International Journal of Innovative
Computing, Information and Control
Volume **15**, Number **3**, June **2019**

 ISSN 1349-4198
pp. **1037–1048**

LEARNING DISCRIMINATIVE FEATURES THROUGH AN INDIVIDUAL'S ENTIRE BODY AND THE VISUAL ATTENTIONAL PARTS FOR PERSON RE-IDENTIFICATION

Tianlong Bao, Binquan Wang, Saleem Karmoshi
Chenglin Liu and Ming Zhu

School of Information Science and Technology
University of Science and Technology of China
No. 96, Jinzhai Road, Baohe District, Hefei 230026, P. R. China
{ btl1991; wbq0556; saleem; lclin }@mail.ustc.edu.cn; mzhu@ustc.edu.cn

Received June 2018; revised November 2018

Abstract. *Person Re-Identification (Re-ID) aims to match a specific person across different camera views, which has wide application in public security and image retrieval. For example, Re-ID can help the police get trajectories of suspects. Re-ID still remains a challenging task due to large variations in illumination, background clutter, occlusion and human pose. In this work, a novel deep learning architecture containing global and attentional branches is proposed to learn discriminative representations of persons in differing contexts for Re-ID. Specifically, the global branch is a traditional deep model that learns global features with the images of a person. The attentional branch uses a low-rank approximation of a bilinear pooling model to learn attentional maps by automatically focusing on the visual attentional parts of an individual. The whole model is trained jointly in an end-to-end method. The features of the entire body and visual attentional parts obtained by the trained model are concatenated as representations of persons. Finally, a generic cosine distance metric is used for the person Re-ID task. Extensive experiments on several benchmark datasets including CUHK01, CUHK03 and Market-1501 demonstrate the effectiveness of our method compared to the current state-of-the-art approaches.*

Keywords: Re-identification, Deep learning, Convolutional neural networks, Attentional pooling

1. **Introduction.** Person Re-Identification (Re-ID) aims to identify a person across differing camera views and has been a controversial research topic in the area of computer vision. The development of deep learning and the availability of large-scale pedestrian datasets [1, 2] have greatly promoted progress in this area in recent years. However, it is still a challenging task due to the large intra-class variations caused by illumination, background clutter, occlusion and human pose. The question of how to extract robust representations has become one of the key points of this task.

Many existing methods solve these challenges by designing low-level hand-crafted features or extracting high-level features by Convolutional Neural Networks (CNNs). Low-level hand-crafted features are typically designed to divide the image of the person into local parts and then encode spatial structural information. For example, images of a person are divided into horizontal stripes in [3]; a structure of color distributions using different parts of the object is used to encode spatial structural information. In [4], chromatic information and structural information are extracted based on the localization of perceptual relevant human parts. With the development of deep learning recently, researchers have

DOI: 10.24507/ijicic.15.03.1037

been dedicated to using CNNs to solve the person Re-ID task. In contrast to low-level hand-crafted features, CNNs favour intrinsic learning of global feature representations. Despite the significant process on deep learning for image classification, performance of the existing deep models directly used on person Re-ID is often unsatisfactory. The main reason for this unsatisfactory performance is the large intra-class variations due to the misalignment of the human body caused by pose variations and occlusions. Researchers try to improve deep model architecture to fit the task by locating body joints in order to align the images of a person or training the model with local patches (horizontal stripes) to obtain local features.

In this paper, we follow the trend of work that applies CNNs to the person Re-ID task. Our aim is to train a deep model that can optimize global and attentional features simultaneously and use a simple metric such as Euclidean or cosine distance for person Re-ID. Therefore, an architecture with two branches (global and attentional) has been designed. The global branch is responsible for extracting features of the entire body so it can best use the discriminative ability of CNNs. In addition, similar to human visual processing, the attentional branch tends to selectively concentrate on a part of the information of the input which can help reduce the impact of misalignment of human body and background clutter.

Instead of designing new CNN architectures, GoogLeNet [5] and ResNet-50 [6] are chosen as base networks to extract global features because they have been widely used in the area of image classification. The task of person Re-ID is similar to fine-grained categorization as the matching process of images in both cases relies on the analysis of fine-texture details and parts that are difficult to localize precisely. Bilinear pooling has been shown to achieve state-of-the-art performance on a variety of fine-grained classification tasks but bilinear features are high dimensional. Fortunately, attentional pooling [7], as a low-rank approximation of the bilinear pooling model, is proposed to extract low dimensional bilinear features. Motivated by the above, the attentional branch in our architecture uses attentional pooling methods to extract spatial-invariant features which focus on the visual attentional parts of images of persons. When this is done, it can reduce the impact of misalignment and background clutter well. Additionally, model size and feature dimension are also reduced. Extensive comparative evaluations demonstrate the superiority of the proposed model on three person Re-ID datasets (CUHK01 [8], CUHK03 [2] and Market-1501 [1]).

The rest of this paper is organized as follows. Related work of Re-ID is discussed in Section 2. Section 3 introduces the framework of the proposed method. Results and analysis of our experiments on three public benchmark Re-ID datasets are shown in Section 4. Section 5 presents the conclusion of our work.

