IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 27, 2019, accepted February 21, 2019, date of publication February 26, 2019, date of current version March 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901764

Omnidirectional Feature Learning for Person Re-Identification

DI WU[1], HONG-WEI YANG[1], DE-SHUANG HUANG[1], (Senior Member, IEEE), CHANG-AN YUAN[2], XIAO QIN[2], YANG ZHAO[3], XIN-YONG ZHAO[3], AND JIAN-HONG SUN[3]

[1]School of Electronics and Information Engineering, Institute of Machine Learning and Systems Biology, Tongji University, Shanghai 201804, China
[2]Science Computing and Intelligent Information Processing of Guangxi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning 530001, China
[3]Beijing E-Hualu Information Technology Co., Ltd., Beijing 100043, China

Corresponding author: De-Shuang Huang (dshuang@tongji.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61520106006, Grant 61732012, Grant 61861146002, Grant 61772370, Grant 61702371, Grant 61672203, Grant 61572447, Grant 61772357, and Grant 61672382, in part by the China Postdoctoral Science Foundation under Grant 2017M611619, and in part by the BAGUI Scholar Program of Guangxi Province of China.

ABSTRACT Person re-identification (PReID) has received increasing attention due to it being an important role in intelligent surveillance. Many state-of-the-art PReID methods are part-based deep models. Most of these models focus on learning the part feature representation of a person's body from the horizontal direction. However, the feature representation of the body from the vertical direction is usually ignored. In addition, the relationships between these part features and different feature channels are not considered. In this paper, we introduce a multi-branch deep model for PReID. Specifically, the model consists of five branches. Among the five branches, two branches learn the part features with spatial information from horizontal and vertical orientations; one branch aims to learn the interdependencies between different feature channels generated by the last convolution layer of the backbone network; the remaining two branches are identification and triplet sub-networks in which the discriminative global feature and a corresponding measurement can be learned simultaneously. All five branches can improve the quality of representation learning. We conduct extensive comparison experiments on three benchmarks, including Market-1501, CUHK03, and DukeMTMC-reID. The proposed deep framework outperforms other competitive state-of-the-art methods. The code is available at *https://github.com/caojunying/person-reidentification*.

INDEX TERMS Person re-identification, deep learning, part feature, triplet model, identification model.

I. INTRODUCTION

As a fundamental task of intelligent surveillance, person re-identification (PReID) aims to re-identify a specific person from multiple camera views. It has been of considerable interest to the computer vision community in recent years. Great progress has been made in PReID, however, the visual appearance of a person may undergo significant variations when facing unpredictable changes in illumination, background clutter as well as person pose, which creates a challenging issue.

In current studies, PReID is resolved from the following two angles: 1) Extracting discriminative descriptors to represent different identities. 2) Learning an effective distance metric to make the relative distance between the inter-classes larger than intra-class.

Benefiting from the considerable development of deep learning in the computer vision community, a large number of deep architecture-based methods have been introduced for PReID. Different from traditional hand-crafted methods, these deep learning-based methods integrate feature and distance metric learning in an end-to-end way. It is worth noting that the most recent state-of-the-art results have been achieved by deep learning-based models. Many of them attempt to learn global pedestrian features. When the pedestrian global features are generated by the deep model, the Euclidean metric is applied to measure the distance between the two pedestrians. However, global feature

The associate editor coordinating the review of this manuscript and approving it for publication was Hugo Proenca.

learning methods have the following drawbacks: a) irrelevant information may be introduced in global features when the pedestrian's body is occluded, b) global features are not sufficiently discriminating to represent the pedestrians with similar appearances. To alleviate these dilemmas, some studies [1], [2] have used predefined horizontal stripes to divide the body into several partitions and separately used the partitions to learn the part features. However, each body partition may locate in different positions in different images due to inaccurate pedestrian detection, pose variations, which causes misalignment problem. Other studies [3]– [5] have introduced pose annotation information to address the alignment problem. Yet, these methods require additional pose estimation procedures. In this paper, similar to the work of Bai *et al.* [6], we learn the part features with spatial information together by using GRU modules rather than separately learn them. Thus, the misalignment problem can be alleviated.

Only few part-based methods consider the spatial contextual information between the different part features. Varior *et al.* [1] first used a recurrent neural network (RNN) to exploit the spatial context information between the extracted sequence features. However, the processes of spatial information learning and feature extraction in this work are separate. Bai *et al.* [6] proposed applying long short-term memory (LSTM) to learn the spatial contextual information between different body parts from head to foot. However, the spatial contextual relationship between body parts from a vertical orientation, i.e., left to right, is ignored. It is noteworthy that almost all part-based models ignore learning part features of a body from the vertical orientation. However, the part features from the vertical orientation may be very useful, especially when the left or right side of the body is occluded. As shown in Figure 1, the left or right side of the bodies are occluded, and using the part features from head to foot can introduce irrelevant information. The part features from left to right (vertical orientation) are useful in this situation. Therefore, we propose adopting a gated recurrent unit (GRU) to simultaneously learn the spatial information between different body parts from head to foot as well as from left to right.

FIGURE 1. Left or right sides of bodies are occluded. Using part features from head to foot may introduce irrelevant information.

Recent studies have shown that the robust property of convolutional neural network (CNN) features can be enhanced by integrating some additional learning modules which help dig the correlations between different feature channels. In this study, apart from part-based features, we also investigate the interdependencies between the different channels to improve the discriminative ability of the learned representations. For this purpose, we use the feature maps generated by the last convolutional layer to input GRU modules to learn the interdependencies between the feature maps.

