient matrix $W^{ap}_{i}$ and $W^{an}_{i}$ where $W^{ap}_{i} = ((X^p_i)^T X^p_i + \beta \cdot I)^{-1} (X^p_i)^T X^a_i$ and $W^{an}_{i} = ((X^n_i)^T X^n_i + \beta \cdot I)^{-1} (X^n_i)^T X^a_i$.

**step 2:** fix $W^{ap}_{i}$ and $W^{an}_{i}$, optimize $\theta$. We only give the gradients of $D_s(X^a_i, X^p_i)$ with respect to $X^a_i$ and $X^p_i$, and the gradients of $D_s(X^a_i, X^n_i)$ with respect to $X^a_i$ and $X^n_i$. 

![Figure 5. Batch hard triplet SFR learning.](image-url)
Algorithm 2 Feature Learning with SFR Embedded Batch Hard Triplet.

\textbf{Input}: Training data $x^n, x^p$ and $x^n$. The parameter of smoothness strength $\beta$ and learning rate $r$. Pre-trained FCN parameter $\theta$. The total of epoch: $T$. $t = 0$.

\textbf{Output}: FCN parameter $\theta$.
\begin{algorithmic}[1]
\State 1: while $t < T$ do
\State 2: Extract multiple spatial feature $X^a$, $X^p$ and $X^n$. And extract global feature $g^a$, $g^p$ and $g^n$.
\State 3: $t + 1 \leftarrow t$
\State 4: Compute the reconstruction error by $L(\theta)$.
\State 5: Update the sparse reconstruction coefficient matrix $W^n$ and $W^i$ using Equation (2).
\State 6: Update the gradients of $\frac{\partial L(\theta)}{\partial \theta}$.
\State 7: Update the parameters $\theta$ by $\theta^{t+1} = \theta^t - r \frac{\partial L(\theta^t)}{\partial \theta^t}$.
\State 8: end while
\end{algorithmic}

$$\begin{align*}
\frac{\partial D(X^a, X^p)}{\partial X^a} &= 2(X^a - X^p W_i^{ap}) \\
\frac{\partial D(X^a, X^p)}{\partial X^p} &= -2(X^a - X^p W_i^{ap}) W_i^{apT} \\
\frac{\partial D(X^a, X^n)}{\partial X^a} &= 2(X^a - X^n W_i^{an}) \\
\frac{\partial D(X^a, X^n)}{\partial X^n} &= -2(X^a - X^n W_i^{an}) W_i^{anT}.
\end{align*}$$ (7)

Then, we use Equation (7) to compute $\frac{\partial L(\theta)}{\partial \theta}$. Clearly, FCN supervised by SFR is end-to-end trainable and can be optimized by standard Stochastic Gradient Descent (SGD).

3.3 Weighted Feature Matching

This subsection will demonstrate the detail of global feature matching, spatial feature reconstruction matching and the weighted fusion of them. Suppose global feature $g_c$ and spatial feature $Y_c$ are generated from subject $c$ in the gallery. So the gallery global feature set and spatial feature set are built as respectively:

\begin{align*}
\text{Global feature set: } G &= \{g_1, g_2, \ldots, g_C\} \\
\text{Spatial feature set: } Y &= \{Y_1, Y_2, \ldots, Y_C\}
\end{align*} \tag{8}

where $g_c \in \mathbb{R}^d$, $Y_c \in \mathbb{R}^{k_c \times d}$, $k_c$ is the number of spatial features. Given an arbitrary-size probe face image, global feature $p$ and spatial feature $X$ are generated respectively. Global feature represents the appearance information of person, we directly use the Euclidean distance: $d_c = \|p - g_c\|_2$ to measure the similarity between two images. Then a distance vector of global feature matching for all the $C$ subjects is denoted as $d = \{d_1, d_2, \ldots, d_C\}$ \tag{9}