2. **Related Work.** Generally, existing works for person Re-ID mainly focus on learning a distance metric or developing a feature extraction method.

In distance metric methods, the purpose is to reduce the intra-class distance while enlarging the inter-class distance in the feature space. For example, Dikmen *et al.* [9] propose a metric learning framework to obtain a robust metric for LMNN (Large Margin Nearest Neighbor) classification with rejection. KISSME (Keep It Simple and Straightforward Metric Learning) [10] is introduced from equivalence constraints based on a statistical inference perspective. Pedagadi *et al.* [11] use LFDA (Local Fisher Discriminant Analysis) for person Re-ID. Xiong *et al.* [12] propose the regularized PCCA (Pairwise Constrained Component Analysis) to maximize the inter-class margin. XQDA (Cross-view Quadratic Discriminant Analysis) [13] is proposed to learn a discriminant low-dimensional subspace.

A part of feature extraction methods employ hand-crafted features such as color and texture histograms which encode spatial structural information. For example, images of persons are divided into horizontal stripes in [3], and then a structure of color distribution using different parts of the object is used to represent pedestrians. In [4], chromatic information and structural information are extracted based on the localization of perceived relevant human parts. The descriptor LOMO (Local Maximal Occurrence) is adopted in [13] to analyze the local occurrence. Zhao *et al.* [14] extract color histogram and SIFT (Scale Invariant Feature Transform) features to learn mid-level filters from automatically discovered patch clusters. In [15], a descriptor is extracted based on mean and covariance information from pixel features in each of the patch and region hierarchies. In [16], interference-ripple images are generated by an iterative process using inverse Sobel filter.

Deep learning based methods can learn high-level features and distance metrics simultaneously in an end-to-end approach. In [2, 17, 18], the Siamese model, which takes a pair of images as input, is used. These types of methods are computation expensive. They need to process pairs consisting of query and every image in the dataset. Moreover, the Siamese model only uses weak Re-ID labels: two images of the same person or not. To fully use strong Re-ID labels, identification models are developed. In [19, 20], pre-trained CNN models are fine-tuned on the target datasets in a classification mode. Xiao *et al.* [21] jointly train a classification model using multiple datasets and propose a new dropout function to deal with the hundreds of classes. To obtain more discriminative features, more sophisticated models are proposed. Wu *et al.* [22] combine CNN embeddings with the hand-crafted features in the FC (Fully Connected) layer. The method used in [23] combines verification and identification models to learn more discriminative pedestrian descriptors. Approaches [24, 25, 26] apply a triplet loss to obtaining the correct order for each probe image and distinguish identities in the feature space. The method used in [27] further designs a quadruplet deep network using a margin-based online hard negative mining process based on the quadruplet loss to make the model output with a larger inter-class variation and a smaller intra-class variation compared to the triplet loss.

Despite the considerable progress of the above mentioned deep learning based methods, the performance is still unsatisfactory because they ignore the large intra-class variations due to misalignment caused by view or pose variations. Researchers try to improve the deep model architecture to suit the task by locating body joints to align images of persons or a training model with local patches (horizontal stripes). In [28], an image of person is divided into horizontal stripes to obtain local features. Zheng *et al.* [29] align pedestrians to a standard pose by detecting 16 body joints to get pose invariant embeddings. These methods, to some extent, reduce the impact of misalignment of the human body, but additional labeled information of body parts is needed.

As we know, the task of person Re-ID is similar to fine-grained classification as the matching process of images in both cases relies on the analysis of fine-texture details and visual attentional parts. Consequently, we consider the person Re-ID task to be a fine-grained recognition problem. Bilinear pooling [30] has achieved state-of-the-art results on a number of fine-grained classification tasks, but bilinear features are high dimensional. The attention module [7], as low-rank approximations of bilinear pooling methods, can be trained without extra supervision while keeping the network size and computational cost nearly the same. In addition, the attention module can automatically focus on specific parts of input relevant to the current task. Doing so may facilitate localization when significant pose variation is present without the need for any part labeling of the training images.

3. Proposed Method.

3.1. **Network architecture.** The proposed architecture is illustrated in Figure 1. The architecture contains global and attentional branches. The base network is a CNN model used as the non-linear embedding functions. The global branch and attentional branch are used to predict the identity of the input image, respectively. In order to discover complementary information to optimize person Re-ID under significant changes in viewing condition, the architecture consisting of global and attentional branches is designed to capture both global and attentional features. Given an input image resized to $224 * 224$, the two branches simultaneously predict the identity of the image. So at training stage, the model is supervised only by the identification label t and no additional body part labeling information is needed. The base network is a CNN model pre-trained on the ImageNet dataset [31]; various CNN models can be considered. In this paper, the GoogLeNet [5] and ResNet-50 [6], which are widely used in many vision tasks, are chosen as the base network. To reduce the model size, facilitate the training process, and learn the global and attentional features simultaneously, the structure of the base network of each branch is the same, and the weights of base network are also shared by global and attentional branches. As far as we know, the attentional method which uses a low-rank approximation of a bilinear pooling model is firstly used to learn attentional maps by automatically focusing on the visual attentional parts of an individual. And the jointly learning of global and attentional branches further improves the performance of person Re-ID.

