Hindawi
Computational Intelligence and Neuroscience
Volume 2019, Article ID 7028107, 12 pages
https://doi.org/10.1155/2019/7028107

Research Article

Multi-Information Flow CNN and Attribute-Aided Reranking for Person Reidentification

Haifeng Sang,[1] Chuanzheng Wang,[1] Dakuo He,[2] and Qing Liu[2]

[1]*School of Information Science and Engineering, Shenyang University of Technology, Shenyang, Liaoning 110870, China*
[2]*College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110004, China*

Correspondence should be addressed to Chuanzheng Wang; qing_0304@outlook.com

Received 7 November 2018; Accepted 8 January 2019; Published 6 February 2019

Academic Editor: Elio Masciari

This paper presents a multi-information flow convolutional neural network (MiF-CNN) model for person reidentification (re-id). It contains several specific multilayer convolutional structures, where the input and output of a convolutional layer are concatenated together on channel dimension. With this idea, layers of model can go deeper and feature maps can be reused by each subsequent layer. Inspired by an image caption, a person attribute recognition network is proposed based on long-short-term memory network and attention mechanism. By fusing identification results of MiF-CNN and attribute recognition, this paper introduces the attribute-aided reranking algorithm to improve the accuracy of person re-id further. Experiments on VIPeR, CUHK01, and Market1501 datasets verify the proposed MiF-CNN can be trained sufficiently with small-scale datasets and obtain outstanding accuracy of person re-id. Contrast experiments also confirm the availability of the attribute-assisted reranking algorithm.

1. Introduction

Person reidentification (re-id) refers to matching and recognizing the identities of pedestrians captured by multicameras with nonoverlapping views, which is significant to improve the efficiency of the security system. Owing to the low resolution of cameras, it is hard to obtain discriminative face features, so the current person re-id methods are mainly based on visual features of pedestrians, such as color and texture [1]. In practice, changes in viewpoint, pose, and illumination among different camera views, as well as partial occlusions and background clutters, pose a great challenge to person re-id [2].

Two principal person re-id methods are feature representation and metric learning. Feature representation seeks to find features with stronger discrimination and better robustness to represent pedestrians. Many kinds of features have been utilized for this, in which appearance features are the simplest and the most popular ones. Color, texture, and shape are the features that can be extracted for human appearance [3]in feature representation, such as HSV color histogram, LBP texture, and Gabor features, and then used for reidentifying people with similarity among pedestrian features. Attribute features are also widely used in person re-id. Common attributes include gender, length of hair, and clothing. These attributes are highly intuitive and understandable descriptors which have proved to be successful in several tasks, such as face recognition and activity recognition [4]. Although attribute features are complicated in terms of extraction and expression, they contain rich semantic information and are more robust to illumination and viewpoint changes. Therefore, the combination of attribute features and low-level features can effectively improve the accuracy of person re-id [5]. The metric learning methods employ the machine learning algorithm to learn a good similarity metric, which makes the feature similarity of the same pedestrian greater than that of different pedestrians.

In recent years, deep learning has shown great success in a variety of tasks in image classification and frequency domain [6], where CNN is particularly outstanding. Compared with the traditional methods, CNN has stronger feature learning ability, and the learned features are more

intrinsically representative to the original data, so it has better performance in extracting image features. Two types of CNN models are commonly employed in the community. The first type is the classification model as used in image classification and the second is the Siamese model using image pairs or triplets as input [7]. Most of the existing public datasets of person re-id only contain thousands of pedestrian image samples; a small number of training samples can easily lead to overfitting, which limits the performance of person re-id model. In addition, the deep neural networks for person re-id are similar in structure; that is, the feature maps extracted by convolutional layer are directly fed into the next convolutional layer [8–12]. Such structure usually ignores the correlation among features of each layer, thus reducing the mobility of feature information to some extent. In the process of back propagation, as the number of layers in neural network deepens, the gradient update information may attenuate in exponential form and cause vanishing gradient problem.

This work proposes to develop a modified deep neural network model for person re-id that could reduce overfitting caused by the lack of training samples. Moreover, this work aims to improve the identify accuracy of person re-id network with assistance of pedestrian attribute recognition. To this end, contribution of this paper is three-fold: first, this paper designs a multi-information flow convolutional neural network (MiF-CNN) to solve the person re-id problem. The network contains a series of multi-information flow convolution structures which connect the input and output of each convolutional layer together, realizes the reuse of features, and enhances the feature information flow and gradient back propagation of the entire network. Second, this paper designs a person attribute recognition network (PARN) based on long-short-term memory (LSTM) network and attention mechanism. The PARN decodes pedestrian visual features extracted by MiF-CNN into attribute features and outputs the attribute words of each person. Third, this paper presents an attribute-aided reranking algorithm which rematches attribute features among samples to aid more positive samples rank higher in rank list so as to improve the identify accuracy further.

The rest of this paper is organized as follows. Section 2 reviews the state of the art for person re-id. Section 3 introduces the details of MiF-CNN. Section 4 shows the principle of the PARN. The proposed attribute-aided reranking algorithm is detailed in Section 5. The experimental results and analysis are given in Section 6. Finally, conclusion and future works are discussed in Section 7.

2. Related Works

The early person re-id methods extract the manually designed features to represent pedestrians. Farenzena et al. divided pedestrian images into multiple areas and extracted three complementary kinds of features, weighted color histograms, maximally stable color regions, and recurrent high-structured patches. Then, match these features and measure the similarity between pedestrian pair [13]. Yang et al. proposed a novel salient color names based color descriptor (SCNCD), which was utilized to guarantee that a higher probability will be assigned to the color name near to the color. Based on SCNCD, color distributions in different color spaces were fused into feature representation for person [14]. Bazzani et al. proposed asymmetry-based HPE descriptor, which accumulated HSV histogram of multiple pedestrian images as a global appearance feature and detected patches portraying highly informative recurrent ingredient in local regions as local feature [15]. Wu et al. designed a novel gradient self-similarity (GSS) feature based on HOG to capture the patterns of pairwise similarities of local gradient patches. The combination of HOG and GSS achieved improvement in person re-id accuracy [16].

