Neurocomputing 337 (2019) 303–314

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-Instance Convolutional Neural Network for multi-shot person re-identification

Xiaokai Liu[a,*], Sheng Bi[a], Xiaorui Ma[b], Jie Wang[a]

[a] Information Science and Technology College, Dalian Maritime University, Dalian 116026, PR China
[b] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, PR China

ARTICLE INFO

Article history:
Received 7 September 2018
Revised 3 December 2018
Accepted 7 January 2019
Available online 1 February 2019

Communicated by dr Yongmin Li

Keywords:
Multi-shot person re-identification
Multi-instance learning
Convolutional Neural Network

ABSTRACT

This paper tackles the challenging problem of multi-shot person re-identification with Convolutional Neural Network (CNN). As no prior information about how importance each instance plays, it is non-trivial to exploit the interaction information shared by the multi-shot images to help identification. Traditional CNN is in single-shot architecture, then how to utilize the interaction information provided by multi-shot images becomes an important problem to solve. Furthermore, as data augmentation methods are not strictly label-preserving, it increases the difficulty to select discriminative instance for CNN training. In this paper, we propose a weakly supervised CNN framework named *Multi-Instance Convolutional Neural Network (MICNN)* to solve the aforementioned problem. We develop two paradigms, i.e., Embedding-Space paradigm and Instance-Space paradigm, which re-formulate the person re-identification problem as a multi-instance verification problem with part-based features extracted by neural network. We respectively devise a specific bag-level loss function which incorporates the characteristics of the multi-instance problem for each paradigm. Experiments show that the proposed IS method outperforms many related state-of-the-art techniques on four benchmark datasets: CUHK03, SYSUm, RAiD and Market-1501.

1. Introduction

Person re-identification, aiming to recognize the same person across a network of cameras with non-overlapping views, is a major task in surveillance applications. This is a fairly challenging issue, since the appearance of the same pedestrian would undergo drastic changes due to camera settings, geometric deformations and photometric variations in different views. Recently, Convolutional Neural Network (CNN) has been successfully applied to person re-identification [1–3]. The success of CNN is the result of its excellent capability to discover intricate structures in high-dimensional data with little manual engineering.

In the literature, most previous works focus on single-shot methods that attempt to retrieve a person by analyzing a single probe image. However, in real-world monitoring systems, it is more general to obtain multiple images for one person even with simple pedestrian trackers. Although to some extent, settings in multi-shot problem is relevant to that in video-based re-identification, with the only difference in frame continuity and the number of instances, multi-shot re-identification is still of great importance in practice, especially when available instances are limited due to hostile environment, heavy occlusion or intentional hiding. In addition, multi-shot setting offers an opportunity to build a more robust model with more informative data, as well as make more accurate measurements by exploiting mutual relations implied in multiple query images. Thus multi-shot methods are more practical and informative to obtain better identification performance for person re-identification.

Although previous works on CNN-based re-identification could be extended to multi-shot problems by randomly sampling one image for each identity and then averaging the scores over all the images with the same identity, as shown in Fig. 1(a), they essentially are single-shot methods, because all the images with the same identity are used separately in model training and image ranking process, and the key information of 'same identity' is actually unused. Such a practice only focuses on separate images while ignoring the interaction information sharing by the multi-shot data with the same identity.

It is non-trivial to exploit the interaction information shared by the multi-shot images in CNN framework, because no prior information about how importance each instance plays is given. Within multiple instances, some are discriminative, while some others are non-informative. We are not able to distinguish them, because we only have bag labels for training, whereas instance labels are not available. Non-informative instances mainly come from the

* Corresponding author.
E-mail address: xkliu@dlmu.edu.cn (X. Liu).

https://doi.org/10.1016/j.neucom.2019.01.076
0925-2312/

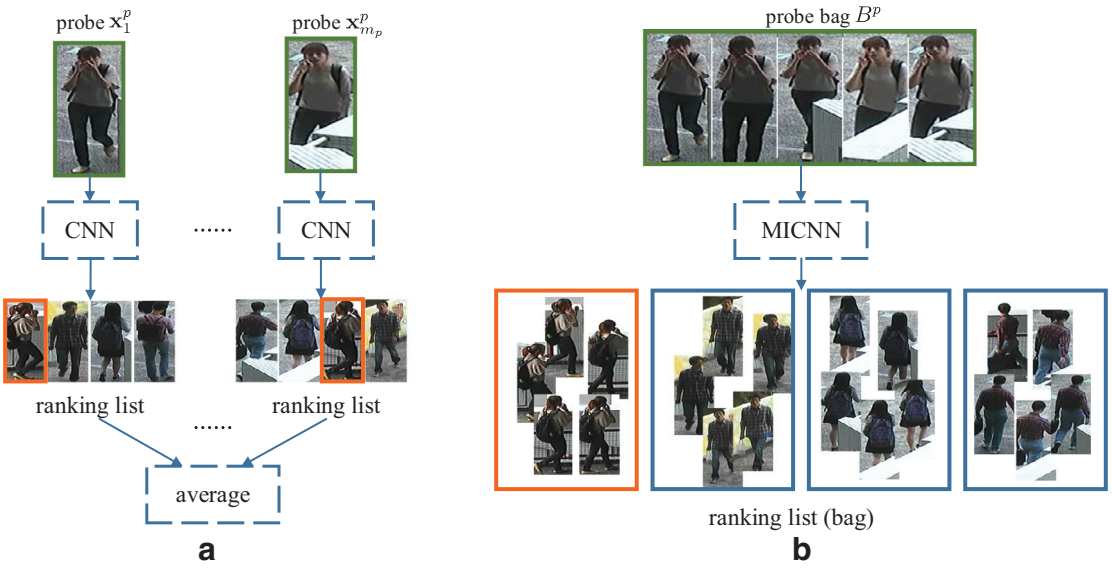


Fig. 1. Illustration of the traditional CNN framework (a) and the proposed multi-instance CNN framework (b). Probe images and bags are marked in green boxes and the real matched images and boxes are marked in orange boxes. Best viewed in color.

following aspects: first, when a person undergoes heavy occlusion, the occluders tend to be regarded as interested targets. Such impostor images would distort the embedding where the pedestrian data are distributed in, thus harm the CNN training. As shown in Fig. 1, within all the 5 probe images in bag B^p, the first two images are complete, while the following three are occluded. If we choose one of the last three in the traditional CNN framework, as in Fig. 1(a), the matching performance would be significantly affected. Second, in person re-identification, data augmentation schemes are widely applied to CNN training [1–3] to reduce the data unbalance and model over-fitting problem. However, the assumption that the data augmentation methods are label-preserving may be violated by indiscriminate transitions, especially when the images are badly detected by human detector or cropped by hand.

The aforementioned properties actually resemble the settings in multi-instance learning, which receives a set of labeled bags instead of individually labeled instances and the instance labels remain unknown. Therefore, we attempt to incorporate multi-instance learning into a deep learning framework and apply it to assist the task of person re-identification. In this paper, we propose a Multi-Instance Convolutional Neural Network framework under two paradigms: the Embedding-Space (ES) paradigm and Instance-Space (IS) paradigm, as shown in Fig. 2. The difference between two paradigms lies in the way the bag-level information is integrated with the CNN strategy. In the IS paradigm, the discriminative information is supposed to lie in the instance-level. Therefore, the network estimates instance probabilities before the last layer and estimates bag probability using a convex loss function. While in ES paradigm, the discriminative information is considered to lie at the bag-level. It explicitly maps multiple outputs from the bag into a fixed-length vector and directly carry out bag-level classification. The contributions of this study can be summarized in two-folds:

(1) To our best knowledge, this is the first study that attempts to formulate person re-identification in a multi-instance learning problem and solve it in deep learning framework. Two paradigms IS and ES are proposed to effectively perform multi-instance learning in an end-to-end way, which takes pairs of bags with various numbers of instances pairs as input and directly output bag labels.

(2) The proposed MICNN-IS method relaxes the requirements for strictly accurate annotations. Instead, it automatically discovers discriminative and none-informative instances in a bag through multi-instance learning, to make up for biases in data augmentation and cater for high appearance variance and occlusions. Accordingly, the proposed MICNN-IS are highly scalable.

Extensive experiments have been conducted to demonstrate that incorporating Multiple Instance Learning (MIL) into deep learning algorithm would fully utilize the potential discriminant information of the training set and achieve better performance on the person re-identification task.

