Hindawi
Journal of Sensors
Volume 2021, Article ID 9463092, 12 pages
https://doi.org/10.1155/2021/9463092

Research Article

Pedestrian Reidentification Algorithm Based on Deconvolution Network Feature Extraction-Multilayer Attention Mechanism Convolutional Neural Network

Feng-Ping An,[1,2] **Jun-e Liu**,[3] **and Lei Bai**[4]

[1] *School of Physics and Electronic Electrical Engineering, Huaiyin Normal University, Huaian, JS 223300, China*
[2] *School of Information and Electronics, Beijing Institute of Technology, Beijing, BJ 100081, China*
[3] *School of Information, Beijing Wuzi University, Beijing, BJ 100061, China*
[4] *Hebei IoT Monitoring Engineering Technology Research Center, North China Institute of Science and Technology, Langfang, HB 065201, China*

Correspondence should be addressed to Feng-Ping An; anfengping@163.com, Jun-e Liu; 2924175349@qq.com, and Lei Bai; zhouxianwei1961@163.com

Received 17 February 2020; Revised 18 July 2020; Accepted 23 November 2020; Published 28 January 2021

Academic Editor: Manuel Aleixandre

Pedestrian reidentification is a key technology in large-scale distributed camera systems. It can quickly and efficiently detect and track target people in large-scale distributed surveillance networks. The existing traditional pedestrian reidentification methods have problems such as low recognition accuracy, low calculation efficiency, and weak adaptive ability. Pedestrian reidentification algorithms based on deep learning have been widely used in the field of pedestrian reidentification due to their strong adaptive ability and high recognition accuracy. However, the pedestrian recognition method based on deep learning has the following problems: first, during the learning process of the deep learning model, the initial value of the convolution kernel is usually randomly assigned, which makes the model learning process easily fall into a local optimum. The second is that the model parameter learning method based on the gradient descent method exhibits gradient dispersion. The third is that the information transfer of pedestrian reidentification sequence images is not considered. In view of these issues, this paper first examines the feature map matrix from the original image through a deconvolution neural network, uses it as a convolution kernel, and then performs layer-by-layer convolution and pooling operations. Then, the second derivative information of the error function is directly obtained without calculating the Hessian matrix, and the momentum coefficient is used to improve the convergence of the backpropagation, thereby suppressing the gradient dispersion phenomenon. At the same time, to solve the problem of information transfer of pedestrian reidentification sequence images, this paper proposes a memory network model based on a multilayer attention mechanism, which uses the network to effectively store image visual information and pedestrian behavior information, respectively. It can solve the problem of information transmission. Based on the above ideas, this paper proposes a pedestrian reidentification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network. Experiments are performed on the related data sets using this algorithm and other major popular human reidentification algorithms. The results show that the pedestrian reidentification method proposed in this paper not only has strong adaptive ability but also has significantly improved average recognition accuracy and rank-1 matching rate compared with other mainstream methods.

1. Introduction

Public security is the prerequisite for the development of the national economy and social stability. With the rapid development of the economy, the people are paying more and more attention to security issues [1]. Therefore, social precautionary measures have gradually been strengthened, and a large number of surveillance cameras have been installed in public gathering places such as traffic arteries, security checkpoints, hotels, schools, and hospitals. They constitute a large-scale distributed monitoring system. Pedestrian reidentification is a key technology in large-scale distributed camera systems. It can quickly and efficiently detect and track target people in large-scale distributed surveillance networks [2–4]. Therefore, pedestrian reidentification technology has great scientific value and broadness in maintaining social stability, improving public safety, and protecting people's lives and property safety [5–7]. Zajdel et al. [8] first proposed related concepts about pedestrian reidentification. Since then, the issue of pedestrian reidentification has attracted increasing attention from academia and industry. In particular, the public release of the 2007 pedestrian reidentification data set VIPeR [9] prompted an increasing number of researchers to publish their research results at major international conferences on computer vision and machine learning (CVPR, ICCV, ECVC, IJCAI, and AAAI), as well as in top journals, such as the International Journal of Computer Vision (IJCV), IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), and IEEE Transactions on Image Processing (TIP) [10, 11]. They have greatly promoted the rapid development of pedestrian reidentification technology. In general, pedestrian reidentification methods can be divided into pedestrian reidentification methods based on feature learning, pedestrian reidentification method based on distance metric learning, and pedestrian reidentification method based on deep learning technology.

Pedestrian reidentification methods based on feature learning mainly include Gheissari et al. [12] proposed a spatiotemporal segmentation algorithm to detect pedestrian dressing changes and stable foreground areas. Their standardized color histograms and edge histograms are calculated to combine them into pedestrian signature invariant features. However, this method has a poor recognition effect. Gheissari et al. [12] proposed extracting the three complementary features of the overall color content of pedestrian body parts, the spatial arrangement of colors to stable areas, and the circulation of highly structured local areas to perform pedestrian reidentification. However, this method has problems such as weak adaptive capability and low efficiency. Zhao et al. [13] proposed a pedestrian reidentification method based on unsupervised saliency learning. However, this method has low feature extraction efficiency. Liao et al. [14] analyzed the frequency of occurrence of local features in the horizontal direction, obtained stable feature representation by maximizing the frequency of occurrence, and then reidentified pedestrians. However, this method has a low recognition effect and efficiency. Layne et al. [15] proposed a pedestrian semantic attribute learning method to select and weight and then perform pedestrian recognition. However, the feature extraction method of this method has the problem of weak extraction ability.

Pedestrian reidentification method based on distance metric learning: Zhang et al. [16] proposed a zero-space transformation method based on kernel functions and completed pedestrian matching in this space. But pedestrian matching is weak. Zhou et al. [17] trained a part of metric learning through online learning to learn the special local metric of semipositive definiteness. The local metric can be adaptively integrated with the global metric to enhance the performance of the model. But this method has low learning efficiency and low effect. Zhou et al. [18] proposed a new type of similarity constraint and studied several constraint generation schemes based on optimal transmission for identification. However, the constraint generation scheme obtained by this method has a low error rate. Li et al. [19] used the relationship between related samples and their adjacent points to construct the neighborhood structure popularity and proposed a neighborhood structured metric learning method to learn this manifold discrimination by adjusting the related domain metric difference. But the measurement method is not effective. Jia et al. [20] proposed a multidistance metric joint learning method. It learns each suitable subdistance metric separately for each different feature and forms the final distance metric through the weighted sum of the submetrics to perform pedestrian recognition. But this method has a low recognition effect and adaptive ability.