Apart from manually designed low-level features, attribute features that represent mid-level semantic information apply to person re-id as well. Compared with low-level descriptors, attributes are more robust to image translations [7]. Layne et al. labeled 15 binary attributes for the VIPeR dataset and trained SVM to detect attributes. They also learned a weighted L2-norm distance metric to fix each attribute and fused them with low-level visual features [17]. Wang et al. predicted complete attribute vector by exploiting both visual feature and marked attributes and obtained the overall ranking list by fusing the rank result from visual features and attribute vectors separately [18]. Chen et al. learned attribute of person by part-specific CNN and merged them with another identification CNN embedding in a triplet structure for person re-id task [19]. Wang et al. proposed a deep neural network that contains an auto-encoder model to learn hidden attributes of person from visual feature in an unsupervised manner, which alleviated the requirement of massive annotation [20].

Deep learning has become popular for solving person re-id problems in recent years. Ahmed et al. present a deep CNN architecture for person re-id. The architecture computed differences in feature values across the two views around a neighborhood of each feature location to add robustness to positional differences in corresponding features of the two input images [12]. Cheng et al. proposed a novel multichannel CNN. After the first layer of CNN, features were divided into four equal parts that aimed to learn features for the respective body part. The proposed CNN was trained with improved triplet loss function [21]. Lin et al. used ResNet [22] as the base network to learn low-level features and attributes jointly, and trained network with combining the person re-ID loss and attribute prediction loss [23]. Yan et al. proposed an attention block which learned par-level attention on different local regions, and integrated the proposed block into existing CNN structures for training with the identify loss [24]. Inspired by above works, this paper proposes a multi-information flow convolutional neural network to extract discriminative pedestrian features. In addition, this paper designs a person attribute recognition network based on LSTM and attention mechanism for improving person re-id results with assistance of attributes recognition.

3. Multi-Information Flow Convolutional Neural Network

The proposed MiF-CNN solves the person re-id problem with classification thought. The overall network structure is shown in Figure 1. The structure includes 2 shallow convolutional layers, 3 multi-information flow convolutional structures with novel connection pattern, fully connected layers, max pooling layers, and classification output layers. Low-level features of pedestrian images are extracted first by 2 shallow convolutional layers. After deeper multi-information flow convolutional structures, MiF-CNN extracted higher level features. The final discriminative pedestrian feature vectors are obtained after reducing dimensions by pooling layers and integrating by fully connected layers.

3.1. Features Extraction. In MiF-CNN, all convolutional filters are 3×3 with stride 1. Batch normalization and ReLU activation function are applied after each convolutional layer. The operation process of convolutional layer can be formulated as

$$\begin{cases} z_j^{(l)} = \sum_i x_i^{(l-1)} \otimes w_j^{(l)}, \\ x_j^{(l)} = \sigma\left(z_j^l\right), \end{cases} \quad l > 1, \tag{1}$$

where $x_i^{(l-1)}$ is the i-th feature map from the $(l-1)$-th convolutional layer, $z_j^{(l)}$ is the convolutional output of $x_i^{(l-1)}$, $x_j^{(l)}$ is the j-th feature map from the l-th convolutional layer, $w_j^{(l)}$ is the filter on the j-th feature map in the l-th convolutional layer, and $\otimes$ represents the convolutional operation. The process of convolutional layers extracting features is that neuron on the j-th feature map in the l-th convolutional layer sum each feature map after connecting and convolution by filter $w_j^{(l)}$, and map the extracted features on j-th feature map in the l-th convolutional layer. $\sigma(\cdot)$ is the ReLU activation function, which is formulated as $\sigma(x) = \max(0, x)$. Because of batch normalization, bias is ignored.

3.2. Multi-Information Flow Convolutional Structure. In this structure, both output and input of the current convolutional layer are concatenated together and fed into the next convolutional layer; i.e., the input of each layer is the connection combination of outputs from all previous layers. The detail of multi-information flow convolutional structures is shown in Figure 2.

This connection pattern makes feature maps of each layers be reused by all subsequent layers in forward propagation process, which makes the whole CNN model learn more feature information of pedestrian images. It can be considered as a special "Data Augmentation" in feature maps so as to enhance the information mobility of the network. In back propagation process, gradient of input in each layer contains derivative of loss function with respect to input, which makes propagation of gradient more effective and network easier to be trained. In multi-information flow convolutional structure, the number of feature maps that each layer outputs is a constant value ρ, so the number of feature maps that l-th layers outputs is $\rho_0 + \rho(l-1)$, where ρ_0 is the number of feature maps in the initial layer. Supposing the feature map of l-th channel in the initial layer is $x_i^{(0)}$, where $i \in (1, \rho_0)$, then the feature map of j-th channel that initial layer outputs can be expressed as

$$z_j^{(1)} = \sum_i x_i^{(0)} \otimes w_j^{(1)}, \tag{2}$$

where $w_j^{(1)}$ is the weight of the initial layer. The output after activation function is

$$a_j^{(1)} = \sigma\left(z_j^{(1)}\right), \tag{3}$$

where $j \in (1, \rho)$. The feature map that the l-th layer outputs can be expressed as

$$\begin{cases} z_q^{(l)} = \sum_p x_p^{(l-1)} \otimes w_q^{(l)}, \\ a_q^{(l)} = \sigma\left(z_q^{(l)}\right), \end{cases} \quad l > 0, \tag{4}$$

where $p \in (1,\ \rho_0 + \rho(l-1))$, $q \in (1, \rho)$. The output of the l-th layer after concatenate operation is

$$x_r^{(l)} = \left[x_p^{(l-1)},\ a_q^{(l)}\right], \quad l > 0, \tag{5}$$

where $r \in (1,\ \rho_0 + \rho \cdot l)$, $[\cdot, \cdot]$ represents the concatenate operation on channel dimension.