2. Related work

2.1. Multi-shot person re-identification

Compared with single-shot re-identification, multi-shot case is more informative, because more complementary information can be extracted for a more robust identity signature. However, constrained by the difficulties of identity ambiguity and lack of appropriate training datasets, multi-shot based methods are still relatively sparse. Bąk et al. [4] formulated multiple appearances in a covariance metric space and selected the most descriptive features for a specific pedestrian. Li et al. [5] formulated multi-shot re-identification as a set based metric learning problem, and proposed to integrate collaboration and learning strategies to enhance the discrimination capability. Lin et al. [6] introduced the multi-instance multi-label learning methods into re-identification, and solved the identification problem using Boost and SVM algorithms. Guo et al. [7] trained an offline ambiguity classifier to recognize and remove ambiguous samples from multi-shot images. Khan and Emond [8] proposed a multiple-appearance model, each of which described the appearance as a probability distribution of a low-level feature. Specifically, video-based setting is a special case of multi-shot tasks except that videos are temporally continuous and could implicitly offer gait characters to increase the discriminative power of the extracted features. Liu et al. [9] took the video of a walking person as input and built a spatio-temporal appearance representation for pedestrian re-identification. McLaughlin et al. [10] introduced a novel recurrent neural network architecture to

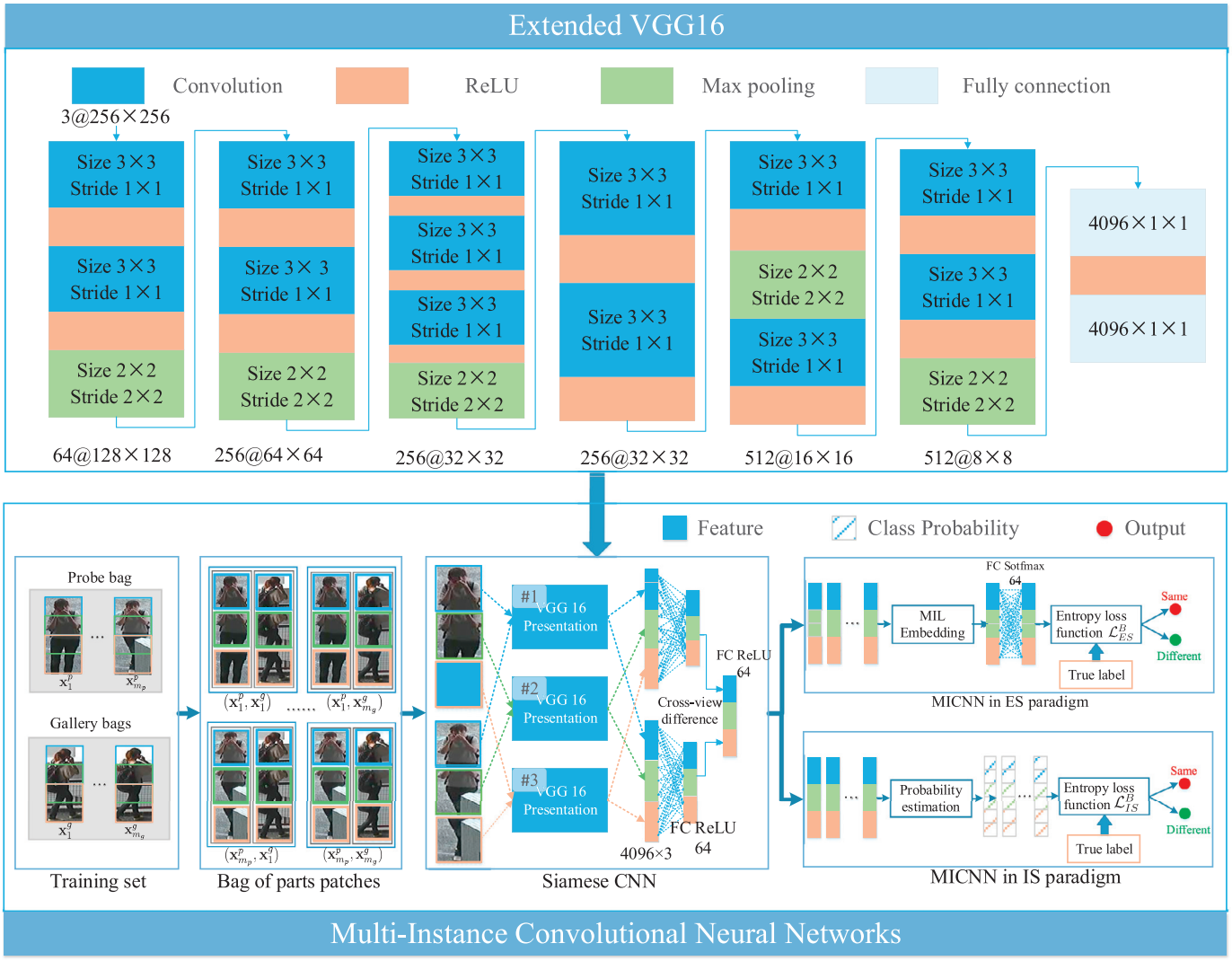


Fig. 2. Illustration of the proposed deep models for multi-instance learning. Output of each unit is listed in the bottom of 'Extended VGG16' part. Best viewed in color.

account for information flow between consecutive frames. Hadjkacem et al. [11,12] introduced a multi-scale video covariance descriptor to tackle the occlusion and uncontrolled changes in video sequences. In this paper, we concentrate on the traditional multi-shot setting, which assumes temporally discontinuous frames. This task is more challenging due to the lack of dependencies between frames.

2.2. CNN for person re-identification

In recent years, CNN achieves excellent performance on image recognition. CNN is typically composed of a stack of convolutional, non-linear, pooling and fully-connected layers, followed by a loss function to specify how training penalizes the deviation between the predicted and true labels, thus update weights when the network is back propagating. It is intended to take advantages of local connections, shared weights, pooling and the use of deep layers to learn high-level representations of natural images. Li et al. [1] built a large scale re-identification dataset CUHK03 to make it feasible to train a deep neural network and make a first attempt to apply CNN to person re-identification. They train the neural network with carefully designed training strategies including dropout, data augmentation, data balancing, and bootstrapping. Yi et al. [13] proposed to jointly learn deep feature and similarity metric in a unified 'Siamese' deep neural network. Ahmed et al. [2] introduced an improved deep learning architecture (IDLA) with two well designed novel layers: a cross-input neighborhood differences layer and a difference summary layer. Shi et al. [3] adopted a moderate positive mining method to account for the irregular distribution in feature space. Although these methods could be applied on multi-shot problem by randomly sampling one image for each identity and then averaging the scores over all the images with the same identity, essentially they all concentrated on single-shot problem, because all the images with the same identity are used separately in model training and image ranking process, the key information of 'same identity' is actually unused. In real-world monitoring systems, such setting amounts to assuming that only one instance in one view could be obtained, without considering the mutual relations between multiple instances with the same identity.

2.3. Multi-instance deep Neural Network

Over ten years ago, Neural Networks, as a supervised learning method, and multi-instance learning, as a typical weakly supervised learning method, has been integrated to help solve the drug activity prediction issue [14,15]. Ramon and Raedt [14] adopted log-sum-exp strategy to calculate bag probabilities from instance probabilities. Zhou and Zhang devised a specific error function

to incorporate the characteristics of multi-instance problem in [15] and improve multi-instance neural networks by feature selection using Diverse Density and PCA in [16]. Recent years, with the revival of deep neural networks, the problem of solving MIL using neural network has been revisited and reinvestigated by several works [17–20]. Wu et al. [17] introduced deep MIL which uses max pooling to find positive instances/patches for image classification and annotation. Yan et al. [18] introduced a multi-stage deep learning framework for image classification and apply it on body-part recognition. The proposed framework learns an image-level classifier to discover local regions for image recognition. Amores [21] focused on bag-level presentation learning and propose a MIL Pooling Layer (MPL) to map multiple outputs from the bag into a fixed-length vector and directly carry out bag-level classification. Sun et al. [20] presented a mathematical formulation for how to incorporate multi-instance learning to deep learning architecture and successfully apply it to object recognition.

According to how the information in the multiple instance data is injected to the deep learning framework, the MIL deep learning methods can be categorized into two categories: instance-space (IS) paradigm and embedding-space (ES) paradigm [21]. In IS paradigm, the discriminative information is considered to lie in the instance-level. Therefore, the network estimates instance probabilities before the last layer and calculates bag probability using a convex loss function. The IS paradigm could automatically extract the most discriminative and none-informative instances in a bag and gives an overall evaluation. This paradigm has been exploited in most previous work: [14,15,17,18,20,21]. While in ES paradigm, the discriminative information is considered to lie at the bag-level. It directly maps multiple outputs from the bag into a fixed-length vector, which summarizes the relevant information within the whole bag and directly carries out bag-level classification. This paradigm has been studied in [19,21]. All these works involve single deep architecture and focus most on recognition problems, while how to solve the MIL problem in a Siamese network [13] and how to exploit both paradigms to help person re-identification problem has not been exploited.

3. Multi-instance CNN for person re-identification

In this section, we will firstly introduce the notations and the problem statement of MIL, then describe the Siamese CNN architecture we use for extracting features, and lastly unify CNN within the multi-instance learning framework, respectively, in two different paradigms: IS and ES.

3.1. Notations and problem statement

In single-shot and fully supervised setting, instance $\mathbf{x}_i$ is manually associated with their true identity label coded by a binary vector $y_i \in \{0, 1\}^C$, where C is the number of the person. In a Siamese network, instance pair $(\mathbf{x}_i, \mathbf{x}_j)$ get the label $t_{ij} = y_i^\top y_j \in \{0, 1\}$, where t_{ij} denotes the equality of labels y_i and y_j, and t_{ij} could equal 1 if and only if $y_i = y_j$. Let the outputs of the Siamese CNN be $h_{ij} = F(\mathbf{W}, \mathbf{x}_i, \mathbf{x}_j) = f(z_{ij}^l)$, where $\mathbf{W}$ is the parameter of the network, l is the number of the layers, and z_{ij}^l is the total weighted sum of inputs in layer l.