It can be known from the above literature analysis that the pedestrian reidentification method based on feature learning needs to manually set feature information. Therefore, it is unreasonable to set the feature information, that is, it does not meet the pedestrian feature information contained in the image itself, resulting in the loss of some feature information. At the same time, it also has a method of extracting feature information, which has a weak adaptive ability and cannot fully extract the pedestrian feature information contained in the image itself. Pedestrian reidentification methods based on distance metric learning are different due to the selected distance metric method. It will affect the effect of subsequent pedestrian reidentification, that is to say, there are certain difficulties in the selection of distance measurement methods. At the same time, this method has the problem of weak adaptive ability, that is, for different images containing pedestrians, the same type of distance measurement method for pedestrian reidentification may lead to different results, which will affect the accuracy of pedestrian recognition. For this reason, related scholars have been researching better pedestrian reidentification methods. Hinton and Salakhutdinov first proposed the concept of deep learning in Science in 2006 [21]. Deep learning technology has been widely used in computer vision [22], image classification [23], video analysis [24], and other fields. In view of the above characteristics of deep learning technology, a deep learning-based pedestrian reidentification method has been formed, mainly including Li et al. [25] first convolutional neural network (CNN) is used in the field of pedestrian reidentification. It released a new database CUHK03 for training of CNNs. Since then, the application of CNN in the field of pedestrian reidentification has attracted increasing attention from researchers. Li et al. [26] designed a multiscale context-aware network for learning effective pedestrian features from the whole and local body parts of pedestrians, using global appearance features and local appearance features to fuse and use them for pedestrian classification tasks. Xiao et al. [27] proposed to learn effective deep features from different data sets and then perform pedestrian recognition. Schumann et al. [28] proposed to use the information contained in the automatically detected attributes to perform pedestrian recognition. Xu et al. [29] proposed an attention pooling network that combined time and space to perform later pedestrian re-identification. Liu et al. [30] proposed a cumulative motion context network that is aimed at using long-term contextual motion information to robustly identify different pedestrians. Zhong et al. [31] introduced a heterogeneous and homogeneous learning method. This method uses camera invariance and domain connectivity constraints for different data sets to generate more robust pedestrian features and then perform pedestrian reidentification. Bak et al. [32] proposed to publish a synthetic pedestrian dataset for indoor and outdoor scenes and developed a lighting condition estimator to improve the accuracy of pedestrian reidentification. Yao et al. [33] proposed a deep learning model based on part loss network to minimize the empirical classification risk of training pedestrian images and the learning risk of invisible pedestrian images, thereby improving pedestrian reidentification. Bai et al. [34] proposed a unified integrated diffusion deep learning framework that can improve the pedestrian recognition effect by imposing additional constraints on the objective function and changing the solver for similarity propagation. Cui et al. [35] proposed a greedy hierarchical unsupervised learning algorithm, which is a generative model causal variable with many hidden layers. It can provide ideas for the training of unsupervised deep learning pedestrian recognition models. Ali and Chaudhuri [36] pointed out that they demonstrate the robustness of the training procedure with respect to the random initialization, the positive effect of pretraining in terms of optimization, and its role as a regularizer. Chen et al. [37] also proposed a fast greedy algorithm to train deep learning models and achieved good results. Han et al. [38] proposed a Spiking Neural Networks. The paper proposes a pretraining scheme using biologically plausible unsupervised learning, namely, Spike-Timing-Dependent-Plasticity (STDP). The STDP-based pretraining with gradient-based optimization provides improved robustness, faster training time, and better generalization compared with purely gradient-based training without pretraining.

In summary, with its powerful learning capabilities, deep learning models provide an effective solution for pedestrian reidentification tasks. However, it is found that the pedestrian recognition method based on deep learning has the following problems in the application process: First, during the learning process of the deep learning model, the initial value of the convolution kernel is usually randomly assigned. It makes the model learning process easily fall into a local optimum. The second is that the model parameter learning method based on the gradient descent method will appear gradient dispersion. The third is that the information transfer of pedestrian reidentification sequence images is not considered. In view of this, a feature extraction model learning method based on a deconvolution network is proposed. First, a feature map matrix can be learned from the original image by using two layers of unsupervised deconvolutional neural networks. Then, the obtained feature map matrix is used as a convolution kernel of the deep convolution network, and the original image is subjected to layer-wise convolution and pooling operations. Finally, the second derivative information of the error function is directly obtained without calculating the Hessian matrix. Based on this information, a momentum coefficient is introduced to improve the convergence of back propagation in the small-batch gradient descent method. It can fine-tune the deep convolutional network to suppress gradient dispersion. At the same time, in order to solve the problem of information transmission of pedestrian reidentification sequence images, a memory network based on attention mechanism was proposed to effectively store image visual information and pedestrian behavior information, respectively. It projects the problem to both the visual memory network and the behavioral memory network to retrieve the multimodal factual basis. Then, a multilayer attention mechanism architecture is proposed to establish the multiround interaction between the problem and the two types of memory networks and solve the problem of information transfer. Based on the above ideas, this paper proposes a pedestrian reidentification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network.

Section 2 of this paper will mainly describe the feature extraction model of the deconvolution network proposed in this paper. Section 3 details the convolutional neural network model of the multilayer attention mechanism proposed in this paper. Section 4 constructs a pedestrian reidentification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network. Section 5 analyzes the case and compares it with the main recognition algorithm. Finally, the full text is summarized and discussed.

2. Feature Extraction Model Based on Deconvolution Network

2.1. Feature Extraction Model Based on Deconvolution Network. This paper uses a deconvolution network to perform unsupervised feature extraction on the original input image. A deconvolution network is a neural network consisting of alternating deconvolution layers, depooling layers, and correction layers [35]. The layered network structure helps to extract the middle and high-level features of the image. In this paper, a hierarchical network model is constructed by stacking two layers of deconvolutional layers, and the feature mapping matrix is directly extracted using efficient optimization techniques. As a neural network with decoder only, the deconvolution network first performs unsupervised feature extraction in each layer of the network in the decoding stage. Subsequently, the convolution kernel convolves the image features and reconstructs a reconstructed image similar to the original image. Each layer of the model reconstructs the next layer of the network. The error of the objective function is the difference between the reconstructed image and the input image. The layers of the network structure are connected in a fully connected manner. After adjusting the parameters through training to obtain the weights of each layer, it can calculate different representations of the input signal, and these representations are recorded as input feature expressions. Assuming that the input and output of the network are the same, the inference operation process of the two-layer stacked deconvolution network model is shown in Figure 1. The mapping operation is to map the error signal from the first layer to the second

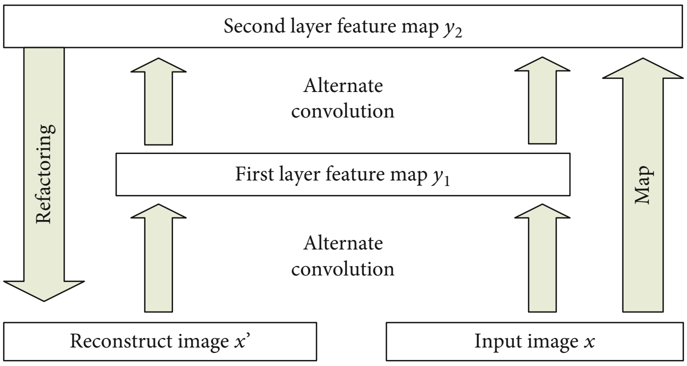

Figure 1: Schematic diagram of the convolution process of the deconvolution network.

layer after convolving the input image x through a convolution kernel. It can get the corresponding feature map. The reconstruction operation starts from the feature map y_2 to the input layer. The specific derivation process will be explained in detail below.