In the process of back propagation, supposing $\Delta x_r^{(l)}$ is the derivative of loss function with respect to $x_r^{(l)}$. Due to $x_r^{(l)}$ containing $x_p^{(l-1)}$ and $a_q^{(l)}$, it produces two parts of gradient as shown below:

$$\begin{cases} \Delta a_q^{(l)} = \Delta x_r^{(l)} \cdot \dfrac{\partial x_r^{(l)}}{\partial a_q^{(l)}}, \\ \Delta x_p^{(l-1)} = \Delta x_r^{(l)} \cdot \dfrac{\partial x_r^{(l)}}{\partial x_p^{(l-1)}}, \end{cases} \tag{6}$$

where $\Delta a_q^{(l)}$ is the gradient of output from the l-th layer after activation function. $\Delta x_p^{(l-1)}$ is the gradient of output from the $(l-1)$-th layer. The gradient of weight in the l-th layer is

$$\begin{cases} \Delta z_q^{(l)} = \Delta a_q^{(l)} \cdot \dfrac{\partial a_q^{(l)}}{\partial z_q^{(l)}} = \Delta x_r^{(l)} \cdot \dfrac{\partial x_r^{(l)}}{\partial a_q^{(l)}} \cdot \sigma'\left(z_q^{(l)}\right), \\ \Delta w_q^{(l)} = \Delta z_q^{(l)} \cdot \dfrac{\partial z_q^{(l)}}{\partial w_q^{(l)}} = \Delta z_q^{(l)} \cdot x_p^{(l-1)}, \end{cases} \tag{7}$$

where $\Delta z_q^{(l)}$ is the gradient of the convolution result in the l-th layer, $\sigma'(z_q^{(l)})$ is the derivative of the activation function with respect to $z_q^{(l)}$, and $\Delta w_q^{(l)}$ is the gradient of weights in the l-th layer. The network utilizes $\Delta w_q^{(l)}$ to update weights of each layer, which is formulated as

$$\left(w_q^{(l)}\right)_{\text{new}} = w_q^{(l)} - \eta \cdot \Delta w_q^{(l)}, \tag{8}$$

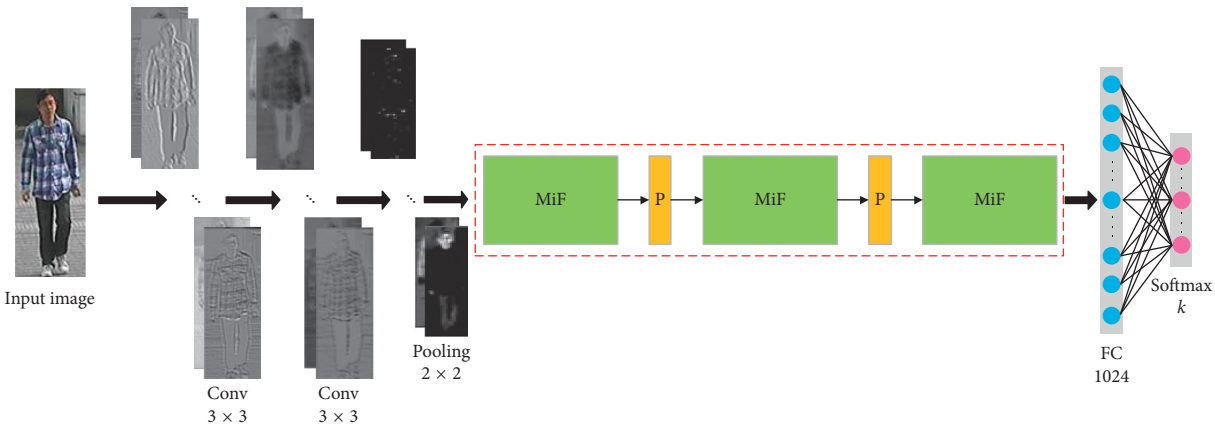


Figure 1: Architecture for MiF-CNN, where green rectangle denotes the multi-information flow convolutional structures, orange rectangle denotes the middle pooling layer, and k is the number of pedestrian categories in the training set.

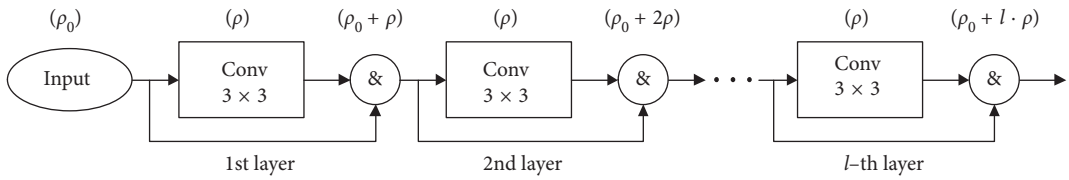


Figure 2: Multi-information flow convolutional structures, where "&" represents concatenate operation on channel dimension.

where η is the learning rate. Gradient keeps back propagation to the $(l-1)$-th layer. The gradient of $x_p^{(l-1)}$ is

$$\Delta x_p^{(l-1)} = \Delta z_q^{(l)} \cdot \frac{\partial z_q^{(l)}}{\partial x_p^{(l-1)}} = \Delta z_q^{(l)} \cdot w_q^{(l)}. \tag{9}$$

As shown in equations (6) and (9), the loss function produces two flows of gradient with respect to outputs of each convolutional layer, which makes error information propagating more effective in network and restrains the vanishing gradient to a certain extent.

The proposed MiF-CNN includes three multi-information flow convolutional structures. Between every two multi-information flow convolutional structures, a middle pooling layer is applied to compares the redundancy features. Hyperparameter ρ is set with a small value. When it comes to a new multi-information flow convolutional structure, ρ is doubled. Such a design makes each convolution layer learn a small quantity of features and reduce the redundant features so as to optimize the efficiency of network. With deeper layers, the network can learn more high-level and complex pedestrian features and improve the final identification accuracy.

3.3. Loss Function. Current deep learning algorithms usually use cross-entropy loss as cost function, which is formulated as

$$L_S = -\frac{1}{M}\sum_{i=1}^{M} y^{(i)} \log \frac{e^{\theta_{y(i)}^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}, \tag{10}$$

where $y^{(i)}$ is the ground truth of pedestrian categories in training set, θ is the parameter of the last fully connected layer, $x^{(i)}$ is the feature vector of training samples, k is the number of pedestrian categories in the training set, and M is the batch size.