Inspired by the loss function in conventional softmax regression [22], which maximizes the concave log-likelihood $\mathcal{L}$ of a logistic discriminant model, the loss function of the CNN could be written as

$$\mathcal{L} = -\sum_{i,j} t_{ij} \log p(t_{ij} = 1|\mathbf{x}_i, \mathbf{x}_j) + (1 - t_{ij}) \log p(t_{ij} = 0|\mathbf{x}_i, \mathbf{x}_j) \quad (1)$$

For multi-instance learning, instances come in 'bag', and the labels are given in bag level. We modify the objective to make it adapted to multi-instance deep model framework. Given a bag of probe images $B_p = \{\mathbf{x}_1^p, \mathbf{x}_2^p, \ldots, \mathbf{x}_{m_p}^p\}$ captured from a camera view and a bag of gallery images $B_g = \{\mathbf{x}_1^g, \mathbf{x}_2^g, \ldots, \mathbf{x}_{m_g}^g\}$ captured from another camera view, where m_p and m_g respectively indicate the number of the instances from a probe/gallery bag and all instances from the same bag ideally share the same identity label, we aim to identify whether the images from the two bags are captured from the same person. Let t_{pg}^B denotes the matching status of the bag pair, which is binary with the form

$$t_{pg}^B = \begin{cases} 1 = \text{if the bags } B_p \text{ and } B_g \text{ are annotated with the} \\ \qquad \text{same identity label} \\ 0 = \text{otherwise} \end{cases} \quad (2)$$

The set of the instance pairs respectively from the two bags is composed by the Cartesian product of the two bags, and denoted as $\Psi_{pg} = B_p \times B_g = \{(\mathbf{x}_i^p, \mathbf{x}_j^g) | \mathbf{x}_i^p \in B_p \wedge \mathbf{x}_j^g \in B_g\}$, to include all potential matching pairs in the training set. Therefore, the loss function with reference to the pair of bags could be written as

$$\mathcal{L} = -\sum_{p,g} t_{pg}^B \log p(t_{pg}^B = 1|B_p, B_g) + (1 - t_{pg}^B) \log p(t_{pg}^B = 0|B_p, B_g) \quad (3)$$

This objective is different from Eq. (1) by adopting a multi-instance learning criterion. Here, each original training identity is treated as a bag consisting multiple instances$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$. It makes bags that share a label closer, and pushes bags that do not share any label apart. For a negative pair of bags, all the pairs of instances that can be made from these two bags are pushed apart since the pair of instances with minimum distance is maximized to optimize the objective function.

3.2. Siamese CNN

In this paper, a 'Siamese' [13] network structure is applied to meet the requirement of pairwise comparison setting and discover the relationships between the instance pairs from two camera views, as shown in Fig. 2. The network has a symmetry structure with two sub-networks which are connected by a difference measuring layer and parameters are shared between two sub-networks, to ensure that both views apply the same filters to extract features. We use a 16-layer 'VGG' model pre-trained on ImageNet [23]. The VGG16 network is selected following Jimenez et al. [24], where it is proven to outperform ResNet50 in image retrieval task. Furthermore, motivated by Zeng and Ji [25], 3 convolutional layers and 2 fully connected layers are extended to the VGG16 model, in order to capture the complex relationships between instances and thus form global representations in bag level. Specifically, all the input images are up-sampled to 512×256 by Bicubic interpolation. Then they are equally split in vertical direction into three 256×256 patches with half overlapping, in order to capture the distinctive features on different body structures. Each patch is charged by a separate extended VGG16 branch. Each branch is constituted of 13 convolutional layers and 5 max pooling layers, and each convolutional layer is followed by a Rectified Linear Unit (ReLU). Note that in order to encourage different emphasis in feature space for different parts, parameter sharing is not performed between branches. In each branch, first two convolutional layer filter the 3-dimensional 256×256 color image with 64 kernels of size $3 \times 3 \times n$ with a stride of 1 pixels, where n is the dimension of the feature maps from previous layer. The resulting feature maps are passed through a max-pooling kernel that halves the width and height of features. The rest unit follow the similar structure, and the detailed layer information is illustrated in the 'Extended VGG16' part of Fig. 2. The outputs of the 3 branches are flattened and concatenated into a 4096 $\times$ 3 vector. Then the vectors are feed into a fully-connected layer, which has 64 neurons

(a) Sample images

	$\mathbf{x}_1^g$	$\mathbf{x}_2^g$	$\mathbf{x}_3^g$	$\mathbf{x}_4^g$	$\mathbf{x}_5^g$
$\mathbf{x}_1^p$	0.19	0.47	0.06	0.73	0.17
$\mathbf{x}_2^p$	0.19	0.32	0.12	0.53	0.11
$\mathbf{x}_3^p$	0.12	0.37	0.11	0.78	0.18
$\mathbf{x}_4^p$	0.54	0.68	0.36	0.94	0.67
$\mathbf{x}_5^p$	0.03	0.06	0.00	0.31	0.02

(b) Similarity matrix of pairs of instances

Fig. 3. Illustration of 'majority principle' in Section 3.4. (a) Sample images from the CUHK03-detected dataset; (b) similarity scores of pairs of instances measured by sotfmax classifier.

and is followed by ReLU activation. After that, the outputs from the two symmetry sub-networks are subtracted from one another and passed through a fully-connected layer with ReLU activation.

3.3. MICNN with ES paradigm

In ES paradigm, the discriminative information is considered to lie at the bag-level. In this paradigm, each bag is treated as a whole entity, and the learning process discriminates between entire bags. As a result, it obtains a discriminative bag-level classifier which makes use of information from the whole bag in order to take a discriminative decision about the class. In this paper, we add a pooling layer, which maps the outputs from multiple instance pairs to a single value, following the last layer of the Siamese CNN architecture. Therefore, the original bag space is mapped to a vectorial embedded space, where the discriminative classifier is learned. The pooling action effectively transforms the original MI problem into a standard supervised learning problem, where each mapped vector has an associated label. This framework is illustrated in Fig. 2 marked with 'MICNN in ES paradigm'.

As for the pooling methods, we employ two statistical measures of central tendency as pooling operations: one is mean pooling [26], which is proven to be effective in medical image analysis; besides, in order to avoid the pitfalls of 'pseudo-similar' phenomenon shown in Fig. 3, we proposed a new trimmed mean (tr-mean) pooling method. Tr-mean involves the calculation of the mean after discarding given parts of samples at the high and low end. Both pooling methods are denoted as:

$$\begin{cases} mean: & h_{pg}^B = \frac{1}{n}\prod_{(\mathbf{x}_i,\mathbf{x}_j)\in B_p\times B_g} h_{ij} \\ tr\text{-}mean: & h_{pg}^B = \dfrac{h_{(a+1)}+h_{(a+2)},\ldots,h_{(n-a)}}{n-2a} \end{cases} \tag{4}$$

where $n=[B_p\times B_g]$ is the cardinality of the potential matching set, $h_{(1)},h_{(2)},\ldots,h_{(n)}$ are the order statistics of the network outputs $\{h_{ij}\}_{(\mathbf{x}_i,\mathbf{x}_j)\in B_p\times B_g}$ and a is a hand-tuned parameter controlling the amount of the values $h_{(\cdot)}$ to be discarded on both ends. We choose to use pooling methods which represent central tendency instead of the popular max-pooling operation, because that the bag pairs with negative labels are susceptible to impostors induced by heavy occlusions, crop errors and improper augmentation operations. Max pooling may mistakenly choose the isolated impostor, without considering the overall tendency.

Through MIL pooling, the bag-level probabilities h_{pg}^B are obtained. As the last layer of the Siamese CNN is activated by a softmax function, the pooled scores could directly represent the probability that a pair of bags are matched or not. Accordingly, we set $p(t_{pg}^B=1|B_p,B_g)=h_{pg}^B$, $p(t_{pg}^B=0|B_p,B_g)=1-h_{pg}^B$, and the loss function with reference to the pairs of bags as Eq. (3) can be rewritten as:

$$\begin{aligned} \mathcal{L}_{ES}^B &= -\sum_{p,g} t_{pg}^B \log p(t_{pg}^B=1|B_p,B_g) \\ &\quad + (1-t_{pg}^B)\log p(t_{pg}^B=0|B_p,B_g) \\ &= -\sum_{p,g} t_{pg}^B \log h_{pg}^B + (1-t_{pg}^B)\log(1-h_{pg}^B) \end{aligned} \tag{5}$$

The gradient of the loss $\mathcal{L}_{ES}^B$ with respect to h_{pg}^B can be derived as:

$$\frac{\partial \mathcal{L}_{ES}^B}{\partial h_{pg}^B} = -\frac{t_{pg}^B}{h_{pg}^B} + \frac{1-t_{pg}^B}{1-h_{pg}^B} \tag{6}$$

With Eqs. (4)–(6), MICNN-ES algorithm for person re-identification can be summarized in Algorithm 1.

Algorithm 1 MICNN-ES training algorithm for person re-identification.

