SCIENCE CHINA
Information Sciences

December 2019, Vol. 62 220102:1–220102:13
https://doi.org/10.1007/s11432-019-2675-3

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

# Uncertainty-optimized deep learning model for small-scale person re-identification

Cairong ZHAO[1*], Kang CHEN[1], Di ZANG[1], Zhaoxiang ZHANG[2], Wangmeng ZUO[3] & Duoqian MIAO[1]

[1]*Department of Computer Science and Technology, Tongji University, Shanghai* 201804*, China;*
[2]*Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001*, China*

Received 25 August 2019/Accepted 5 September 2019/Published online 15 November 2019

**Abstract** In recent years, deep learning has developed rapidly and is widely used in various fields, such as computer vision, speech recognition, and natural language processing. For end-to-end person re-identification, most deep learning methods rely on large-scale datasets. Relatively few methods work with small-scale datasets. Insufficient training samples will affect neural network accuracy significantly. This problem limits the practical application of person re-identification. For small-scale person re-identification, the uncertainty of person representation and the overfitting problem associated with deep learning remain to be solved. Quantifying the uncertainty is difficult owing to complex network structures and the large number of hyper-parameters. In this study, we consider the uncertainty of pedestrian representation for small-scale person re-identification. To reduce the impact of uncertain person representations, we transform parameters into distributions and conduct multiple sampling by using multilevel dropout in a testing process. We design an improved Monte Carlo strategy that considers both the average distance and shortest distance for matching and ranking. When compared with state-of-the-art methods, the proposed method significantly improve accuracy on two small-scale person re-identification datasets and is robust on four large-scale datasets.

**Keywords** person re-identification, uncertainty analysis, deep learning

**Citation** Zhao C R, Chen K, Zang D, et al. Uncertainty-optimized deep learning model for small-scale person re-identification. Sci China Inf Sci, 2019, 62(12): 220102, https://doi.org/10.1007/s11432-019-2675-3

## 1 Introduction

In this study, the person re-identification task is to match pictures of a target pedestrian taken from different perspectives among a set of gallery pictures. The probe can be one or multiple pictures and the gallery pictures are captured by more than two sets of cameras. Many factors contribute to uncertainty in this typical computer vision problem. First, video quality various among cameras. Second, surveillance cameras must deal with different scene information, i.e., location, illumination, and weather conditions. Third, occlusion objects as well as the pedestrians' pose and their clothing can change.

The availability of large-scale person re-identification datasets, such as Market1501 [1], CUHK03 [2], DukeMTMC [3], and MSMT17 [4], benefits research on large-scale video surveillance networks that can be used for person re-identification. Currently, mainstream research in this area is focused on three different methods. (1) Methods to extract pedestrian descriptors based on image features [5–10] focus on robust and reliable pedestrian representation models that can describe pedestrian image color, texture, shape, as well as spatial information. (2) Methods based on metric learning [11–13] rely on machine

* Corresponding author (email: zhaocairong@tongji.edu.cn)

 

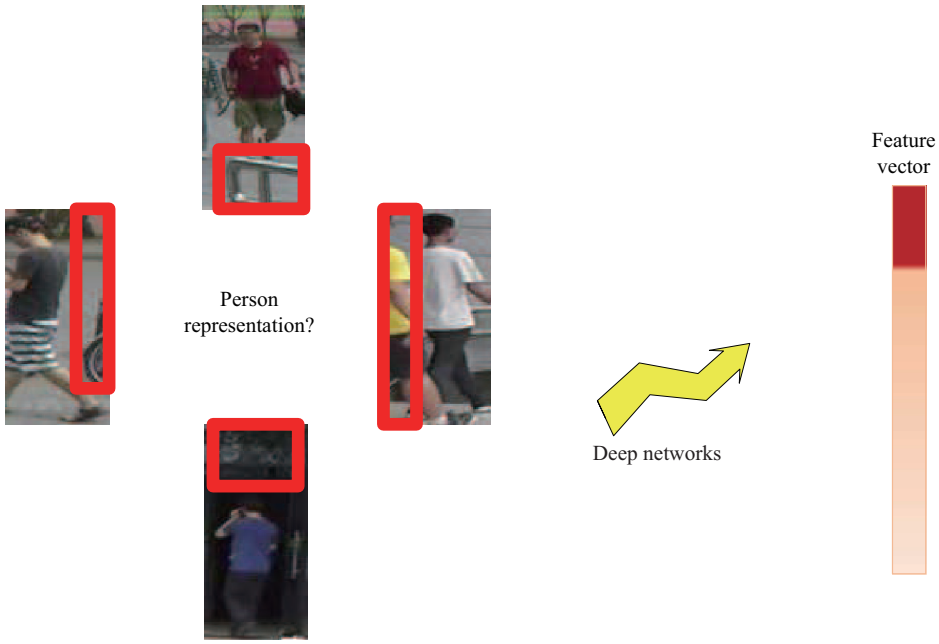


**Figure 1** (Color online) Typical examples of uncertainty in person representation.

learning and train a group of new spaces to optimize a pedestrian feature similarity measurement model. (3) Methods based on end-to-end deep learning architecture [14–21] to build an end-to-end platform for training and testing. Such methods employ an appropriate feature-extraction network and loss function, and process the input image directly to obtain final identification results.