Assume that the ith original input image is $x^{(i)}$. It consists of K_0 color channels to form $x^{(i)} = \{x_1^{(i)}, \cdots, x_{K_0}^{(i)}\}$. By deconvolving the feature map ${y^{(i)}}_k$ of the hidden layer and the convolution kernel $f_{k,s}$, K linear sums are obtained. This process will yield the corresponding s-th color channel map of the image. Specifically

$$\sum_{k=1}^{K} y_k^{(i)} \oplus f_{k,s} = x_s^{(i)}. \tag{1}$$

In the formula, ${y^{(i)}}_k$ is a $(N_r + H - 1) \times (N_s + H - 1)$ -dimensional feature map, and each input image has a corresponding feature map. ⊕ represents the deconvolution operator. $f_{k,s}$ is a convolution kernel of $H \times H$ size. ${x^{(i)}}_s$ is an $N_r \times N_r$ image. To satisfy the uniqueness of the solution, a regularization sparsity term is introduced into the feature map ${y^{(i)}}_k$. Therefore, the one-sample objective function is as follows:

$$T_1\left(x^{(i)}\right) = \frac{\tau}{2} \sum_{t=1}^{K_0} \left\| y_k^{(i)} \oplus f_{k,t}^1 - x_t^{(i)} \right\|_2^2 + \sum_{k=1}^{K} \left| y_k^{(i)} \right|^p. \tag{2}$$

In the formula, ${f_{k,s}}^1$ represents the convolution kernel of the first layer. p is the norm of the regularization sparse term. Generally, set $p = 1$. The value of τ is 1, which plays the role of balancing the image $x^{(i)}$ and the sparse feature map ${y^{(i)}}_k$. The reconstruction phase of the deconvolution network follows the top-down direction to convolve the feature map of the hidden layer to synthesize the input image, which is different from the algorithm principles of sparse self-coding and deep belief networks. It only obtains the feature description of the hidden layer of the input image by optimizing the objective function T_1.

Suppose the sample set containing N examples is $x = \{x^{(1)}, \cdots, x^{(N)}\}$. The overall objective function of the second layer can be formed by stacking single-sample objective functions, specifically

$$T_2\left(x^{(i)}\right) = \frac{\tau}{2} \sum_{i=1}^{N} \sum_{s=1}^{K_0} \left\| \sum_{k=1}^{K} g_{k,s}^2 \left(y_{k,2}^{(i)} \oplus f_{k,s}^2 - y_{k,1}^{(i)} \right) \right\|_2^2 + \sum_{i=1}^{N} \sum_{k=1}^{K} \left| y_{k,2}^{(i)} \right|^p. \tag{3}$$

In the formula, ${y^{(i)}}_{k,1}$ and ${y^{(i)}}_{k,2}$, respectively, represent the first and second layer feature maps. ${g^2}_{k,s}$ is an element in a fixed binary matrix, which plays the role of connecting feature maps.

The training process of the deconvolution network can be divided into two stages. In the first stage, given the convolution kernel ${f_{k,s}}^2$, the feature map ${y^{(i)}}_{k,2}$ is obtained so that the objective function $T_2(x)$ is minimized. In the second stage, given the feature map ${y^{(i)}}_{k,2}$, the convolution kernel ${f_{k,s}}^2$ is solved so that the objective function $T_2(x)$ is minimized. In the specific training process, an auxiliary objective function ${T_2}^{'}(x)$ is first introduced to avoid falling into a local optimum. Then, the minimum value of the auxiliary objective function ${T_2}^{'}(x)$ is solved. It makes the auxiliary variable ${\xi^{(i)}}_{k,2}$ and the feature map ${y^{(i)}}_{k,2}$ approximate. The auxiliary objective function ${T_2}^{'}(x)$ is defined as

$$\begin{aligned} T_2'(x) = {} & \frac{\tau}{2} \sum_{i=1}^{N} \sum_{s=1}^{K_0} \left\| \sum_{k=1}^{K} g_{k,s}^2 \left(y_{k,2}^{(i)} \oplus f_{k,s}^2 - y_{k,1}^{(i)} \right) \right\|_2^2 \\ & + \frac{\beta}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} \left\| y_{k,2}^{(i)} - \xi_{k,2}^{(i)} \right\|_2^2 + \sum_{i=1}^{N} \sum_{k=1}^{K} \left| \xi_{k,2}^{(i)} \right|^p, \end{aligned} \tag{4}$$

where β is a continuous variable. It slowly increases from a very small initial value until the feature map ${y^{(i)}}_{k,2}$ and the auxiliary variable ${\xi^{(i)}}_{k,2}$ finally approach. In the stage of obtaining the feature map ${y^{(i)}}_{k,2}$, assuming that the auxiliary variable ${\xi^{(i)}}_{k,2}$ is given, the derivative of the objective function ${T_2}^{'}(x)$ to the feature map ${y^{(i)}}_{k,2}$ is

$$\frac{\partial T_2'(x)}{\partial y_{k,2}^{(i)}} = \tau \sum_{s=1}^{K_0} \left(F_{k,s}^2\right)^T \left(\sum_{k=1}^{K} F_{k,s}^2 y_{k,2}^{(i)} - y_{k,1}^{(i)}\right) + \beta\left(y_{k,2}^{(i)} \text{-} \xi_{k,2}^{(i)}\right). \tag{5}$$

In the formula, if ${g^2}_{k,s} = 1$, then ${F^2}_{k,s}$ is equivalent to a sparse convolution matrix convolved with ${f^2}_{k,s}$. If ${g^2}_{k,s} = 0$, then ${F^2}_{k,s}$ is a zero matrix.

The feature map ${y^{(i)}}_{k,2}$ is updated as follows: first, given i, let formula (5) be 0, and find ${y^{(i)}}_{k,2}$. That is to solve the following $K(N_r + H - 1) \times (N_s + H - 1)$-dimensional linear system:

$$A \begin{pmatrix} y_{1,2}^{(i)} \\ \vdots \\ y_{k,2}^{(i)} \end{pmatrix} = \begin{pmatrix} \sum_{s=1}^{K} \left(F_{1,s}^2\right)^T y_{s,1}^{(i)} + \frac{\beta}{\tau} \xi_{1,2}^{(i)} \\ \vdots \\ \sum_{s=1}^{K} \left(F_{K,s}^2\right)^T y_{s,1}^{(i)} + \frac{\beta}{\tau} \xi_{K,2}^{(i)} \end{pmatrix}. \tag{6}$$

Among them,

$$A = \begin{pmatrix} \sum_{s=1}^{K_0} \left(F_{1,s}^2\right)^T F_{1,s}^2 + \frac{\beta}{\tau} I & \cdots & \sum_{s=1}^{K_0} F_{1,s}^2 \left(F_{K,s}^2\right)^T \\ \vdots & \vdots & \vdots \\ \sum_{s=1}^{K_0} \left(F_{K,s}^2\right)^T F_{1,s}^2 & \cdots & \sum_{t=1}^{K_0} \left(F_{K,s}^2\right)^T F_{1,s}^2 + \frac{\beta}{\tau} I \end{pmatrix}. \tag{7}$$

Then, the conjugate gradient descent method is used to obtain the optimal solution of formula (6).