Input: Pairs of bags $\{(B_p,B_g)\}_{pg}$, where $p=1,\ldots,n_p$, $g=1,\ldots,n_g$ and $B_p=\{\mathbf{x}_1^p,\mathbf{x}_2^p,\ldots,\mathbf{x}_{m_p}^p\}$, $B_g=\{\mathbf{x}_1^g,\mathbf{x}_2^g,\ldots,\mathbf{x}_{m_g}^g\}$ and related labels $\{t_{pg}^B\}_{pg}$.
Output: CNN parameters $\mathbf{W}$.
for n = 1 **to** N_T (number of iterations) **do**
  for p = 1 **to** n_p **do**
    for g = 1 **to** n_g **do**
      1. Augment the data by performing random translation and mirroring.
      2. Get the prediction scores $\{h_{ij}\}_{(\mathbf{x}_i,\mathbf{x}_j)\in\Psi_{pg}}$ of the pairs of instances in the pairs of bags (B_p,B_g) with the Siamese CNN using feedforward propagation.
      3. Get the embedded score $h_{p,g}^B$ of the pairs of bags (B_p,B_g) with the MIL pooling principles in Eq. (4).
      4. Get the loss $\mathcal{L}_{ES}^B$ with Eq. (5) and the gradients $\frac{\partial \mathcal{L}_{ES}^B}{\partial h_{pg}^B}$ with Eq. (6).
      5. Update CNN parameters using stochastic gradient descent algorithm according to chain rule:
        $\mathbf{W}' = \mathbf{W} - \alpha\frac{\partial \mathcal{L}_{ES}^B}{\partial \mathbf{W}} = \mathbf{W} - \alpha\frac{\partial \mathcal{L}_{ES}^B}{\partial \mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{W}}$
    end for
  end for
end for

The motivation of this paradigm can be interpreted from two folds: (1) in training stage, instance features in bottom layer can be aggregated to infer labels of bag pairs and receive better supervision by exploiting the potential of the training set; and (2) in testing stage, we can average multiple probabilities of bag pairs to get a more robust bag label. This network directly learns bag representation and produces better bag identification accuracy. The pooling layer introduced in this section works in harmony with

data augmentation to fully explore the potential discriminative information of training sets. However, although ES paradigm could explicitly ensemble multiple instances in feature level, all instances in a bag are assumed to contribute equally to the bags label. The power to reduce the impression of outliers takes effects by averaging the features of all the instances. In this case, due to the lack of instance-level contribution factors, such brainless ensemble pattern tends to neutralize some isolated discriminative instances, leading to information loss. Furthermore, when the percentage of the outliers passes a certain proportion, the bag label would face the risk of being dominated by the outliers. Therefore, we introduce a new Instance-Space paradigm, where instances within a bag remain independent during the parameter updates, and the information interaction takes effects in the process of maximizing the bag-level loss function.

3.4. MICNN with IS paradigm

In IS paradigm, the discriminative information is considered to lie in the instance-level. Therefore, the network estimates instance probabilities before the last layer and calculates bag probability using a convex loss function. This framework is illustrated in Fig. 2 marked with 'MICNN in IS paradigm'. The likelihood is defined over bags, as true instance labels are supposed to be unknown during training. Therefore we need to express $p(t^B_{pq}=1|B_p, B_g)$ and $p(t^B_{pq}=0|B_p, B_g)$, the probabilities of the pair of bags (B_p, B_g) having the same identity or not, in terms of the instances in the bag. In traditional multi-instance learning settings, a bag is positively labeled if at least one instance in it is positive, and is negatively labeled if all instances in it are negative. However, this setting does not work in the multi-shot re-identification, because the pairs of bags with negative labels are susceptible to impostors caused by pose variation, illumination changes and occlusions. Only considering isolated pairs without considering the relevant information in rest instances may lead to wrongly choosing the impostors. As shown in Fig. 3, the probe bag and the gallery bag are from different pedestrians. Under the comprehensive action of illumination variation and posture coincidence, the similarity of $\mathbf{x}^p_4$ and $\mathbf{x}^g_4$ measured by softmax classifier reaches up to 0.94, which is a 'pseudo-similar' phenomenon. The 'one vote support' mode in traditional multi-instance learning tends to decide as 'positive' in this situation, without considering the rest similarity scores by other instance pairs. Therefore in this paper, we reconsider the problem with a 'majority principle', in which all matching probability of the pairs of instance from two bags are jointly measured to balance the contributions of the majority to the final bag label decision. If given the probabilities of the pairs of instances $p(t_{ij}=c|\mathbf{x}_i, \mathbf{x}_j)$, $c\in\{0, 1\}$, the probability that the pairs of bags are matched or not is defined as:

$$p(t^B_{pg}=c|B_p, B_g)=\prod_{(i,j)\in\Psi_{pg}} p(t_{ij}=c|x_i, x_j) \tag{7}$$

Different from the typical MIL assumption that the most discriminative instance pair dominates the classification probability, this bag-level probability states that all instances in a bag contribute to the bag label. And the probabilities of the pairs of instances (x_i, x_j) is respectively defined as:

$$\begin{cases} p(t_{ij}=0|x_i, x_j)=\exp(-\lambda h_{ij}) \\ p(t_{ij}=1|x_i, x_j)=\exp(-\lambda(1-h_{ij})) \end{cases} \tag{8}$$

where λ is a constant positive value controlling the affinity between h_{ij} and $p(t_{ij}|x_i, x_j)$. The probability that the instance pair is not matched is set to be inversely proportional to the output of the Siamese CNN h_{ij}, as h_{ij} in the neural network represents the probability that the instance pair is matched. The reason why we define the probability in the form of exponential function is to simplify the calculation of the gradient below. Therefore, in IS paradigm, with Eqs. (7) and (8), optimization of Eq. (3) is equivalent to minimize the following function:

$$\begin{aligned} \mathcal{L}^B_{IS} &= -\sum_{p,g} t^B_{pg}\log p(t^B_{pg}=1|B_p, B_g) \\ &\quad +(1-t^B_{pg})\log p(t^B_{pg}=0|B_p, B_g) \\ &= -\sum_{p,g} t^B_{pg}\log\prod_{(i,j)\in\Psi_{pg}} p(t_{ij}=1|\mathbf{x}_i, \mathbf{x}_j) \\ &\quad +(1-t^B_{pg})\log\prod_{(i,j)\in\Psi_{pg}} p(t_{ij}=0|\mathbf{x}_i, \mathbf{x}_j) \\ &= \sum_{p,g} t^B_{pg}\log\prod_{(i,j)\in\Psi_{pg}}\exp(\lambda(1-h_{ij})) \\ &\quad +(1-t^B_{pg})\log\prod_{(i,j)\in\Psi_{pg}}\exp(\lambda h_{ij}) \\ &= \sum_{p,g}\sum_{(i,j)\in\Psi_{pg}}\lambda(t^B_{pg}+h_{ij}-2t^B_{pg}h_{ij}) \end{aligned} \tag{9}$$

And the gradients will be

$$\frac{\partial\mathcal{L}^B_{IS}}{\partial h_{ij}}=\lambda(1-2t^B_{pg}) \tag{10}$$

Given all the equations above, MICNN-IS algorithm for person re-identification can be summarized in Algorithm 2.

Algorithm 2 MICNN-IS training algorithm for person re-identification.

Input: Pairs of bags $\{(B_p, B_g)\}_{pg}$, where $p=1,\ldots,n_p$, $g=1,\ldots,n_g$ and $B_p=\{\mathbf{x}^p_1, \mathbf{x}^p_2,\ldots,\mathbf{x}^p_{m_p}\}$, $B_g=\{\mathbf{x}^g_1, \mathbf{x}^g_2,\ldots,\mathbf{x}^g_{m_g}\}$ and related labels $\{t^B_{pg}\}_{pg}$.
Output: CNN parameters $\mathbf{W}$.
for n = 1 **to** N_T (number of iterations) **do**
  for p = 1 **to** n_p **do**
    for g = 1 **to** n_g **do**
      1. Augment the data by performing random translation and mirroring.
      2. Get the prediction scores $\{h_{ij}\}_{(\mathbf{x}_i,\mathbf{x}_j)\in\Psi_{pg}}$ of the pairs of instances in the pairs of bags (B_p, B_g) with the Siamese CNN using feedforward propagation.
      3. Get the probabilities $p(t_{ij}|\mathbf{x}_i, \mathbf{x}_j)$ of the pairs of instances $(\mathbf{x}_i, \mathbf{x}_j)\in(B_p, B_g)$ with Eq. (8).
      4. Get the loss $\mathcal{L}^B_{IS}$ with Eq. (9) and the gradients $\frac{\partial\mathcal{L}^B_{IS}}{\partial h_{ij}}$ with Eq. (10).
      5. Update CNN parameters using stochastic gradient descent algorithm according to chain rule:
        $\mathbf{W}'=\mathbf{W}-\alpha\frac{\partial\mathcal{L}^B_{IS}}{\partial\mathbf{W}}=\mathbf{W}-\alpha\frac{\partial\mathcal{L}^B_{IS}}{\partial\mathbf{h}}\frac{\partial\mathbf{h}}{\partial\mathbf{W}}$
    end for
  end for
end for

The loss function in Eq. (9) synthesizes the contributions of all the instance pairs. We assume that in a pair of bag with the same identity, the majority pairs of instances are well matched with only minority pairs undesirably or badly matched and vice verse for the bag pairs with different identity. Within a pair of bags, the instance pairs with high probability to be correctly matched are counted more in the loss function and drive the updates of network coefficients during the backward propagation. Accordingly, the learned CNN is expected to have high responses on discriminative instance pairs. To sum up, ES and IS diagrams mainly differs in two aspects: first, ES runs in feature-level fusion mode, where instance-level

features are equally contributed and indiscriminately embedded, while IS performs in decision-level fusion mode, where the contribution of each instance differs and are automatically evaluated in the training process. Secondly, ES maps the original bag space to a vectorial embedding space, and explicitly gets a bag-level feature, accordingly degenerate to a standard supervised learning space. However, IS keeps all the instances within a bag independent, and acts as a weakly supervised learning method.