In large-scale person re-identification, accuracy has been greatly improved with the application of deep learning methods. However, in many cases, the number of training samples is not sufficient. For small-scale person re-identification, deep learning methods face serious overfitting problems. The influence of uncertainties in pedestrian representation is magnified. In addition, the noisy data and inaccurate labels become more problematic with an insufficient number of training samples. These problems lead to inadequate optimization of network hyperparameters, which results in the deviation of the final trained model and affects re-identification accuracy.

Figure 1 illustrates a typical problem in small-scale person re-identification caused by uncertainty. For most deep learning methods, we treat some uncertain factors, such as occlusion, background, and illumination, as parts of the pedestrian representation. When the number of samples becomes half or even a quarter of the original, these noises will become saliency features during the matching process and may lead to error matching.

In this study, we aim to deal with the overfitting problem and uncertain pedestrian representation in small-scale person re-identification. The primary contributions of this study are as follows: (1) We use a multilevel dropout method in both training and testing to generalize parameters and obtain multiple distance matrices when testing. (2) We design an improved Monte Carlo strategy that considers the mean value and the closest person representation to obtain the final distance matrix for ranking. (3) Experiments on different scales of datasets demonstrate that the proposed method significantly improves the person re-identification recognition rate on small-scale datasets and remains robust on large-scale datasets.

## 2 Related work

Our research refers to many existing person re-identification studies, particularly those that involve deep learning methods. In addition, we refer to some existing methods for small-scale person re-identification and some studies on Bayesian deep learning.

Typical person re-identification systems [22, 23] focus on mining robust and characteristic descriptors for images and creating a distance space or embedding kernels to measure the similarity between different images. Inspired by the implementation of deep learning, mainstream research is focused on constructing deep convolutional neural network (CNN) architectures and designing functions to optimize CNN networks.

Deep person re-identification architectures primarily involve two components. The first is the mainstream CNN backbone that is designed to learn features, which has developed from a shallow convolutional network to today's more popular backbone networks (ResNet [24], Inceptions [25]), that are pretrained using the ImageNet [26] dataset. Such backbone networks are widely used in computer vision problems and can be easily fine-tuned for person re-identification tasks. The improvements usually lie in alignment, an attention model, or divide images into blocks. Overall, the backbone network aims to extract global and semantic features as descriptors. Ahmed et al. [27] presented the first research on simultaneously learning features and a corresponding distance metric that measures the Euclidean distance between two embedding features in a cross neighborhood layer. Zhang et al. [28] used the shortest path method to align the horizontal part of pedestrian pictures. Sun et al. [29] proposed a part-based convolutional and a refined part pooling (PCB-RPP) method, in which the pedestrian image is evenly partitioned and combined in the horizontal direction, and the parts are aligned using attention mechanisms. Wang et al. [30] proposed the multiple granularity network (MGN), which learns discriminative features with multiple granularity by combining multiple block schemes for pedestrian images.

The second component of person re-identification architectures aims to design proper functions to penalize error matching and optimize network parameters through back propagation [31, 32]. Zheng et al. [33] trained a Siamese network with pairwise input. The cross entropy and Siamese loss jointly optimize the shared parameters. Hermans et al. [34] proved that triplet loss, particularly hard triplet loss, outperforms other optimization methods for deep person re-identification models and significantly improves the baseline. Zhong et al. [35] proposed a k-order derivative coding method to re-rank images in a gallery. This re-ranking method is widely used in deep methods for large scale person re-identification.

For small-scale person re-identification, there are relatively few deep learning methods, most of which focus on how to expand the training data [36–39]. Chen et al. [40] proposed a cross-domain architecture that could use an auxiliary training set. Zheng et al. [41] generated some unlabeled training data using a generative adversal network (GAN). Wei et al. [4] proposed a person transfer GAN to eliminate differences in camera styles between different datasets.

There have been some pioneering efforts in the area of uncertainty in small-scale tasks and deep learning [42–46]. Xu et al. [47] reduced the uncertainty in a face recognition task by preprocessing the training data and provided a theorem that determines the upper bound of the number of useful training samples. Blundell et al. [48] discussed weight uncertainty and assigning weight with Gaussian distribution. Gal et al. [49] proposed a Bayesian neural network and used variational inference to solve the posterior distribution.

Inspired by these studies, we find that there is a common way to solve such problems, i.e., changing parameters from discrete values to probability distributions. Thus, we use a multilevel dropout method to generalize parameters and an improved Monte Carlo strategy for end-to-end training and testing.

## 3 Proposed method

Person re-identification aims to find images of the same identity with a probe image from a set of gallery images. For small-scale datasets, data are usually not sufficient to train a deep model. The overfitting and uncertain problems are serious. Thus, we design an uncertain-optimized deep model for small-scale person re-identification.

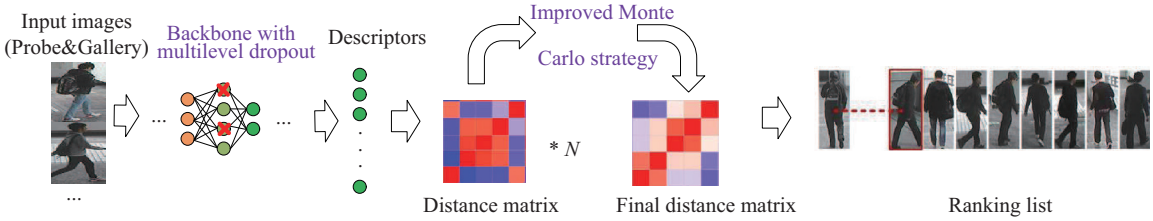


**Figure 2** (Color online) Overall framework when testing.

### 3.1 Overall framework

First, we describe the overall framework. Figure 2 illustrates the proposed framework when testing. We use GoogleNet (inceptionv1) with multilevel dropout as the backbone. Our approach differs from conventional approaches in that we use multilevel dropout in both train and test process. In the test processing, for each probe, we repeatedly send the probe image to a network with multilevel dropout and obtain more than one distance matrix. Then, we introduce an improved Monte Carlo strategy to obtain the final distance matrix. For each probe, we can obtain the final ranking list by sorting the distances.