However, in practice, when using cross-entropy loss merely, if the quality of extracted features is not good enough, it will lead to intraclass distance being greater than interclass distance. Aiming at this problem, Wen et al. proposed center loss in 2016 [25]. Combination of cross-entropy loss and center loss can enhance the discrimination and generalization ability of the network. Center loss is defined as follows:

$$L_C = \frac{1}{2M}\sum_{i=1}^{M} \left\| x_i - c_j \right\|_2^2, \tag{11}$$

where c_j is the center of the j-th pedestrian feature and x_i is the feature vector of pedestrian. Center loss minimizes the distance between feature and its center in order to reduce the intraclass distance. Center c_j is updated with equation (12):

$$\begin{cases} \Delta c_j^t = \dfrac{\sum_{i=1}^{M} \delta(y_i = j) \cdot \left(c_j^t - x_i\right)}{1 + \sum_{i=1}^{M} \delta(y_i = j)}, \\ c_j^{t+1} = c_j^t - \beta \cdot \Delta c_j^t, \end{cases} \tag{12}$$

where β is the update rate, $\delta(y_i = j)$ is 1 if prediction equals to ground truth, otherwise is 0. That is to say, center is updated only when network predicts correctly.

4. Person Attributes Recognition Network

The rank list of MiF-CNN is shown in Figure 3. In the incorrect identification results of rank 1 (pedestrian B and pedestrian C), there is a big difference in attribute features between the top-ranking negative samples and the query image, including gender, clothing, whether carrying handbag or not, and so on. Hence, this paper recognizes person attributes for improving accuracy of person re-id.

Based on Encoder-Decoder idea, Xu et al. [26] proposed a neural network model that can learn and generate the content of images. The model utilized CNN as encoder to extract features of images which were then fed into a recurrent neural network (RNN) for decoding into language captions of images. Inspired by that, this paper presents a person attribute recognition network (PARN) with LSTM and attention mechanism. The proposed PARN takes pedestrian features that are extracted by MiF-CNN as input and outputs the attributes information of pedestrian images. The architecture of PARN is demonstrated in Figure 4.

4.1. Input of PARN. In PARN, the input is the feature maps that are before the last fully connected layer in the MiF-CNN. The input feature is split into n feature vectors, each of which corresponds to a part of the image. Each feature vector is a D_X-dimensional vector which is represented as

$$X = \{x_1, x_2, ..., x_n\}, \quad x_i \in \mathbb{R}^{D_X}. \tag{13}$$

Referring to the natural language processing method, PARN transforms words in person attribute labels into word embedding. As a part of the input of PARN, each word embedding is a D_Y-dimensional vector:

$$y = \{y_1, y_2, ..., y_m\}, \quad y_i \in \mathbb{R}^{D_Y}, \tag{14}$$

where m is the number of attribute words in each pedestrian image and y_i is the word embedding corresponding to each attribute word.

For attribute words, common one-hot encoding considers each word as an individual, which ignores the correlation among words. However, word embedding represents each word as a continuous dense vector, which makes those correlative words closer in space.

4.2. Attention Mechanism. Attention mechanism has been widely used in natural language processing and computer vision. By measuring the correlation between the output and different parts of the input, attention mechanism gives different weights to different parts of the input, enabling the network to use more important feature information for prediction and reduce the dimension of input data [27].

In practice, a certain attribute of pedestrian is only corresponding to a certain part of the image. For example, when recognizing whether a pedestrian is wearing a hat, people only pay attention to the area above the head of the pedestrian usually, instead of other areas irrelevant to the attribute. Therefore, before feeding the image features into the LSTM network, attention mechanism is introduced to calculate the correlation between different positions of image features and the hidden state of LSTM at the previous time. The schematic diagram of attention mechanism is shown in Figure 5.

At time t, the fully connected layer and tanh function are used to integrate the information of the input feature vector and the hidden state of LSTM at the previous time, which is formulated as

$$v_i = \tanh\left(W_{\text{att}X} x_i + W_{\text{att}h} h_{t-1}\right), \tag{15}$$

where $W_{\text{att}X}$ and $W_{\text{att}h}$ represent the fully connected layer weights of x_i and h_{t-1}, respectively. The attention weights α_i is obtained by supplying softmax function on score vector v_i:

$$\alpha_i = \text{softmax}\left(W_{\text{att}v} v_i\right), \tag{16}$$

where $W_{\text{att}v}$ represents the fully connected layer weights of v_i. The output Z is the weighted sums of all α_i:

$$Z = \sum_{i=1}^{n} \alpha_i x_i. \tag{17}$$

Feature Z highlights the local features that are helpful to predict and suppress the other local features that make small contribution to prediction, which reduces the dimension of features to some extent and makes LSTM network focus on the part of input features that have greater correlation with prediction while recognizing person attributes so that improves the efficiency and prediction accuracy of the network.

4.3. LSTM. Normal RNN updates network parameters with back propagation, which easily suffers from the vanishing gradient when there is a long gap between relevant information and the current position to predict. LSTM network solves this problem well because of its own structure advantage. The architecture of LSTM unit is shown in Figure 6.

Here, c is the cell state for storing and transferring information, f is the forget gate which decides what should be abandoned in the cell state, i is the input gate which decides what should be stored in the cell state, g represents a vector of new candidate values that should be added to the cell state, o is the output gate which decides what parts of the cell state should be output to the next time, h is the hidden state of LSTM, x is the input of LSTM, and t denotes the current time.

In PARN, at any time t, the input of LSTM consists of two parts: word embedding y_{t-1} at the previous time and image features Z_t at the current time. The operation processing of LSTM in PARN is formulated as

$$\begin{cases} f = \text{sigmoid}\left(W_f \left[h_{t-1}, y_{t-1}, Z_t\right] + b_f\right), \\ i = \text{sigmoid}\left(W_i \left[h_{t-1}, y_{t-1}, Z_t\right] + b_i\right), \\ g = \text{sigmoid}\left(W_g \left[h_{t-1}, y_{t-1}, Z_t\right] + b_g\right), \\ o = \text{sigmoid}\left(W_o \left[h_{t-1}, y_{t-1}, Z_t\right] + b_o\right), \\ c_t = f \odot c_{t-1} + i \odot g, \\ h_t = o \odot \tanh\left(c_t\right), \end{cases} \tag{18}$$

Figure 3: Identification results of MiF-CNN (green bounding box represents positive samples in gallery).