In the process of obtaining the auxiliary variable ${\xi^{(i)}}_{k,2}$. Assuming that the feature map ${y^{(i)}}_{k,2}$ is fixed, then the optimal problem of ${\xi^{(i)}}_{k,2}$ is transformed into a 1D optimal problem for the feature map. If $p = 1$, the solution expression of the auxiliary variable ${\xi^{(i)}}_{k,2}$ is

$$\xi_{k,2}^{(i)} = \max\left(\left|y_{k,2}^{(i)}\right| - \frac{1}{\beta}, 0\right) \frac{y_{k,2}^{(i)}}{\left|y_{k,2}^{(i)}\right|}. \tag{8}$$

The convolution kernel ${f_{k,s}}^2$ can be updated according to the gradient descent method, specifically

$$\frac{\partial T_2'(x)}{\partial f_{k,s}^2} = \tau \sum_{i=1}^{N} \sum_{s=1}^{K_0} \left(y_{k,2}^{(i)}\right)^T \left(\sum_{k=1}^{K} y_{k,s}^2 Z_{k,2}^{(i)} f_{k,s}^2 - y_{s,1}^{(i)}\right). \tag{9}$$

In the formula, Z is a convolution matrix similar to F.

2.2. Network Training Process. The training sample set $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(N)}, y^{(N)})\}$ contains N samples, and the convolutional layer uses the input image or the feature map of the previous layer. Convolution operation with the convolution kernel, it will get the output characteristics. The feature map is convolved with multiple sets of convolution kernels to obtain corresponding output features. Its feature map calculation formula is as follows:

$$x_{j'}^{l} = f\left(\sum_{j \in M^l} x_j^{l-1} * f_{jj'}^{l}\right). \tag{10}$$

In the formula, l is the number of layers of the convolution layer. $x_{j'}^{l}$ represents the j'th output feature map of the l th layer. $f_{jj'}^{l}$ is a convolution kernel connecting the j-th feature map of the l-1th layer and the j'-th feature map of the l-th layer. M^l is the number of feature maps of the l-1 layer. $f(.)$ represents the nonlinear activation function ReLU. $*$ is a convolution operator. The distribution of the input data changes after the convolution operation. To solve the internal covariate shift caused by changes in data distribution, this paper introduces the BN layer to deal with the data distribution. The data passing through the BN layer is equivalent to PCA dimensionality reduction. That is, the correlation between features is reduced. The data mean and standard deviation are normalized such that the feature mean of each dimension is 0 and the standard deviation is 1. In network structure design, the BN layer is usually placed after the convolution operation before the activation function. Therefore, the calculation formula (10) of the forward transfer convolution layer is changed as follows:

$$x_{j'}^{l} = f\left(BN\left(\sum_{j \in M^l} x_j^{l-1} * f_{jj'}^{l}\right)\right). \tag{11}$$

The pooling layer combines all nonoverlapping subregions of the feature map of the convolutional layer. It downsamples a large matrix into a matrix with a smaller dimension. This downsampling can reduce the amount of calculation and prevent overfitting after processing. Its output feature map is

$$x_{j'}^{l} = f\left(\beta_{j'}^{l} \operatorname{down}\left(x_{j'}^{l-1}\right)\right). \tag{12}$$

In the formula, $\operatorname{down}(.)$ represents the downsampling function. β represents the downsampling coefficient. Similarly, after the pooling layer, the BN layer is used to process the feature map, and it is placed after the pooling operation before the activation function. Therefore, the calculation formula (12) of the forward transfer pooling layer is transformed into

$$x_{j'}^{l} = f\left(BN\left(\beta_{j'}^{l} \operatorname{down}\left(x_{j'}^{l-1}\right)\right)\right). \tag{13}$$

Deep learning model training uses the back-propagation layer-by-layer training mechanism, and the training parameter is mainly the convolution kernel f. Let ${z'}_{m}^{(i)}$ denote the label corresponding to the i-th sample in the m-th dimension. $z_m^{(i)}$

represents the output corresponding to the i-th sample in the m-th dimension, then the cost function is

$$J(f) = \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M}\left({z'}^{(i)}_{m} - z^{(i)}_{m}\right). \tag{14}$$

In the formula, M represents the total number of categories. The MB-SGD method is used to update the weights, that is

$$f^{l}_{jj'}(t) = f^{l}_{jj'}(t-1) - \frac{1}{N}\eta\sum_{i=1}^{N}\frac{\partial J}{\partial f^{l}_{jj'}(t-1)}. \tag{15}$$

In the formula, t represents the current time. η represents the learning rate. During the training of the network using MB-SGD, when the gradient changes its holding direction, the error surface will have different curvatures along different directions. It is easy to cause the points on the surface to oscillate from side to side with the continuous decline of the gradient. It causes the gradient to fail to converge to a minimum. For this reason, a case is considered in which MB-SGD retains both the gradient vector information at the previous time and the network parameter update value at the previous time. Then find the second derivative of the error function. This information not only estimates the gradient of the cost function surface at a point but also estimates the surface curvature. After calculating the curvature and obtaining the second derivative information, the weight update formula is

$$\nabla f^{l}_{jj'}(t) = \frac{\nabla f^{l}_{jj'}(t-1)}{\nabla f^{l}_{jj'}(t-2) - \nabla f^{l}_{jj'}(t-1)}\nabla f^{l}_{jj'}(t-2). \tag{16}$$

In the formula, $\nabla f^{l}_{jj'}(t-1)$ represents the gradient function at time $(t-1)$. If the step size in the weight update formula (16) grows too fast, it will easily cause the convergence process to diverge. To this end, the abovementioned defects are overcome by introducing a momentum coefficient μ, that is,

$$\Delta f^{l}_{jj'}(t-1) > \mu\Delta f^{l}_{jj'}(t-2). \tag{17}$$

Therefore,

$$\Delta f^{l}_{jj'}(t-1) = \mu\Delta f^{l}_{jj'}(t-2). \tag{18}$$

According to the above analysis, formula (15) is modified as

$$f^{l}_{jj'}(t) = f^{l}_{jj'}(t-1) - \frac{1}{N}\eta\sum_{i=1}^{N}\mu\frac{\partial J}{\partial f^{l}_{jj'}(t-2)}. \tag{19}$$