### 3.2 Multilevel dropout

The existing methods [27, 28] usually position the dropout layer after the fully connected layer and use it only for training to avoid overfitting. Assuming that the network parameter can be expressed as $\omega = (W_i)_{i=1}^{L}$, $L$ is the number of convolutional layers. Random discarding can be seen as adding binary distributions on each node of the network. We only consider weights when the number of convolutional layer is 1. This changes the parameters from discrete values to probability distributions. In our method, a dropout operation is executed over multilevel convolution layers in both the training and testing. We multiply $W_i$ with a Bernoulli random distribution $z_{i,j}$. The output of the $i$th layer $W_i^*$ can be expressed as follows:

$$W_i^* = W_i \cdot (z_{i,j})_{j=1}^{K_i}, \tag{1}$$

$$z_{i,j} = \begin{cases} 1, & p_i, \\ 0, & 1 - p_i, \end{cases} \quad i = 1, \ldots, L_i; \quad j = 1, \ldots, K_{i-1}, \tag{2}$$

where $z_{i,j}$ is a random Bernoulli distribution with $p_i$ being the probability. $W_i$ is the weight matrix to be optimized, and the " $\cdot$ " operator is matrix dot. $j = 1, \ldots, K_i$ is the index of kernels in a CNN. In back propagation, we use the same binary variable values when transferring derivatives and optimizing the weight parameters. We make copies of each sample and put them in different batches for training. Therefore, the weights under Bernoulli distribution will be learned and updated during the training process. With multilevel dropout, the output of a deep network will not depend on local features, which means that the uncertain person representations will not be saliency descriptors. In this way, the feature descriptors will lose a certain amount of information. This is exactly what we expect to avoid the impact of uncertainty factors. However, there is a certain probability that significant information will be discarded. Thus, forward propagation is executed multiple times when testing to ensure the existence of saliency descriptors. These descriptors are used to calculate the distance matrix. Because we have more than one distance matrix, we improved the Monte Carlo strategy to calculate the final distance matrix.

### 3.3 Improved Monte Carlo strategy

When we need to solve random variables with probability distribution, the process can be simulated using a Monte Carlo method. The random variables are network parameters. For each instance of forward propagation, we obtain a distance matrix that represents the similarity between image descriptors. Typically, the Monte Carlo strategy takes the mean distance as the final distance, which represents the similarity between probe and gallery. This process can be seen as the estimation of distance from bootstrap sampling. We improved this strategy by considering the shortest distance. The shortest

distance means that after eliminating some descriptors, the remaining representations of two person images are closest to each other. In theory, when the uncertain descriptor is discarded, the shortest distance is more likely to be obtained. We obtain the final distance $d^*(P, G_i)$ according to the following formula:

$$d(P, G_i) = (x_P - x_{G_i})^{\mathrm{T}} (x_P - x_{G_i}), \tag{3}$$

$$d^*(P, G_i) = \lambda \frac{1}{N} \sum_{k=1}^{N} d_k(P, G_i) + (1 - \lambda) \min(d_k(P, G_i), k \sim (1 - N)), \tag{4}$$

where $d(P, G_i)$ is the Euclidean distance between probe $P$ and person $G_i$ in gallery. $x_p$ and $x_{G_i}$ are the descriptor of person $P$ and $G_i$. $N$ is the number of repetitions and $k$ denotes each of the repetition. $\frac{1}{N}\sum_{k=1}^{N} d_k(P, G_i)$ is Monte Carlo distance and $\min(\cdot)$ is the minimum distance of $N$ times. $\lambda$ represents the tradeoff parameter between the Monte Carlo distance and the minimum distance. Finally, we can obtain the ranking list $L^*(P, G_i)$ by sorting the final distance in ascending order.

### 3.4 Algorithm

The complete proposed method is summarized in Algorithm 1.

**Algorithm 1** Uncertainty-optimized testing process

Input: Probe and Gallery: $P, G_i$, number to repetitions $N$, trade off parameter $\lambda$;
Output: Ranking list $L^*(P, G_i)$;
$t = 0$;
**while** $t < N$ **do**
  $t \Leftarrow t + 1$;
  **for** all input images $P, G_i$ **do**
    Compute feature embedding $x_i$ by forward propagation (multilevel dropout);
    Compute $d(P, G_i)$ by Euclidean distance of $x_p$, $x_{G_i}$;
    $d^*(P, G_i) += d(P, G_i)$;
    **if** $d(P, G_i) < d_{\min}(P, G_i)$ **then**
      $d_{\min}(P, G_i) = d(P, G_i)$;
    **end if**
  **end for**
  $d^*(P, G_i) = \lambda/n \times d^*(P, G_i) + (1 - \lambda) \times d_{\min}(P, G_i)$;
**end while**
$L^*(P, G_i) = \mathrm{sort}(d^*(P, G_i))$ for each $P$.