Some studies [7]– [9] have shown that the combination of identification and triplet losses can help promote the discriminative ability of the learned deep representations. Triplet loss forces the relative distances between the negative pairs to be larger than that of the positive pairs. The training target of triplet loss is similar to its test manner; however, it uses weak label information. Identification loss treats the PReID tasks as a multiclass classification issue, and thus, it makes full use of annotation information. However, the training goal of the identification loss is inconsistent with its testing mode. Therefore, the two types of loss functions can complement each other by making use of their advantages. In this study, we also adopt this hybrid strategy. Unlike previous works, we use online instance matching loss (OIM) [10], rather than the commonly utilized softmax loss, to supervise the identification subnetworks. The generalization ability of softmax loss from the training set to the test set may weaken when the dataset contains a large number of identities, and each identity has a limited number of instances. One possible reason for this is that the softmax loss needs to learn too many discriminative functions with limited instances for each identity. Thus, the classifier matrix cannot be fully learned at each backpropagation stage. Compared to the softmax loss, OIM loss is nonparametric, and therefore, the gradients are directly performed on the features rather than on the classifier matrix. Our previous work [11] also proves the effectiveness of the combination of OIM and triplet losses. As shown in Figure 2, we use the OIM losses to train the four identification subnetworks, i.e., ***Branch 1*∼*Branch 4***. Moreover, the triplet loss is applied to the pooled features for learning a corresponding similarity measurement. In the test phase, we choose feature f_{OIM} as the final pedestrian descriptor for the DukeMTMC-reID and Market-1501 datasets. For the CUHK03 dataset, we use feature f_{Trip} as the final descriptor.

In summary, the contributions of this study are:

1) We propose using GRU to learn the omnidirectional correlation information for a pedestrian body. The proposed model considers the spatial information between different body parts from head to foot and also from left to right.
2) We propose learning the interdependencies between channels to promote the discriminative ability of the learned descriptors.
3) The proposed architecture chooses suitable features as final descriptors based on the characteristics of the datasets. Our model outperforms other competitive state-of-the-art methods on the widely used PReID datasets.

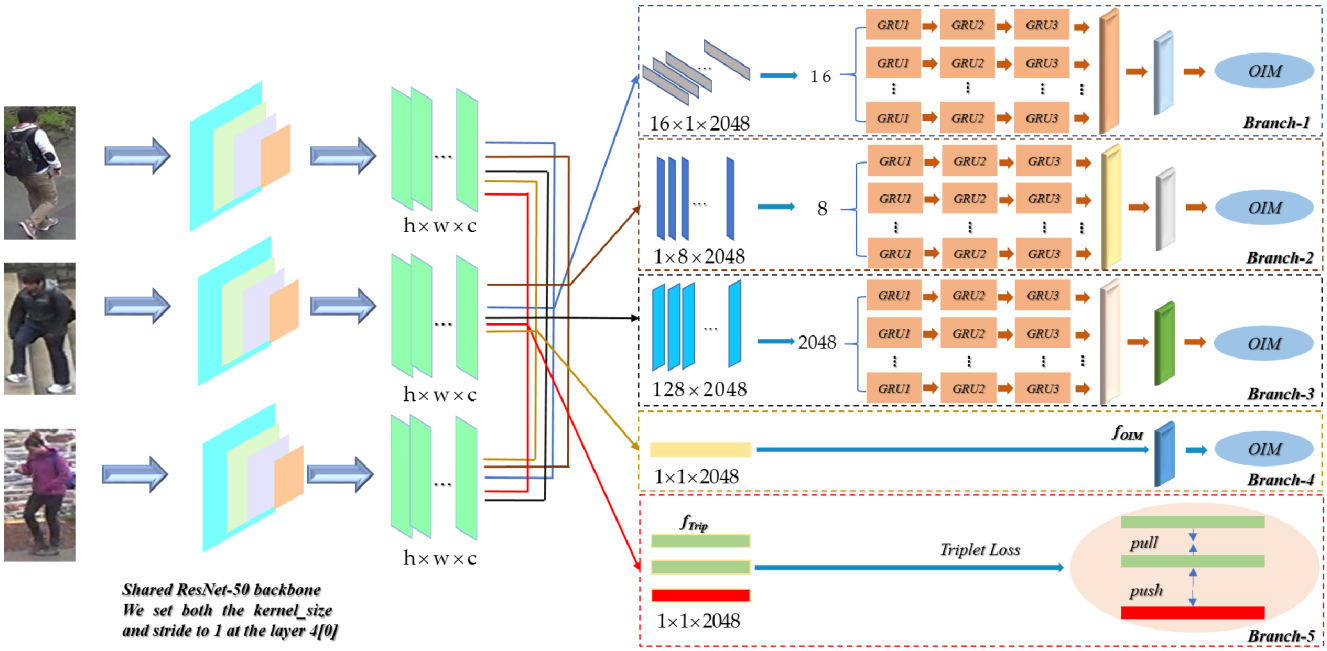

FIGURE 2. Illustration of the proposed deep model. We use pre-trained ResNet-50 [12] as the backbone network and set both the kernel size and stride to 1 at layer 4 [0] of the ResNet-50 model. The architecture contains five branches. The above two branches learn the part-based features of the body from horizontal and vertical orientations. The middle branch learns the interdependencies between different channels. The bottom two branches simultaneously learn the global features and a corresponding measurement for PReID. The whole deep architecture is supervised by four OIM loss functions and one triplet loss function. The image is best viewed in color.

