Deep Learning Research With an Expectation-Maximization Model for Person Re-Identification

FEI ZHOU, WENFENG CHEN, AND YANI XIAO

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Wenfeng Chen (s180101088@stu.cqupt.edu.cn)

ABSTRACT

In existing person re-identification methods based on deep learning, the extraction of good features is still a key step. Some efforts divide the image of a person into multiple parts to extract more detailed information from semantically coherent parts but ignore their correlation with each other. Others adopt self attention to reallocate weights of pixels for learning the association between different regions. This association can improve the accuracy of the person re-identification task, but the features obtained by this type of algorithm have high redundancy, which is not conducive to the expression of feature information. In order to address the above challenges, we propose a feature extraction method based on a novel attention mechanism which combines the expectation maximization (EM) algorithm and non-local operation. We embed the attention module into the ResNet50 backbone network. The attention module captures the correlation between different regional features through non-local operation and then reconstructs these features through the EM algorithm. In addition, we divide the network into a global branch and a local branch, where the global branch extracts the complete features, and the local branch uses the Batch DropBlock method to erase a portion of the features to achieve feature diversity. Finally, extensive experiments validate the superiority of the proposed model for person re-ID over a wide variety of state-of-the-art methods on three large-scale benchmarks, including DukeMTMC-ReID, Market-1501 and CUHK03.

INDEX TERMS

Person re-identification, deep learning, attention, non-local, expectation maximization, Batch DropBlock.

I. INTRODUCTION

Person re-identification (re-ID), also known as cross-border tracking, is one of the most important research endeavors in the field of computer vision, serving vital roles in video surveillance, intelligent security, pedestrian authentication, human-computer interaction and other fields. There are some differences between the recognition task and the detection task. The detection task [1], [2] devotes more attention to the category of the object, while the recognition task devotes more attention to the ID of each target. The purpose of person re-ID is to retrieve images of a person with the same identity as the query person from the large-scale person gallery in scenes with different perspectives, time and places acquired by multiple non-overlapping cameras. There are also many works on person identification technology from different sources, such as human faces, fingerprints, gaits, and movements. Compared with face recognition, the person re-identification scene is closer to the real environment, but it is more difficult and challenging to achieve under the dramatic variations with respect to illumination, occlusion, resolution, human pose, view angle, clothing, and background. In previous studies [3]–[5], the authors use gait and movement behaviors as dynamic features for person identification. These methods do not require high image resolution, so they can identify a person at a greater distance. Compared with gait and motion identification, the feature extraction process of person re-identification is simpler, because it only needs to extract static features of a person, such as clothing, hairstyle and backpack, etc. Current person re-ID research only uses cropped or detected person images to match query images. Therefore, extracting a rich representation for each input picture is one of the goals for person re-ID in the labeled dataset.

In addition, the person images captured by cameras in various areas have different backgrounds, so an effective mechanism is needed to direct attention to the person.
A small number of attention deep learning models for person re-ID have been recently developed for reducing the negative effects of poor detection and human pose change. Nevertheless, attention models trained by some deep learning methods generally have a high degree of redundancy. Others only focus on the attention selection of spatial or channel domain of features. Additionally, they often consider only the local feature information whilst ignoring the relationship between them.

Therefore, in order to alleviate the negative impact of the above problems on model performance, we propose a technique to capture inter-regional associations and reduce redundancy between features. The remainder of this paper is divided into the following sections. Related works are described in Section II, and the contribution of this study is described in Section III. The framework of the whole network and the details of the attention module are explained in Section IV. In Section V, we conducted multiple experiments on commonly used datasets to prove the effectiveness of the proposed approach. Finally, the conclusions of this study are presented in Section VI.

II. RELATED WORKS

At present, there are two main types of person re-ID research: feature-based [6] learning methods and measure-based [7] learning methods. Feature-based learning methods treat person re-ID as a classification task, which classifies a person with the same ID. Therefore, the main task of this kind of method is to learn more recognizable features from each person ID image, reducing the difficulty of classification. Measure-based learning methods measure the semantic similarity between the embedded features of the model by mapping the high-dimensional image of a person to the low-dimensional feature space. These methods can reduce the intra-class distance and increase the inter-class distance between features. In this paper, we focus on the former type of methods. Conventional methods describe features by manually designing feature descriptors. In recent years, with the development of deep learning [8], [9], methods based on deep convolutional networks can learn more discriminative features than manual features. In [10], [11], a ResNet-based network was proposed to extract the global features of the whole body and to measure the similarity of the extracted features. Among these methods, several efforts [12]–[15] locate a number of body parts first and then input each part into convolutional networks to extract discriminative part feature representations. In [15], the Part-based Convolutional Baseline (PCB) network was proposed, which divides the feature map into six parts horizontally on the conv-layer and forces each part to satisfy individual identification loss to acquire rich part feature representations. The Multiple Granularity Network [16] (MGN) is one of the best performing person re-ID networks among the part-based approaches. MGN uses multi-branch structure to combine global features and local features as the final feature representation. On local branches, MGN horizontally divides the feature maps into 2 or 3 parts for learning part-level features. These part-based methods can achieve very promising results due to the mining of more detailed features. However, these models require relatively well-aligned body parts for the same person ID in order to reduce external noise.