### 3.5 Understanding our approach from Bayesian perspective

This subsection explains the working of our method in the Bayesian perspective. Following the Bayesian approach, it is assumed that there is an ideal function $f$ between inputs $X$ and their labels $Y$. From person re-identification datasets, we have the posterior distribution $p(f|X, Y)$. The process of pedestrian matching proceeds as follows:

$$p(y^*|x^*, X, Y) = \int p(y^*|f^*)\, p(f^*|x^*, \omega)\, p(\omega|X, Y) \mathrm{d}f^* \mathrm{d}\omega, \tag{5}$$

where $y^*$ is the probe label, and $x^*$ are galleries. From re-identification datasets, we have person $X$ and label $Y$. $f^*$ can be considered as a set of transformations in deep networks. By using $p(\omega|X, Y)$ and following an approximate inference [50], we define a simpler distribution $q(\omega)$ as the approximate of $p(\omega|X, Y)$. Then, we obtain

$$p(y^*|x^*, X, Y) \approx \int p(y^*|x^*, \omega)\, q(\omega) \mathrm{d}\omega. \tag{6}$$

In order to match our multiple dropout layers, we set $q(\omega)$ as a Bernoulli distribution, which is a fairly weak approximation with no additional parameters. The sampling operation on $q(\omega)$ is the same as the

**Figure 3** (Color online) Example images from the CUHK01 dataset.

multilevel dropout operation on the parameters of the $i$th layer $(W_i)_{i=1}^{L}$. This is why we use multilevel dropout in the test process. The generation of over-fitting is partly attributed to the fixed value of the parameters of the neural networks. When we apply multilevel dropout to give Bernoulli distribution to parameters, the probability solution can be obtained by a Monte Carlo strategy. By following Monte Carlo approach we can approximate the integral as follows:

$$p\left(y^{*}|x^{*},X,Y\right)\approx\frac{1}{N}\sum_{k=1}^{N}p\left(y^{*}|x^{*},\widehat{\omega}_{k}\right),\quad\widehat{\omega}_{i}\sim q(\omega). \tag{7}$$

In small-scale person re-identification, we should consider both the mean value obtained by the Monte Carlo strategy, and the closest pedestrian representation. Consequently, we improve the Monte Carlo strategy as described in Subsection 3.3.

## 4 Experiments

We carry out experiments on some mainstream person re-identification datasets. These datasets have different scales and the results of our uncertainty-optimized deep learning model are analyzed in Subsection 4.4.

### 4.1 Datasets

The following datasets are used in our experiment.

CUHK01 [2]. This dataset contains 3884 pictures of 971 identities collected from two camera views. Each identity has two images from two camera views. We use 486 identities for testing in the experiment. The remaining pedestrian images are used for training. Figure 3 is example images in CUHK01.

VIPeR [51]. This dataset contains 632 identities collected from two camera views where the illumination conditions differ significantly. This dataset contains 632 images. We use 316 identities for training and 316 identities for testing. Figure 4 is example images in VIPeR.

**Figure 4** (Color online) Example images from the VIPeR dataset.

CUHK03 [2]. CUHK03 contains 13164 well pre-cut pictures from 1360 pedestrians. The images are captured from six different pairs of camera views. The dataset is well-divided, and we shuffle the training sets. For each epoch, we use approximately 1500 pictures from 150 different identities for the training.

Market-1501 [1]. Market-1501 contains 32668 detected person boxes from 1501 identities. The images are collected from six cameras, one of which is a low-pixel camera. We use 12936 pictures as a training set and the other 19732 images are used for testing. In addition, this dataset contains some labeled detection errors, which are not used in our methods.

DukeMTMC [3]. DukeMTMC contains more than 2000000 manually annotated frames from more than 2000 identities. The pictures are captured by eight synchronized cameras with more than 7000 single camera trajectories. This dataset can be used for both image person re-identification and video person re-identification.

MSMT17 [4]. MSMT17 is a new large-scale dataset that contains 126441 bounding boxes of 4101 pedestrians. The bounding boxes are obtained by Faster R-CNN [52] and validated manually. The images in MSMT17 are captured by 15 groups of camera in a campus over a long period of time.

### 4.2 Evaluation matrix

According to our method, we obtain the final ranking lists in the test process. Cumulative matching characteristic (CMC) curves [53] are used to evaluate our ability to find pedestrians from top $k$ similar matches. The mean average precision (mAP) [54] is used to estimate the ability to distinguish pedestrians. Rank-$k$ means that we find the target in $k$-most similar gallery pictures.

### 4.3 Implementation

The architecture of our backbone in Table 1 is the multilevel dropout added to the convolutional layer with different dropout ratios. We train our model with a combination of hard mining triplet loss and cross entropy loss. We select this training process with reference to Zheng et al. [33] and Herman et al. [34]. We do not use tricks like alignment or blocking on person images. On one hand, we do not want the network to be too large. On the other hand, a simple network can better reflect the practicability of our method on small-scale data.