Moreover, the spatial feature matching presented above not only capture the spatial layout information of local feature, but it also achieves spatial feature matching without alignment. Therefore, it is robust to pose/view variations and person deformation. Meanwhile, such multi-scale spatial feature representation benefits scale inconsistency. Spatial feature reconstruction can always search similar spatial features from multi-scale spatial feature pool to reconstruct probe spatial feature with minimum error. The spatial feature reconstruction distance is represented as

$$r_c = D_s(X, Y_c) = tr(\sqrt{W_c})/k_c$$ \tag{10}

where $W_c = (Y_c^T Y_c + \beta I)^{-1} Y_c^T X$, and $M = X - Y_c W_c$. Then, a distance vector for all the $C$ subjects is denoted as $r = [r_1, r_2, \ldots, r_C]$ \tag{11}

To improve the retrieve accuracy, we combine the two distance vectors. The final distance vector can be written as

$$s = \alpha \cdot d + (1 - \alpha) \cdot r$$ \tag{12}

where $\alpha$ is a weight for regulating the effect of global feature matching and spatial feature reconstruction. Finally, the identity of the probe image can be determined by $\hat{c} = \arg \min_{c} s_c$, where $s_c$ is the $c^{th}$ entry of $s$.

4 Experiments

To verify the performance as well as the generalization ability of proposed method, this section includes several experiments in the order of person re-identification, partial person re-identification and partial face recognition.

4.1 Implementation Details and Evaluation Protocol

Our implementation is based on the publicly available code of PyTorch. All models in this paper are trained and tested on Linux with GTX TITAN X GPU. In the training term, all training samples are all re-scaled to $256 \times 128$, thus $8 \times 4$ spatial features are generated by FCN. No data augmentation method is used for training samples. Besides, we set margin $m = 0.3$ and $\beta = 0.001$ because it can achieve the best performance. With regard to the batch hard triplet SFR function, one batch consists of 32 subjects, and each subject has 4 different images. Therefore, each batch returns 128 groups of hard triples. The model is trained with 400 epochs and the learning rate is shown in Fig. 6.

For performance evaluation, we employ the standard metrics as in most person ReID literatures, namely the cumulative matching cure (CMC) and the mean Average Precision (mAP). To evaluate our method, we re-implement the evaluation code provided by [49] in Python.
Table 1

Performance comparison on Market1501 and CHUK03. R1: rank-1. mAP: mean Accuracy Precision.

| Method                | Market1501 | CHUK03 |
|-----------------------|-------------|--------|
|                       | single query | multiple query | Labeled | Detected |
|                       | R1 | mAP | R1 | mAP | R1 | mAP | R1 | mAP |
| Part-based            |     |      |     |      |     |      |     |      |
| Spindle (CVPR17) [47] | 76.50 | -   | -   | -   | -   | -   | -   | -   |
| MSCAN (CVPR17) [17]  | 80.31 | 57.53 | 86.79 | 66.70 | -   | -   | -   | -   |
| DLPAP (CVPR17) [48]  | 81.00 | 63.40 | -   | -   | -   | -   | -   | -   |
| AlignedReID (Arxiv17) [46] | 91.80 | 79.30 | -   | -   | -   | -   | -   | -   |
| PCB (Arxiv17) [39]   | 92.30 | 77.40 | -   | -   | 61.30 | 57.50 | -   | -   |
| Mask-guided           |     |      |     |      |     |      |     |      |
| SPReID (CVPR18) [16] | 92.54 | 81.34 | -   | -   | -   | -   | -   | -   |
| MGCAM (CVPR18) [36]  | 83.79 | 74.33 | -   | -   | 50.14 | 50.21 | 46.71 | 46.87 |
| MaskReID (Arxiv18) [28] | 90.02 | 75.30 | 93.32 | 82.29 | -   | -   | -   | -   |
| Pose-guided           |     |      |     |      |     |      |     |      |
| PDC (ICCV17) [37]    | 84.14 | 63.41 | -   | -   | -   | -   | -   | -   |
| PABR (Arxiv18) [38]  | 90.20 | 76.00 | 93.20 | 82.70 | -   | -   | -   | -   |
| Pose-transfer (CVPR18) [21] | 87.65 | 68.92 | -   | -   | 33.80 | 30.50 | 30.10 | 28.20 |
| PN-GAN (Arxiv17) [29] | 89.43 | 72.58 | -   | -   | -   | -   | -   | -   |
| PSE (CVPR18) [31]    | 87.70 | 69.00 | -   | -   | 27.30 | 30.20 | -   | -   |
| Attention-based       |     |      |     |      |     |      |     |      |
| DuATM (CVPR18) [34]  | 91.42 | 76.62 | -   | -   | -   | -   | -   | -   |
| HA-CNN (CVPR18) [19] | 91.20 | 70.70 | 93.80 | 82.80 | 44.40 | 41.00 | 41.70 | 38.60 |
| AACN (CVPR18) [43]   | 85.90 | 66.87 | 89.78 | 75.10 | -   | -   | -   | -   |
| Baseline (ResNet-50+tri) | 88.18 | 73.85 | 92.25 | 80.96 | 62.14 | 58.47 | 60.43 | 54.24 |
| DSR (CVPR18) [13]    | 91.26 | 75.62 | 93.45 | 82.44 | -   | -   | 61.78 | 56.87 |
| SFR (ours)            | 93.04 | 81.02 | 94.84 | 85.47 | 67.29 | 61.47 | 63.86 | 58.97 |