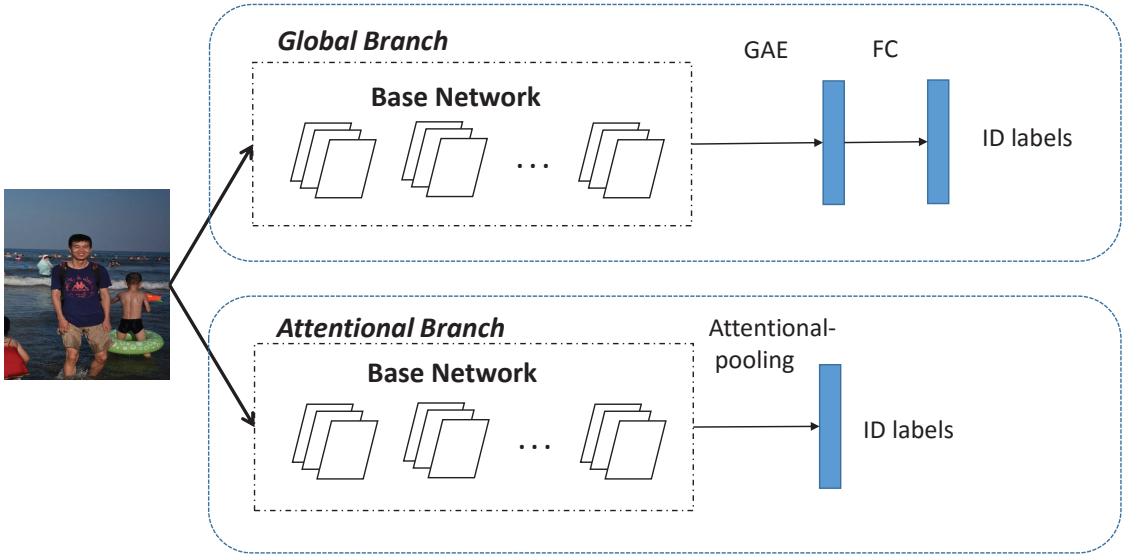


FIGURE 1. The proposed architecture for person Re-ID. 'GAE' and 'FC' denote the operation of global average pooling and a full connection. The base network is a CNN model pre-trained on the ImageNet dataset used as the non-linear embedding functions. The global branch and attentional branch are used to predict the identity of the input image, respectively.

3.1.1. *Global branch.* The global branch learns the global features from the entire image. In order to fine-tune the base network on a new dataset, the last fully connected layer of the pre-trained CNN model is replaced with a new one. The new fully connected layer projects the output of the base network to an N-dim vector where N is the unique person number in the training set. A loss function is then added to normalize the output. Similar

to conventional CNN models, the cross-entropy loss for identity prediction is used, which is:

$$\hat{p} = softmax(\theta_G(M)), \quad \text{where } M = \theta_B(I) \tag{1}$$

$$loss_G(M, t, \theta_G) = \sum_{i=1}^{k} - p_i \log(\hat{p_i}) \tag{2}$$

where *softmax* function is a classifier that gives a non-linear variant of multinomial logistic regression, I is the input image, t is the target class of I, k is the unique person number, θ_B and θ_G denote the parameters of base network and global branch respectively, and M is the feature map generated by the base network. The predicted probability is $\hat{p}$, and p_i is the target probability. $p_t = 1$, $p_i = 0$ while $i \neq t$.

3.1.2. *Attentional branch.* The attentional branch is the low rank of bilinear pooling responsible for learning discriminative features from visual attentional parts.

Suppose $M \in R^{h*w*c}$ is the output feature map of the base network, where h, w and c indicate the height, width and number of feature channels. Bilinear pooling forms a global image descriptor by calculating:

$$Bi(M) = \sum_{i \in [1, h*w]} m_i m_i^T \tag{3}$$

where $m_i \in R^{c*1}$ is the feature vector at a specific location. A holistic representation of the image with dimension $c * c$ is $Bi(M)$. This computation can be written in matrix notation as $X^T X = \sum_{i \in [1, h*w]} m_i m_i^T$ where $X \in R^{hw*c}$ is a matrix by reshaping M in terms of the third mode. Typically, one then reshapes this representation into a vector as $x \in R^{c^2*1}$ and learns a linear classifier (a mapping matrix with size $c^2 * k$) to generate a classification result (to simplify the explanation, we let $k = 1$). From the above, we can see the bilinear representations and the weights are high dimensional. To reduce the parameter and computation cost of the proposed architecture, the low rank approximation of bilinear pooling as in [7] is used to replace the bilinear pooling in our method.

The original mapping matrix of linear classifier is a vector $w \in R^{c^2*1}$ while $k = 1$. If it is re-written as a $c * c$ matrix W, then the calculation of classification result can be written using the trace operator as follows:

$$class(X) = w^T x = Tr\left(X^T X W^T\right) \tag{4}$$

where $x \in R^{c^2*1}$, $X \in R^{hw*c}$, $W \in R^{c*c}$.