We use the pytorch [55] framework to implement our model. We shuffle the dataset and randomly order the images without any data augmentation. We use the cosine distance for feature measurement. We set the batch size to 64. The learning rate is 0.001 at the beginning with a 0.9 learning rate decay over 1000 iterations. The weight decay is 0.0002, and the momentum parameter is 0.95. We train our model for 50000 iterations within 5 hours on an NVIDIA 1080Ti GPU. When testing with the Monte

**Table 1** Structure of our backbone network

| Name | Patch size/stride | Output size | #1×1 | #3×3 reduce | #3×3 | #5×5 reduce | #5×5 | pool + proj | Dropout ratio |
|---|---|---|---|---|---|---|---|---|---|
| Input | – | 3×224×224 | – | – | – | – | – | – | – |
| Conv1/Relu | 7×7/2 | 64×112×112 | – | – | – | – | – | – | 0.1 |
| Pool1 | 3×3/2 | 64×56×56 | – | – | – | – | – | Max | – |
| Conv2/Relu | 3×3/1 | 192×56×56 | – | – | – | – | – | – | 0.1 |
| Pool2 | 3×3/2 | 192×28×28 | – | – | – | – | – | Max | – |
| Inception 3a | – | 256×28×28 | 64 | 96 | 128 | 16 | 32 | Max+32 | 0.1 |
| Inception 3b | – | 480×28×28 | 128 | 128 | 192 | 16 | 32 | Max+64 | 0.1 |
| Pool3 | 3×3/2 | 480×14×14 | – | – | – | – | – | Max | – |
| Inception 4a | – | 512×14×14 | 192 | 96 | 208 | 16 | 48 | Max+64 | 0.2 |
| Inception 4b | – | 512×14×14 | 160 | 112 | 224 | 24 | 64 | Max+64 | 0.2 |
| Inception 4c | – | 512×14×14 | 128 | 128 | 256 | 24 | 64 | Max+64 | 0.2 |
| Inception 4d | – | 512×14×14 | 112 | 144 | 288 | 32 | 48 | Max+64 | 0.2 |
| Inception 4e | – | 512×14×14 | 256 | 160 | 320 | 32 | 128 | Max+64 | 0.3 |
| Pool4 | 3×3/2 | 832×7×7 | – | – | – | – | – | Max | – |
| Inception 5a | – | 832×7×7 | 256 | 160 | 320 | 32 | 128 | Max+128 | 0.3 |
| Inception 5b | – | 1024×7×7 | 384 | 192 | 384 | 48 | 128 | Max+128 | 0.3 |
| Pool5 | 7×7/1 | 1024×1×1 | – | – | – | – | – | Average | – |
| fc | – | 1024 | – | – | – | – | – | – | 0.3 |

Carlo classification strategy, we set the number of repetitions to 100 and the tradeoff parameter to 1 ($N = 100$, $\lambda = 0.5$).

### 4.4 Experimental results

Our uncertainty-optimized deep learning model is trained and tested on both small-scale and large-scale datasets. We discuss the experimental results on different datasets in detail in this subsection.

#### 4.4.1 *Results on CUHK01 dataset*

CUHK01 is a small-scale dataset. To improve matching accuracy, some state-of-the-art methods often merge CUHK03 and CUHK01 datasets to increase the number of training samples. For our method, we use CUHK01 dataset samples directly to train the deep learning model. We contrast with the existing mainstream methods for person re-identification, and give the result as follows:

As can be seen from Figure 5 and Table 2, obtaining highly accurate small-scale person re-identification results using deep learning methods is difficult, particularly with complex neural networks, such as PCB [29] and MGN [30]. When there are insufficient training samples, the multi-granularity features increase overfitting, and it is difficult for the networks to converge. The baseline network has relatively good result (55.2% Rank-1 accuracy). After the introduction of multiple dropout and the improved Monte Carlo strategy, the accuracy increases by 7.1%.

#### 4.4.2 *Results on VIPeR dataset*

VIPeR is a very small person re-identification dataset. Existing deep learning methods do not work well. Experimental results obtained using the proposed method are shown in Figure 6 and Table 3.

It is obvious that the proposed deep learning method achieves relatively good results with the VIPeR dataset.

#### 4.4.3 *Results on Large-scale datasets*

We also conduct experiments on four large-scale person re-identification datasets. The performance of our uncertainty-optimized method on large-scale datasets is shown in Table 4. Here, we use Resnet50 as a baseline for comparison.