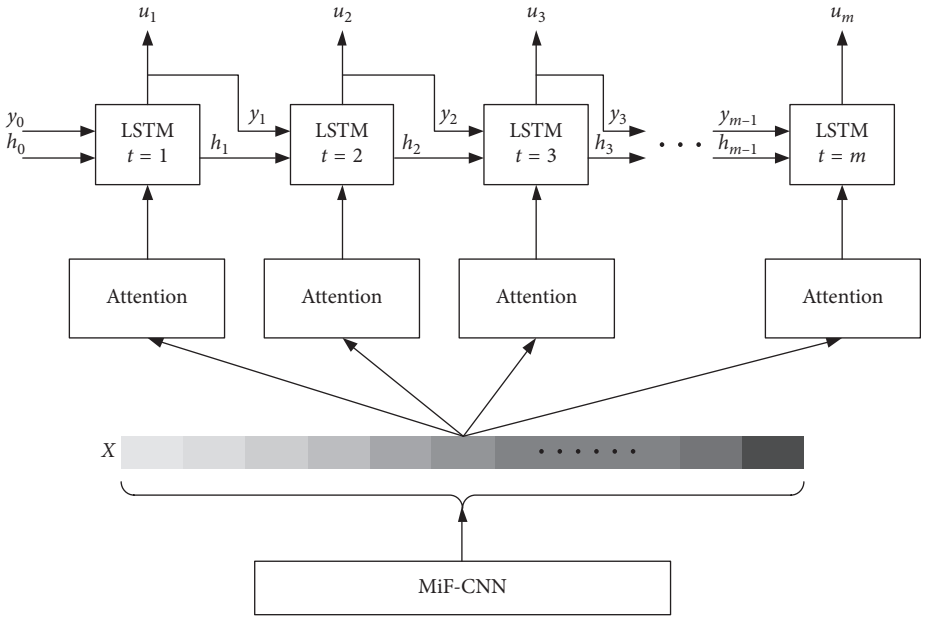


Figure 4: Architecture of PARN.

where W and b represent the weights and biases of each gate, respectively. Sigmoid function outputs a number between 0 and 1 to decide the quantity scale of values that go through these gates. Tanh function is the activation function of input. This design makes LSTM give different weights for information at different times so that it can choose what part of information to remember or forget in a long-term sequence.

The time step of LSTM in PARN is set as the number of attributes of each pedestrian image m. The output hidden state h_t is calculated by a fully connected layer and softmax function at each time step to predict pedestrian attribute words.

It is worth noting that, at each time step, the input word embedding of LSTM is the one that LSTM learned at the previous time step, which takes advantage of the LSTM when solving the long-term sequence problem. Because of the

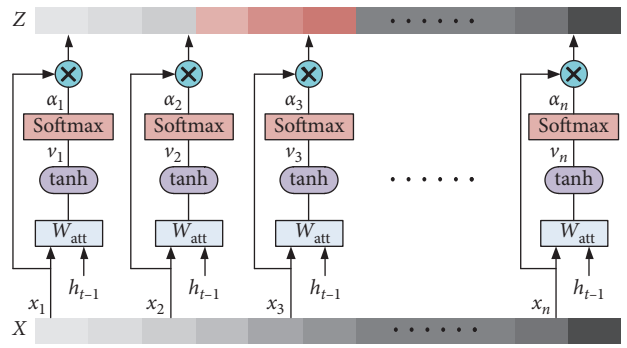


Figure 5: Schematic diagram of attention mechanism.

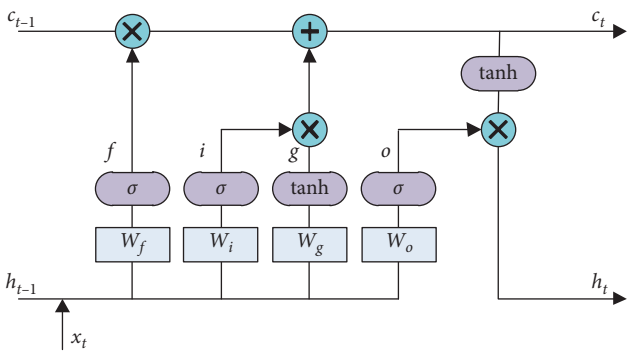


Figure 6: Architecture of the LSTM unit.

correlation among attributes of pedestrian, like the pedestrian owning attribute "*male*" generally own attribute "*short hair*" and attribute "*pants*," LSTM can utilize the previous attribute result to predict the current attribute more accurately.

4.4. Network Optimization. The loss function of PARN includes two parts, one is the cross-entropy between prediction and ground truth of network, which is expressed as

$$L_{\mathrm{sa}} = \sum_{i=1}^{m} u' \log\left(\frac{e^{\theta_{ai}^{T} u}}{\sum_{j=1}^{n_w} e^{\theta_{aj}^{T} u}}\right), \tag{19}$$

where u is the prediction value of the network, u' is the ground truth of pedestrian labels, m is the time step of LSTM, and n_w is the total number of attribute words in each dataset. As shown in equation (19), the loss of a single image in single epoch is the cumulated loss value after m iterations. The other one is the loss function of attention mechanism as described in [40]:

$$L_{\alpha} = \frac{1}{2n}\sum_{i=1}^{n}\left(1 - \sum_{i=1}^{m} \alpha_i^j\right)^2, \tag{20}$$

where α_i^j represents the attention weights of the image feature x_i and n is the number of parts in the image feature. The object function of PARN to optimize is

$$J = -\frac{1}{M}\sum_{i=1}^{M}\left(L_{sa}^{i} + \lambda \cdot L_{\alpha}^{i}\right), \tag{21}$$

where M is the batch size and λ is the rate of L_{α} in the object function.

5. Attribute-Aided Reranking Algorithm

Given a probe image q_0 and gallery image set $G = \{g_1, g_2, ..., g_N\}$ including N pedestrian images, the initial similarity distance between q_0 and g_i is computed with the Euclidean distance as

$$d(q_0, g_i) = \sqrt{\sum_{i=1}^{N}\left(x_{q_0} - x_{g_i}\right)^2}, \tag{22}$$

where x_{q_0} and x_{g_i} represent the features of q_0 and g_i, respectively, that extracted by MiF-CNN. The initial rank list $R_0 = \{g_1^{0'}, g_2^{0'}, ..., g_N^{0'}\}$ is obtained by sorting similarity distance $d(q_0, g_i)$ in ascending order, where $d(q_0, g_i^{0'}) < d(q_0, g_{i+1}^{0'})$. If the top-1 gallery $g_1^{0'}$ is the positive sample, it means rank 1 is correct. If the positive sample is not the top-1 gallery but the top-5 gallery, it means rank 5 is correct.