Assume the matrix $W = ab^T$, where $a, b \in R^{c*1}$, which is a rank-1 approximation of W. Then the classification result can be calculated by:

$$class_{attention}(X) = Tr\left(X^T X b a^T\right) = Tr\left(a^T X^T X b\right) = a^T X^T X b = (Xa)^T Xb \tag{5}$$

Equation (5) shows that single channel attentional map of the same spatial resolution as the last feature map can be obtained using the linear classifier (Xa and Xb), and the final classification result can be calculated as the inner product between two attentional maps defined over all $h * w$ spatial locations. The formula can be extended to include multi-class classification. The linear classifier a remains unchanged while b extends from $b \in R^{c*1}$ to $b \in R^{c*k}$ to generate the k attentional map (Xb), where k is number of the classes.

In this paper, the attentional pooling is implemented by simple convolutional operations. The detailed implementation is shown in Figure 2. M is the output feature map of the base network. The single channel attentional map $\left(A \in R^{h*w*1}\right)$ is generated through operation $Conv(1, 1, 1)$, while the k channel attentional maps $\left(B \in R^{h*w*k}\right)$ are generated

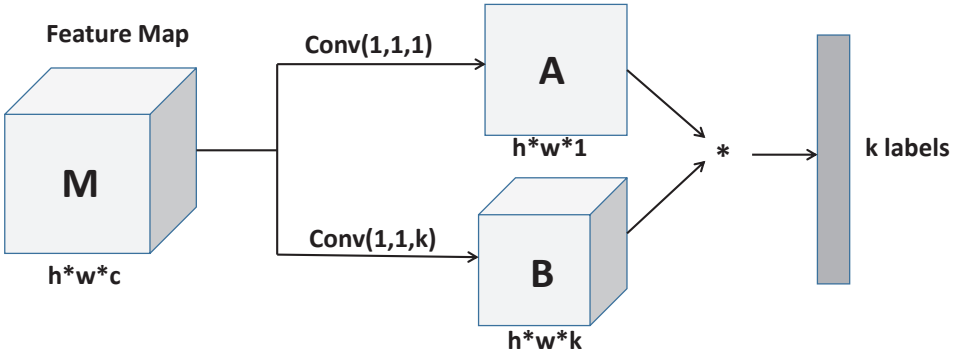


FIGURE 2. Implementation of our approach to attentional pooling. M is the feature map generated by base network with size $h*w*c$, $Conv(h,w,n)$ is the convolutional operation where h, w are the height and width of filter size, and n is the number of filters.

through operation $Conv(1,1,k)$. $A^T * B$ generates a k-dim vector. Similar to the global branch, the cross-entropy loss for identity prediction is used.

3.2. **Model training and testing.**

3.2.1. *Model training.* The training images are first resized to $256 * 256$. Next, the mean image is computed from all the training images and then is subtracted from all the images. In order to prevent over-fitting, all the images are randomly cropped to $224 * 224$ and mirrored horizontally to increase data diversity. Furthermore, the training dataset is shuffled and used in a random order. The standard Stochastic Gradient Descent (SGD) optimization algorithm [32] is used to perform the batch-wise learning of global and attentional branches. Supposed the parameter is θ and the loss function is $J(\theta)$, SGD computes the gradient of θ using a batch of training samples. The update is given by:

$$\theta = \theta - \alpha \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J\left(\theta; x^{(i)}, y^{(i)}\right)}{\partial \theta} \tag{6}$$

where α is the learning rate, m is the batch size, $x^{(i)}$ is the input images and $y^{(i)}$ is the label. The loss function used in our paper is defined as Equation (2). Specifically, the weights of global and attentional branches are updated using gradients produced by each branch respectively. Next, gradients are added together to update the base network because the weights of the base network are shared. The optimization processes of the two branches are synchronized by using SGD with identity supervision. The joint learning process can avoid divergence between the two branches and help enhance the correlated complementary learning capability.

3.2.2. *Model testing.* Once the model is learned, we can adopt an efficient method to extract activation in the intermediate layer to represent images of persons. The weights of base network are shared, so features are extracted by only activating one fine-tuned model. Once given an image, we first feed forward the image to the base network, and then the global and attentional descriptors are concatenated to represent the input:

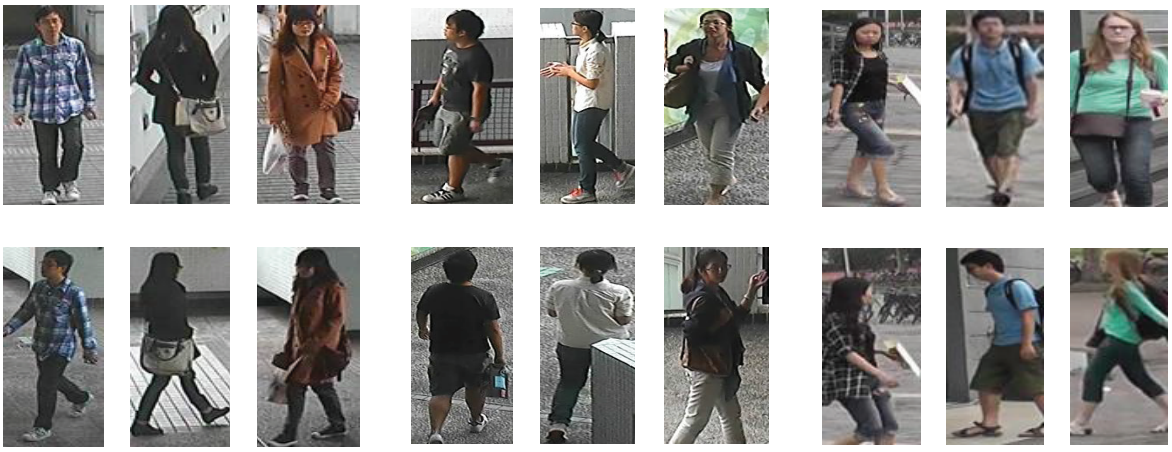
$$descriptor(I, \theta) = \left[GAE(M); M^T A\right] \tag{7}$$

where I is the input image, θ denotes the parameters of the base network, M is the feature map generated by the base network, and A is the attentional map produced by the attentional branch as shown in Figure 2. Moreover, a $2c$-dim feature is obtained to represent I where c is the number of feature channels of M. Once the features for the

gallery sets are obtained, they are stored off-line. Given a query image, its feature is first extracted online. The cosine distances between the query and all the gallery features are sorted in ascendant order to obtain the final ranking results. The probabilities of true matches of query images indicate the correctness of the combined features for person Re-ID task.

4. **Experiment.** In this section, we conducted an evaluation of our proposed method. The experimental results of our approach are compared with base CNNs and the state-of-the-art methods on three benchmark datasets.

4.1. **Datasets.** To evaluate the effectiveness of our approach, experiments are carried on public datasets: CUHK01 [8], CUHK03 [2] and Market-1501 [1]. Some examples of three datasets are shown in Figure 3.



(a) CUHK01 (b) CUHK03 (c) Market-1501

FIGURE 3. Image samples of the person Re-ID datasets. Images in the same column are from the same person across two views. (a) CUHK01; (b) CUHK03; (c) Market-1501.

4.1.1. *CUHK01.* The CUHK01 dataset contains images of 971 identities from two disjointed camera views collected in the CUHK campus. One camera captures images of persons with frontal and back views, and the other one has more variations in pose and viewpoint. Each identity has two samples per camera view. We used 485 randomly chosen identities for training and the other 486 for testing.

4.1.2. *CUHK03.* CUHK03 dataset contains 14,097 cropped images of 1,467 identities captured with 6 surveillance cameras. Each identity is observed by two disjointed camera views. The author provides two kinds of bounding boxes: CUHK03-labeled and CUHK03-detected with manually cropped bounding boxes and automatically detected ones by Deformable Part Model (DPM) accordingly. We provided results for both versions. According to the dataset setting, the dataset is partitioned into a training set of 1,367 persons and a testing set of 100 persons. As in previous works, we adopted the single-shot setting.

4.1.3. *Market-1501.* Market-1501 contains 32,668 annotated bounding boxes of 1,501 identities detected by the DPM, which is closer to the realistic setting. Images of each annotated identity are captured by at least two cameras so that a cross-camera search can be performed. Following the setting of the dataset, the training set contains 12,936 cropped images of 751 identities, the testing set contains 19,732 gallery images of 750

identities and 3,368 query images. Similar to experiments on CUHK03, we adopted the single-shot setting on Market-1501.

4.2. **Evaluation protocol.** Typically, the Cumulated Matching Characteristics (CMC) curve is used to evaluate the performance of person Re-ID on CUHK01 and CUHK03. Mean Average Precision (mAP) is additionally shown on Market-1501. The CMC curve shows the probability that a query identity appears in different-sized candidate lists. The mAP is the mean value of the average precision of all queries, which considers both precision and recall.

4.2.1. *Comparison with the CNN baseline.* The proposed architecture consists of a base network (GoogLeNet and ResNet-50) and two branches. To validate the effectiveness of the proposed method compared with base CNNs, pre-trained GoogLeNet and ResNet-50 models on the ImageNet dataset are first fine-tuned on the target dataset. ResNet-50 is deeper than GooggLeNet. Then the output of the last pooling layer is extracted as deep representations for person Re-ID. We follow the method in [33] to fine-tune the base CNNs on the target dataset to predict the person identities. As shown in Table 4, we obtained 73.5% and 75.8% rank-1 accuracy by GoogLeNet and ResNet-50, respectively, on Market-1501. The advantage of using deep learning algorithms is their ability to automatically extract useful representative features during the training phase. To show the superiority of deep learning methods, we also compare the performance of person Re-ID with traditional methods such as KISSME and XQDA. Note that using the base CNNs alone exceeds many previous methods, which proves the superiority of deep learning in dealing with person Re-ID problems. As can be seen in Table 2, Table 3 and Table 4, the performance of ResNet-50 is better than GoogLeNet, which proves that deeper networks can get better features.