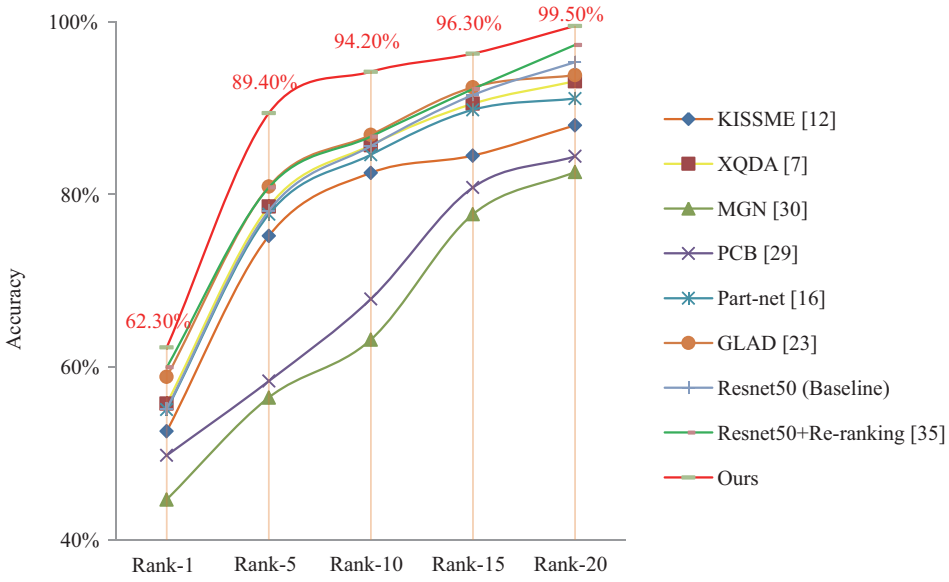


**Figure 5** (Color online) CMC curves on CUHK01 dataset.

**Table 2** Performance comparison on CUHK01 dataset

| Method | Rank-1 (%) | Rank-5 (%) | Rank-10 (%) | Rank-15 (%) | Rank-20 (%) |
|---|---|---|---|---|---|
| KISSME [12] | 52.6 | 75.2 | 82.5 | 84.5 | 88.0 |
| XQDA [7] | 55.8 | 78.6 | 85.7 | 90.5 | 93.1 |
| MGN [30] | 44.7 | 56.5 | 63.2 | 77.7 | 82.6 |
| PCB [29] | 49.8 | 58.4 | 67.9 | 80.8 | 84.4 |
| Part-net [16] | 55.1 | 77.7 | 84.6 | 89.8 | 91.1 |
| GLAD [23] | 58.9 | 80.9 | 86.9 | 92.4 | 93.8 |
| Resnet50 (Baseline) | 55.2 | 78.1 | 85.6 | 91.5 | 95.3 |
| Resnet50+Re-ranking [35] | 60.0 | 80.8 | 86.7 | 92.2 | 97.3 |
| Ours | 55.2 | 89.4 | 94.2 | 96.3 | 99.5 |

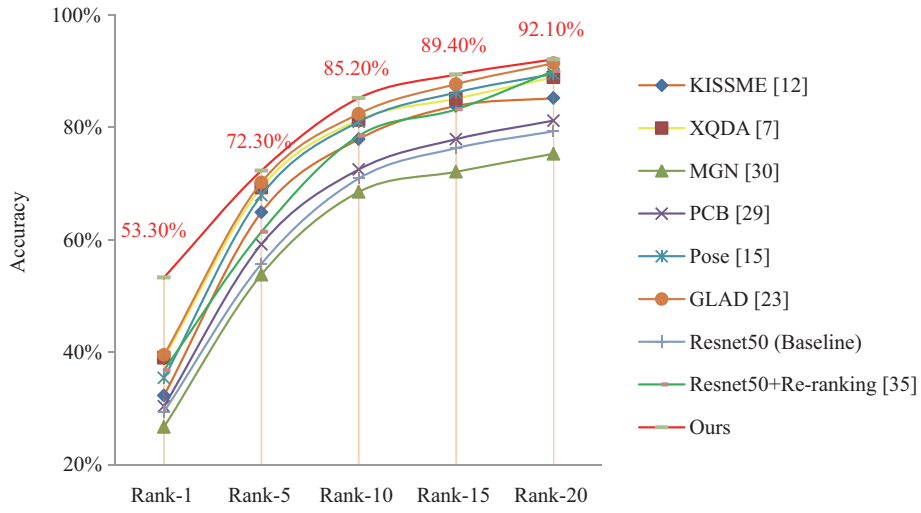


**Figure 6** (Color online) CMC curves on VIPeR dataset.

In Figure 7 the blue and red lines indicate the Rank-1 accuracy of our method and the baseline variation with decrease in training samples on four large-scale datasets.

It can be seen from Table 4 that the accuracy is close to the baseline, which means that the proposed

**Table 3** Performance comparison on VIPeR dataset

| Method | Rank-1 (%) | Rank-5 (%) | Rank-10 (%) | Rank-15 (%) | Rank-20 (%) |
|---|---|---|---|---|---|
| KISSME [12] | 32.3 | 64.9 | 77.9 | 83.8 | 85.2 |
| XQDA [7] | 39.0 | 69.3 | 81.3 | 85.1 | 88.9 |
| MGN [30] | 26.7 | 53.8 | 68.5 | 72.1 | 75.3 |
| PCB [29] | 30.4 | 59.2 | 72.5 | 77.9 | 81.2 |
| Pose [15] | 35.4 | 67.9 | 81.0 | 86.2 | 89.5 |
| GLAD [23] | 39.5 | 70.2 | 82.4 | 87.7 | 91.4 |
| Resnet50 (Baseline) | 29.4 | 55.7 | 70.9 | 76.3 | 79.3 |
| Resnet50+Re-ranking [35] | 36.8 | 61.4 | 78.5 | 83.2 | 90.0 |
| Ours | 53.3 | 72.3 | 85.2 | 89.4 | 92.1 |

**Table 4** Performance on large-scale datasets

| Dataset | Rank-1 (%) | Rank-5 (%) | Rank-10 (%) | Rank-20 (%) | mAP (%) |
|---|---|---|---|---|---|
| Market1501 (Ours) | 86.2 | 94.6 | 97.1 | 83.8 | 67.8 |
| Market1501 (Baseline) | 82.3 | 89.9 | 95.4 | 97.9 | 60.3 |
| CUHK03 (Ours) | 80.4 | 92.5 | 94.3 | 97.0 | 59.6 |
| CUHK03 (Baseline) | 78.3 | 93.3 | 97.0 | 98.7 | 61.1 |
| DukeMTMC (Ours) | 76.9 | 84.5 | 87.5 | 90.2 | 62.0 |
| DukeMTMC (Baseline) | 73.5 | 78.6 | 81.8 | 85.1 | 57.4 |
| MSMT17 (Ours) | 68.4 | 78.8 | 82.6 | 88.4 | 40.2 |
| MSMT17 (Baseline) | 68.3 | 81.4 | 85.9 | 92.5 | 45.6 |