The attribute-aided reranking algorithm reranks the rank list R_0 according to the attribute feature similarity between q_0 and g_i so that more positive gallery gets higher ranking in R_0, which could improve the performance of person re-id. To be specific, when rank 5 is correct but rank 1 is not, the proposed algorithm distributes the initial feature score s_{f} for top-5 gallery $g_1^{0'}$~$g_5^{0'}$ according to their ranking in R_0, the higher the ranking, the higher the s_{f}. Then, the proposed algorithm distributes attribute score s_{a} for $g_1^{0'}$~$g_5^{0'}$ according to the number of their attributes that are the same as q_0, the more same attributes, the higher s_{a}. The total score of each gallery is $s = s_{\mathrm{f}}(1-\gamma) + s_{\mathrm{a}}\gamma$, where γ is the score weight. The reranking rank list R_0^* is obtained by reranking $g_1^{0'}$~$g_5^{0'}$ in descending order of their total score s. For M query images, the whole processing of the attribute-aided reranking algorithm is shown as Algorithm 1.

6. Experiments

This paper evaluates the attribute recognition accuracy of PARN on person re-id public datasets first and then compares our results with other person attribute recognition methods. Then, this paper evaluates the performance of MiF-CNN on three people re-id datasets and the availability of the proposed attribute-aided reranking algorithm for improving the accuracy of person re-id. The analysis of experiment results is also given after comparing our results with the state-of-the-art person re-id methods.

6.1. Datasets and Evaluation Protocol. This paper evaluates the proposed methods on three challenging person re-id

public datasets including VIPeR [28], CUHK01 [29], and Market1501 [30].

VIPeR Dataset. It contains 1264 images of 632 persons. Because of the lack of image samples, low-resolution, a great change in pose, view, and illumination, it is one of the hardest difficult person re-id datasets. In experiment, the VIPeR dataset is randomly split into two parts: images of 316 persons for training and the images of remaining 316 persons for testing.

CUHK01 Dataset. It consists of 3884 images of 971 persons which are captured by two cameras with different views. 485 pedestrians are randomly chosen for training and other 486 pedestrians for testing.

Market1501 Dataset. It is one of the largest person re-id datasets that include 32268 images of 1501 persons. Each person has a number of images that are captured by six cameras with different views. The dataset is divided into two parts: 12936 images of 751 persons for training and 19732 images of 750 persons for testing.

Evaluation Protocol. For the single-shot datasets VIPeR and CUHK01, cumulative match characteristic (CMC) is used to record the ranks of correct identify [31]. For the multishot dataset Market1501, apart from CMC, mean Average Precision (mAP) is utilized to evaluate the performance of the proposed methods.

6.2. Experimental Results on Attribute Recognition. The attributes in PARN refer to pedestrian identity-level attributes. For the VIPeR dataset, attributes include "*gender*," "*length of hair*," "*lower clothing*," "*upper clothing*," "*backpack or not*," and "*carrying anything or not*." For the CUHK01 dataset, attributes include "*gender*," "*length of hair*," "*backpack or not*," "*handbag or not*," "*color of upper*," and "*color of lower*." For the Market1501 dataset, attributes contain "*gender*," "*length of hair*," "*wearing hat or not*," "*lower clothing*," "*backpack or not*," "*handbag or not*," "*length of sleeve*," and "*length of lower clothing*." Each person in each dataset owns an attribute word list as shown in Figure 7.

This paper evaluates the attribute recognition accuracy of PARN on the Market1501 dataset and compares our results with outstanding person attribute recognition method APR [22]. The experiment results are reported in Table 1, where "*L.slv*" represents the "*length of sleeve*" and "*L.low*" represents the "*length of lower clothing*."

It can be observed from Table 1 that PARN obtains superior recognition accuracy of 8 attributes, especially "length of hair," "handbag or not," and mean accuracy are higher than APR, whereas accuracy of other attributes is closer with APR. Comparison with APR demonstrates the outstanding attribute recognition performance of PARN which plays an important role in improving the accuracy of person re-id.

6.3. Experimental Results on Person re-id. This paper evaluates the performance of the proposed methods on three datasets. The experimental results of the proposed methods and other state-of-the-art methods are presented in Table 2.

As shown in Table 2, the proposed MiF-CNN model obtains a great performance among state-of-the-art methods, and on the contrary, comparison between MiF and MiF + PARN demonstrates that the proposed attribute-aided reranking algorithm is helpful to increase the person re-id accuracy. The improvement effect is especially obvious on VIPeR and CUHK01 datasets, where the identification accuracy of rank 1, rank 5, and rank 10 improves by 6.02%, 6.33%, and 2.22% and 4.94%, 4.94%, and 2.26%, respectively. The proposed MiF-CNN model with the attribute-aided reranking algorithm gets the best rank 5 accuracy on the VIPeR dataset, the best rank 1 and rank 5 accuracy on the CUHK01 dataset, and the best mAP on the Market1501 dataset among various methods.

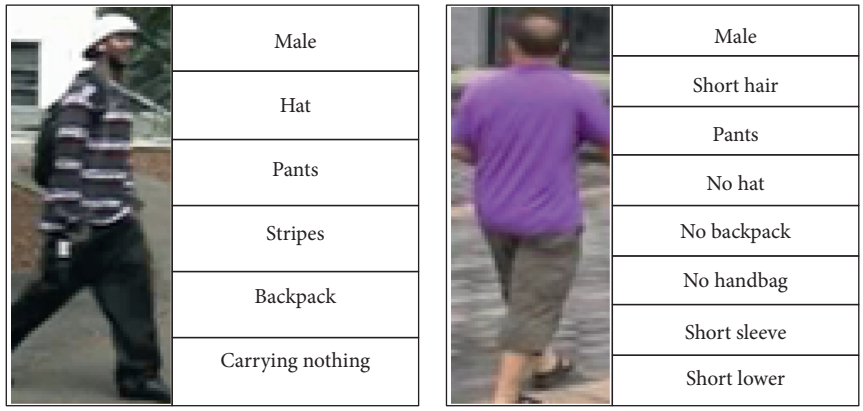
6.4. Analysis of Experimental Results. Because of the small quantity of training samples in VIPeR and CUHK01 datasets, a deep CNN model is hard to be trained and easily suffers from overfitting. To handle this, the FT-JSTL + DGD method based on deep CNN learned deep features from multiple domains jointly by merging all the datasets together and fine-tuned the pretrained model on VIPeR and CUHK01 separately. The structure and hyperparameters of the deep CNN model in the M3TCP method need to be adjusted manually for adapting the scale of different datasets so as to overcome the training problem of the deep neural network.