4. Details for training and testing

In this paper, the person re-identification task is regarded as a binary classification problem. The learning process is conducted on a training subset and a validation subset. Each time we randomly select 15% identities from the original training set as validation set and apply 5-fold cross-validation to determine all hyper-parameters. Training data are constituted by pairs of bags containing images captured from two views, and are labeled as positive (same) or negative (different). The data are augmented by random translation and horizontally mirroring as also did in [2]. The model performs forward propagation on the dataset and computes the output and loss. Backpropagation is then used to compute the gradients, and network weights are updated. Gradient descent algorithm is applied to update the weight of the network. A validation set is used to evaluate intermediate models and select the one that has maximum performance.

Even with the data augmentation, the scale of the bag-level data (equal to the number of the identities) for person re-identification is still not enough to guarantee a well-trained CNN model. Therefore, we pre-train the Siamese CNN on the Market-1501 dataset in a softmax classification framework, and fine turn the parameters on the other datasets in the proposed MICNN framework. We randomly select one image from each bag as input and data augmentation including horizontal flipping and random clipping is also applied. In the following experiments, the re-identification results from the Siamese CNN framework with the Extended VGG16 network are used as the baseline. In test stage, to bridge the gap between single-shot and multi-shot settings and make a more fair comparison, similarity scores evaluated on the same identity are averaged to explicitly integrate multiple view information.

5. Experiments

Extensive experiments are conducted on four multi-shot datasets designed for person re-identification: CUHK03 [1], SYSUm [7], RAiD [27] and Market-1501 [28]. The proposed MICNN is implemented in Python with Keras [29] programming interface, and all of our experiments are running on a PC with Intel(R) i7-4790K CPU (4.00GHZ) and NVIDIA GeForce GTX 1080 8GB GPU. The softmax pre-train on Market-1501 dataset takes roughly 2 h to converge, and with the pre-trained model, training on the other datasets are more efficient and only takes about 1 hours. In comparison, the proposed MICNN algorithms only take approximately 40 min to converge. In this section, we begin with the description of the datasets, experimental protocol and parameters settings. Then we report the evaluation on four datasets (six in total, with two versions in CUHK03 and two camera combinations in RAiD). Finally, we compare the proposed method with the state-of-the-art methods on all the datasets.

5.1. Datasets and settings

(1) Datasets: The evaluations of the proposed methods are carried out on four challenging datasets: CUHK03 [1], SYSUm [7], RAiD [27] and Market-1501 [28]. The CUHK03 dataset contains 14,096 images of 1467 pedestrians,[1] with 4.5 images per identity in each view. Both manually labeled pedestrian bounding boxes and automatically detected ones by a pedestrian detector are provided. We report results on both versions of the data. SYSUm contains 502 individuals in an uncontrolled scenario, for a total of 48,892 samples. RAiD is a 4 camera dataset, with 2 indoor and 2 outdoor. 43 pedestrians are included in the dataset, resulting in 6920 images. All data in this dataset are videos, a special case of multi-shot setting with the only difference of continuous frames. This dataset is designedly selected to demonstrate the effectiveness of the proposed method on videos. The Market-1501 dataset is one of the largest publicly available datasets for human re-identification. The dataset contains 32,668 annotated bounding boxes of 1501 identities. Images of each identity are captured by at most six cameras.

(2) Experimental protocol: For fair comparison, we use the same protocol with previous works in training and test phases. For the CUHK03 dataset, we follow the protocol used in [1,2], and randomly divide 1467 identities into non-overlapping train (1267), test (100), and validation (100) sets. For the Market-1501 dataset, we use the protocol that the authors provided in [28], and use 751 identities in training, and use 750 identities in test. For RAiD, we use two camera pairs 1–3 and 1–4 for evaluation (denoted as RAiD(1–3) and RAiD(1–4)). SYSUm and RAiD are evaluated with the same training protocol as in [30], in which each time half of the pedestrians were selected randomly to form the training set, and the remaining pedestrian images were used to form a test set. The instances in the same bag are randomly selected from the same identity group. The number of the instances in a bag are drawn from a Gaussian distribution, where the mean μ and standard deviation σ is decided by the maximum numbers of images of each identity. In all the experiments, we set $\mu = 5$ and $\sigma = 3$. The reason for this setting is explained in the next section. Each experiment in this paper is repeated 10 times and the average results are reported.

(3) Parameter settings: In the MICNN-IS paradigm, the constant parameter λ, which controls the affinity between the output of CNN and the probability of pairs of instances, is set to 0.2–0.4 by performing 5-fold validation process. In the MICNN-ES paradigm, the parameter a is set to 1 considering that there are limited amount of impostors. In both MICNN training phases, the learning rate α is initially set to 0.02, and decreases with the policy $\alpha^{new} = \alpha * \gamma$, whenever the loss of the validation set stops decreasing through 3 epochs. The decrease factor γ is set to 0.3 and the lower bound of the learning rate is set to 0.0001 in all the experiments.

5.2. Evaluation of the proposed methods

(1) Evaluation of the MICNN methods: To evaluate performance of the proposed MICNN methods, we carry out experiments on all the datasets, and compare the results of the proposed methods with the single-shot 'softmax' baseline. Performances are presented in Fig. 4. Compared with VGG16 baseline, both MICNN paradigms achieve significant improvements. ES-trmean performs constantly better than ES-mean on all the datasets, owing to its ability to reduce the effect of the impostors by discarding values on both ends. IS paradigm performs overall best, and achieves 12.14%, 14.65%, 8.9%, 14.75%, 16.67% and 10.05% improvements over softmax at rank-1 on CUHK03 labeled, CUHK03 detected, Market-1501, SYSUm, RAiD(1–3) and RAiD(1–4). As we can see, imposing multi-shot interaction improves the overall performance. Note that

[1] The statistic data in [1] is slightly different from the data we obtained from http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html. We re-state the data here.

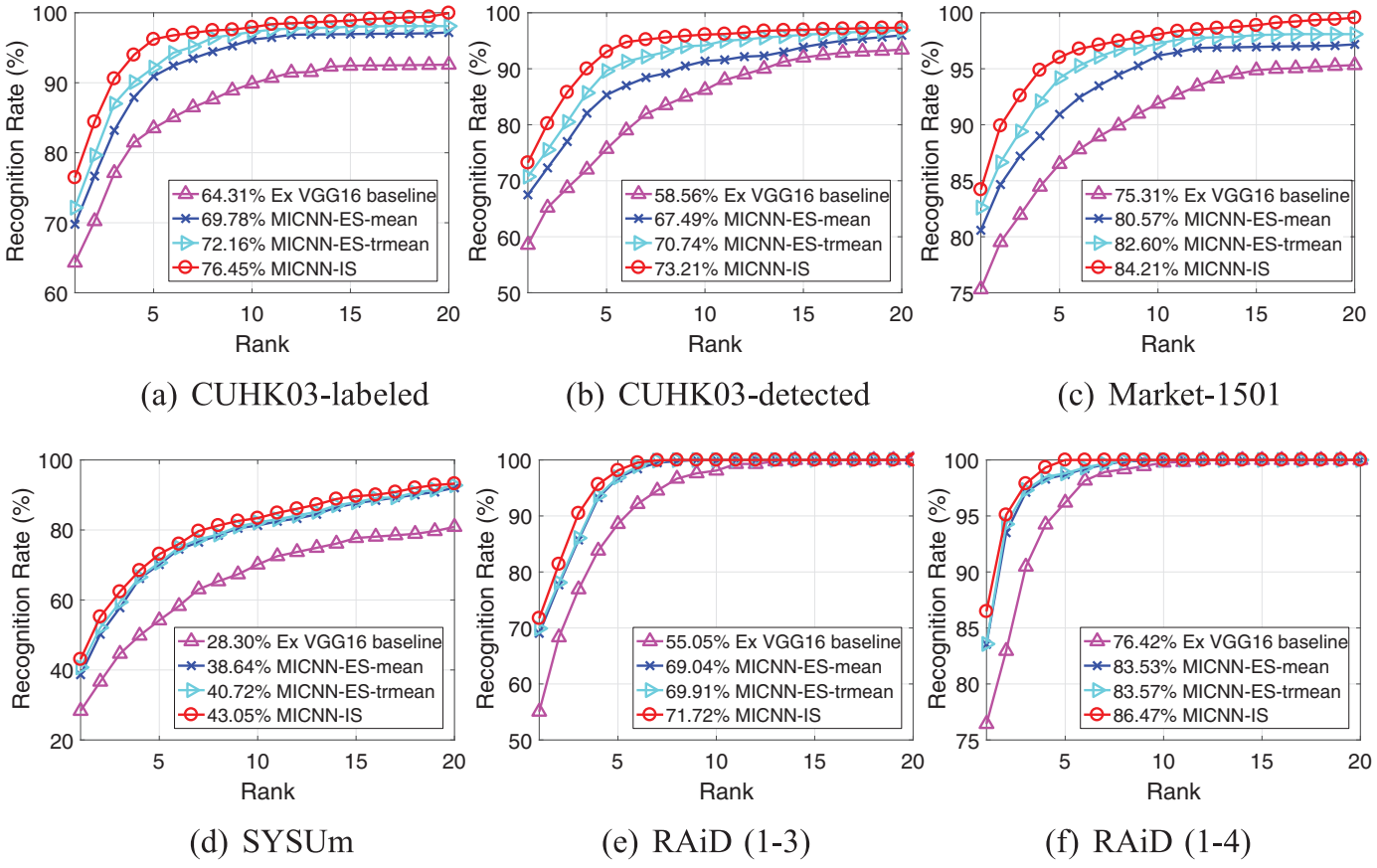


Fig. 4. CMC curves of the proposed paradigms compared with Extended VGG baseline on CUHK03-labeled (a), CUHK03-detected (b), Market-1501 (c), SYSUm (d), RAiD (1–3) (e) and RAiD (1–4) (f) datasets. Rank-1 recognition rate is marked in front of the ranker name. Best viewed in color.