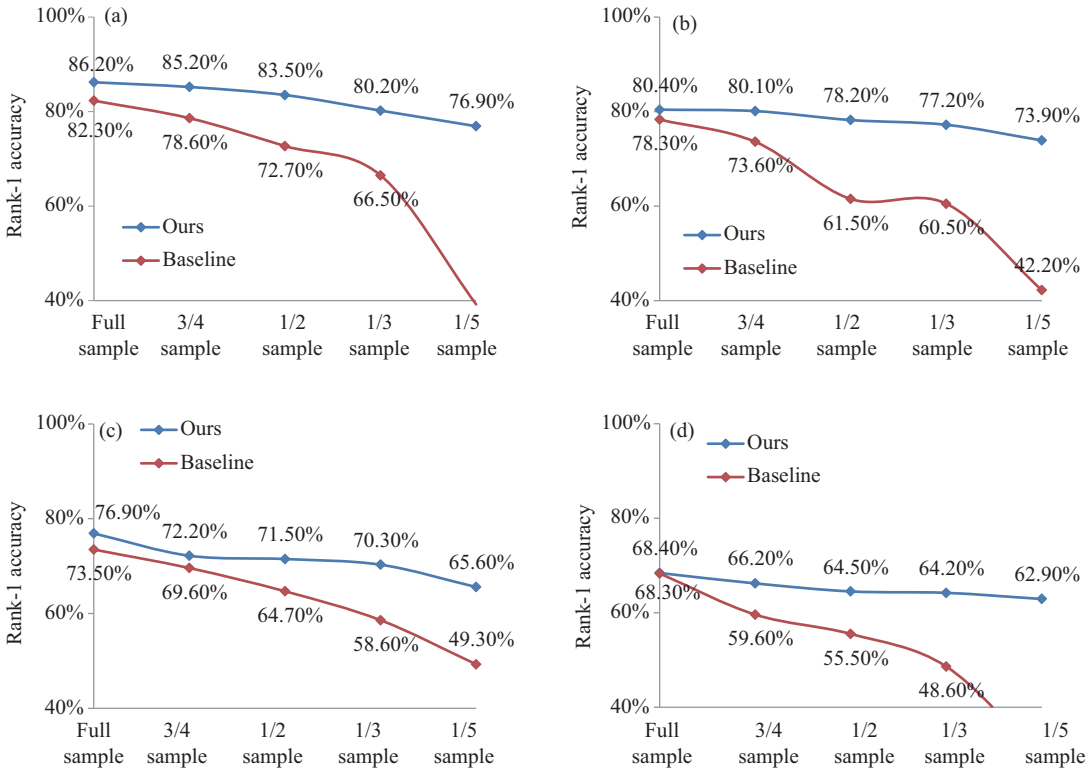


**Figure 7** (Color online) Robustness of the proposed model to the number of samples on large-scale datasets of (a) Market1501, (b) CUHK03, (c) DukeMTMC, and (d) MSMT17.

method also works on large-scale datasets. Typically, when there is sufficient training data, deeper and more complex networks can obtain better results. The network structure used in the proposed method is not complicated, and the accuracy did not reach the top level. However, considering both accuracy and

consumption of network computing resources, the proposed method still has some advantages.

Obviously, our method is more applicable to small-scale datasets, which are more common in practice. Therefore, we perform experiments where we gradually reduce the number of training sets on four large-scale datasets to test the robustness of the proposed model relative to the number of samples. Here, we also use Resnet50 as a baseline for comparison. The results are shown in Figure 7. As can be seen, as the number of training samples decreases, the accuracy of the baseline model drops dramatically while the proposed method is obviously more stable and accurate. This is because the multilevel dropout layers and improved Monte Carlo strategy reduce the impact of uncertainties and the Bayesian approach can prevent overfitting to a certain extent.

## 5 Conclusion

In this study, we implemented a multilevel dropout method and an improved Monte Carlo strategy to solve the overfitting problem and reduce the uncertainty in person representation. In this way, a deep learning method can be applied to small-scale person re-identification. The proposed method significantly improves the recognition rate of person re-identification on small-scale datasets and also works on large-scale datasets. The end-to-end network structure and robustness under small data conditions make our method more suitable for practical application.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant Nos. 61673299, 61203247, 61573259, 61573255, 61876218), Fundamental Research Funds for the Central Universities, and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR). The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions.

**References**

1 Zheng L, Shen L Y, Tian L, et al. Scalable person re-identification: a benchmark. In: Proceedings of IEEE International Conference on Computer Vision, 2016. 1116–1124
2 Li W, Zhao R, Xiao T, et al. DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 152–159
3 Gou M, Karanam S, Liu W, et al. DukeMTMC4ReID: a large-scale multi-camera person re-identification dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1425–1434
4 Wei L H, Zhang S L, Gao W, et al. Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 79–88
5 Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2008. 262–275
6 Ma B P, Su Y, Jurie F. Local descriptors encoded by fisher vectors for person re-identification. In: Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2012. 413–422
7 Matsukawa T, Okabe T, Suzuki E, et al. Hierarchical gaussian descriptor for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1363–1372
8 Pala F, Satta R, Fumera G, et al. Multimodal person reidentification using RGB-D cameras. IEEE Trans Circuits Syst Video Technol, 2016, 26: 788–799
9 Bai S, Tang P, Torr P H S, et al. Re-ranking via metric fusion for object retrieval and person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 740–749
10 Yu R, Zhou Z C, Bai S, et al. Divide and fuse: a re-ranking approach for person re-identification. 2017. ArXiv: 1708.04169
11 Davis J V, Kulis B, Jain P, et al. Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning. New York: ACM, 2007. 209–216
12 Köstinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012. 2288–2295
13 Xiong F, Gou M, Camps O, et al. Person re-identification using kernel-based metric learning methods. In: Proceedings of the European Conference on Computer Vision, 2014. 1–16
14 Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification. In: Proceedings of the European Conference on Computer Vision, 2016. 791–808