Figure 7. Examples of person images (a) CUHK03 (b) Market1501 (c) Duke.

4.2 Person Re-identification

4.2.1 Datasets

Three person re-identification datasets: Market1501 [49], CHUK03 [52] and DukeMTMC-reID [53] are used for evaluate the proposed SFR.

Market1501 has 12,936 training and 19,732 testing images with 1,501 identities in total from 6 cameras. Deformable Part Model (DPM) is used as the person detector. We follow the standard training and evaluation protocols in [49] where 751 identities are used for training and the remaining 750 identities for testing.

CHUK03 consists of 13,164 images of 1,467 subjects captured by two cameras from CHUK campus. Both manually labelled and DFM detected person bounding boxes are provided. We adopt the new training/testing protocol [52] proposed in since it defines a more realistic and challenging ReID task. In particular, 767 identities are used for training and the remaining 700 identities are used for testing.

DukeMTMC-reID is the subset of Duke Dataset [30], which consists of 16,522 training images from 702 identities, 2,228 query images and 1,7,661 gallery images from the other identities. It provide manually labelled person bounding boxes. Here, we follow the setup in [53].

The examples of the three datasets are shown in Fig. 7. And we set $\alpha = 0.7$ in all person re-identification experiments.

4.2.2 Results

Results on Market1501. Comparisons between SFR and 17 state-of-the-art approaches of four categories (part-based model, mask-guided model, pose-guided model and attention-based model) published after 2017 on Market1501 [49] are shown in Table 1. We conduct the single and multiple query experiments, respectively [49]. The results suggest that the proposed SFR achieves the competitive performance on all evaluation criteria under single and multiple query settings.

It is noted that: (1) The gaps between our results and baseline model (ResNet-50+Triplet) are significant: SFR increases from 88.18% to 93.04% under single query setting, and from 92.25% to 94.84% under multiple query setting, which fully suggests that spatial feature with alignment-free reconstruction is more effective than only using global feature matching. (2) Benefit from batch hard triplet spatial reconstruction (BHTSR) and pyramid pooling, SFR outperforms our previous work DSR [13] by 1.78%, 1.39% at the Rank 1 accuracy under single query setting, respectively. BHTSR can learn more discriminative local feature and the
pyramid pooling avoids the influence of scale variations of the detected person. (3) Our SFR achieves the best performance at the Rank 1 accuracy. Contributed by exact human semantic parsing, SPReID [16] achieves the competitive accuracy. However, SPReID relies on excellent human semantic parsing model in a extreme extension and would fail to address arbitrary-size person patch. (4) Although mask and pose estimation provide external cues to improve the performance of person re-identification compared to other methods without using external cues, the overusing of external cues easily result in unstable of these methods due to partial occlusions and the missing of person component. (5) Performance differences among these existing approaches are mainly come from input size (e.g., 224 × 224).