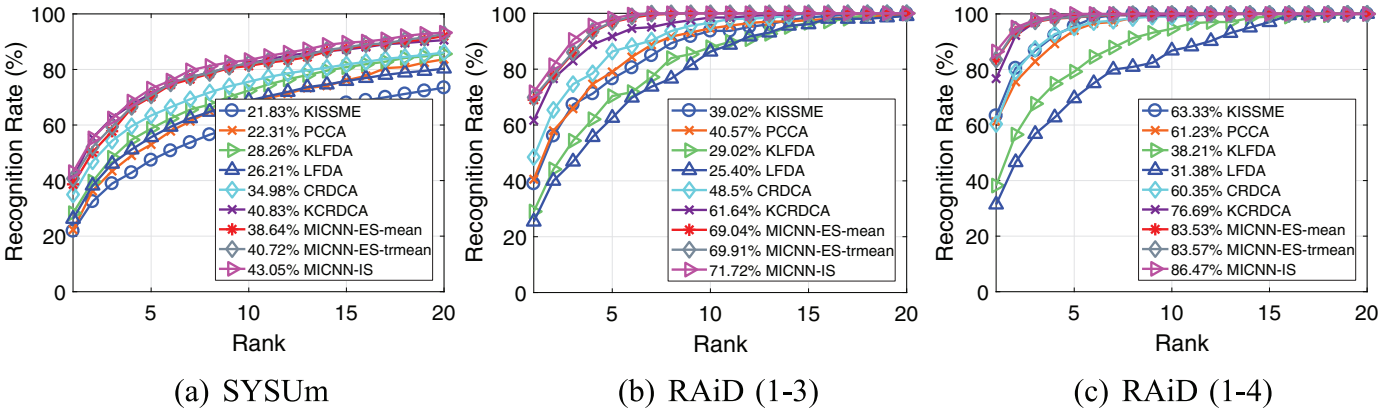


Fig. 5. Comparison to the state-of-the-art on CUHK03-labeled (a), CUHK03-detected (b), Market-1501 (c), SYSU (d), RAiD (1–3), (e) and RAiD (1–4) (f). Rank-1 recognition rate is marked in front of the ranker name. Best viewed in color.

in all the datasets except SYSUm, CMC curves in IS paradigm begin to level off at a high level (over 95%) since rank-10. In SYSUm dataset, the scale of the gallery set (251) and the illumination changes are extremely large, both of which increase the difficulties to recognize a probe pedestrian within the gallery set. Compared with ES paradigm, IS overcomes the disadvantage of indiscriminate ensemble. All the instances manage to interact on each other, and drives the instances pairs with high probability to be correctly matched count more in the loss function during the update of the networks, accordingly achieves superior performance over ES paradigm.

(2) Sensitivity to scale of bags: In multi-shot person re-identification, the required scale of bags (which means the number of the instances with the same identity in one view contained in a bag) is a crucial aspect and approaches with small bag scales are preferred. Because in training stage, ting labeled data from camera pairs is complex and time consuming, and also in realistic monitoring systems, the available instances may be limited due to the hostile environment and intentional hiding. We conduct experiments to discover the relationship between the recognition rate and the scale of the bags. The experiments are conducted using MICNN-IS on three datasets: SYSUm, RAiD(1–3) and RAiD(1–4), all of which have large scales of bags. Corresponding number of instances are randomly selected from a bag, and each experiment is repeated 10 times to get the average recognition rate. Fig. 6 presents the rank-1 recognition rates with bag scales varying from 2 to 10. Unsur-

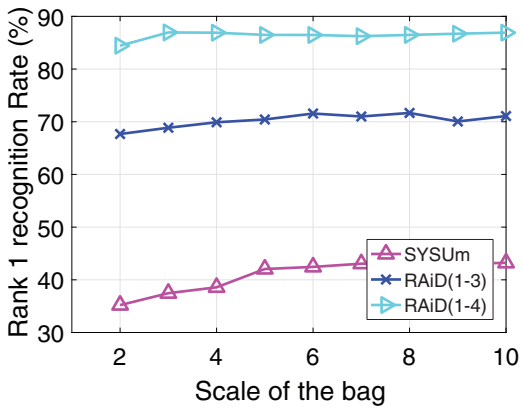


Fig. 6. Rank 1 recognition rate with bag scales varying from 2 to 10 on two large bag-scale datasets: SYSUm and RAiD.

prisingly, the rank-1 recognition rate presents increase trend with slight fluctuations. However, rising trends begin to level off since 2 to 6. For example on SYSUm Dataset, as the bag scales rising from 2 to 6, the increase of the recognition rate is notable and then begins to level off since 6. The rank-1 matching rate achieves 43.05% when bag scales are 6, only 1.17% less than training with bag scales 10 (44.22%). This indicates that to some extent, the proposed algorithm relaxes the restriction to scale of bags.

5.3. Comparison with state-of-the-art results

For CUHK03, we compare our methods with four traditional methods: Keep It Simple and Straightforward (KISSME) [31], Logistic Discriminant Metric (LDM) [32], NULL [34], Supervised Smoothed Method (SSM) [38] and several deep learning methods: Filter Pairing neural network (FPNN) [1], Improved Deep Learning Architecture (IDLA) [2] and Embedding Deep Metric (EDM) [3], Gaussian of Gaussian (GoG) [35], Cross-View Discriminative (CVD) [36], Multi-Task Deep Net (MTDNet) [37], in which SSM gained the previously best performance on CUHK03. The key rank-*r* identification rates are shown in Table 1. The newly proposed deep learning based methods IDLA and EDM outperform the traditional approaches and the originally proposed deep learning method FPNN by a large margin. The proposed MICNN-IS achieves significant improvements over the previously best results in EDM on CUHK03. Overall, the MICNN-IS method achieves 76.45% at rank-1, 96.21% at rank-5, 98.73% at rank-10, and 99.96% at rank-20 on 'labeled' version, and achieves 73.21% at rank-1, 93.04% at rank-5, 96.80% at rank-10, and 97.52% at rank-20 on 'detected' version. Compared with the state-of-the-art methods, the proposed MICNN-IS performs better on the 'detected' version than on the 'labeled' version. This indicates that the proposed method increases robustness towards outliers coming from detection misalignment. Note that although the ES-mean and ES-tr-mean obtain slightly inferior results than SSM, they achieve significant improvements over all other comparison methods.

As for SYSUm, RAiD(1–3), and RAiD(1–4), we compare the proposed methods with four single-shot algorithms: KLFDA [39], PCCA [40], KISSME [31], LFDA [41] and four multi-shot oriented methods: NCR [27], CAM [33], CVDCA [30], KCVDCA [30], in which KCVDCA achieved the previously best performance on all the datasets. The CMC curves are presented in Fig. 5(d)–(f) and the key rank-r identification rates are shown in Table 2. The proposed MICNN-IS method achieves overall better performance than the previous state-of-the-art methods on all the datasets, especially on RAiD(1–3) and RAiD(1–4), where the recognition rates shoot up to 71.7% and 86.5% at rank-1 compared with KCVDCA, respectively raise 10.1% and 9.8%. As both camera pairs 1–3 and 1–4 are indoor–outdoor combinations, illumination variations are extremely large. Note that the proposed method achieves good results on video-based dataset RAiD. The experimental results indicate that the proposed methods are robust to illumination changes, owing to the CNN feature extraction and multiple instance settings.

For Market-1501, as it is one of the most commonly used large scale ReID dataset, we compare our approach with a series of state-of-the-art methods, and list the results in Table 3. Experimental results show that our method achieves superior performance on both rank-1 and mean average precision (mAP). The mAP is an important criterion to measure the performance of the multi-shot approaches. Although given multiple instance in the input, they are all regarded as a whole in the proposed architecture, thus mAP is not available. Therefore, when we calculate mAP on Market-1501 dataset, we use multi-instance setting as aforementioned to train the deep model, and choose one of the equivalent branches to get the feature of test instance, and apply Euclidean distance to obtain the ranking list of the candidate images. It is exactly verified that the improvement of the performance is not just from the average of the incoming instances, but learning a more robust model by exploiting the discriminative and non-informative instances.

We visualize image pairs with their similarity score $p(x_i, x_j)$ predicted by MICNN-IS to explore the motivation of IS paradigm in Fig. 7. All images are selected from the training set of Market-1501 dataset. Instances from the same person with different qualities are

Table 1
Comparison of CMC ranking rate (%) on the CUHK03 dataset (in both labeled and detected versions). Best results for each rank are emphasized in bold font. '–' means that the result is not reported.