In contrast, the proposed MiF-CNN has an immobile structure and can be trained on smaller datasets directly without any fine tuning. In spite of the deep structure, the model can converge quickly and well. MiF-CNN without the attribute-aided reranking algorithm outperforms FT-JSTL + DGD and M3TCP by 0.27% and 13.17%, respectively, at rank 1 accuracy on the CUHK01 dataset. It also outperforms FT-JSTL + DGD by 2.22% at rank 1 accuracy on the VIPeR dataset, which indicates that the proposed MiF-CNN model has an excellent ability of reducing overfitting and generalization and an outstanding performance in person re-id.

Figure 8 demonstrates the rank list of MiF and MiF + PARN. It can be seen that the MiF-CNN with the attribute-aided reranking method enables the lower ranked positive sample in rank list of MiF-CNN obtain a higher ranking so as to improve the accuracy of person re-id, which verifies the effectiveness of the attribute-aided reranking algorithm for improving performance of person re-id model.

7. Conclusion

This paper studies and discusses the training problem of deep neural network in person re-id task, and the using of pedestrian attributes for further improving the accuracy of



(a) (b)

Figure 7: Attribute label of the pedestrian in datasets.

Input: Initial rank list R_0
Output: The reranking rank list R^*
(1) **for** $i = 1, 2, \ldots, M$ **do**
(2) **if** rank 1 correct
(3) $R^* = R_0$
(4) **end if**
(5) **elif** rank 5 correct
(6) Distributing initial feature score s_f for top-5 gallery $g_1^{i'} \sim g_5^{i'}$ according to their ranking in R_0
(7) Distributing attribute score s_a for $g_1^{i'} \sim g_5^{i'}$ according to the number of their attributes that are the same as q_i
(8) $s = s_f(1-\gamma) + s_a\gamma$
(9) Reranking $g_1^{i'} \sim g_5^{i'}$ in descending order of their total score s and obtain R^*
(10) **end if**
(11) **elif** rank 10 correct
(12) Changing $g_1^{i'} \sim g_5^{i'}$ to $g_1^{i'} \sim g_{10}^{i'}$, repeat 6–10
(13) **end if**
(14) **else**
(15) Changing $g_1^{i'} \sim g_5^{i'}$ to $g_1^{i'} \sim g_M^{i'}$, repeat 6–10
(16) **end if**
(17) **end for**

Algorithm 1: The attribute-aided reranking algorithm.

Table 1: Recognition accuracy of different attributes in PARN and APR (%).

Method	Gender	Hair	Clothes	Hat	Backpack	Handbag	L.slv	L.low	Mean
APR	85.78	83.46	91.36	88.21	86.32	76.01	94.12	92.64	87.23
PARN	84.93	79.10	89.35	96.67	83.86	85.18	92.84	88.45	87.55

person re-id. The proposed MiF-CNN model realizes the reuse of feature maps and gradient information, which enhances the feature mobility of network and improves the efficiency of gradient propagation. The designed person attribute recognition network uses an attention mechanism to measure the correlation between input feature maps and the hidden state of LSTM at previous time for reducing the dimension of features. It also employs LSTM to decode image features into pedestrian attributes. On the basis of the attribute recognition results, the attribute-aided reranking algorithm is presented, which rematches attribute features among samples to aid more positive samples rank higher in rank list so as to improve the identify accuracy further.

The experimental results on three public person re-id datasets indicate the outstanding performance of MiF-CNN model in person re-id. The attribute-aided reranking algorithm makes a major contribution to improve the accuracy of person re-id. In the future, more improvement and optimization will be done on pedestrian attribute recognition network. Moreover, the caption of person images or videos

Figure 8: Rank list comparison between MiF and MiF + PARN.

Table 2: Comparison of various methods with the proposed methods on three datasets (%).

Methods	VIPeR			CUHK01			Market1501			
	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10	mAP
DNS [32]	42.28	71.46	82.94	64.98	84.96	89.92	67.96	—	—	41.89
DLDA [33]	44.11	72.59	81.66	67.12	89.45	91.68	48.15	—	—	29.94
FT-JSTL + DGD [11]	38.6	—	—	66.60	—	—	—	—	—	—
PDC [34]	51.27	74.05	84.18	—	—	—	**84.14**	**92.73**	**94.92**	63.41
K-means-CNN [35]	46.50	69.30	80.70	53.50	82.50	91.20	—	—	—	—
Spindle [9]	**53.80**	74.10	83.20	—	—	—	76.90	91.50	94.60	—
PersonNet [36]	—	—	—	71.14	90.07	**95.00**	37.21	—	—	18.57
CSBT [37]	36.60	66.20	**88.30**	51.20	76.30	91.80	42.90	—	—	20.30
SDH-CNN [38]	—	—	—	—	—	—	58.12	68.50	80.82	48.20
M3TCP [21]	—	—	—	53.70	84.30	91.00	—	—	—	—
DM^3 [39]	42.70	74.30	85.10	49.70	77.30	86.10	75.80	89.10	92.40	53.20
MiF	40.82	68.35	81.01	66.87	85.39	89.51	78.92	86.46	89.28	66.00
MiF + PARN	46.84	**74.68**	83.23	**71.81**	**90.33**	91.77	79.39	88.95	91.36	**66.46**

"MiF" represents the MiF-CNN method, and "MiF + PARN" represents the MiF-CNN with the attribute-aided reranking method.

can be obtained by the LSTM model with natural language process ideas which make person re-id methods more significant.

Data Availability

The (attributes recognition accuracy on the Market1501 dataset and person re-id accuracy on VIPeR, CUHK01, and Market1501 datasets) data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61773105), the Natural Science Foundation of Liaoning Province (no. 20170540675), and the Scientific Research Project of Liaoning Educational Department (no. LQGD2017023).