II. RELATED WORK

With the development of deep learning technology, especially convolutional neural networks (CNN), deep feature learning by CNN has become a frequent method in the PReID domain. Most of the deep learning-based structures are based on the following models: identification-based models, verification-based models and triplet-based models. Identification-based models regard PReID as a multi-classification task. Verification-based models use paired images as input and output a value to estimate whether the paired images are the same pedestrian. Triplet-based models aim at making the distances between the same person images as small as possible while making the distances between different person images as large as possible. Some approaches combine two types of models mentioned above. Chen *et al.* [7] designed a multi-task deep architecture that integrates the verification and triplet losses to take advantage of the two losses. Wang *et al.* [8] analyzed the advantages and limitations of single-image representation (SIR) and cross-image representation (CIR) in the PReID community. They proposed pairwise comparison and triplet comparison formulations to simultaneously learn the CIR and SIR. Qian *et al.* [9] introduced a multi-scale deep architecture which contains verification and classification subnets for PReID. In this study, we employed OIM and triplet losses to jointly supervise the training of the proposed deep model.

Recently, there have been some studies [13]– [23] that adopted a part-based strategy to learn discriminative features for PReID. Cheng *et al.* [21] used a global convolution layer to obtain the global deep features and then divided the global features into four equal individual branches to learn the part-based deep features. Finally, they concatenated the global and part-based features to compose the final deep descriptors. To address the misalignment problem in person PReID, Zhao *et al.* [22] introduced a CNN-based attention model which utilizes the similarity information of paired person images to learn the body part for matching. In [23], a harmonious attention convolutional neural network was introduced to simultaneously learn representations and PReID selection in an end-to-end way. More specifically, they combined hard attention and soft attention to learn the region-level and pixel-level parts of the person image. However, those part-based methods possess several limitations as follows. a) Using the fused global and part features to retrieve, thus increasing the memory and time cost. b) Ignoring the correlation between different part features. c) Ignoring the information about body parts from the vertical orientation.

There also exist studies that attempt to exploit the relationship between channels. Hu *et al.* [24] introduced a mechanism that can integrate the network to perform feature recalibration. Through the recalibration operation, the informative features can be emphasized and the less useful features can be suppressed at the same time. In this paper, different from [24], we adopt GRU to directly learn the interdependencies between channels.

III. PROPOSED METHOD

The framework of the proposed deep model is shown in Figure 2. Our purpose is to learn omnidirectional correlation information of a pedestrian body. As seen, the backbone

network adopted in this study is a pre-trained ResNet-50 [12] model. We reset the kernel size and stride to 1 at layer 4 [0] of ResNet-50, because the higher spatial resolution before global pooling contains more detail information of pedestrian features. The proposed network consists of five branches. It has been proven that the head, upper body and lower body as body part features learning can promote the discriminative ability of the learned pedestrian descriptors. Similar to other part-based feature learning models, we use branch-1 to learn the part features from the horizontal orientation (from head to foot). As mentioned above, almost all of the part-based methods ignore learning the part features of the body from the vertical orientation. The vertical orientation features may be useful when the left or right side of the body is occluded. Therefore, we utilize branch-2 to learn the part-based features from the vertical orientation. In addition to part-based features learning, we also introduce branch-3 to exploit the interdependencies between channels. The introduction of this branch can enhance the robust ability of the learned descriptors. Similar to our previous work [11], branch-4 and branch-5 are used to learn global features and a similarity measurement, respectively.

A. PART-BASED FEATURES LEARNING

The usual part-based approaches divide the pedestrian body into several horizontal stripes. Then, each stripe is used to input an independent subnetwork to learn part features from the horizontal orientation. However, these part-based methods ignore the spatial correlation between different stripes. Moreover, the part features learning from the vertical orientation is usually overlooked. To overcome these limitations, in this study, we use a recurrent neural network with GRU cells to learn the spatial information between different local features both from vertical and horizontal orientations. As shown in Figure 2, we use the backbone network to extract deep features with a dimension of $16 \times 8 \times 2048$. Then, the deep features are implemented with average pooling operations with 1×8, 1×16 kernel sizes in branch-1 and branch-2, respectively. As shown in Figure 3, each pooled deep feature represents a corresponding receptive field from vertical orientation or horizontal orientation. We then use the pooled deep features as sequence features to feed into three-layer bidirectional GRU modules, and thus, the GRU modules in branch-1 and branch-2 simultaneously learn the part features with spatial information from horizontal and vertical orientations. We separately concatenate the outputted GRU features in branch-1 and branch-2 as the final part features. Finally, each of the two concatenated GRU features is used to input the two independent OIM layers to perform multi-classification tasks.

B. EXPLOITING THE RELATIONSHIP BETWEEN CHANNELS

The model attempts to exploit the interdependencies between different channels as well. As shown in Figure 4, we convert each of the feature channels with the size of 16×8 to a 128-dimension vector, and thus, each vector represents

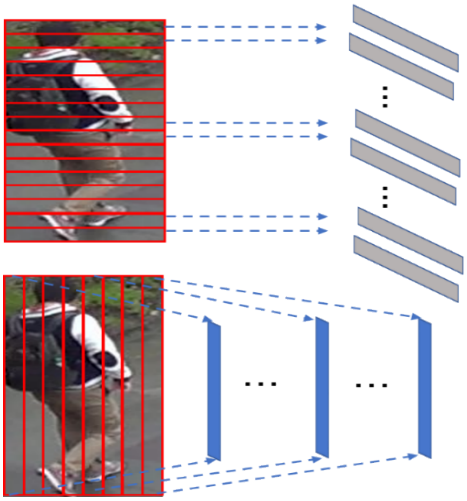

FIGURE 3. Each pooled deep feature vector represents a corresponding receptive field of the pedestrian body from horizontal orientation or vertical orientation.