**Performance differences among these existing approaches.**

Internal cues easily result in unstable of these methods due to overusing of external cues, the overusing of external cues easily result in unstable of these methods due to partial occlusions and the missing of person component. (5) Performance differences among these existing approaches mainly come from input size (e.g., 224 × 224, 256 × 128 and 384 × 192), baseline model (e.g., AlexNet, VGGNet, ResNet, and Inception) and algorithms themselves.

**Results on CUHK03.** We only list the results of those methods that use the new training/testing protocol [52]. Table 6 shows results on CUHK03 when detected person bounding boxes and manually labeled bounding boxes are respectively used for both training and testing. The proposed method SFR get 65.86% and 63.86% accuracies while using manually labeled bounding boxes and detected bounding boxes by DPM, respectively. From the results shown in Table 1, we can find that our proposed method SFR outperforms the previous best method PCB [39] implemented by deep learning with multiple parts by 2.56% at Rank 1 using detected person bounding boxes. It is also noted that: (1) SFR performs much better than mask-guided model: MGCAM [36], pose-guided models: Pose-transfer [21] and PSE [31], and attention-based model: HA-CNN [19].

Clear gaps are shown between our method SFR and these state-of-the-art methods. The Rank 1 performance of SFR is 16.00% higher using either labeled or detected person images than others. The results fully suggest that the advantage of SFR is more pronounced. (2) Training with BHTSR and multi-scale spatial representation with pyramid pooling performs better than DSR trained with single-scale spatial feature and the pairwise loss function. Similar results are also observed using the mAP metric.

**Results on DukeMTMC-reID.** Person Re-ID results on DukeMTMC-reID [53] are given in Table 2. This dataset is challenging because the person bounding box size varies drastically across different camera views, which naturally suits the proposed SFR with multi-scale representation. Except for Spindle Net [47], MSCAN [17], DLPAP [48], AlignedReID [46], MGCAM [36] and PDC [37], other comparison methods have reported the results on DukeMTMC-reID. The results show that SFR is 0.40% and 0.27% higher than the second best ReID model: SPReID [16] at the Rank 1 and mAP metrics respectively. Besides, SFR beats the previous work DSR by 2.40% and 2.51% at the Rank 1 and mAP metrics, respectively, which indicates that multi-scale representation using pyramid pooling can cope with scale variations to some extent.

**Influence of weight $\alpha$.** Similarity measure between two images is achieved by combining global feature matching and spatial feature reconstruction. We set the value of $\alpha$ by from 0 to 1 at the stride of 0.1. Similarity distance only contains spatial feature reconstruction when $\alpha = 0$, and similarity distance only contains spatial feature reconstruction when $\alpha = 1$. Spatial feature reconstruction performs much better than global feature matching by 3.95%, 1.82%, 1.07%, 0.93% and 3.95% on Market1501 under single query and multiple query setting, CHUK03 using labeled and detected person images, and DukeMTMC-reID, respectively. It shows that spatial feature reconstruction is more effective.
by discovering detail information of the persons. It is worth to note that fusion of global feature matching and spatial feature reconstruction performs better than single distance measure, which suggests that global feature matching incorporated with spatial feature reconstruction is able to improve the performance of ReID. From the results in Fig. 8, SFR achieves the best performance when we set $\alpha = 0.7-0.9$, indicating that spatial feature reconstruction is of more importance than global feature matching.

### 4.3 Partial Person Re-identification

#### 4.3.1 Datasets

**Partial REID** 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, backgrounds and different types of severe occlusions. The examples of partial persons in the Partial REID dataset are shown in Fig. 9(a). The region in the red bounding box is the partial person image. The testing protocol can be found in the open code.