Various works [17]–[22] have studied attention mechanisms, which help neural networks mine more discriminative features and suppress useless features. In the person re-ID task, the attention mechanism mainly focuses on learning meaningful information for the task, enhancing the representation ability of features, and reducing the harmful effects of useless information, such as background and blocking. In [23], SENet learned the relations between feature channels and filtered out the features of the channel with the greatest response. In [24], spatial and frequency domain information is selectively fused to classify targets. In [25], the Mixed High-Order Attention Network employed high-order statistics for learning attention information. In addition, self-attention-based methods have also been used in computer vision tasks. The self-attention mechanism usually focuses on all positions of the feature map and takes its weighted average value in the embedded space as the response of the corresponding position in the image. For the self-attention method, non-local operation [26] can effectively capture the relationships between various parts of a person’s body, which is ignored by the previous methods. However, these methods, which involve the calculation of the correlation between each pixel and others in the image, will cause a sharp increase in the number of model parameters and are computationally intensive. In the past few years, some approaches have utilized various dropout-based data augmentations to enlarge the dataset, which is beneficial for discovering rich features. For example, Batch Drop-Block [27] achieved high performance by partially erasing the extracted features and jointly training with the original features. However, these previous studies only considered extracting features from images in various ways and did not consider that the redundancy generated in this process would harm the performance of the model.

Therefore, in this study, we present a method which combines the expectation maximization (EM) [28] algorithm and non-Local operation to resolve the above problem.

III. CONTRIBUTIONS

The main contributions of this paper are summarized as follows:

- To obtain the relationships between distant body parts in the person image, we propose a method using non-local operations to model the correlation between pixels.

- Inspired by the expectation maximization [28] algorithm, we use this algorithm to map the extracted features into some of the most representative feature descriptors, and use them to reconstruct the original features. This process can effectively filter out redundant features and useless information. In addition, we combine this process and non-local methods into an attention module. The input and output dimensions of the module are the same and can be easily inserted into the CNN.
In this study, we propose a novel two-branch lightweight architecture for discovering rich features of the person re-ID task. Specifically, we adopt the idea of Batch DropBlock, developing a feature erasing branch on the basis of the global branch, to ensure the diversity of features.

IV. EM ALGORITHM-BASED RE-ID NETWORK

A. OVERALL PROCEDURE OF PROPOSED METHOD

In this section, we describe in detail the proposed method used in this study. Figure 1 shows the overall network architecture, which includes a backbone network, a global branch (orange colored arrows) and a local branch (blue colored arrows). There are many options for the backbone network, among which VGG, Inception, ResNet and their improved versions are the most commonly used. The number of VGG layers can reach 19, but VGG has two disadvantages, namely: 1. excessive network parameters, which consume substantial disk space; 2. training is very slow. Inception has 22 layers, and it uses convolution kernels with different sizes in the same layer, which is beneficial to obtain richer feature information. The gradient disappearance problem of VGG and Inception will become very serious after the network depth increases. ResNet uses shortcut connection, which effectively relieves the problem of gradient disappearance, helps to increase the depth of the network and improves the accuracy of the model. Based on the above considerations, we use the residual convolutional neural network ResNet-50 [29] as the backbone network for feature extraction.

B. FEATURE EXTRACTION

Table 1 describes the detailed calculation of the layer parameters in the proposed network. Firstly, we convert the original image through a series of preprocessing steps. We resize the image into the shape of 384×128×3 and use it as an input to the whole network. In Table 1, convolution sets are expressed as (Conv1, Conv2_x, Conv3_x, Conv4_x, Conv5_x), where x represents the frequency of convolution. In order to describe the data flow more concisely, we no longer list BN and ReLU separately after Stage 1. In fact, they appear in every convolution process. ‘‘Up’’ and ‘‘dilated’’ in Table 1 represent upsampling and dilated convolution, respectively.

Due to the existence of the 3 × 3, 7 × 7 convolution kernel, feature maps gradually become smaller in the process of extracting features, which is not conducive for the attention module to capture detailed correlations of the feature map in the deep network. To this end, we modified the convolutions in Stage 4 of ResNet-50 to dilated convolutions [30] for a larger receptive field. It can be seen from Table 1 that the final feature maps of Stage 3 and Stage 4 have the same spatial size, 48 × 16 × 512, 48 × 16 × 2048, respectively, where the difference is that the number of output channels of Stage 4 is four times as large as that of Stage 3. Larger feature maps are beneficial for obtaining richer human body information. The network is divided into two branches after the feature map output by Stage 4 of ResNet-50 passes through a bottleneck block.
TABLE 1. Description of the structure of the proposed network.