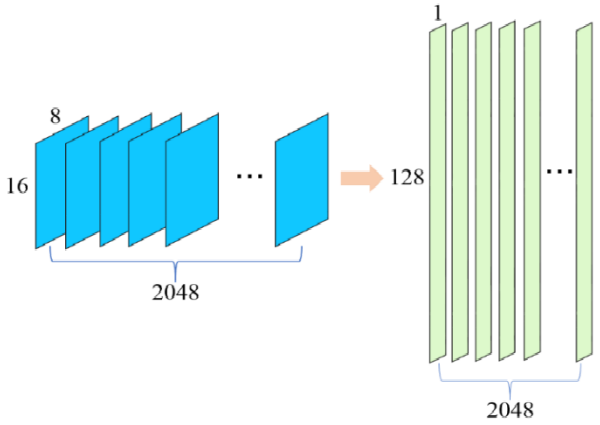

FIGURE 4. Each 16×8 feature map is converted to a 128-dimension vector.

a channel. Then, the 2048 vectors (channels) are used as the sequence features for feeding into the three-layer bidirectional GRU to learn the relationship between them. Finally, the outputs of GRU are also concatenated as a final descriptor to input the OIM layer. Through learning the interdependencies between these feature channels, we can improve the discriminative ability of the learned representations.

C. LOSS FUNCTIONS

1) IDENTIFICATION LOSS

In this study, we employ the online instance matching loss (OIM) to train the identification subnetworks. OIM loss is nonparametric, and thus, the gradients can be directly performed on descriptors without any classifier matrix transforming the operations. The loss keeps a circular queue and a lookup table (LUT) to save the unlabeled and labeled features, respectively. The probability that pedestrian M is

recognized as identity j can be written as:

$$p_j = \frac{exp(\delta_j^T M/\sigma)}{\sum_{i=1}^{L} exp(\delta_i^T M/\sigma) + \sum_{k=1}^{Q} exp(\mu_k^T M/\sigma)} \quad (1)$$

where μ_k^T and δ_i^T are the transpositions of the k-th identity of the circular queue and the j-th column of the LUT, respectively. The higher the temperature σ, the softer the probability distribution. Q and L represent the queue size and column number for the LUT, respectively.

The probability of being re-identified to the j-th ID in the circular queue can be formulated as:

$$R_j = \frac{exp(\mu_j^T M/\sigma)}{\sum_{i=1}^{L} exp(\delta_i^T M/\sigma) + \sum_{k=1}^{Q} exp(\mu_k^T M/\sigma)} \quad (2)$$

The purpose of OIM is to maximize the log-likelihood:

$$L_{oim} = E_M \left[\log p_n\right] \quad (3)$$

2) TRIPLET LOSS

For branch-5, we use the batch-hard triplet loss [25] as the loss function. We randomly sample P identities (classes) and then choose K images for each identity. Among these PK images, only the hardest positive and negative samples are selected for each anchor to form the triplet units. Given a triplet unit (x_a, x_p, x_n), the triplet function can be defined as:

$$L_{trp} = \frac{1}{PK}\sum_{i=1}^{PK}[thre + max\, d(F_w(x_a), F_w(x_p)) - min\, d(F_w(x_a), F_w(x_n))]_+ \quad (4)$$

where $thre$ is a margin, $[\mathrm{x}]_+ = max(\mathrm{x}, 0)$, $F_w(x)$ is the deep descriptor produced by the deep model for image x, and w represents the parameters of the backbone network. $d(x, y)$ denotes the Euclidean distance between x and y.

3) OVERALL LOSS FUNCTION

The whole architecture is supervised by one triplet loss and four OIM losses. The final loss function is redefined as:

$$L = \lambda_1 L_{oim_1} + \lambda_2 L_{oim_2} + \lambda_3 L_{oim_3} + \lambda_4 L_{oim_4} + \lambda_5 L_{trp} \quad (5)$$

where λ_i $(i = 1, 2, 3, 4, 5)$ represent the trade-off parameters for different branches, L_{oim_i} $(i = 1, 2, 3, 4)$ denote the OIM losses for different identification subnetworks.

IV. EXPERIMENTS

The proposed deep model is evaluated on three widely used person re-identification datasets (i.e., CUHK03, Market-1501 and DukeMTMC_reID). Furthermore, we provide the comparison results of several baseline configurations.

A. DATASETS AND PROTOCOLS

The brief descriptions of three PReID datasets used in this study are presented as follows:

1) CUHK03 DATASET

The dataset is one of the largest person ReID datasets and contains 13164 images of 1360 identities. All identities are taken from six camera views, and each pedestrian is captured by two cameras. This dataset provides two settings. One automatically annotated by a detector and the other manually annotated by a human. Between the two settings, the former is closer to practical scenarios. We evaluate the proposed model on the two settings.

2) MARKET-1501 DATASET

This dataset consists of 32643 annotated boxes of 1501 persons. Each pedestrian is collected by at least two cameras and at most six cameras from the front of a supermarket. The boxes of pedestrians are captured by the deformable part model (DPM) detector.

3) DukeMTMC-reID DATASET

The dataset created for image-based person ReID is a subset of the DukeMTMC dataset. It consists of 36411 pedestrian images that belong to 1812 identities taken from eight high-resolution surveillance equipment cameras. Among these 1812 pedestrians, 1404 of them are captured by more than two camera views, and the rest of them are regarded as distractor identifications.