15 Zheng L, Huang Y J, Lu H C, et al. Pose invariant embedding for deep person re-identification. 2017. ArXiv: 1701.07732

16 Cho Y J, Yoon K J. PaMM: pose-aware multi-shot matching for improving person re-identification. 2017. ArXiv: 1705.06011

17 Lin Y T, Zheng L, Zheng Z D, et al. Improving person re-identification by attribute and identity learning. 2017. ArXiv: 1703.07220

18 Geng M Y, Wang Y W, Xiang T, et al. Deep transfer learning for person re-identification. 2016. ArXiv: 1611.05244

19 Jin H B, Wang X B, Liao S C, et al. Deep person re-identification with improved embedding and efficient training. In: Proceedings of IEEE International Joint Conference on Biometrics (IJCB). New York: IEEE, 2017. 261–267

20 Zhu J Q, Zeng H Q, Du Y Z, et al. Joint feature and similarity deep learning for vehicle re-identification. IEEE Access, 2018, 6: 43724–43731

21 Imani Z, Soltanizadeh H. Histogram of the node strength and histogram of the edge weight: two new features for RGB-D person re-identification. Sci China Inf Sci, 2018, 61: 092108

22 Liao S C, Hu Y, Zhu X Y, et al. Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 2197–2206

23 Wei L H, Zhang S L, Yao H T, et al. Glad: global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017. 420–428

24 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778

25 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of IEEE International Conference on Computer Vision, 2015. 1–9

26 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255

27 Ahmed E, Jones M, Marks T K. An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3908–3916

28 Zhang X, Luo H, Fan X, et al. Alignedreid: surpassing human-level performance in person re-identification. 2017. ArXiv: 1711.08184

29 Sun Y F, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision, 2018. 480–496

30 Wang G S, Yuan Y F, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of 2018 ACM Multimedia Conference on Multimedia Conference. New York: ACM, 2018. 274–282

31 Bai S, Bai X, Tian Q. Scalable person re-identification on supervised smoothed manifold. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2530–2539

32 Yu R, Dou Z Y, Bai S, et al. Hard-aware point-to-set deep metric for person re-identification. In: Proceedings of the European Conference on Computer Vision, 2018. 188–204

33 Zheng Z D, Zheng L, Yang Y. A discriminatively learned CNN embedding for person re-identification. ACM Trans Multim Comput Commun Appl, 2017, 14: 13

34 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. 2017. ArXiv: 1703.07737

35 Zhong Z, Zheng L, Cao D L, et al. Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1318–1327

36 Wu L, Hong R C, Wang Y, et al. Cross-entropy adversarial view adaptation for person re-identification. IEEE Trans Circ Syst Video Tech, 2019. doi: 10.1109/TCSVT.2019.2909549

37 Liu Z, Wang Y H, Li A N. Hierarchical integration of rich features for video-based person re-identification. IEEE Trans Circuits Syst Video Technol, 2018. doi: 10.1109/TCSVT.2018.2883995

38 Zhu Z, Huang T T, Shi B G, et al. Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 2347–2356

39 Hou R B, Ma B P, Chang H, et al. VRSTC: occlusion-free video person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7183–7192

40 Chen W H, Chen X T, Zhang J G, et al. A multi-task deep network for person re-identification. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017

41 Zheng Z D, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 3754–3762

42 Bui T, Hernández-Lobato D, Hernandez-Lobato J, et al. Deep Gaussian processes for regression using approximate expectation propagation. In: Proceedings of International Conference on Machine Learning, 2016. 1472–1481

43 Gal Y, Ghahramani Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference. 2015. ArXiv: 1506.02158

44 Kwon J, Lee K M. Adaptive visual tracking with minimum uncertainty gap estimation. IEEE Trans Pattern Anal

Mach Intell, 2016, 39: 18–31
45 Shen F M, Yang Y, Zhou X, et al. Face identification with second-order pooling in single-layer networks. Neurocomputing, 2016, 187: 11–18
46 Li Z C, Tang J H. Weakly supervised deep matrix factorization for social image understanding. IEEE Trans Image Process, 2017, 26: 276–288
47 Xu Y, Fang X, Li X, et al. Data uncertainty in face recognition. IEEE Trans Cybern, 2014, 44: 1950–1961
48 Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural networks. 2015. ArXiv: 1505.05424
49 Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of International Conference on Machine Learning, 2016. 1050–1059
50 Minka T P. A Family of Algorithms for Approximate Bayesian Inference. Cambridge: Massachusetts Institute of Technology, 2001
51 Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), 2007. 3: 1–7
52 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149
53 Bolle R M, Connell J H, Pankanti S, et al. The relation between the ROC curve and the CMC. In: Proceedings of the 4th IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), 2005. 15–20
54 Cormack G V, Lynam T R. Statistical precision of information retrieval evaluation. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006. 533–540
55 Ketkar N. Introduction to pytorch. In: Deep Learning With Python. Berkeley: Apress, 2017. 195–208