2.3. Basic Steps of the Algorithm

2.3.1. Use a Deconvolution Network to Extract the Feature Map Matrix

(1) Use formula (1) to perform a bottom-up convolution mapping operation on the input image data g. It can get the convolution feature map Z

(2) Start from the feature map Z of the second layer and perform the convolution operation from top to bottom. Use formula (8) to get the auxiliary variable $\xi^{(i)}{}_{k,2}$

(3) According to the obtained auxiliary variable $\xi^{(i)}{}_{k,2}$, and then use formula (6) to obtain the feature map $y^{(i)}{}_{k,2}$

(4) After multiple iterations of the above steps, use formula (9) to update the convolution kernel $f^{2}{}_{k,s}$, and finally output the mapping matrix f

2.3.2. Use Deep Learning Models to Classify Pedestrian Images

(1) Use the feature map matrix f extracted by the two-layer stacked deconvolution network as the convolution kernel. Convolution operation is performed on the image data according to the model formula (11)

(2) The obtained convolution feature map is downsampled using formula (13)

(3) After going through the layers of the model in sequence, use formula (19) to perform supervised fine-tuning on the deep learning model. It can update model parameters

3. Multilayer Attention Mechanism Convolutional Neural Network Model

3.1. Multilayer Attention Memory Network Encoder. This paper treats pedestrian images as a retrieval problem, that is, given an image I, a title c. The former $t-1$ rounds pedestrian information $\{(q_1, a_1), \cdots, (q_{t-1}, a_{t-1})\}$ and the current question q_t. The goal of this task is to rank the N candidate answers related to the question q_t and make the ranking of the correct answers as high as possible.

For questions that need to be answered q_t, first, it is entered into the standard long-short term memory (LSTM) to calculate its vector expression $V_{q_t} \in \mathbb{R}^{d}$, that is,

$$V_{q_t} = \mathrm{LSTM}^{q}(q_t). \tag{20}$$

In the formula, all problems share the encoder LSTM^{q}.

For image I, the feature map f_I is extracted from the original image I through the deconvolution network proposed in Section 2, that is

$$f_I = \mathrm{OCNN}(I). \tag{21}$$

After the above model training converges, the final pooling layer is taken out as the feature $f_I \in \mathbb{R}^{512\times7\times7}$ of image I. This pooling layer retains the spatial information of the original picture. The number of regions divided by the image is 7×7, and 512 is the size of the feature vector of each region. Then, it is reshaped to $f_I{}' \in \mathbb{R}^{512\times 49}$. A single-layer perceptron is then used to convert the 49 feature vectors into new feature vectors with the same dimensions as the problem V_{q_t}, specifically

$$M^I = \tanh\left(W_I f_I{}' + b_I\right). \tag{22}$$

$W_I \in \mathbb{R}^{d\times 512}$ is the transformation weight matrix. d is the size of the transformed feature vector. It is consistent with the V_{q_t} size. The obtained result $M_I \in \mathbb{R}^{d\times 49}$ is the visual memory of the i^{th} column vector $m^I{}_i$ in the visual memory corresponding to the picture area i.

First, the question-and-answer pairs (q_i, a_i) are stitched together to form a question-and-answer fact QA_i. Then, input the image title and the question and answer facts into the long-term and short-term memory model to calculate their vector expression, that is

$$m_i^H = \text{LSTM}^f(QA_i), i = 1, \cdots, t-1, \tag{23}$$

$$QA_0 = c. \tag{24}$$

All the pedestrian facts share the encoder LSTM^f. QA_i is the input pedestrian image sequence. $m_i^H \in \mathbb{R}^d$ is the output of the last layer. This method treats m_i^H as a vector representation of the actual behavior of pedestrians. Through the above coding method, all visual fact memories $M^H = \{m_1^H, \cdots, m_{t-1}^H$ related to the problem q_t can be obtained.

This paper proposes a network structure with a multilayered attention mechanism, while extracting enough useful information from the cross-modal behavioral fact memory and visual fact memory. The model has information transmission and reasoning problems, so it can answer the question q_t more accurately.

The vector of a given problem q_t represents V_{q_t} and the two memory banks M^I and M^H associated with it. First, the problem V_{q_t} is projected into the historical behavior memory through the following formula to retrieve the behavior facts related to the problem:

$$u_0 \equiv V_{q_t}, \tag{25}$$

$$s_i = u_{j-1} \cdot m_i^H, j = 1, \cdots, r, \tag{26}$$

$$u_{j-1} = u_{j-1} + \sum_{i=0}^{t-1} a_i m_i, a_i = \frac{\exp(s_i)}{\sum_{i=0}^{t-1}\exp(s_i)}. \tag{27}$$

In the formula, r represents the total number of rounds of the attention mechanism. s_i is a similarity measure between u_{j-1} and m_i^H. The result u_{j-1} is calculated each time by projecting the problem to the behavior memory. This method treats it as the latest expression of the problem q_t containing behavioral facts. Then, the model in Section 2 is used to continue projecting it into the visual factual behavior memory to retrieve related visual behaviors. The specific formula is as follows:

$$h = \tanh\left(W_{I,h} M^I \oplus \left(W_u u_{j-1} + b_h\right)\right), \tag{28}$$

$$p^I = \text{soft max}\left(W_p h + b_p\right), \tag{29}$$

$$u_j = u_{j-1} + \sum_{i=0}^{49} p_i^I m_i^I, \tag{30}$$

where $W_{I,h}$, $W_u \in \mathbb{R}^{k\times d}$. $b_h \in \mathbb{R}^k$. $\oplus$ represents the addition between a matrix and a vector. h is the output of the single-layer neural network obtained earlier. $W_p \in \mathbb{R}^{I\times k}$ and $p^I \in \mathbb{R}^{49}$ represent the projection probability value between each picture area and $u_j - 1$. By projecting visual memory, output u_j will be obtained. This output contains both behavioral and visual factual information related to problem q_t. Then, the next round of cross-modal attention mechanism training will continue to be performed through formulas (24)–(30) until the preset upper limit r of the total rounds is reached. This technique can solve the problem of information transfer and reasoning in the deep learning pedestrian reidentification algorithm.

After r-round alternate projection calculations of behavior and visual fact memory, the u_r obtained contains fact information sufficient to support answering the question q_t. Then, u_r is inputted to the single-layer neural network to obtain the final output of the encoder. The calculation formula is

$$e_t = \tanh\left(W_e u_r + b_e\right). \tag{31}$$

In the formula, W_e and b_e are the weights and biases of the fully connected neural network, respectively. e_t is the final encoded output about the problem q_t.

3.2. Generative and Discriminant Decoders.

Both the generator decoder and the discriminant decoder take e_t as input. It decodes to get the final correct answer a_t.