4) EVALUATION METRICS

Cumulative match characteristic (CMC) is used as the evaluation protocol for each dataset. Moreover, we report the mean average precision (mAP) for the DukeMTMC-reID and Market-1501 datasets. Both the evaluations on the three datasets are performed under a single query setting. Furthermore, we report the re-ranking results based on the *k-reciprocal* encoding [26].

B. IMPLEMENTATION DETAILS

The model is performed on the PyTorch framework. We discard the fully-connected layer of the pre-trained ResNet-50 model and reset the kernel size and stride to 1 at layer 4[0]. For branch-1 and branch-2, we set the hidden units of the GRU modules to 256, and the hidden units of the GRU module for branch-3 is set to 128. The training images are resized to 256×128. We use random erasing and random horizontal flipping as data augmentation for training. We set P and K to 16 and 8, respectively. The training epochs are set to 150. We use the adaptive moment estimation (Adam) as the optimizer for the model. The weight decay and initial learning for Adam are set to 5×10^{-4} and 2×10^{-4}, respectively. The learning rate is then set according to the following update rules:

$$lr = \begin{cases} 2 \times 10^{-4} & if\ epoch \leq 100 \\ 2 \times 10^{-4} \times (0.001^{((epoch-100)/50)}) \end{cases} \quad (6)$$

We set all λ_i $(i = 1, 2, 3, 4, 5)$ to 1. The margin $thre$ is set to 0.5. During the test phase, the images are also resized

TABLE 1. Comparison results on the Market-1501 dataset. '-' means no available reported results.

	Single Query			
Method	mAP	Rank-1	Rank-5	Rank-10
DNS [27]	35.6	61.0	-	-
Gated Siamese [28]	39.5	65.8	-	-
HydraPlus-Net [29]	-	76.9	90.9	-
CNN+DCGAN [30]	56.2	78.0	-	-
Embedding [38]	59.8	79.5	92.	95.2
SVD-Net [31]	62.1	82.3	91.3	94.5
Deep Transfer [32]	65.5	83.7	-	-
TriNet [25]+RK	81.0	86.6	93.3	-
CamStyle [39]+RK	71.5	89.4	-	-
HA-CNN [23]	75.7	91.2	-	-
PCB [33]	77.4	92.3	97.2	98.2
Deep-Person [6]	79.6	92.3	-	-
PCB+RPP[33]	81.6	93.8	97.5	98.5
SPReID [37]	83.3	93.6	97.5	98.4
Our	85.4	94.7	**98.6**	**99.0**
+re-ranking [26]	**93.8**	**95.9**	97.7	98.5

to 256 × 128. We use the feature f_{OIM} as the final descriptor to retrieve for the DukeMTMC-reID and Market-1501 datasets. As to the CUHK03 dataset, the feature f_{Trip} is used for retrieval. Note that we alone use the global features for retrieval, because we found that the performance of only adopting global feature for retrieval is almost as good as the fused features. We speculate this is because the global features already can pay attention to the contextual information between different local bodies of person with the help of the part-based branches. Thus, integrating the local feature with contextual information may not help to obviously promote the performance of global feature. Besides, applying the fused feature for retrieval is time consuming, especially when facing the large gallery set. Hence, we only use the global features to retrieve.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

We compare the proposed architecture with the state-of-the-art methods, including DNS [27], Gated Siamese [28], HydraPlus-Net [29], CNN+DCGAN [30], SVD-Net [31], Deep Transfer [32], TriNet [25], PCB [33], Deep-Person [6], PCB+RPP [33], Part-Aligned [22], JLML [16], GOG [34], OIM [10], PAN [35], ACRN [36], SPReID [37] and so on. Among these comparison methods, SPReID achieves the highest performance in the CVPR 2018. The comparison details are presented as follows:

1) EVALUATION ON THE MARKET-1501 DATASET

For this dataset, we use 12936 images of 751 identities for training. The rest of the 750 identities with 19732 images are used for testing. The experimental results of our model against fourteen models on Market-1501 are presented in Table 1. We observe that our model outperforms all competing methods, which demonstrates the effectiveness of the proposed method. Specifically, our model achieves mAP = 85.4%, Rank-1 accuracy = 94.7%, Rank-5 accuracy = 98.6% and Rank-10 accuracy = 99.0% under single query mode. When adopting re-ranking, the mAP and Rank-1 accuracy further achieve to 93.8% and 95.9%, respectively.

TABLE 2. Comparison results on the CUHK03_labeled dataset. '-' means no available reported results.

Method	Rank-1	Rank-5	Rank-10
Siamese LSTM [1]	57.30	80.10	88.30
GOG [34]	67.30	91.00	96.00
Quadruplet [40]	74.47	96.62	98.95
Embedding [38]	83.40	97.10	98.70
Deep Transfer [32]	84.10	-	-
Part-Aligned [22]	85.4	97.6	99.4
MLS Deep [41]	87.5	97.8	99.4
Deep-Person [6]	91.5	99.0	99.5
Deep CRF [42]	90.2	98.5	-
HydraPlus [29]	91.8	98.4	99.1
SPReID [37]	94.2	99.0	99.5
Our	95.0	99.3	99.5
+re-ranking [26]	**97.7**	**99.5**	**99.7**

TABLE 3. Comparison results on the CUHK03_detected dataset. '-' means no available reported results.