| Layer name          | frequency | of filters | Filter size (stride) | Output size   |
|---------------------|-----------|------------|---------------------|---------------|
| Input layer         | ×1        |            |                     | 384×128 × 3   |
| Stage 1             |           |            |                     |               |
| Conv1               | ×1        | 64         | 7×7(2)              | 192×64×64     |
| BN                  | ×1        |            |                     |               |
| ReLU                | ×1        |            |                     |               |
| Max pool            | ×1        |            | 3×3(2)              | 96×32×64      |
| Stage 2             |           |            |                     |               |
| Conv2_x             | 64        | 1×1(1)     | 96×32×64            |
| Stage 3             |           |            |                     |               |
| Conv3_x             | ×4        | 128        | 1×1(1)              | 96×32×128     |
|                     |           | 128        | 3×3(2)              | 48×16×128     |
|                     |           | 512        | 1×1(1)              | 48×16×512     |
| Stage 4             |           |            |                     |               |
| Up                  | 256       | 1×1(1)     | 48×16×256           |
| Conv4_x (dilated)   | ×6        | 256        | 96×32×256           |
|                     |           | 1024       | 48×16×1024          |
|                     |           | 512        | 1×1(1)              | 48×16×512     |
| Up                  | 512       |            | 96×32×512           |
| Conv5_x (dilated)   | ×3        | 512        | 3×3(2)              | 48×16×512     |
|                     |           | 2048       | 1×1(1)              | 48×16×2048    |
| Bottleneck          | ×1        | 512        | 1×1(1)              | 48×16×512     |
|                     |           | 512        | 3×3(1)              | 48×16×512     |
|                     |           | 2048       | 1×1(1)              | 48×16×2048    |
| GAP (Global Branch)/GMP (Local Branch) | ×1 | 2048 | 1×1×2048 |

The global branch provides a global feature representation across the network structure and also supervises the training of the local branch. It calculates the average value of all pixels on each feature map through a global average pooling (GAP) layer to produce a 2048-dimensional feature vector, and then passes through a feature dimension reduction module to reduce the dimensionality of the feature vector. The feature dimension reduction module consists of a 1×1 convolution layer, batch normalization (BN) layer, and a ReLU activation layer. The final feature dimension is reduced to 512, which is jointly trained with triplet loss, label smooth cross entropy loss and center loss.

The local branch uses a Batch DropBlock module to erase a certain percentage of the feature map after the bottleneck layer. Different from the ordinary random erase method, Batch DropBlock erases the same area of the features which come from the same batch for finding semantic correlations. We use the global max pooling (GMP) layer instead of the global average pooling layer to produce a 2048-dimensional max feature vector. GMP can identify significant areas of unerased features here. The local branch feature is reduced to 512 dimensions after dimensionality reduction. Similar to the global branch, we use the triplet loss, label smooth cross entropy loss and center loss to jointly train the local branch features.

Finally, the 512-dimensional feature vectors from the two branches are concatenated as the final 1024-dimensional embedding feature for the inference phase of the person re-ID task.

C. ATTENTION MODULE

The convolutional neural network is the mainstream method for person re-ID tasks. In order to extract rich feature information from the receptive field, it stacks multiple convolutional layers to obtain a larger receptive field. However, it is difficult for the convolution to capture the dependencies of distant areas on the feature map, which is exactly what the task needs. Non-local operation quickly captures the long-distance dependencies by directly calculating the correlations between two positions on the feature map. It can be formulated as:

\[ y_i = \frac{1}{C(x)} \sum_{j} f(x_i, x_j) g(x_j). \]  

where \( x \) represents the input feature vector, function \( f(\cdot, \cdot) \) calculates the similarity between pixel \( i \) and \( j \), function \( g \) calculates the representation of the feature map on the pixel \( f \) and \( C(x) \) is the normalization factor. The final \( y_i \) represents the weighted average of all pixels, except pixel \( i \), which are transformed by the function \( g \), and the weight is the normalized similarity function \( f(\cdot, \cdot) \).

The attention module in this paper introduces a second-order statistical covariance instead of a simple dot-product operation to calculate the similarity between pixels, because covariance can obtain richer inter-regional...
information than dot-product operation. However, the covariance inevitably results in some redundant features, which will affect the expression of effective features. In order to reduce the redundancy of the features while retaining the valid feature information, we adopt the reconstruction idea to map the features in high-dimensional space to the low-dimensional manifold. We then reconstruct the features through a few of the most representative feature descriptors to obtain more compact low-rank features than the original features. The reconstruction process can filter redundant features and effectively restore valid information.

In this paper, we use the EM algorithm to reconstruct redundant features. Only a small portion of the complex high-dimensional data has a beneficial effect on the person re-ID task, so this information deserves more attention. The EM algorithm is a heuristic iterative method. It aims to maximize the posterior probability or likelihood value for latent variables models, and then obtain the optimal estimation of parameters. Suppose that $X = \{x_1, x_2, \ldots, x_N\}$ is the collection of feature information obtained, which consists of $N$ observed samples, where each feature point $x_i$ has its corresponding latent variable $z_i$. $(X, Z)$ is the complete data and its likelihood function is $\ln p(X, Z|\theta)$, where $\theta$ is the set of all parameters in the model. In fact, the knowledge of latent variables in $Z$ comes from the posterior distribution $p(Z|X, \theta)$. The EM algorithm maximizes the likelihood value of $\ln p(X, Z|\theta)$ by two steps, separately calculating expectation and maximizing expectation, which is given by:

$$\max_{\theta} E_\theta \ln p(X, Z|\theta).$$

Before the iteration, we firstly select the initial value of the parameter $\theta^{(0)}$. We denote $\theta^{(i)}$ as the estimated value of the $i$-th iteration parameter, and in step E, we calculate the posterior distribution of the latent variable $Z$ according to $\theta^{(i)}$ at the $(i + 1)$-th iteration. Then we calculate the expectation of the complete data likelihood $Q(\theta, \theta^{(i)})$, as following in formula (3),

$$Q(\theta, \theta^{(i)}) = E_\theta [\ln p(X, Z|\theta)]|X, \theta^{(i)}] = \sum_Z \ln p(X, Z|\theta) p(Z|X, \theta^{(i)}).$$

where $p(Z|X, \theta^{(i)})$ represents the conditional probability distribution of the latent variable while feature information $X$ and $i$-th parameter estimation $\theta^{(i)}$ are given. Step M updates the parameters by maximizing the expectations obtained in step E to obtain the $(i + 1)$-th parameter estimate value $\theta^{(i+1)}$, which is given by formula (4),

$$\theta^{(i+1)} = \arg \max_\theta Q(\theta, \theta^{(i)}).$$

The EM algorithm executes step E and step M repeatedly until the convergence state is reached. In this paper, we regard the attention information in the features as latent variables in the EM algorithm. We firstly estimate the attention information, then maximize the likelihood based on the observed features and attention information to obtain the current model parameters (feature descriptors). After a few iterations, the EM algorithm reaches the convergence state when the change of model parameters becomes very small. At this time, the feature descriptors are the most representative, and the features reconstructed with attention information are more robust than those learned with convolutional networks.

As shown in Figure 2, we divide the proposed attention module into two stages: F and B. Because there are certain relationships between the various parts of the human body, Stage F introduces second-order statistical covariance to capture the relations between long-distance regions on the feature map. Given the input feature map $X \in R^{h \times w \times c}$, $h$ and $w$ are the height and width of the feature map, while $c$ represents the number of channels. We compress the spatial dimension of $X$ to the single dimension to become $X \in R^{h w \times c}$. We then build two functions of $\theta(x)$ and $g(x)$ by a $1 \times 1$ convolution of batch normalization layer and ReLU activation layer to obtain feature maps with shape $hw \times c'r$, where $r$ is a reduction factor of the feature channel. The covariance matrix is then calculated with $\theta(x)$, as following in formula (5),

$$\sum = \theta(x) \tilde{T} \theta(x)^T.$$  

where $\tilde{T} = \frac{1}{c'r^2}(I - \frac{1}{c'r^2})$. $I$ is the identity matrix with shape $c'r \times c'r$, and the shape of $\tilde{T}$ is $hw \times hw$. We adopt $\frac{1}{c'r}$ as the scaling factor for the covariance matrix, then normalize the covariance matrix by applying $softmax$. We multiply the normalized result by $g(x)$ to obtain $X'$ as:

$$X' = softmax\left(\frac{\sum}{\sqrt{c'r}}\right) g(x).$$
While modeling the second-order statistics, Stage F also produces substantial redundant feature information, which negatively impacts the person re-ID task. To this end, we introduce the expectation maximization algorithm to reconstruct the features of Stage F with a small number of feature descriptors, so as to obtain the effective features of low redundancy. Stage B consists of three steps: probability estimation (E), likelihood maximization (M) and feature reconstruction. We regard the mapping matrix $Z$ as the latent variables in the model and $K$ feature descriptors as model parameters. The input given in Stage B is the feature map $X' \in R^{hw \times c'r}$, and the initial values of descriptors are $\mu \in R^{k \times c'r}$. Step E updates the mapping matrix (attention map) $Z$, which is formulated as:

$$Z = \text{softmax}(\lambda X' (\mu^T)). \quad (7)$$

where $\lambda$ is used as a hyperparameter to control the distribution of $Z$, and the default value is 1. Step M updates the descriptors $\mu$ by maximizing the complete data likelihood, where $\mu$ is counted as the weighted average of $X'$. The $k$-th descriptor is formulated as:

$$\mu_k = \frac{\sum_{r=1}^{hw} \sum_{n=1}^{hw} z_{r,n} X_r'}{\sum_{n=1}^{hw} z_{r,n}}. \quad (8)$$

Step E and step M alternately execute $T$ times until $\mu$ and $Z$ reach the state of approximate convergence. The final $\mu$ and $Z$ are used to reconstruct $X'$ to obtain $X''$ as:

$$X'' = Z \mu. \quad (9)$$

Finally, we use $1 \times 1$ convolution to restore the dimension of $X''$ from $hw \times c'r$ to $h \times w \times c$, and add it to the original feature map $X$ to obtain $X'''$, which is formulated as:

$$X''' = X + X''. \quad (10)$$

The pseudocode of this attention module is provided as Algorithm 1.

**Algorithm 1 Processing Phase of Attention Module**

**Input:** The original feature map, $X \in R^{h \times w \times c}$; Set of feature descriptors, $\mu$; Number of iterations, $T$; Reduction factor of channel, $r$; $T$.