3.2.1. Generating Decoder.

Given the problem q_t and the encoder's final encoding expression e_t, its corresponding answer a_t is generated by the following formula:

$$h_0 \equiv e_t, \tag{32}$$

$$h_i = \text{LSTM}^g(h_{i-1}, x_{i-1}), i = 1, \cdots, |a_t|, \tag{33}$$

$$p_i = \text{soft max}\left(W^g h_i + b^g\right). \tag{34}$$

In the formula, LSTM^g is a generative LSTM decoding network. h^i is the output of the LSTM^g at time i. x_i is the vector expression corresponding to the i-th word of the answer at. The length of $a_t\,|$ is $|a_t\,|$. p_i is the probability distribution of the word. During training, this method maximizes the probability of generating the correct answer a_t. In the test and evaluation phase, the probability of generating each candidate answer is calculated first, and then all candidate answers are sorted in descending order by probability.

3.2.2. Discriminant Encoder. The given question q_t and the encoder eventually encode e_t, and the candidate answers $\{a_1^{'}, \cdots, a_N^{'}\}$ related to the question. First, each candidate answer is encoded by the following LSTM. It will get the corresponding vector expression, specifically

$$h_{a_i'} = \mathrm{LSTM}^d\left(a_i'\right), i = 1, \cdots, N. \qquad (35)$$

In the formula, LSTM^d is the encoder of all candidate answers. And they share weights, $h_{a_i'}$ is the final encoded output of the ith candidate answer a_i'. The similarity s_i between the vector e_t and $h_{a_i'}$ is calculated by the dot product similarity, and the calculation formula is

$$s_i = e_t \cdot h_{a_i'}. \qquad (36)$$

Then, all the obtained similarities $\{s_1, \cdots, s_N\}$ are stitched together and input to a softmax classifier, and then the posterior probability of all candidate answers is calculated. The specific formula is

$$p_a = \text{soft max}\left(s_1 \Theta s_2 \Theta \cdots \Theta s_N\right). \qquad (37)$$

In the formula, Θ stands for splicing operation. p_a is the probability distribution of candidate answers. During the training process, this method maximizes the discrimination probability of the correct answer a_t. In the verification and testing stages, the candidate answers are directly sorted from large to small by the posterior probability of the answers, so as to obtain accurate pedestrian recognition.

4. Pedestrian Reidentification Algorithm Based on Deconvolution Network Feature Extraction-Multilayer Attention Mechanism Convolutional Neural Network

Based on the contents of Section 2 and Section 3, this section proposes a pedestrian reidentification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network. First, the feature map matrix is obtained from the original image through the deconvolution neural network proposed in Section 2 of this paper. Then, it uses the obtained feature map matrix as the convolution kernel of the deep convolution network and performs layer-wise convolution and pooling operations on the original image. In this process, a momentum coefficient is introduced to improve the convergence of backpropagation and achieve the purpose of suppressing the gradient dispersion phenomenon. At the same time, a memory network based on attention mechanism is proposed to effectively store image visual information and behavior information. Then propose a multilayer attention mechanism architecture. It can solve the problem of information transmission. The basic idea of the proposed pedestrian reidentification algorithm is shown in Figure 2. The basic steps of the algorithm are as follows:

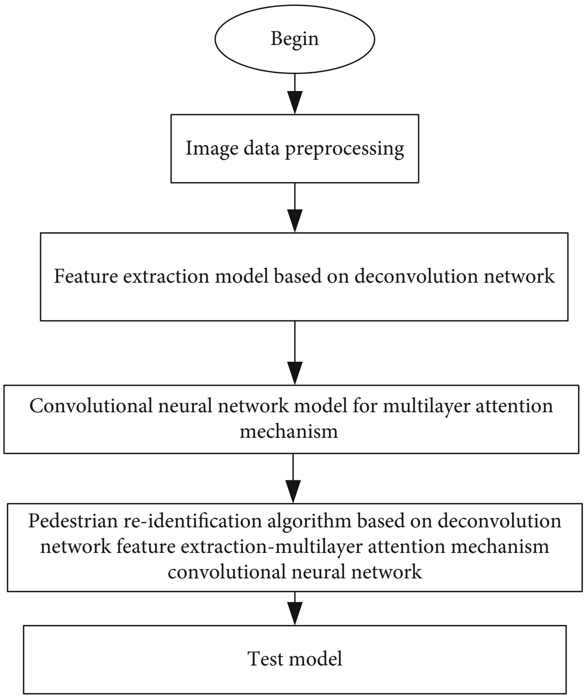
Begin
Image data preprocessing
Feature extraction model based on deconvolution network
Convolutional neural network model for multilayer attention mechanism
Pedestrian re-identification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network
Test model

Figure 2: Pedestrian recognition algorithm framework based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network.

(1) First, perform preprocessing such as denoising and enhancement on pedestrian images

(2) In order to better solve the problem of extracting the feature information of pedestrian images from deep learning models, the feature map matrix is obtained from the original image through the deconvolution neural network proposed in Section 2 of this paper. It uses the obtained feature map matrix as the convolution kernel of the deep convolution network and then performs convolution and pooling operations. It can get richer feature information

(3) In order to solve the problem of information transmission of pedestrian images by deep learning network structure, this paper proposes a memory network based on attention mechanism to effectively store image visual information and behavior information, respectively. Based on this, a multilayer attention mechanism architecture is proposed to solve the information transfer problem of deep learning models

(4) Combining the method of step (3) with step (2), a pedestrian reidentification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network is proposed, and the algorithm is used to perform related instances Analysis, and comparison and analysis with other major epidemic re-identification methods

The test model referred to in Figure 2 is based on the establishment of a pedestrian reidentification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network. It fine-tunes the pedestrian reidentification algorithm through a large number of pedestrian re-identification videos and images, and it can achieve a test model that meets the required recognition accuracy. The test model referred to in Figure 2 is based on the establishment of a pedestrian reidentification algorithm based on deconvolution network feature extraction-multilayer attention mechanism convolutional neural network. It fine-tunes the pedestrian reidentification algorithm through a large number of pedestrian reidentification videos and images, and it can achieve a test model that meets the required recognition accuracy.

5. Example Analysis

5.1. Experimental Evaluation Criteria. To better evaluate the effect of pedestrian reidentification, this paper uses the rank-1 matching rate and mean average precision, which are commonly used performance evaluation indicators in pedestrian reidentification tasks. The rank-1 matching rate considers pedestrian search as a sorting and positioning problem. The mAP indicator is used to evaluate the performance of the pedestrian search model and can reflect the effect of pedestrian recognition. A detailed description of the relevant indicators can be found in [35].

5.2. Market1501 Data Set Experiment

5.2.1. Data Set Description. The Market1501 dataset is the earliest proposed large-scale pedestrian reidentification dataset [34]. The training set contains 750 pedestrians, for a total of 12,936 pedestrian images. The test set contains 751 pedestrians with a total of 19732 pedestrian images. Compared with other test datasets, the Market1501 dataset is larger and more pedestrian-rich, and the interclass and intraclass changes of the Market1501 dataset are more complex.