Dataset	CUHK03-labeled				CUHK03-detected			
Rank-*r*	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
KISSME [31]	14.17	37.46	52.20	69.38	11.70	33.44	48.07	64.85
LDM [32]	13.51	37.13	52.70	70.13	10.92	31.90	47.01	65.00
FPNN [1]	20.65	50.94	67.01	83.00	19.89	49.41	64.79	81.14
IDLA [2]	54.74	86.50	93.88	98.10	44.96	76.01	83.47	93.15
EDM [3]	61.32	88.90	96.44	99.94	52.09	82.88	91.78	97.17
CAM [33]	–	–	–	–	53.5	–	–	–
NULL [34]	62.55	90.05	94.80	98.10	54.70	84.75	94.80	95.20
GOG [35]	67.3	91.0	96.0	–	65.5	88.4	93.7	–
CVD [36]	69.6	91.0	96.9	98.9	–	–	–	–
MTDNet [37]	74.68	95.99	97.47	–	–	–	–	–
SSM (fusion) [38]	**76.63**	94.59	97.95	–	72.70	92.40	96.05	–
MICNN-ES-mean	69.48	90.92	96.21	97.18	67.49	85.32	91.37	95.47
MICNN-ES-tr-mean	72.16	92.17	97.25	98.14	70.74	89.62	94.18	96.91
MICNN-IS	76.45	**96.21**	**98.73**	**99.96**	**73.21**	**93.04**	**96.80**	**97.52**

Table 2
Comparison of CMC ranking rate (%) on the SYSUm dataset and RAiD dataset (with both (1–3) and (1–4) camera pairs). Multi-instance (MI) and non Multi-instance methods are divided into different groups. Best results for each rank are emphasized in bold font.

Dataset		SYSU-m				RAiD(1–3)				Raid(1–4)			
Rank-r		$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
Non-ML	KLFDA [39]	21.8	47.4	61.1	73.4	39.0	76.6	93.7	100.0	63.3	95.6	100.0	100.0
	PCCA [40]	22.3	53.0	68.1	83.7	40.6	79.0	94.6	99.0	61.2	94.1	99.5	100.0
	KISSME [31]	28.3	58.7	72.4	87.5	29.0	70.4	87.8	98.4	38.2	79.2	94.6	100.0
	LFDA [41]	26.2	55.6	68.8	80.3	25.4	62.6	86.3	99.0	31.4	69.6	86.8	100.0
MI	CVDCA [30]	35.0	63.4	75.6	86.1	48.5	86.3	96.6	100.0	60.4	95.1	99.0	100.0
	NCR [27]	–	–	–	–	67.1	82.3	93.8	98.6	68.2	86.4	98.6	100.0
	CAM [33]	36.8	–	–	–	–	–	–	–	–	–	–	–
	KCVDCA[30]	40.8	71.4	82.2	90.6	61.6	91.7	98.5	100.0	76.7	99.5	99.5	100.0
MICNN-ES-mean		38.6	70.1	81.3	92.0	69.0	96.8	100.0	100.0	83.5	98.7	100.0	100.0
MICNN-ES-tr-mean		40.7	70.7	82.2	92.8	69.9	96.9	100.0	100.0	83.6	98.8	100.0	100.0
MICNN-IS		**43.1**	**73.1**	**83.4**	**93.2**	**71.7**	**98.1**	**100.0**	**100.0**	**86.5**	**100.0**	**100.0**	**100.0**

Table 3
Comparison of CMC ranking rate (%) and mAP (%) on the Market-1501 dataset. Best results for each rank are emphasized in bold font. '–' means that the result is not reported.

Methods	Rank-1	mAP
NULL [34]	71.6	46.0
CCAFA [42]	71.8	45.5
Spindle [43]	76.9	–
LSRO [44]	78.1	56.2
CVD [36]	80.3	59.7
ACRN [45]	83.6	62.6
SVDNet [46]	82.3	62.1
MCTM [47]	83.8	74.3
DaF [48]	82.3	72.4
MICNN-ES-mean	80.5	72.9
MICNN-ES-tr-mean	82.6	73.4
MICNN-IS	**84.2**	**75.6**

shown in the same row. The first image in each row is randomly selected from one bag, and all the others are its paired images and are ranked in descending order by similarity scores. It is easy to find that images with deformity, superposition or blur tend to get lower similarity scores and would contribute less in bag-level identification. We can observe that such images always score below 0.7, and we mark them in green rectangle. Moreover, some impostor images, which may be caused by annotation errors, heavy occlusions, or extreme superposition, can provide little information in bag-level identification. Such images always score below 0.2, and we mark them in red rectangle. Especially some hard images including two or more bodies in the center would cause ambiguity, and the network can hardly discriminate which one is the right target.

Some representative top-10 ranking results on Market-1501 dataset are shown in Fig. 8. As can be seen, most of the false positives are due to the high clothing similarities among pedestrians. Even for a human, it is difficult to verify if it is a true or false positive. The false negatives are mainly caused by partial occlusion (line 2), misalignment (line 3, 4), view point differences (line1, 4) and accessory mismatch (line 2, the false negative takes a backpack in back while the query person does not). This issues are expected to be resolved with view point prediction and accessory recognition, and would be explored in our future work.

Fig. 7. Samples with their predicted similarity probability $p(x_i, x_j)$ from Market-1501 dataset. Samples with low quality are marked in green rectangle, and samples that are discovered to be none-informative (impostors) are marked in red rectangle. Best viewed in color.

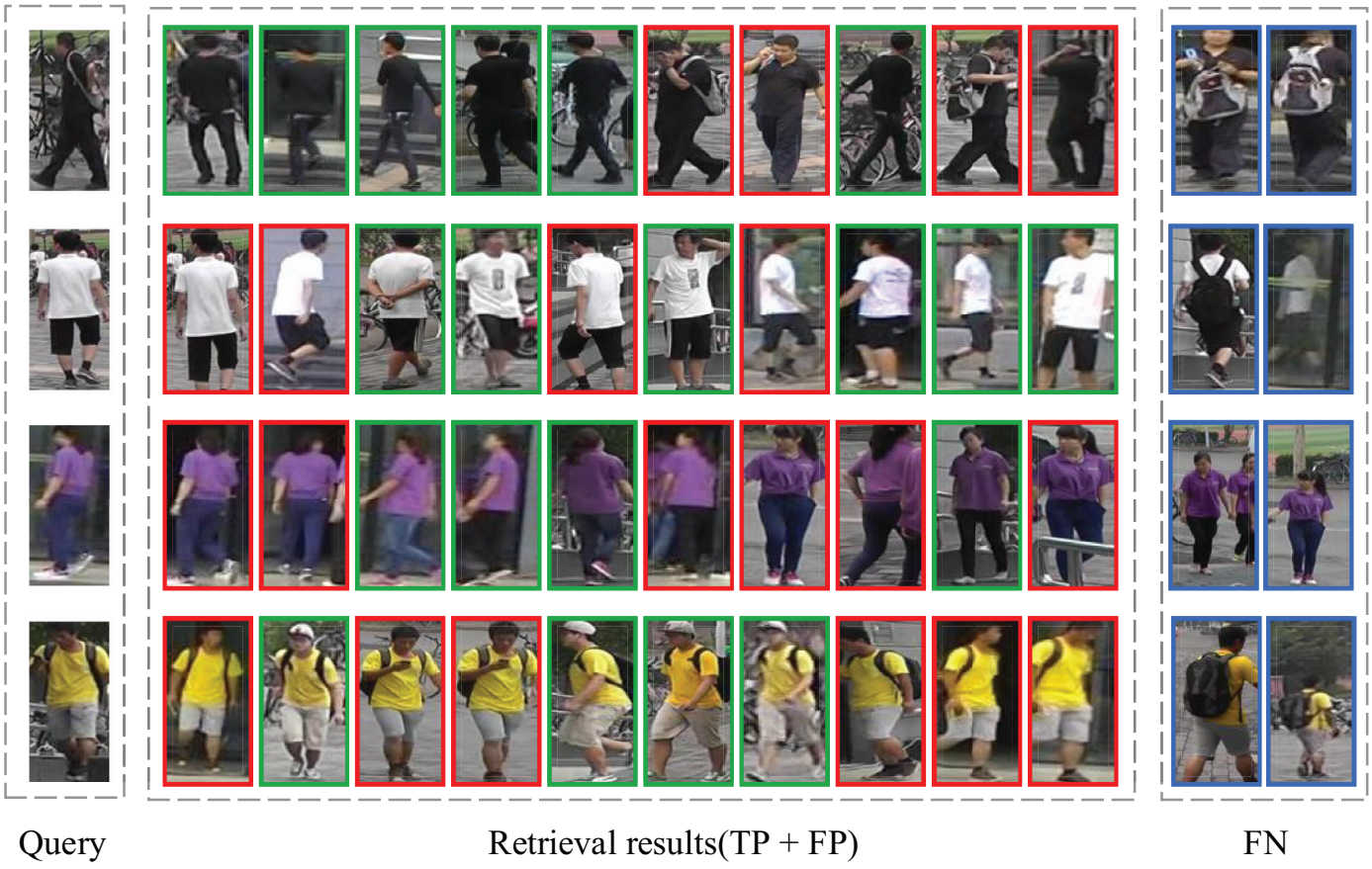


Fig. 8. Qualitative results of the proposed method on Market-1501 dataset, where retrieval number is set to be 10. Query images are on the left side. True positives (TP), false positives (FP), and false negatives (FN) are respectively labeled in red, green and blue boxes.