**Output:** Features processed by the attention module, $X''$.

1. Initial $T = \frac{1}{c/r}(I - \frac{1}{r})$, $\mu \in R^{k \times c'r}$, $\lambda = 1$;
2. $X \in R^{hw \times c} \leftarrow X \in R^{h \times w \times c}$;
3. $\theta(x) \in hw \times c/r, g(x) \in hw \times c/r \leftarrow \text{Conv1} \times 1, \text{BN}, \text{ReLU}$;
4. Covariance $\sum \theta(x)T \theta(x)^T \leftarrow \theta(x)$;
5. $X' = \text{softmax}(\lambda X' (\mu^T)) \leftarrow X' \in R^{hw \times c/r}, \mu \in R^{k \times c/r}$;
6. for $t = 1$ to $T$ do
7. \hspace{1em} E: $Z=\text{softmax}(\lambda X' (\mu^T)) \leftarrow X' \in R^{hw \times c/r}, \mu \in R^{k \times c/r}$;
8. \hspace{1em} M: $\mu_k = \frac{\sum_{r=1}^{hw} \sum_{n=1}^{hw} z_{r,n} X_r'}{\sum_{n=1}^{hw} z_{r,n}} \leftarrow Z$;
9. \hspace{2em} end for
10. $X'' = Z \mu \leftarrow \text{final } Z, \mu$;
11. $X'' \in R^{h \times w \times c} \leftarrow X'' \in R^{hw \times c/r}$;
12. $X''' = X + X''$;

to a class, so we use the cross entropy loss for optimization. Cross entropy describes the distance between two probability distributions, defined as:

$$L_{id} = -\sum_{k=1}^{K} \ln(p(x))q(k). \quad (12)$$

where $k \in 1, 2, \ldots, K$ represents the person category, $p(k)$ represents the probability that the input image belongs to category $k$, and $q(k)$ is the actual probability.

The center loss reduces the distance between similar samples, formulated as

$$L_{center} = \frac{1}{2N} \sum_{i=1}^{N} ||x_i - c||_2^2. \quad (13)$$

where $c$ is the sample center. The final loss is the weighted sum of the above three loss functions, namely,

$$L_{total} = L_{triplet} + \gamma_i L_{id} + \gamma_c L_{center}. \quad (14)$$

where $\gamma_i$, $\gamma_c$ are weighting factors.

**V. EXPERIMENTS**

**A. DATASETS AND EVALUATION CRITERIA**

For better evaluation, we selected three large-scale person re-ID benchmarks: Market-1501 [36], DukeMTMC-ReID [37] and CUHK03 [38].

The Market-1501 dataset includes 1,501 identities collected by six cameras at Tsinghua University. The training set includes 12,936 training images with 751 different identities. Gallery and query sets have 19,732 and 3,368 images, respectively, with another 750 identities.

The DukeMTMC-reID dataset includes 16,522 training images of 702 identities, 2,228 query and 17,661 gallery images of another 702 identities.

**B. LOSS FUNCTION**

We adopt multiple loss functions to jointly train the feature vectors of two branches, which are soft margin batch-hard triplet loss [33], smooth label cross entropy loss [34] and center loss [35]. Triplet loss is formulated as:

$$L_{triplet} = \frac{1}{P} \sum_{j=1}^{P} \sum_{a=1}^{K} \left[ \max_{p=1, \ldots, K} D(f_0(x'_a), f_0(x'_p)) \right]^{hardest \ positive} - \min_{j=1, \ldots, P} D(f_0(x'_a), f_0(x'_n)). \quad (11)$$

where $D(f_0(x'_a), f_0(x'_p))$ and $D(f_0(x'_a), f_0(x'_n))$ represent the distance between the anchor sample and the positive samples and negative samples, respectively, and $m$ is the margin for loss.

The person re-ID problem can be transformed into a classification problem: that is, each person ID is equivalent
The CUHK03 dataset contains 14,096 labeled images and 14,097 detected images of 1,467 identities captured by two camera views. The labeled set contains 7,368 training, 1,400 query and 5,328 gallery images, respectively. The detected CUHK03 set includes 7,365 training images, 1,400 query images and 5,332 gallery images. The dataset parameters are shown in Table 1.

To evaluate the performance, we use the cumulative matching characteristic (CMC) and mean average precision (mAP) metrics. The horizontal coordinate of the CMC curve is the Rank value, among which Rank-1, Rank-5, Rank-10, etc. are commonly used. The vertical coordinate is the matching rate, which represents the probability of finding the correct ID in the first k matching results. It can be calculated as follows: if the top k candidate samples from the returned list contain correctly matched images, then the accuracy rate is 1; otherwise, the accuracy rate is 0.

The second indicator, mAP, regards the person re-ID problem as an object retrieval problem and can evaluate the performance of the whole model. mAP is calculated on the basis of precision (P) and average precision (AP). P refers to the proportion of the query image in the retrieved image. AP measures the quality of the trained model with a single category, and mAP measures the quality of the model with all categories. AP is formulated as:

$$AP = \frac{1}{N} \sum_{k=1}^{N} k / M_k.$$  

where k represents the number of query images, $M_k$ represents the total number of images when the k-th matching image of the query picture is retrieved, and N indicates the number of matching items of the query image. The mAP is the average of the AP, formulated as

$$mAP = \frac{1}{Q} \sum_{k \in Q} AP(k).$$

where Q is the total number of images in the query set.