**Partial-iLIDS** is a simulated partial person dataset based on iLIDS [50]. The iLIDS contains a total of 238 images of 119 people captured by multiple non-overlapping cameras. Some images in the dataset contain people occluded by other individuals or luggage. Fig. 9(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 to construct a gallery set.

#### 4.3.2 Results

The designed Fully Convolutional Network (FCN) is trained with Market1501. We follow the standard training protocols in [49], where 751 identities are used for training the FCN model. For comparison, multi-task sparse representation (MTSR) proposed for partial face modeling, ambiguity-sensitive matching and sliding window matching (AMC-SWM) are considered. Besides, Resizing model is also used for comparison, in which person images in the gallery and probe set are all resized to $256 \times 128$. And then 2,048-dimension feature vector is extracted by FCN followed by global average pooling (GAP).

**Single-Shot Experiments (N=1).** Single-shot experiment means that only one image per person exists in the probe set. Table 3 shows the single-shot experimental results. We find the results on Partial REID and Partial-iLIDS are similar. The proposed method SFR outperforms Resizing model, MTSR, and AMC-SWM. It is noted that: (1) The gaps between SFR and Resizing model are significant: SFR increases from 43.87% to 66.20% and from 26.87% to 63.87% at Rank 1 accuracy on Partial REID and Partial-iLIDS, respectively. SFR takes full advantage of FCN that operate in a sliding-window manner and outputs feature maps without deformation. Such results justifies the fact that the person image deformation would significantly affect the recognition performance. For example, resizing the upper part of a person image to a fixed-size would cause the entire image to be stretched and deformed. (2) AMC-SWM achieves comparable result because local features in AMC combined with global features in SWM makes it robust to occlusions and view/pose various. However, features of non-automatic learning in AMC-SWM make it not as well as SFR performs. (3) Spatial feature reconstruction combined with global feature matching ($\alpha = 0.7$ in Partial REID and $\alpha = 0.6$ in Partial-iLIDS) performs much better than global
Table 5
Performance comparison under multi-shot setting on Partial REID.

| Rank Score | N=2 | N=3 |
|------------|-----|-----|
|             | 1   | 3   | 1   | 3   |
| Resizing model | 46.67 | 67.33 | 45.33 | 67.67 |
| MTRC [20]       | 33.67 | 49.67 | 39.33 | 57.67 |
| AMC+SWM [51]    | 40.67 | 58.67 | 44.67 | 61.33 |
| Baseline (ResNet-50+tri) | 59.00 | 83.33 | 61.33 | 84.33 |
| DSR (CVPR18) [13] | 69.67 | 88.33 | 78.33 | 88.00 |
| SFR (ours)      | 73.33 | 91.33 | 81.33 | 92.67 |

| Rank Score | N=4 | N=5 |
|------------|-----|-----|
|             | 1   | 3   | 1   | 3   |
| Resizing model | 46.00 | 68.67 | 46.33 | 68.67 |
| MTRC [20]       | 42.33 | 61.33 | 47.67 | 63.67 |
| AMC+SWM [51]    | 47.67 | 66.33 | 50.33 | 70.67 |
| Baseline (ResNet-50+tri) | 60.00 | 84.00 | 61.33 | 84.67 |
| DSR (CVPR18) [13] | 79.67 | 91.33 | 81.00 | 90.67 |
| SFR (ours)      | 82.67 | 96.00 | 86.33 | 91.33 |

Figure 10. Rank-1 curve as a function of the weight $\alpha$ on Partial REID and Partial-iLIDS.

We conducted another interesting experiment, where we exchange gallery set and probe set. So the gallery set and probe set contain partial person images and holistic person images, respectively. Table 4 shows the experimental result under single-shot settings. Experimental results show that the proposed SFR also performs much better than Resizing model, MTSR, and AMC-SWM and it is also effective when the gallery set only contains partial person images. Furthermore, compared to the results in Table 5, partial person images exist in the gallery set would influence the performance to some extent.