Table 1. Evaluation on the CUHK01 dataset (matching accuracy in %)

Method	rank-1	rank-5	rank-10	rank-20
LMNN [9]	13.5	31.3	42.3	54.1
KISSME [10]	29.4	57.7	72.4	86.1
LOMO+XQDA [13]	63.2	83.9	90.1	94.2
DeepReID [2]	27.9	58.2	73.5	86.3
ImprovedDeep [18]	47.5	72.3	80.1	83.9
GoogLeNet	67.2	86.3	89.6	92.1
ResNet-50	70.5	89.1	91.8	93.6
Ours(GoogLeNet)	70.1	88.6	90.2	92.4
Ours(ResNet-50)	73.6	90.8	92.5	94.1

Table 2. Evaluation on the CUHK03-labeled dataset (matching accuracy in %)

Method	rank-1	rank-5	rank-10	rank-20
LOMO+XQDA [13]	52.2	82.2	92.1	96.2
ImprovedDeep [18]	54.7	86.5	93.8	98.1
GoogLeNet	74.3	95.5	96.1	98.9
ResNet-50	80.2	97.1	98.9	99.4
Ours(GoogLeNet)	77.4	96.5	97.3	99.2
Ours(ResNet-50)	82.9	98.1	99.2	99.6

TABLE 3. Evaluation on the CUHK03-detected dataset (matching accuracy in %). '−' means no reported result is available.

Method	rank-1	rank-5	rank-10	rank-20
KISSME [10]	11.7	33.3	48.0	−
LOMO+XQDA [13]	46.3	78.9	88.6	93.2
DeepReID [2]	19.9	49.3	64.7	−
ImprovedDeep [18]	44.9	76.0	83.5	93.1
GatedSiamCNN [34]	61.8	80.9	88.3	−
GoogLeNet	65.8	86.7	91.2	94.4
ResNet-50	71.5	91.5	95.9	97.3
Ours(GoogLeNet)	70.2	90.3	94.6	96.9
Ours(ResNet-50)	75.6	95.2	97.1	98.5

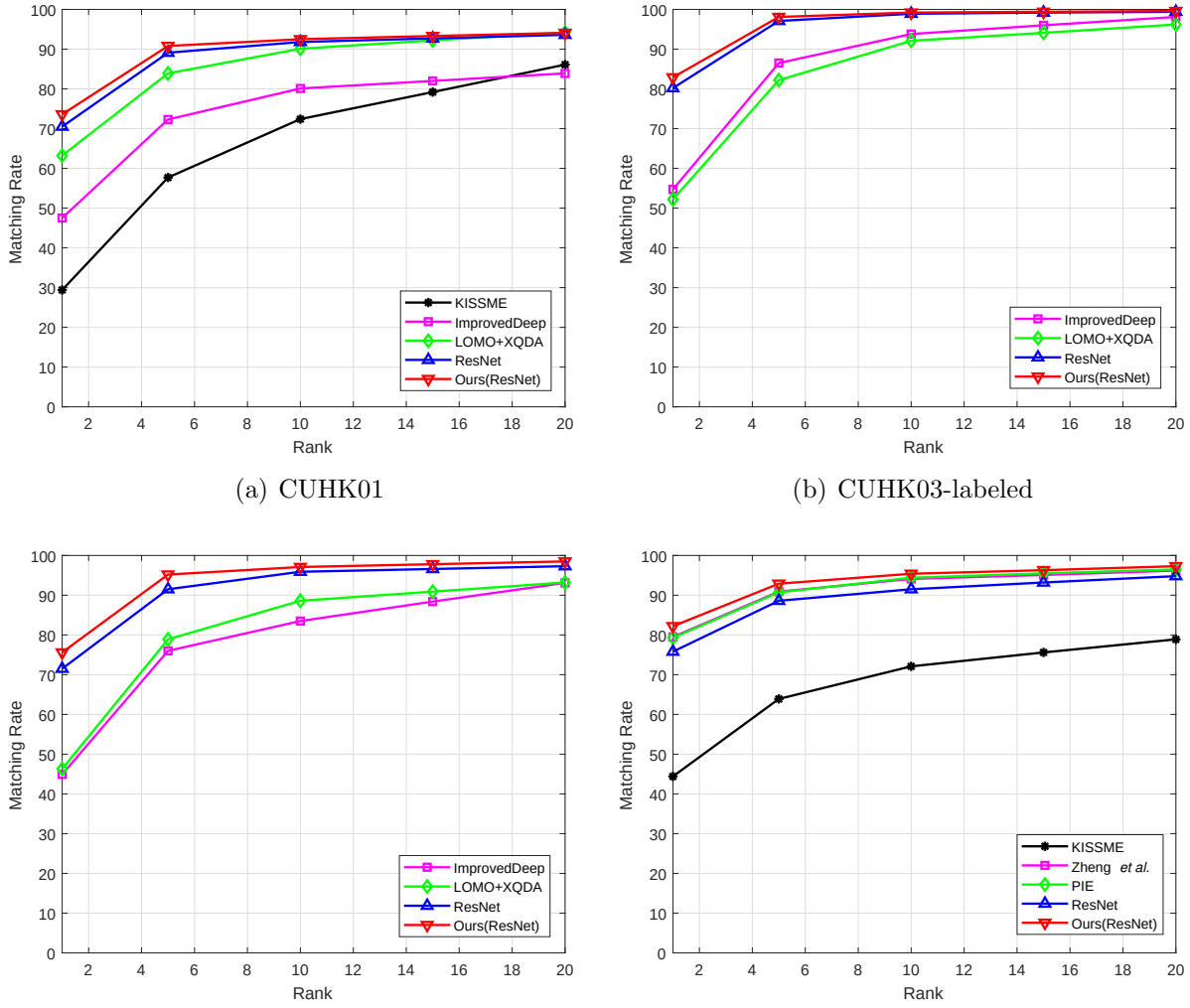
TABLE 4. Evaluation on the Market-1501 dataset (matching accuracy in %). '−' means no reported result is available.