Method	Rank-1	Rank-5	Rank-10
DNS [27]	54.7	84.7	94.8
GOG [34]	65.5	88.4	93.7
Gated Siamese [28]	68.1	88.1	94.6
SVD-Net [31]	81.8	95.2	97.2
Part-Aligned [22]	81.6	97.3	98.4
JLML [16]	80.6	96.9	98.7
MLS Deep [41]	86.4	97.5	99.1
Deep CRF [42]	88.8	97.2	-
Deep-Person [6]	89.4	98.2	99.1
Our	92.0	98.9	**99.5**
+re-ranking [26]	**97.8**	**99.1**	99.4

2) EVALUATION ON THE CUHK03 DATASET

For the CUHK03 dataset, we select 1160 identities for training and 100 identities for validation as well as 100 identities for testing. Table 2 and Table 3 show the comparison results on the labeled and detected datasets, respectively. From Table 2 and Table 3, we see that the performances of our model are superior to all other state-of-the-art models on both the two datasets with different settings. Our proposed model yields Rank-1 accuracy = 95.0%, Rank-5 accuracy = 99.3% on the CUHK03_labeled dataset and Rank-1 accuracy = 92.0%, Rank-5 accuracy = 98.9% on the CUHK03_detected dataset. Furthermore, the proposed method obtains 97.7% Rank-1 accuracy on the CUHK03_labeled dataset and 97.8%

TABLE 4. Comparison results on the DukeMTMC-reID dataset. '-' means no available reported results.

Method	mAP	Rank-1	Rank-5	Rank-10
LOMO [43]	17.0	30.7	-	-
DCGAN [30]	47.1	67.6	-	-
PAN [35]	51.5	71.5	-	-
OIM [10]	47.4	68.1	-	-
Embedding [38]	49.3	68.9	-	-
SVD-Net [31]	56.8	76.7	86.4	89.9
TriNet [25]	53.5	72.4	-	-
ACRN [36]	51.9	72.5	84.7	-
Deep CRF [42]	69.5	84.9	-	-
Deep-Person [6]	64.8	80.9	-	-
SPReID [38]	73.3	85.9	92.9	94.5
Our	74.6	86.7	93.4	95.7
+re-ranking [26]	**86.9**	**89.6**	**94.3**	**95.9**

Rank-1 accuracy on the CUHK03_detected dataset with the help of re-ranking.

3) EVALUATION ON THE DukeMTMC-reID DATASET

The 16522 images of 702 persons of this dataset are used for training, and the remaining 702 persons are divided into query images and gallery images. As presented in Table 4, the proposed model achieves mAP = 74.6%, Rank-1 accuracy = 86.7%, Rank-5 accuracy = 93.4% and Rank-10 = 95.7%, which outperforms the compared state-of-the-art methods and further verifies the effectiveness of the proposed method. Similarly, the results of mAP and Rank-1 accuracy are improved to 86.9% and 89.6% by utilizing the re-ranking, respectively.

D. ABLATION ANALYSIS AND DISCUSSIONS

To evaluate the effectiveness of each component in the proposed model, we design several baseline models with different configurations for comparison. The details of the comparison experiments are presented below:

TABLE 5. The effectiveness of GRU-based part features learning.

Model	mAP	Rank-1	Rank-5
Our-G	82.1	92.8	98.3
Our-G-B1	83.5	93.6	98.2
Our	**85.4**	**94.7**	**98.6**

1) EFFECT OF GRU-BASED PART FEATURES LEARNING

In this setting, we first discard the GRU-based part features learning branches, i.e.,Branch-1 and Branch-2, and train the network with the remaining branches. We name this configuration model Our-G. Then, we add Branch-1 to Our-G and name it Our-G-B1. Finally, Branch-2 is added to Our-G-B1. As depicted in Table 5, among the three configurations, Our-G obtains the worst results. When we add Branch-1 to Our-G, the performance of Our-G is obviously promoted. The model that integrates the two branches achieves the best performance among the three configurations, which demonstrates the effectiveness of the GRU-based part features learning branches.

TABLE 6. The effectiveness of exploiting the relationship between channels.

Model	mAP	Rank-1	Rank-5
Our/$_{\text{channels}}$	84.1	93.8	98.4
Our	**85.4**	**94.7**	**98.6**

TABLE 7. The effectiveness of the feature selection strategy.

dataset	Feature	mAP	Rank-1
CUHK03_labeled	f_{OIM}	88.4	91.2
	f_{Trip}	92.2	95.0
Market-1501	f_{OIM}	85.4	94.7
	f_{Trip}	82.1	92.6
DukeMTMC-reID	f_{OIM}	74.6	86.7
	f_{Trip}	71.3	84.4

2) EFFECT OF EXPLOITING THE RELATIONSHIP BETWEEN CHANNELS

In this setting, Branch-3 is discarded. We name this discarded model Our/$_{\text{channels}}$. From Table 6, we observe that the performance of Our/$_{\text{channels}}$ is enhanced when introducing Branch-3, which proves that the interdependencies between different channels learning can improve the discriminative ability of the learned deep descriptors.

3) PERSON DESCRIPTOR CHOICE

As mentioned above, we use the feature f_{Trip} as the person descriptor to re-identify on the CUHK03_labeled dataset. For the Market-1501 and DukeMTMC-reID datasets, we choose the feature f_{OIM} for retrieval. To validate the selection strategy, we also use the feature f_{Trip} for retrieval on the Market-1501 and DukeMTMC-reID datasets and use the feature f_{OIM} for retrieval on the CUHK03_labeled dataset. As shown in Table 7, when using feature f_{OIM} as the descriptor for CUHK03, the accuracies of mAP and Rank-1 obviously decrease. We speculate this is because the CUHK03 dataset contains a large number of identities and each identity only has a limited number of image samples (average 9.6). Using such scale datasets to train the multi-classification branch-3 may lead to an overfitting issue, so we use the feature f_{Trip} as the final representation for the CUHK03 dataset. Different than [6], compared with feature f_{Trip}, f_{OIM} achieves higher results on both the Market-1501 and DukeMTMC-reID datasets. This indicates that OIM loss may achieve better performance in a relatively larger PReID dataset in our deep architecture. Both the results demonstrate the effectiveness of our chosen strategy.