Multi-shot experiments (N>1). Multi-shot means that multiple person images per subject exist in the gallery set. The results are shown in Table 5. Similar results are obtained in the single-shot experiment, all approaches achieve significant improvement compared to the single-shot experiment. Specifically, the results show that multi-shot setup helps to improve the performance of SFR since it can increase from 66.20% to 73.33%, 81.33%, 82.67% and 86.33% on Partial REID dataset at Rank 1 accuracy, respectively.

Influence of weight $\alpha$. Similarity measure between two images are achieved by combining global feature matching and spatial feature reconstruction. We set the value of $\alpha$ by from 0 to 1 at the stride of 0.1. Similarity distance only contains global feature matching distance when $\alpha = 0$, and similarity distance only contains spatial feature reconstruction when $\alpha = 1$. The results are shown in Fig. 10, we can find that SFR achieves the best rank-1 accuracy under single-shot setting on Partial REID (66.20%) and Partial-iLIDS (63.87%) when $\alpha = 0.7$ and $\alpha = 0.6$, respectively. For multi-shot experiments, we find that SFR performs much better than global feature matching, which can improve more than 10.00% at the Rank 1 accuracy. It shows that spatial feature reconstruction is more effective by discovering detail information of the persons.

4.4 Partial Face Re-identification

4.4.1 Dataset

CASIA-NIR-Distance [12] database is a newly proposed partial face database, which contains 4,300 face images from 276 subjects. Half of them contains the entire facial region of the subject. Partial face images are captured by cameras under near-infrared illumination with subject presenting the different arbitrary region of the face. Besides, the variations of presented partial face images in CASIA-NIR-Distance database include imaging at different distances, views, scales, and illumination. Fig. 11(second row) shows some examples of partial faces in the CASIA-NIR-Distance database and the acquisition device.

Partial LFW, another simulated partial face database based on LFW database [15], is used for evaluation. LFW database contains 13,233 images from 7,749 individuals. Face images in LFW have large variations in pose, illumination, and expression, and may be partially occluded by other faces of individuals, sunglasses, etc.

4.4.2 Result

VGGFace [27] model is used as base model. The fully-connected layers are discarded to evolve into a Fully Convolutional Network (FCN). Close-set experiments are conducted on the CASIA-NIR-Distance and Partial LFW datasets, containing images of 276 and 1,000 subjects respectively. One image per subject (N=1) is selected to construct the gallery set and one different image per subject is used.
between a probe patch and each face image in a gallery is 

Although I2C does not require alignment, the similarity the required alignment step limits the practical applications ing first aligns the partial faces and then computes the simi-

descriptor [2] and LBP [1] histogram for partial face match-

I2C). (2) RPSM method based on SIFT [22] descriptor, SURF with keypoint-based algorithms (MKDSRC-GTP , RPSM, and could represent a partial face more robustly in comparison takes full advantages of local and global information, which al-
databases are 96.74% and 46.30%, which clearly shows that the rank-1 matching accuracies achieved on the two datasets suggests that combining global feature matching and place a huge advantage over the existing partial face recognition approaches.

5 Discussion

The experiments on person, partial person and partial face re-identificaton datasets unveil the extensibility of our approach. On each datasets, the proposed approach, SFR, always outperforms other state-of-the-art approaches including part-based model, mask-guided model, pose-guided model and attention-based model. This is anticipated as these methods require either alignment or external cues, which extremely leads these approaches to poor stability due to relying on segmentation or pose estimation. On the contrast, that SFR relies on both global feature and spatial feature masks it alignment-free, more robust to scale various and external cues unnecessary.

Also, SFR embedded model is able to achieve remarkable performance without requiring fixed-size input image, which is demanded in AMC-SWM, MTRC and Resizing model. In the form of dictionary learning, SFR is designed for matching a pair of images of different sizes, which makes the model free to address re-id problems of partial images with arbitrary-sizes.