References

[1] L. Li and J. Guo, "Pedestrian re-identification based on reinforced deep feature fusion," *Information Technology*, vol. 42, no. 7, pp. 15–19, 2018, in Chinese.

[2] J. Dai, Y. Zhang, H. Lu, and H. Wang, "Cross-view semantic projection learning for person re-identification," *Pattern Recognition*, vol. 75, pp. 63–76, 2018.

[3] T. T. T. Pham, T.-L. Le, H. Vu et al., "Fully-automated person re-identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method," *Image and Vision Computing*, vol. 59, pp. 44–62, 2017.

[4] P. DaoDao and H. J. Lee, "Deep multi-task network for learning person identity and attributes," *IEEE Access*, vol. 6, pp. 60801–60811, 2018.

[5] J. Li, L. Zhuo, J. Zhang, F. Li, and H. Zhang, "A survey of person re-identification," *Acta Automatica Sinica*, vol. 44, no. 9, pp. 1554–1568, 2018, in Chinese.

[6] L. Wu, Y. Wang, J. Gao et al., "Deep adaptive feature embedding with local sample distributions for person re-identification," *Pattern Recognition*, vol. 73, pp. 275–288, 2018.

[7] L. Li, Y. Yang, and A. G. Hauptmann, "Person re-identification: past, present and future," 2016, http://arxiv.org/abs/1610.02984.

[8] D. Wu, S. J. Zheng, C. A. Yuan, and D. S. Huang, "A deep model with combined losses for person re-identification," *Cognitive Systems Research*, vol. 54, pp. 74–82, 2018.

[9] H. Zhao, M. Tian, S. Sun et al., "Spindle net: person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1077–1085, Honolulu, HI, USA, 2017.

[10] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 28, Honolulu, HI, USA, July 2017.

[11] T. Xiao, H. Li, W. Ouyang et al., "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1249–1258, Las Vegas, NV, USA, June 2016.

[12] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3908–3916, Boston, MA, USA, June 2015.

[13] M. Farenzena, L. Bazzani, A. Perina et al., "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2360–2367, San Francisco, CA, USA, June 2010.

[14] Y. Yang, J. Yang, J. Yan et al., "Salient color names for person re-identification," in *Proceedings of Computer Vision-ECCV 2014*, pp. 536–551, Zurich, Switzerland, September 2014.

[15] L. Liao, M. Cristani, A. Perina et al., "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 898–903, 2012.

[16] S. Murino, R. Laganière, and P. Payeur, "Improving pedestrian detection with selective gradient self-similarity feature," *Pattern Recognition*, vol. 48, no. 8, pp. 2364–2376, 2015.

[17] R. Layne, T. M. Hospedales, S. Gong et al., "Person re-identification by attributes," in *Proceedings of the 23th British Machine Vision Conference (BMVC)*, vol. 2, no. 3, p. 83, Surrey, UK, September 2012.

[18] Z. Wang, R. Hu, Y. Yu, C. Liang, and W. Huang, "Multi-level fusion for person Re-identification with incomplete marks," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1267–1270, Brisbane, Australia, March 2018.

[19] Y. Chen, S. Duffner, A. Stoian et al., "Deep and low-level feature based attribute learning for person re-identification," *Image and Vision Computing*, vol. 79, pp. 25–34, 2018.

[20] Z. Dufour, X. Bai, M. Ye, and S. Satoh, "Incremental deep hidden attribute learning," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 72–80, Seoul, South Korea, October 2018.

[21] D. Cheng, Y. Gong, S. Zhou et al., "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1335–1344, Las Vegas, NV, USA, June 2016.

[22] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[23] Y. Lin, L. Zheng, Z. Zheng et al., "Improving person re-identification by attribute and identity learning," 2017, http://arxiv.org/abs/1703.07220.

[24] Y. Yan, B. Ni, J. Liu, and X. Yang, "Multi-level attention model for person re-identification," *Pattern Recognition Letters*, 2018, In press.

[25] Y. Wen, K. Zhang, Z. Li et al., "A discriminative feature learning approach for deep face recognition," in *Proceedings of 14th European Conference*, pp. 499–515, Amsterdam, Netherlands, October 2016.

[26] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2048–2057, Lille, France, July 2015.

[27] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, 2018.

[28] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of European Conference on Computer Vision Lecture Notes in Computer Science*, pp. 262–275, Marseille, France, October 2008.

[29] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proceedings of the 2008 Asian Conference on Computer Vision (ACCV)*, pp. 31–44, Daejeon, Korea, November 2012.

[30] L. Zheng, L. Shen, L. Tian et al., "Scalable person re-identification: a benchmark," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1116–1124, Boston, MA, USA, December 2015.

[31] W. Li, R. Zhao, T. Xiao et al., "DeepReID: deep filter pairing neural network for person re-identification," in *Proceedings of the 2014 Computer Vision and Pattern Recognition (CVPR)*, pp. 152–159, Columbus, OH, USA, June 2014.

[32] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition (CVPR)*, pp. 1239–1248, Las Vegas, NV, USA, June 2016.

[33] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on Fisher networks: a hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, pp. 238–250, 2017.

[34] C. Su, J. Li, S. Zhang et al., "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the 2016 International Conference on Computer Vision (ICCV)*, pp. 3980–3989, Venice, Italy, October 2017.

[35] Z. Zhao, B. Zhao, and F. Su, "Person re-identification via integrating patch-based metric learning and local salience learning," *Pattern Recognition*, vol. 75, pp. 90–98, 2018.

[36] L. Wu, C. Shen, and A. Hengel, "Personnet: person re-identification with deep convolutional neural networks," 2016, http://arxiv.org/abs/1601.07255.

[37] J. Chen, Y. Wang, J. Qin et al., "Fast person re-identification via cross-camera semantic binary transformation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5330–5339, Honolulu, HI, USA, July 2017.

[38] L. Wu, Y. Wang, Z. Ge et al., "Structured deep hashing with convolutional neural networks for fast person re-identification," *Computer Vision and Image Understanding*, vol. 167, no. 10, pp. 63–73, 2018.

[39] Z. Hu, R. Hu, C. Chen, Y. Yu et al., "Person reidentification via discrepancy matrix and matrix metric," *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 3006–3020, 2018.

[40] D. Jiang, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, http://arxiv.org/abs/1409.0473.