4) WEIGHTS OF LOSSES SETTING

To reveal the effects of trade-off parameters change on the performance of the proposed deep architecture, we set five

different sets of values for these parameters, which are listed as follows:

Setting A: Equal treatment of the five losses, i.e., $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 1$.

Setting B: Pay more attention to optimize the global triple branch, i.e., $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.1$ and $\lambda_5 = 1$.

Setting C: Pay more attention to optimize the global OIM branch, i.e., $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_5 = 0.1$ and $\lambda_4 = 1$.

Setting D: Pay more attention to optimize the GRU-based branches, i.e., $\lambda_1 = \lambda_2 = \lambda_3 = 1$ and $\lambda_4 = \lambda_5 = 0.1$.

Setting E: Pay more attention to optimize the global feature learning branches, i.e., $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$ and $\lambda_4 = \lambda_5 = 1$.

TABLE 8. Effects of different trade-off parameters setting.

	λ_1	λ_2	λ_3	λ_4	λ_5	mAP	Rank-1
Setting A	1	1	1	1	1	**85.4**	**94.7**
Setting B	0.1	0.1	0.1	0.1	1	83.7	93.2
Setting C	0.1	0.1	0.1	1	0.1	84.1	93.9
Setting D	1	1	1	0.1	0.1	83.8	92.8
Setting E	0.1	0.1	0.1	1	1	83.0	93.0

We implement the comparison experiments on the Market-1501 dataset. As shown in Table 8, we can observe that the setting A achieves the highest performance among the five different sets of values. Therefore, we recommend using setting A as trade-off parameters to train the proposed architecture.

TABLE 9. Testing memory and feature extraction time cost.

	Memory Cost	Time Cost (128 images)
Our	3424 MiB	0.6272s
PCB [33]	10157 MiB	0.7040s

5) TESTING MEMORY AND FEATURE EXTRACTION TIME COST

To further verify the effectiveness of the proposed method, we present the comparison of testing memory and feature extraction time cost between the proposed method and PCB. Both the batch sizes of the two methods are set to 128. We implement the comparison experiments on a Nvidia TITAN Xp card. The comparison results are shown in Table 9. From the Table 9, we can see that both the memory and time cost of the proposed method less than the PCB. More specifically, the proposed method only spends 3423 mebibyte for testing, and the average time for feature extraction is 0.0049 seconds per image.

V. CONCLUSION

In this study, we propose an omnidirectional feature learning deep model for person re-identification. The proposed model can learn part representations with spatial information from vertical and horizontal orientations. To further improve the discriminative ability of the learned descriptors, we use a GRU module to exploit the relationship between channels. Moreover, the triplet loss and OIM loss are employed to learn a similarity measurement and global features at the same time. Extensive experiments on the CUHK03, Market-1501 and DukeMTMC-reID datasets show that the proposed deep architecture achieves state-of-the-art results.

REFERENCES

[1] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 135–153.

[2] H. Yao, S. Zhang, Y. Zhang, J. Li, and T. Qi, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, to be published.

[3] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 420–428.

[4] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1077–1085.

[5] L. Zheng, Y. Huang, H. Lu, and Y. Yang. (2017). "Pose invariant embedding for deep person re-identification." [Online]. Available: https://arxiv.org/abs/1701.07732

[6] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu. (2017). "Deep-person: Learning discriminative deep features for person re-identification." [Online]. Available: https://arxiv.org/abs/1711.10658

[7] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Proc. AAAI*, 2017, pp. 3988–3994.

[8] F. Wang, W. Zuo, L. Liang, D. Zhang, and Z. Lei, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1288–1296.

[9] X. Qian, Y. Fu, Y. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5409–5418.

[10] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3376–3385.

[11] D. Wu, S.-J. Zheng, C.-A. Yuan, and D.-S. Huang, "A deep model with combined losses for person re-identification," *Cogn. Syst. Res.*, vol. 54, pp. 74–82, May 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[13] H. Shi *et al.*, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 732–748.

[14] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Proc. IEEE Int. Conf. Adv. Video Signal Surveill.*, Aug. 2017, pp. 2993–3003.

[15] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 135–153.

[16] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. IJCAI*, 2017, pp. 2194–2200.

[17] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7398–7407.

[18] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3980–3989.

[19] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3741–3750.

[20] S. Zhou, J. Wang, S. Rui, Q. Hou, Y. Gong, and N. Zheng, "Large margin learning in set-to-set similarity comparison for person reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 593–604, Mar. 2018.

[21] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[22] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3239–3248.

[23] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 2285–2294.
[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[25] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737
[26] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3652–3661.
[27] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.
[28] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 791–808.
[29] X. Liu *et al.*, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 350–359.
[30] 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.*, Jul. 2017, pp. 3774–3782.
[31] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3820–3828.
[32] M. Geng, Y. Wang, T. Xiang, and Y. Tian. (2016). "Deep transfer learning for person re-identification." [Online]. Available: https://arxiv.org/abs/1611.05244
[33] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. (2017). "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)." [Online]. Available: https://arxiv.org/abs/1711.09349
[34] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1363–1372.
[35] Z. Zheng, L. Zheng, and Y. Yang. (2017). "Pedestrian alignment network for large-scale person re-identification." [Online]. Available: https://arxiv.org/abs/1707.00408
[36] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1435–1443.
[37] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.
[38] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1, p. 13, 2016.
[39] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.
[40] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1320–1329.
[41] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2335–2344.
[42] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8649–8658.
[43] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.