Nevertheless, the proposed approach also has a drawback. Compared to global feature matching, it costs more computational consumption for SFR while solving reconstruction coefficients. Therefore, we are considering the acceleration of the proposed approach as our future work.

6 Conclusions

In this paper, we have proposed a novel approach called Spatial Feature Reconstruction (SFR) to get rid of the fixed-size input limitation. The proposed spatial feature reconstruction method provides a feasible scheme to reconstruct the probe spatial feature map linearly from a gallery spatial map. Besides, pyramid pooling layer combined with global pooling layer reduces the influence of scale various, which avoids the alignment step in many other approaches. Furthermore, we embedded SFR into batch hard triplet loss function to learn more discriminative features for minimiz-

The proposed SFR is compared against the existing partial face algorithms including MRCNN [12], MKDSRC-GTP [20], RPSM [41], I2C [14], and DFM [11]. MKDSRC-GTP, RPSM and DFM are implemented using the source codes provided by authors. I2C is implemented by ourselves according to the paper [14]. Table 6 and Table 7 show the performance of the proposed SFR algorithm on the CASIA-NIR-Distance and Partial LFW datasets, respectively. The rank-1 matching accuracies achieved on the two databases are 96.74% and 46.30%, which clearly shows that our algorithm performs much better than those traditional algorithms for partial face recognition. The reasons are analyzed as follows: (1) Multi-scale spatial feature in our SFR takes full advantages of local and global information, which could represent a partial face more robustly in comparison with keypoint-based algorithms (MKDSRC-GTP, RPSM, and I2C). (2) RPSM method based on SIFT [22] descriptor, SURF descriptor [2] and LBP [1] histogram for partial face matching first aligns the partial faces and then computes the similarity of the partial face and a gallery face image. However, the required alignment step limits the practical applications of RPSM and the same story happens in MRCNN either. (3) Although I2C does not require alignment, the similarity between a probe patch and each face image in a gallery is computed by the instance-to-class (I2C) distance with the sparse constraint. Similar to [14], [41], MKDSRC-GTP simply uses local features and this leads to poor performance. From these perspectives, the characteristics of alignment-free property and more distinctive and robust descriptions in SFR contribute to the improvement of partial face recognition and place a huge advantage over the existing partial face recognition approaches.

### Table 6

| Method       | Rank Score 1 | Rank Score 3 | Rank Score 5 | Rank Score 10 |
|--------------|--------------|--------------|--------------|---------------|
| MKDSRC-GTP [20] | 83.81        | 85.25        | 86.69        | 89.21         |
| RPSM [41]    | 77.70        | 80.22        | 82.37        | 86.69         |
| I2C [14]     | 71.94        | 75.18        | 78.06        | 83.81         |
| MRCNN [12]   | 85.97        | 88.13        | 89.93        | 93.17         |
| DFM [11]     | 94.96        | 96.40        | 97.84        | 98.55         |
| SFR          | 96.74        | 97.46        | 98.55        | 99.64         |

### Table 7

| Method       | Rank Score 1 | Rank Score 3 | Rank Score 5 | Rank Score 10 |
|--------------|--------------|--------------|--------------|---------------|
| MKDSRC-GTP [20] | 1.10         | 3.70         | 5.60         | 8.40          |
| I2C [14]     | 6.80         | 8.30         | 11.20        | 14.60         |
| MRCNN [12]   | 24.70        | 28.60        | 31.24        | 35.47         |
| DFM [11]     | 27.30        | 34.40        | 39.20        | 47.58         |
| SFR          | 46.30        | 59.30        | 65.50        | 70.90         |
[44] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentional spatial-temporal pooling networks for video-based person re-identification. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[45] 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.

[46] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[47] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1077–1085, 2017.

[48] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[50] 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)*, 2011.

[51] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015.

[52] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv preprint arXiv:1707.00408*, 2017.

[53] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[54] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6776–6785, 2017.