Method	rank-1	rank-5	rank-10	rank-20	mAP
KISSME [10]	44.4	63.9	72.1	78.9	20.7
XQDA [13]	43.8	−	−	−	22.2
GatedSiamCNN [34]	65.8	−	−	−	39.5
Zheng *et al.* [23]	79.5	90.9	94.1	96.2	59.8
PIE [29]	79.3	90.7	94.4	96.5	55.9
GoogLeNet	73.5	86.9	90.2	93.5	48.1
ResNet-50	75.8	88.6	91.5	94.8	52.4
Ours(GoogLeNet)	78.5	90.1	93.2	95.6	58.6
Ours(ResNet-50)	82.2	92.9	95.4	97.3	60.1

Our model consisting of global and attentional branches further improves these baselines on the benchmark dataset. Similar to the fine-tuning process of base CNNs, the proposed model is also pre-trained on the ImageSet dataset first and subsequently fine-tuned on the target dataset to predict the person identities with two branches simultaneously. Then Equation (7) is used to extract the descriptors of images of persons. In Table 1, the rank-1 accuracy 70.1% and 73.6% are achieved on CUHK01 using GoogLeNet and ResNet-50 with two designed branches respectively. Improvements of 2.9% and 3.1% are obtained compared with base CNNs. Similar to the result on CUHK01, rank-k accuracy is also improved to some extent compared with base CNNs on CUHK03, Market-1501 as shown in Table 2, Table 3 and Table 4.

As can be clearly seen from Figure 4, the CMC curves of experimental results show significant improvements over base CNNs on all three datasets. These results show that our method can work with different networks on all three benchmark datasets and improve their results. The global branch can learn the global features of individual's entire body while the attentional branch can learn the attentional features of the visual attentional parts of persons. And the combination of global and attentional features helps the network to learn more discriminative features. So bettter performance is achieved compared with base networks.

4.2.2. *Comparison with the state-of-the-art methods.* We compare our method with other state-of-the-art algorithms in terms of rank-k accuracy on CUHK01 and CUHK03 (both



(a) CUHK01 (b) CUHK03-labeled

(c) CUHK03-detected (d) Market-1501

FIGURE 4. CMC curves of different methods on the (a) CUHK01, (b) CUHK03-labeled, (c) CUHK03-detected and (d) Market-1501 dataset

labeled and detected versions). For the Market-1501 dataset, the mAP value is additionally compared. The results on CUHK01, CUHK03-labeled, CUHK03-detected and Market-1501 are reported in Table 1, Table 2, Table 3 and Table 4. Our model with ResNet-50 as base network yields 73.6%, 82.9% and 75.6% rank-1 accuracy on CUHK01, CUHK03-labeled, CUHK03-detected datesets, 82.2% rank-1 accuracy and 60.1% mAP on Market-1501 respectively. The above results show that the deep model with global and attentional branches introduced in this paper obtains competitive performance compared with previously published methods on the CUHK01, CUHK03 and Market-1501 datasets.

5. **Conclusion.** In this work, we proposed a novel deep learning architecture consisting of global branch and attentional branch for person Re-ID. Our method benefits from both base CNNs and bilinear pooling. Discriminative representations are extracted through the entire body and visual attentional parts of persons. Specifically, the global branch is a base CNN model fine-tuned on the target dataset that learns the features of an entire body. The attentional branch uses a low-rank approximation of the bilinear pooling model to extract the representation of visual attentional body parts of a person, which

can significantly reduce the effect of misalignment. The whole model is trained jointly without extra supervision in an end-to-end way. Finally, the representations of the entire body and the visual attentional parts are combined together for the person Re-ID task. Extensive experiments on three public benchmark datasets show the effectiveness of our method against other state-of-the-art approaches. Our future research concentrates on two fronts: extracting more accurate features according to human body joints or increasing the size of the dataset to train deeper networks.

Acknowledgment. We appreciate the support of subtask of New Generation Broadband Wireless Mobile Communication Network Key Project under Grant No. 2017ZX03001019-004.