5.2.2. Experimental Process. The network in this experiment is pretrained on the ImageNet ILSVRC dataset and then fine-tuned on the Market1501 dataset. The distance interval parameter is set to 1. The number of training samples in the minibatch is set to 256. Each minibatch contains 64 randomly selected pedestrians, and each pedestrian corresponds to 8 randomly selected images. The fully connected layer is initialized with randomly generated values. The size of all input images is uniformly adjusted to 256×256. The default network learning rate is 0.001. The learning rate of the fully connected layer is 0.001. The final learned feature dimension is 256. 256-dimensional features are more conducive to large-scale pedestrian reidentification in practical applications. In addition, feature extraction uses the feature extraction model of the deconvolution network proposed in Section 2 of this paper. For the input region of each time step, the ROI pooling layer is applied on its conv4-3 convolutional feature map to normalize all feature maps to the same size of $28 \times 28 \times 1024$. For querying pedestrian images, this paper uses the same method to extract their $28 \times 28 \times 1024$ convolution features. These feature maps are then fed back into the architecture mentioned in Section 3 of this article. This paper uses Theano deep learning framework to implement the deep learning model mentioned in this paper. The basic configuration of this experimental platform is as follows: NVIDIA GeForce GTX GPU, Intel i7-7600 CPU, memory is 128GB.

5.2.3. Analysis of Experimental Results. The pedestrian reidentification algorithm proposed in this paper and other major popular people reidentification algorithms were used to identify the data set selected in this experiment. The specific results are shown in Table 1.

In general, according to Table 1, it can be seen that the algorithm proposed in this paper has better recognition performance than the traditional machine learning pedestrian recognition algorithm proposed in [36] and the deep learning pedestrian recognition algorithm proposed in [37, 38]. It proves the advantages of the proposed algorithm. Specifically, the traditional machine learning method proposed in [36] has the lowest mAP and rank-1 matching rate after pedestrian reidentification, which is 7% and 10% lower than other deep learning pedestrian reidentification methods, respectively. It shows that traditional machine learning pedestrian recognition methods are the worst of the categories listed above. The mAP indicators obtained by the deep learning pedestrian recognition methods proposed in [37, 38] are higher than 60%, and the rank-1 indicators are higher than 80%. They are each more than 7% higher than traditional machine learning methods. It shows that the deep learning pedestrian reidentification method proposed in [37, 38] has a significant improvement over traditional pedestrian reidentification methods. This is mainly because deep learning models can better train the experimental data and obtain better image segmentation models. The mAP value and rank-1 obtained by the method proposed in this paper are the highest of all methods, reaching 69.54% and 88.93%, respectively. It shows that the method proposed in this paper is highly adaptive to pedestrian images. This is mainly because the method proposed in this paper is more optimized than the deep convolution network feature extraction and multilayer attention mechanism theory introduced in the deep learning pedestrian reidentification model proposed in [37, 38]. The method proposed herein can better adapt to pedestrian images.

5.3. CUHK03-NP Data Set Experiment

5.3.1. Data Set Description. The CUHK03 dataset consists of 767 identities and 700 identities, respectively. In the test, this paper randomly selects an image from each camera as a query for each identity and uses other images to construct a data set. It can guarantee that both cameras have selected each query identifier, and it can realize cross-mirror search.

5.3.2. Experimental Process. In this experiment, the network was first pretrained on the ImageNet ILSVRC dataset and then fine-tuned on the CUHK03-NP dataset. The distance interval parameter is set to 1. The number of training samples in minibatch is set to 128. Each minibatch contains 32 randomly selected pedestrians, and each pedestrian corresponds to 4 randomly selected images. The fully connected layer is initialized with randomly generated values. The size

Table 1: Comparison of different methods of Market1501 dataset.

Type of method	mAP (%)	Rank-1 (%)
[36] (metric learning)	55.41	70.16
[37] (HSM-deep learning)	62.38	80.05
[38] (KISS+-deep learning)	63.72	81.11
Method of this paper	69.54	88.93

Table 2: Comparison of different methods of CUHK03-NP dataset.

Type of method	mAP (%)	Rank-1 (%)
[36] (metric learning)	62.19	63.32
[39] (optimization RCNN)	75.45	76.12
[40] (AACN)	79.99	80.50
Method of this paper	83.57	84.96

of all input images is uniformly adjusted to 512×512. The default network learning rate is 0.001. The learning rate of the fully connected layer is 0.001. The finally learned feature dimension is 256. The 256-dimensional feature is more conducive to large-scale pedestrian reidentification in practical applications. In addition, feature extraction uses the feature extraction model of the deconvolution network proposed in Section 2 of this article. For the input region of each time step, the ROI pooling layer is applied on its conv4-3 convolution feature map. It normalizes all feature maps to the same size of $14 \times 14 \times 512$. For querying pedestrian images, this paper uses the same way to extract their convolution features whose size is $14 \times 14 \times 512$. These feature maps are fed back to the architecture mentioned in Section 3 of this article. This paper uses Theano deep learning framework to implement the deep learning model mentioned in this article. The basic configuration of this experimental platform is as follows: NVIDIA GeForce GTX GPU, Intel i7-7600 CPU, memory is 128GB.

5.3.3. Analysis of Experimental Results. In order to better the advantages of the method mentioned in this article, the method mentioned in this article, the traditional machine learning method, the random initialization deep learning method (Recurrent convolutional neural network), and other mainstream deep learning methods are used to pedestrian the CUHK03-NP dataset identify. The specific results are shown in Table 2.

In general, according to Table 2, the pedestrian recognition effect of the proposed method is improved to a certain extent compared to the traditional machine learning method proposed in [36] and the deep learning method proposed in [39, 40]. It proves the advantages of the proposed algorithm. Specifically, the traditional machine learning method proposed in [36] has the lowest mAP and rank-1, which is 12% and 13% lower than other deep learning pedestrian reidentification methods, respectively. It shows that the traditional machine learning method has not been able to obtain a better recognition effect. The random initialization method proposed in [39] has lower mAP and rank-1 indicators than the optimized deep learning method proposed in [40]. It shows that the random initialization method cannot get the pedestrian recognition results well. The mAP value and rank-1 obtained by the method proposed in this paper are the highest among all methods, reaching 83.57% and 84.96%, respectively. It shows that the method proposed in this paper has better generalization ability than the deep learning method proposed in [39, 40]. The reason why it has better generalization ability is that the method proposed in this paper introduces the deconvolution network feature extraction and multilayer attention mechanism theory into the deep learning model. The introduction of these theories can better generalize different pedestrian images.