DI WU received the B.S. degree from Zhengzhou University, China, in 2013, and the M.S. degree from the Zhengzhou Institute of Light Industry, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering, Institute of Machine Learning and Systems Biology, Tongji University, China. His research focuses on deep learning and image processing.

HONG-WEI YANG received the B.S. degree from the College of Electronic Information Engineering, Anhui University, China, in 2018. He is currently pursuing the M.C.S. degree with the School of Electronic and Information Engineering, Tongji University, Shanghai, China. From 2017 to 2018, he focused on computer vision issues. His research interests include the person re-identification, image detection, image segmentation, and the detection and segmentation of liver lesions.

DE-SHUANG HUANG (SM'96) received the B.Sc. degree in electronic engineering from the Institute of Electronic Engineering, Hefei, China, in 1986, the M.Sc. degree in electronic engineering from the National Defense University of Science and Technology, Changsha, China, in 1989, and the Ph.D. degree in electronic engineering from Xidian University, Xi'an, China, in 1993. During 1993–1997, he was a Postdoctoral Research Fellow with the Beijing Institute of Technology and also with the National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences (CAS), Beijing, China. In 2000, he joined the Institute of Intelligent Machines, CAS. From 2000 to 2001, he was a Research Associate with The Hong Kong Polytechnic University. In 2003, he visited The George Washington University, Washington, DC, USA, as a Visiting Professor. In 2004, he was the University Fellow with Hong Kong Baptist University. From 2005 to 2006, he was a Research Fellow with The Chinese University of Hong Kong. In 2006, he was a Visiting Professor with Queen's University, Belfast, U.K. In 2007, 2008, and 2009, he was a Visiting Professor with Inha University, South Korea. In 2011, he joined Tongji University as a Chaired Professor, where he is currently the Director of the Institute of Machines Learning and Systems Biology. He has published over 180 journal papers. He has published the book entitled *Systematic Theory of Neural Networks for Pattern Recognition* (in Chinese), in 1996, which received the Second-Class Prize of the 8th Excellent High Technology Books of China, and another two books entitled *Intelligent Signal Processing Technique for High Resolution Radars* (in Chinese), in 2001, and *The Study of Data Mining Methods for Gene Expression Profiles* (in Chinese), in 2009. His current research interests include bioinformatics, pattern recognition, and machine learning. He is a Senior Member of the International Neural Networks Society. He is a Fellow of the International Association of Pattern Recognition. He received the Hundred Talents Program from CAS.

CHANG-AN YUAN received the Ph.D. degree in computer application technology from Sichuan University, China, in 2006. He is currently a Professor with Guangxi Teachers Education University. His research interests include computational intelligence and data mining.

XIAO QIN received the M.S. degree in computer application technology from Guangxi Teachers Education University, China, in 2009, where she is currently an Assistant Professor. Her research interests include computational intelligence and data mining.

YANG ZHAO received the Ph.D. degree in computer application technology from Northeastern University, Shenyang, China, in 2001. He was a Senior Researcher with Lucent Bell Labs (for two years), a Postdoctoral and Senior Technical Manager with China Unicom, Hong Kong (for two years), the Chief Engineer with the Beijing Research Institute, China Telecom (for six years), and the Director of the Network Systems Department, China State Shipbuilding Corporation Information Technology Center (for four years). He has undertaken several national and ministerial research projects. He has been the Chief Engineer (Professorate Senior Engineer) and the Vice President of the Centre Research Institute, Beijing E-Hualu Information Technology Co., Ltd., since 2014. He is a Review Expert of the National Ministry of Science and Technology in the field of communications and information security for the projects of the China Torch Program and the Innovation Fund, and an Expert of the State-Owned Assets Supervision and Administration Commission of the State Council in the field of information security.

XIN-YONG ZHAO received the Ph.D. degree in engineering, majoring in transportation planning and management, from the Harbin Institute of Technology, Harbin, China, in 2008. He was the Deputy Office Director of Computer with the Traffic Management Research Institute, Ministry of Public Security (for one year), an Office Director (for two years), an Assistant of the Director of the Institute (assistant section level, for three years), and the Deputy Director of the Institute (section level, for six years). He concurrently served as the Executive Deputy Director for the National Road Traffic Management Engineering Laboratory (for four years), as the General Manager for Wuxi Huatong Company (for five years), and as the General Manager for Beijing Guotong Company (for eight years), in 2006. He held a temporary position with Wuhan Traffic Management Bureau, and transferred to the Deputy Director of the Road Traffic Safety Research Center, Ministry of Public Security, in 2009 (deputy bureau level). He is currently the Executive Vice President and a Researcher of Beijing E-Hualu Information Technology Co., Ltd. He is the National Expert of the New Century Millions of Talent Projects National Candidates, the Special Allowance Expert of the State Council and the Ministry of Public Security, an Expert of the Ministry of Science and Technology International Cooperation Project, the Safety Production Expert of the State Administration of Work Safety, the Talent of the Top 100 Technology Talents of Beijing, in 2016, and the Talent of the Zhongguancun High-Level Leading Talent Gathering Project, in 2017.

JIAN-HONG SUN received the master's degree from the Xi'an Highway College (currently Chang'an University), in 1997. He is currently the Vice President of Beijing E-Hualu Information Technology Co., Ltd., the President of the Academy of Central Research, and a Senior Engineer.