REFERENCES

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, Scalable person re-identification: A benchmark, *Proc. of the IEEE International Conference on Computer Vision*, pp.1116-1124, 2015.

[2] W. Li, R. Zhao, T. Xiao and X. Wang, DeepReID: Deep filter pairing neural network for person re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.152-159, 2014.

[3] I. Kviatkovsky, A. Adam and E. Rivlin, Color invariants for person reidentification, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.35, no.7, pp.1622-1634, 2013.

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino and M. Cristani, Person re-identification by symmetry-driven accumulation of local features, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.2360-2367, 2010.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov and A. Rabinovich, Going deeper with convolutions, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[6] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[7] R. Girdhar and D. Ramanan, Attentional pooling for action recognition, *Advances in Neural Information Processing Systems*, pp.33-44, 2017.

[8] W. Li, R. Zhao and X. Wang, Human reidentification with transferred metric learning, *Asian Conference on Computer Vision*, Berlin, Heidelberg, pp.31-44, 2012.

[9] M. Dikmen, E. Akbas, T. S. Huang and N. Ahuja, Pedestrian recognition with a learned metric, *Asian Conference on Computer Vision*, Berlin, Heidelberg, pp.501-512, 2010.

[10] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth and H. Bischof, Large scale metric learning from equivalence constraints, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.2288-2295, 2012.

[11] S. Pedagadi, J. Orwell, S. Velastin and B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.3318-3325, 2013.

[12] F. Xiong, M. Gou, O. Camps and M. Sznaier, Person re-identification using kernel-based metric learning methods, *European Conference on Computer Vision*, Cham, pp.1-16, 2014.

[13] S. Liao, Y. Hu, X. Zhu and S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.2197-2206, 2015.

[14] R. Zhao, W. Ouyang and X. Wang, Learning mid-level filters for person re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.144-151, 2014.

[15] T. Matsukawa, T. Okabe, E. Suzuki and Y. Sato, Hierarchical Gaussian descriptor for person re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1363-1372, 2014.

[16] T. Hiraoka, Generation of interference-ripple images by inverse Sobel filter, *ICIC Express Letters*, vol.12, no.5, pp.409-415, 2018.

[17] D. Yi, Z. Lei, S. Liao and S. Z. Li, Deep metric learning for person re-identification, *The 22nd International Conference on Pattern Recognition (ICPR)*, pp.34-39, 2014.

[18] E. Ahmed, M. Jones and T. K. Marks, An improved deep learning architecture for person re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.3908-3916, 2015.

[19] L. Zheng, Y. Yang and A. G. Hauptmann, Person re-identification: Past, present and future, *arXiv preprint arXiv:1610.02984*, 2016.
[20] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang and Q. Tian, Mars: A video benchmark for large-scale person re-identification, *European Conference on Computer Vision*, Cham, pp.868-884, 2016.
[21] T. Xiao, H. Li, W. Ouyang and X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1249-1258, 2016.
[22] S. Wu, Y. C. Chen, X. Li, A. C. Wu, J. J. You and W. S. Zheng, An enhanced deep feature representation for person re-identification, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1-8, 2016.
[23] Z. Zheng, L. Zheng and Y. Yang, A discriminatively learned CNN embedding for person re-identification, *ACM Trans. Multimedia Computing, Communications, and Applications (TOMM)*, vol.14, no.1, 2017.
[24] S. Z. Chen, C. C. Guo and J. H. Lai, Deep ranking for person re-identification via joint representation learning, *IEEE Trans. Image Processing*, vol.25, no.5, pp.2353-2367, 2016.
[25] S. Ding, L. Lin, G. Wang and H. Chao, Deep feature learning with relative distance comparison for person re-identification, *Pattern Recognition*, vol.48, no.10, pp.2993-3003, 2015.
[26] F. Wang, W. Zuo, L. Lin, D. Zhang and L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1288-1296, 2016.
[27] W. Chen, X. Chen, J. Zhang and K. Huang, Beyond triplet loss: A deep quadruplet network for person re-identification, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol.2, 2017.
[28] W. Li, X. Zhu and S. Gong, Person re-identification by deep joint learning of multi-loss classification, *arXiv preprint arXiv:1705.04724*, 2017.
[29] L. Zheng, Y. Huang, H. Lu and Y. Yang, Pose invariant embedding for deep person re-identification, *arXiv preprint arXiv:1701.07732*, 2017.
[30] T. Y. Lin, A. RoyChowdhury and S. Maji, Bilinear CNN models for fine-grained visual recognition, *Proc. of the IEEE International Conference on Computer Vision*, pp.1449-1457, 2015.
[31] J. Deng, W. Dong, R. Socher, L. Li, K. Li and F. Li, ImageNet: A large-scale hierarchical image database, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.248-255, 2009.
[32] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, pp.1097-1105, 2012.
[33] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang and Q. Tian, Person re-identification in the wild, *arXiv:1604.02531*, 2017.
[34] R. R. Varior, M. Haloi and G. Wang, Gated Siamese convolutional neural network architecture for human re-identification, *European Conference on Computer Vision*, Cham, pp.791-808, 2016.