6. Conclusion

Due to the deep learning-based pedestrian reidentification method, the deep learning model learning process tends to fall into a local optimum, the model parameter learning method will appear gradient dispersion, and the information transfer of pedestrian reidentification sequence images is not considered. Therefore, this paper first uses a two-layer unsupervised deconvolution neural network to learn a feature map matrix from the original image and uses it as a convolution kernel for a deep convolutional network. Then, layer-by-layer convolution and pooling operations, which can be used to suppress gradient dispersion, are performed. At the same time, to solve the problem of information transmission of pedestrian reidentification sequence images, this paper proposes a memory network based on an attention mechanism to effectively store visual image information and pedestrian behavior information. Questions are then projected to both visual memory networks and behavioral memory networks to retrieve multimodal factual evidence, which can solve the problem of information transfer of the model. Based on these techniques, this paper proposes a pedestrian reidentification algorithm based on a deconvolutional network feature extraction-multilayer attention mechanism convolutional neural network.

The experimental results of the Market1501 and CUHK03-NP data set show that the deep learning-based pedestrian reidentification method proposed in this paper has the most satisfactory recognition effect. Not only is it far better than traditional machine learning pedestrian recognition methods but it is also a great improvement compared with other deep learning pedestrian recognition methods. The pedestrian recognition method proposed in this paper can obtain the best recognition results for the following reasons. First, the deep learning method proposed in this paper can better solve the problem of local optimization of deep learning models. The second is that the proposed method solves the gradient dispersion phenomenon of deep learning models. Third, the deep learning method proposed in this paper can better solve the problem of information transmission of pedestrian recognition images.

Data Availability

The data used to support the findings of this study are included within the paper.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Feng-Ping An did the conceptualization, funding acquisition, methodology, and wrote the original draft; Jun-e Liu did the software, supervision, and wrote the review and editing; Lei Bai did the validation and wrote the review and editing.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (No. 61701188), China Postdoctoral Science Foundation funded project (No. 2019M650512), Natural Science Foundation of Jiangsu Province (No. BK20201479), Scientific, technological innovation service capacity building—high-level discipline construction (city level), Beijing Intelligent Logistics System Collaborative Innovation Center (No. BILSCIC-2019KF-22), and Hebei IoT Monitoring Engineering Technology Research Center funded project (No. IOT202004).

References

[1] Y. Lin, L. Zheng, Z. Zheng et al., "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[2] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4500–4509, 2019.

[3] S. Salehian, P. Sebastian, and A. B. Sayuti, "Framework for pedestrian detection, tracking and re-identification in video surveillance system," in *2019 IEEE international conference on signal and image processing applications (ICSIPA)*, pp. 192–197, Kuala Lumpur, Malaysia, Malaysia, 2019.

[4] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 608–617, Long Beach, CA, 2019.

[5] F. Ma, X. Y. Jing, X. Zhu, Z. Tang, and Z. Peng, "True-color and grayscale video person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 115–129, 2019.

[6] Y. Sun, Q. Xu, Y. Li et al., "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 393–402, Long Beach, CA, 2019.

[7] A. Nanda, D. S. Chauhan, P. K. Sa, and S. Bakshi, "Illumination and scale invariant relevant visual features with hypergraph-based learning for multi-shot person re-identification," *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 3885–3910, 2019.

[8] W. Zajdel, Z. Zivkovic, and B. J. A. Krose, "Keeping track of humans: have I seen this person before?," in *Proceedings of the 2005 IEEE international conference on robotics and automation*, pp. 2081–2086, Barcelona, Spain, Spain, 2005.

[9] M. E. Monroe, N. Tolić, N. Jaitly, J. L. Shaw, J. N. Adkins, and R. D. Smith, "VIPER: an advanced software package to support high-throughput LC-MS peptide identification," *Bioinformatics*, vol. 23, no. 15, pp. 2021–2023, 2007.

[10] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3318–3325, Portland, America, 2013.

[11] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Vrstc: occlusion-free video person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7183–7192, Long Beach, CA, 2019.

[12] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2, pp. 1528–1535, New-York, America, 2006.

[13] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3586–3593, Portland, America, 2013.

[14] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2197–2206, Boston, America, 2015.

[15] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," *Bmvc*, vol. 2, no. 3, pp. 8–17, 2012.

[16] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1239–1248, Las Vegas, America, 2016.

[17] J. Zhou, P. Yu, W. Tang, and Y. Wu, "Efficient online local metric adaptation via negative samples for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2420–2428, Hawaii, America, 2017.

[18] J. Zhou, B. Su, and Y. Wu, "Easy identification from better constraints: multi-shot person re-identification from reference constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5373–5381, Salt Lake City, UT, 2018.

[19] W. Li, Y. Wu, and J. Li, "Re-identification by neighborhood structure metric learning," *Pattern Recognition*, vol. 61, pp. 327–338, 2017.

[20] J. Jia, Q. Ruan, G. An, and Y. Jin, "Multiple metric learning with query adaptive weights and multi-task re-weighting for person re-identification," *Computer Vision and Image Understanding*, vol. 160, pp. 87–99, 2017.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[22] B. Khagi, G. R. Kwon, and R. Lama, "Comparative analysis of Alzheimer's disease classification by CDR level using CNN, feature selection, and machine-learning techniques," *International Journal of Imaging Systems and Technology*, vol. 29, no. 3, pp. 297–310, 2019.

[23] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5934–5938, Calgary, AB, Canada, 2018.

[24] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, and M. Tiede, "Hybrid convolutional neural networks for

articulatory and acoustic information based speech recognition," *Speech Communication*, vol. 89, pp. 103–112, 2017.

[25] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159, Columbus, America, 2014.

[26] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 384–393, Hawaii, America, 2017.

[27] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1249–1258, Las Vegas, America, 2016.

[28] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–28, Hawaii, America, 2017.

[29] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proceedings of the IEEE international conference on computer vision*, pp. 4733–4742, Venice, Italy, 2017.

[30] H. Liu, Z. Jie, K. Jayashree et al., "Video-based person re-identification with accumulative motion context," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2788–2802, 2017.

[31] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–188, Munich, Germany, 2018.

[32] S. Bak, P. Carr, and J. F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 189–205, Munich, Germany, 2018.

[33] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019.

[34] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 740–749, Long Beach, CA, 2019.

[35] H. Cui, Z. Wu, G. Wu, X. Xu, Y. You, and Y. Fang, "Convolutional neural networks for electrical endurance prediction of alternating current contactors," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 9, pp. 1785–1793, 2019.

[36] A. TMF and S. Chaudhuri, "Maximum margin metric learning over discriminative nullspace for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 122–138, Munich, Germany, 2018.

[37] K. Chen, Y. Chen, C. Han, N. Sang, and C. Gao, "Hard sample mining makes person re-identification more efficient and accurate," *Neurocomputing*, vol. 382, pp. 259–267, 2020.

[38] H. Han, M. C. Zhou, X. Shang, W. Cao, and A. Abusorrah, "KISS+ for rapid and accurate pedestrian re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, pp. 1–10, 2020.

[39] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," *Computer Vision and Pattern Recognition*, pp. 3367–3375, 2015.

[40] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," *Computer Vision and Pattern Recognition*, pp. 2119–2128, 2018.