In addition, we utilized the receiver operating characteristic (ROC) curves and equal error rate (EER) for performance evaluation. The ROC curve is plotted with (true positive rate) TPR against the (false positive rate) FPR, where TPR is on the y-axis and FPR is on the x-axis. We could evaluate a binary classification model with different classification thresholds and then obtain the ROC curve. FPR and TPR are formulated as:

$$FPR = \frac{FP}{TN + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

where FP (false positive) represents a positive predicted value which is actually false. TP (true positive) represents a positive predicted value which is actually true. FN (false negative) represents a negative predicted value which is actually true. TN (true negative) represents a negative predicted value which is actually true.

The equal error rate (EER) is the point on the ROC curve that corresponds to an equal probability of misclassifying a positive or negative sample. This point is obtained by intersecting the ROC curve with a diagonal of the unit square.

**TABLE 2. Introduction to the experimental datasets.**

| Dataset      | Time  | ID Number | Picture Number | Camera Number |
|--------------|-------|-----------|----------------|---------------|
| Market1501   | 2015  | 1501      | 32668          | 6             |
| DukeMTMC-reID| 2017  | 1812      | 36441          | 6             |
| CUHK03       | 2014  | 1467      | 13164          | 10            |

In the experiment, we use the cross-validation method to prevent the model from overfitting. The detailed contents of the cross-validation method are to first divide the dataset into a training set and a test set, and then divide a small part from the training set as a validation set to fine-tune the model. The person IDs in the training set and the test set do not overlap. After every 10 epochs of training, the program will use the validation set to verify the performance of the model and fine-tune it. In the testing phase, we use the test set to verify the final performance of the model. In this way, we can obtain a proper model with good generalization performance.

**C. EXPERIMENTAL ANALYSIS**

We evaluate the effectiveness of the proposed network on three datasets. The attention module is divided into two stages, F and B, where F represents Non-Local with covariance, and B represents the process of feature reconstruction with the EM algorithm. Figure 3 shows the attention map generated by the EM algorithm in the attention module. It can be seen that the EM algorithm guides the attention of the model to the person’s body, which ignores the disturbance caused by the background information. The feature descriptor $\mu$ generated in the iterative process is orthogonal, which ensures that the redundancy of the features is minimized. Experimental verification shows that the highest accuracy
TABLE 3. Comparison of splitting experiments of attention module on DukeMTMC-reID and Market501 datasets.

| Method       | Dataset          | DukeMTMC-reID | Market1501 |
|--------------|------------------|---------------|------------|
|              |                  | mAP | Rank -1 | Rank -5 | mAP | Rank -1 | Rank -5 |
| ResNet50     |                  | 76.9 | 87.9    | 94.6    | 85.9 | 94.8    | 98.0    |
| ResNet50 + F |                  | 78.0 | 88.9    | 94.7    | 86.2 | 95.0    | 98.2    |
| ResNet50 + B |                  | 78.6 | 89.4    | 95.3    | 87.0 | 95.1    | 98.3    |
| ResNet50 + F + B |            | **78.8** | **89.9** | **95.6** | **87.1** | **95.4** | **98.5** |

TABLE 4. Comparison of splitting experiments of attention module on CUHK03 dataset.

| Method       | Dataset          | Labeled | Detected |
|--------------|------------------|---------|----------|
|              |                  | mAP | Rank -1 | Rank -5 | mAP | Rank -1 | Rank -5 |
| ResNet50     |                  | 67.5 | 69.8    | 84.1    | 63.1 | 66.0    | 88.4    |
| ResNet50 + F |                  | 69.1 | 70.7    | 85.6    | 66.5 | 69.4    | 90.3    |
| ResNet50 + B |                  | 70.3 | 72.3    | 86.6    | 67.2 | 69.2    | 90.5    |
| ResNet50 + F + B |            | **70.9** | **72.4** | **87.8** | **67.3** | **69.9** | **90.3** |

FIGURE 3. Attention visualization effect diagram generated by step E in the EM algorithm.

FIGURE 5. The CMC curve of the attention part on the labeled CUHK03 dataset.

FIGURE 4. The CMC curve of the attention part on the DukeMTMC and Market1501 datasets.

FIGURE 6. The CMC curve of the attention part on the detected CUHK03 dataset.

can be achieved when the number of feature descriptors $K$ is 160 and the number of EM iterations $T$ is 3.

We compared the impact of different parts of the attention module on network performance. We can see from Table 3 that both Stage F and Stage B offer a certain degree of accuracy improvement on the network. For the DukeMTMC-reID dataset, our model with only F improved the performance by 1.1% and 1.0% for mAP and Rank-1 metrics, respectively. This proves that the feature information extracted after the introduction of covariance has stronger expression ability than the original features. When the base model only added Stage B, mAP and Rank-1 increased by 1.7% and 1.5% as compared with the baseline, respectively. These results show that the features reconstructed with the EM algorithm have certain validity for model optimization. Our model with a complete attention module achieved better...
results than the split module: 78.8% for mAP and 89.4% for Rank-1. Similar to the results on the DukeMTMC-reID dataset, the use of the attention module on Market1501 and CUHK03 also provides an improvement in accuracy as compared with the baseline. Tables 3 and 4 show the detailed comparisons on the above three datasets.