6. Conclusion

In this paper, the person re-identification task is re-formulated as a multi-instance verification problem, with part-based features extracted by neural network, to meet the requirement of more realistic real-world applications. To the best of our knowledge, it is the first attempt on addressing the person re-identification problem under such a challenging setting. Two types of integration fashions are exploited, resulting in two different types of MICNN paradigms. We respectively devise a specific bag-level loss function which incorporates the characteristics of the multi-instance problems for each paradigm. Regarding the experimental study, the MICNN in IS paradigm has been verified to be a more efficient approach, as well as having the potential to exploit the discriminative and non-informative instance pairs for better verification performance.

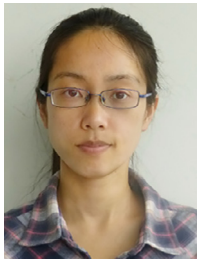
Acknowledgments

X. Liu is supported in part by the National Natural Science Foundation of China under grant No. 61802044, and the Fundamental Research Funds for the Central Universities under grant No. 3132018184. J. Wang is supported in part by Liaoning Province Natural Science Foundation under grant No. 20180520026.

References

[1] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: Deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[2] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: Proceedings of the Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.
[3] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S.Z. Li, Embedding deep metric for person re-identification: a study against large variations, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 732–748.
[4] S. Bąk, G. Charpiat, E. Corvee, Learning to match appearances by correlations in a covariance metric space, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 806–820.
[5] W. Li, J. Li, L. Zhu, Multiple-shot person re-identification via fair set-collaboration metric learning, Neurocomputing 242 (2017) 15–27.
[6] Y. Lin, F. Guo, L. Cao, J. Wang, Person re-identification based on multi-instance multi-label learning, Neurocomputing 217 (2016) 19–26.
[7] C.C. Guo, S.Z. Chen, J.H. Lai, X.J. Hu, S.C. Shi, Multi-shot person re-identification with automatic ambiguity inference and removal, in: Proceedings of the International Conference on Pattern Recognition, 2014, pp. 3540–3545.
[8] F.M. Khan, F. Emond, Multi-shot person re-identification using part appearance mixture, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2017, pp. 605–614.
[9] K. Liu, W. Zhang, R. Huang, A spatio-temporal appearance representation for video-based pedestrian re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3810–3818.
[10] N. McLaughlin, J.M. del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1325–1334.
[11] B. Hadjkacem, W. Ayedi, M. Abid, H. Snoussi, Multi-shot human re-identification using a fast multi-scale video covariance descriptor, Eng. Appl. Artif. Intell. 65 (C) (2017) 60–67.
[12] B. Hadjkacem, W. Ayedi, M. Abid, H. Snoussi, Multi-shot human re-identification for the security in video surveillance systems, in: Proceedings of the 2018 IEEE Twenty-seventh International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, IEEE, 2018, pp. 203–208.
[13] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: Proceedings of the International Conference on Pattern Recognition, 11, 2014, pp. 34–39.
[14] J. Ramon, L. Raedt, Muti instance neural network, in: Proceedings of the ICML Workshop on Attribute-Value and Relational Learning, 2000, pp. 53–60.
[15] Z. Zhou, M. Zhang, Neural networks for multi-instance learning, in: Proceed-

ings of the International Conference on Intelligent Information Technology, 2002, pp. 455–459.
[16] M.L. Zhang, Z.H. Zhou, Improve multi-instance neural networks through feature delection, Neural Process. Lett. 19 (1) (2004) 1–10.
[17] J. Wu, Y. Yu, C. Huang, K. Yu, Deep multiple instance learning for image classification and auto-annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3460–3469.
[18] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D.N. Metaxas, X.S. Zhou, Multi-instance deep learning: discover discriminative local anatomies for body part recognition, IEEE Trans. Med. Imag. 35 (5) (2016) 1332–1343.
[19] X. Wang, Y. Yan, P. Tang, X. Bai, W. Liu, Revisiting multiple instance neural networks, Pattern Recognition 74 (2018) 15–24.
[20] M. Sun, T.X. Han, M.-C. Liu, A. Khodayari-Rostamabad, 23rd International Conference on Pattern Recognition, 2016.
[21] J. Amores, Multiple instance classification: review, taxonomy and comparative study, Artif. Intel. 201 (2013) 81–105.
[22] M. Guillaumin, J.J. Verbeek, C. Schmid, Multiple instance metric learning from automatically labeled bags of faces, in: Proceedings of the European Conference on Computer Vision, ECCV, 2010.
[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[24] A. Jimenez, J.M. Alvarez, X. Giró, Class-weighted convolutional features for visual instance search, Proceedings of the 28th British Machine Vision Conference. London, 2017, pp. 1–12.
[25] T. Zeng, S. Ji, Deep convolutional neural networks for multi-instance multi-task learning, in: Proceedings of the IEEE International Conference on Data Mining, 2015, pp. 579–588.
[26] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, E.I.-C. Chang, Deep learning of feature representation with multiple instance learning for medical image analysis, in: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1626–1630.
[27] A. Das, A. Chakraborty, A.K. Roy-Chowdhury, Consistent re-identification in a camera network, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 330–345.
[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015.
[29] F. Chollet, et al., Keras, 2015, (https://github.com/fchollet/keras).
[30] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, P. Yuen, An asymmetric distance model for cross-view feature mapping in person re-identification, IEEE Trans. Circ. Syst. Video Technol. 26 (2016) 2588–2603.
[31] M. Kostinger, M. Hirzer, Large scale metric learning from equivalence constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Rhode Island, 2012, pp. 2288–2295.
[32] M. Guillaumin, J. Verbeek, C. Schmid, L.J. Kuntzmann, Is that you? Metric learning approaches for face identification, in: Proceedings of the IEEE International Conference on Computer Vision, Kyoto, 2009.
[33] H.-X. Yu, A. Wu, W.-S. Zheng, Cross-view asymmetric metric learning for unsupervised person re-identification, in: Proceedings of the IEEE Conference on Computer Vision, Venice, Italy, 2017.
[34] X. Lin, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1239–1248.
[35] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical Gaussian descriptor for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1363–1372.
[36] A. Borgia, Y. Hua, E. Kodirov, N.M. Robertson, Cross-view discriminative feature learning for person re-identification, IEEE Trans. Image Process. 27 (11) (2018) 5338–5349.
[37] W. Chen, X. Chen, J. Zhang, K. Huang, A multi-task deep network for person re-identification, in: Proceedings of the AAAI, 2017.
[38] S. Bai, X. Bai, Q. Tian, Scalable person re-identification on supervised smoothed manifold, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3356–3365.
[39] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: Proceedings of the European conference on computer vision, Zurich, 2014, pp. 1–16.
[40] A. Mignon, F. Jurie, PCCA: a new approach for distance learning from sparse pairwise constraints, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012, pp. 2666–2672.
[41] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013, pp. 3318–3325.
[42] Y.-C. Chen, X. Zhu, W.-S. Zheng, J.-H. Lai, Person re-identification by camera correlation aware feature augmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 392–408.
[43] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the Computer Vision and Pattern Recognition, 2017.
[44] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: Proceedings of the the International Conference on Computer Vision, 2017.
[45] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: Proceedings of the the Computer Vision and Pattern Recognition Workshops, 2017, pp. 1435–1443.
[46] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3820–3828.
[47] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, in: Proceedings of the Computer Vision and Pattern Recognition Workshops, 2018, pp. 1179–1188.
[48] R. Yu, Z. Zhou, S. Bai, X. Bai, Divide and fuse: a re-ranking approach for person re-identification, in: Proceedings of the the British Machine Vision Conference, 2017.



Xiaokai Liu received her B.S. degree from Dalian Maritime University (DMU), Dalian, China, in 2010, and Ph.D. degree from Dalian University of Technology (DUT), Dalian, China, in 2011, both in Information and Communication Engineering. She is currently a lecturer at Dalian Maritime University. She used to be a visiting researcher working with Professor Ming-Hsuan Yang in University of California, Merced, from 2013 to 2015. Her research interests include person re-identification, image retrieval, semantic segmentation, computer vision and machine learning.

Sheng Bi received the B.E. degrees in Marine Radio Navigation in 1986, M.E. degrees in Marine Engineering in 1989 and Ph.D. degrees in Information and Communication Engineering in 2007 all from Dalian Maritime University. He is currently a Professor at Dalian Maritime University. He is the dean of the Information Science and Technology College. His research interests include image processing, communication, and embedded system application in marine navigation.

Xiaorui Ma received the B.E. degree in 2008 from School of Mathematics and Statistics, Lanzhou University (LZU), P.R. China. Now she is a doctoral candidate in the School of Information and Communication Engineering of Dalian University of Technology (DUT), PR China. Her research interests include remote sensing image classification and machine learning

Jie Wang (M'12, SM'18) received his B.S. degree from Dalian University of Technology, Dalian, China, in 2003, M.S. degree from Beihang University, Beijing, China, in 2006, and Ph.D degree from Dalian University of Technology, Dalian, China, in 2011, all in Electronic Engineering. He is currently a Professor at Dalian Maritime University. He used to be an Associate Professor at Dalian University of Technology from 2014 to 2017. He was a visiting researcher with University of Florida from 2013 to 2014. His research interests include wireless localization and tracking, radio tomography, wireless sensing, wireless sensor networks, cognitive radio networks, and machine learning. He serves as an Associate Editor for IEEE Transactions on Vehicular Technology. He served as a Guest Editor for International Journal of Distributed Sensor Networks for a special issue on wireless localization in 2015.