Figures 4, 5, and 6 show the CMC curve of the attention module on the three datasets: DukeMTMC-reID, Market501 and CUHK03. From these diagrams, we can see that the F and B stages of the attention module can respectively improve the performance of the network, and the complete attention module after the fusion of the two stages can enable the greatest improvement.

Figures 7 shows the ROC and EER curves of the attention module on the above datasets. Experimental results show that the proposed model performs better than the original model, which indicates that the model with our approach can learn more discriminative deep features.

We compared the proposed model with the original ResNet50 with respect to model complexity. Since our network structure has one more branch as compared with the original ResNet50, the time complexity of the model has increased to a certain extent. In addition, it can be seen from Table 5 that after the proposed attention module is added, the FLOPs and parameters number of the model do not increase significantly. In the groups of Table 6, Table 7 and Table 8, we respectively compared our work with some state-of-the-art methods proposed in the recent years over the popular benchmark datasets DukeMTMC-ReID, Market1501 and CUHK03. All reported results do not apply any re-ranking techniques. From these tables, we observe that our proposed method achieves substantially superior performance over all compared methods on the benchmark of DukeMTMC-ReID and Market1501. For DukeMTMC-ReID, Table 6 shows that our proposed approach obtained 78.8% mAP and 89.9% Rank − 1 accuracy. Compared with HA-CNN (CVPR18), which also researches the attention mechanism, our method improved mAP by 15% and Rank − 1 by 9.4%. Compared with BFE (ICCV19) and MGN (ACM MM18), which are also multi-branch, our attention-based method has also achieved great improvement. Similar to the comparison with DukeMTMC-reID, our model achieved slightly better results on Market1501 as compared to baseline methods, such
as MHN (ICCV19) and MGN (ACM MM18). As shown in Table 7, our model achieved 87.1% for mAP and 95.4% for Rank — 1, respectively. Compared with MHN-6, which introduces 6-level attention information, our method improved mAP by 2.1% and Rank — 1 by 0.3%. On the CUHK03 dataset, our method can also obtain a good result as compared with some classic methods.

VI. CONCLUSION

In this paper, we propose a multi-branch person re-ID method based on the fusion attention mechanism to mine effective information of features from the perspective of feature extraction. Instead of normal attention, such as channel attention and spatial attention, we use covariance in non-local operation to capture the correlations between features, then reconstruct features with expectation-maximization operation, which computes more compact feature descriptors by iteratively executing the EM algorithm. The reconstructed features are low-rank and have more discriminative and robust representations. In the network framework of this paper, we adopt a two-branch structure. In addition, we add a feature erasing branch, which can obtain more features as compared with the global branch. This branch applies batch erasing strategy to erase the same area of the same batch features during training for learning detailed features. Extensive experiments show that the proposed approach achieves state-of-the-art performance on popular re-ID datasets, including Markets1501, DukeMTMC-reID and CUHK03. In addition, the results obtained showed that the our proposed attention module can improve model performance.

In future research, we will first evaluate the performance of a model trained by one domain on another domain, and then study the method of improving the generalization ability of the model. Additionally, we plan to use surveillance cameras to collect person images, and study unsupervised methods to further improve model performance with these unlabeled images.

REFERENCES

[1] T. D. Pham, C. Park, D. T. Nguyen, G. Batchuluun, and K. R. Park, “Deep learning-based fake-banknote detection for the visually impaired people using visible-light images captured by smartphone cameras,” *IEEE Access*, vol. 8, pp. 63144–63161, 2020.
[2] R. Naqvi, M. Arsalan, G. Batchuluun, H. Yoon, and K. Park, “Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor,” *Sensors*, vol. 18, no. 2, p. 456, Feb. 2018.
[3] G. Batchuluun, R. A. Naqvi, W. Kim, and K. R. Park, “Body-movement-based human identification using convolutional neural network,” *Expert Syst. Appl.*, vol. 101, pp. 56–77, Jul. 2018.
[4] G. Batchuluun, H. S. Yoon, J. K. Kang, and K. R. Park, “Gait-based human identification by combining shallow convolutional neural network-stacked long short-term memory and deep convolutional neural network,” *IEEE Access*, vol. 6, pp. 63164–63186, 2018.
[5] G. Batchuluun, Y. Kim, J. Kim, H. Hong, and K. Park, “Robust behavior recognition in intelligent surveillance environments,” *Sensors*, vol. 16, no. 7, p. 1010, Jun. 2016.
[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.
[7] A. Mignon and F. Jurie, “PCCA: A new approach for distance learning from sparse pairwise constraints,” in *Proc. IEEE/Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2666–2672.
[8] G. Batchuluun, N. R. Baek, D. T. Nguyen, T. D. Pham, and K. R. Park, “Region-based removal of thermal reflection using pruned fully convolutional network,” *IEEE Access*, vol. 8, pp. 75741–75760, 2020.
[9] D. T. Nguyen, J. K. Kang, T. D. Pham, G. Batchuluun, and K. R. Park, “Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence,” *Sensors*, vol. 20, no. 7, p. 1822, Mar. 2020.
[10] W. Song, Q. Zhao, C. Chen, Z. Gan, and F. Liu, “Survey on pedestrian re-identification research,” *CAAI Trans. Intell. Syst.*, vol. 12, no. 6, pp. 770–780, 2017.
[11] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou, “Towards good practices on building effective CNN baseline model for person re-identification,” *CoRR*, vol. abs/1807.11042, pp. 145–152, 2018.
[12] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Deep metric learning for person re-identification,” in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 34–39.
[13] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in *Computer Vision (ECCV)*, Cham, Switzerland: Springer, Sep. 2018, pp. 501–518.
[14] S.-L. Chung, Y. Xue, S.-C. Chien, and R.-S. Chan, “Improved part-aligned deep features learning for person re-identification,” in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS)*, Sep. 2019, pp. 1–8.
[15] X. Hu, Z. Jiang, X. Guo, and Y. Zhou, “Person re-identification by deep learning multi-part information complementary,” in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 848–852.
[16] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.
[17] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, “Hydraplus-net: Attentive deep features for pedestrian analysis,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.
[18] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[19] S. Li, S. Bak, P. Carr, and X. Wang, “Diversity regularized spatiotemporal attention for video-based person re-identification,” in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 369–378.
[20] B. Chen, W. Deng, and J. Hu, “Mixed high-order attention network for person re-identification,” in *Proc. IEEE/ACMV Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.
[21] T. Guo, D. Wang, Z. Jiang, A. Men, and Y. Zhou, “Deep network with spatial and channel attention for person re-identification,” in *Proc. IEEE Int. Conf. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.
[22] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, “Second-order non-local attention networks for person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3760–3769.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” in Proc. Adv. Neural Inf. Process. Syst., Jun. 2015, pp. 2017–2025.

[24] D. T. Nguyen, T. D. Pham, G. Batchuluun, H. S. Yoon, and K. R. Park, “Artificial intelligence-based thyroid nodule classification using information from spatial and frequency domains,” J. Clin. Med., vol. 8, no. 11, p. 1976, Nov. 2019.

[25] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7794–7803.

[27] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, “Batch DropBlock network for person re-identification and beyond,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3691–3701.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. Roy. Stat. Soc., B (Methodol.), vol. 39, no. 1, pp. 1–22, 1977.

[29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[30] Y. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in Proc. 4th Int. Conf. Learn. Representations (ICLR), San Juan, Puerto Rico, Y. Bengio and Y. LeCun, Eds., May 2016, pp. 530–543.

[31] M. Lin, Q. Chen, and S. Yan, “Network in network,” in Proc. Neural Evol. Comput., 2013, pp. 1–10.

[32] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.

[33] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Mar. 2017, pp. 2516–2526.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2818–2826.

[35] Y. Yu and V. Koltun, “Latent feature learning and its applications,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1341–1344.

[36] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by GAN improve the person re-identification baseline in vitro,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3774–3782.

[37] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 152–159.

[38] D. Chen, S. Zhang, W. OuYang, J. Yang, and Y. Tai, “Person search by separated modeling and a mask-guided two-stream CNN model,” IEEE Trans. Image Process., vol. 29, pp. 4669–4682, 2020.

[39] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in Proc. 15th Eur. Conf. Comput. Vis., Munich, Germany, Sep. 2018, pp. 501–518.

[40] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, “Mancs: A multi-task attentional network with curriculum sampling for person re-identification,” in Proc. 15th Eur. Conf. Comput. Vis., Munich, Germany, Sep. 2018, pp. 384–400.

[41] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, “Towards rich feature discovery with class activation maps augmentation for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., May 2019, pp. 1389–1398.

[42] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, “Batch DropBlock network for person re-identification and beyond,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2019, pp. 3690–3700.

[43] P. Fang, J. Zhou, S. Roy, L. Petersson, and M. Harandi, “Bilinear attention networks for person retrieval,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2019, pp. 8029–8038.

[44] D. Organisciak, D. Sakkos, E. S. L. Ho, N. Aslam, and H. P. H. Shum, “Unifying person and vehicle re-identification,” IEEE Access, vol. 8, pp. 115673–115684, 2020.

[45] X. Chang, T. Hospedales, and T. Xiang, “Multi-level factorisation net for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2109–2118.

[46] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, B. Harriaran, and K. Weinberger, “Resource aware person re-identification across multiple resolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., May 2018, pp. 8042–8051.

FEI ZHOU was born in Hubei, China, in 1977. He received the Ph.D. degree in communication and information system from the University of Electronic Science and Technology of China, Chengdu, China, in 2006.

He is currently a Professor with the Chongqing High Performance Broadband Network Innovation Team, Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include wireless positioning, communication image processing.

WENFENG CHEN was born in Jiangxi, China, in 1995. He received the B.S. degree in electronic information science and technology from Hainan Normal University, Haikou, China, in 2018. He is currently pursuing the degree with the Chongqing High Performance Broadband Network Innovation Team, Chongqing University of Posts and Telecommunications, Chongqing, China.

His current research interests include object detection and person re-identification.

YANI XIAO was born in Chongqing, China, in 1996. She received the B.S. degree in information engineering from the Chongqing University of Posts and Telecommunications, Chongqing, in 2018, where she is currently pursuing the degree with the Chongqing High Performance Broadband Network Innovation Team.

Her current research interest includes person re-identification.