Collaborative Networks for Person Verification

Yihao Zhang, Wenmin Wang, Jinzhuo Wang
School of Electronics and Computer Engineering, Peking University
ethanyhzhang@pku.edu.cn,wangwm@ece.pku.edu.cn,cr7or9@163.com

ABSTRACT

This paper considers the person verification problem in video surveillance systems. The goal is to verify whether or not a given pair of human body images belong to the same identity. For this purpose, we propose a method of collaborative networks which contains two kinds of novel agents. Specifically, one is implemented by an improved siamese network (iSN) and the other is employed as a deep discriminative network (DDN). The iSN explores the *commonness* and *difference* properties of pairwise feature vectors to enhance the robustness for person verification. Instead, the DDN learns to discriminate the difference of input images from the original difference space, without individual feature extraction. Both of the networks capture the correlation of the input and determine whether they are the same or not. Moreover, we introduce a collaborative learning strategy to fuse them into a unified architecture. Extensive experiments are conducted on four person verification datasets, including CUHK01, CUHK03, PRID2011 and QMUL GRID. We obtain competitive or superior performance compared to state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Matching**; *Object identification*; *Object recognition*;

KEYWORDS

Person Verification; Collaborative Networks; Improved Siamese Network; Deep Discriminative Network; Collaborative Learning

1 INTRODUCTION

Concerns about personal information security is rising with the popularity of multimedia. Under such circumstance, person verification can exert a great help to many problems, e.g. identity spoofing. Biometrics cues like DNA, face, fingerprint, iris, *etc.* are useful for person verification, but they cannot be obtained easily, especially in video surveillance systems. Since large camera networks are increasingly deployed in squares, streets, airports, banks, schools, *etc.*, the demand for person verification in automated video surveillance is also rising. What we need to do is verify whether the same person appears in different videos. From the viewpoint of identification, we think person verification is similar to person re-identification. Although some efforts have been made during these years, there still remains many challenges to be resolved, namely: 1) Different settings of camera parameters; 2) Illumination affected by various angles and imaging time; 3) Alignment problems caused by pose changing and different shooting angles; 4) Body occlusion and background cluttering; 5) Similarity across different identities. Fig. 1 shows some typical challenges in different datasets. To tackle the above issues, traditional methods of person verification or re-identification mainly focus on two parts, feature extraction and metric learning. Feature representation is fundamental, which is the basis of metric learning. The quality of the extracted features might impact the performance a lot. It requires to be discriminative, representative and robust under various conditions. For many years, hand-crafted features have been used for person verification and also gain state-of-the-art results. But they have limitations in considering all the dataset issues at the same time. On the other hand, a good metric is needed to enlarge the variations of inter-class and narrow that of intra-class, in case of mismatch of those in similar appearance.

Recently, deep learning methods are on the rise in computer vision. The success of their application in classification

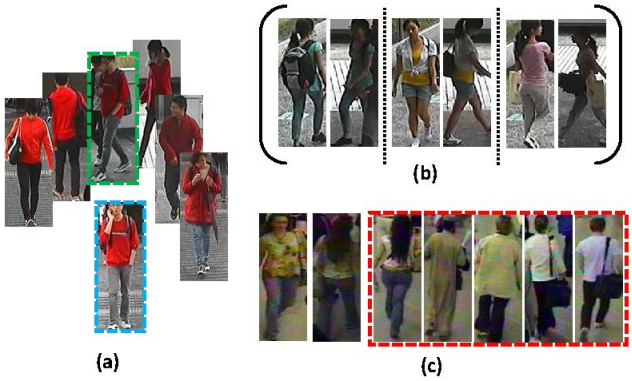


Figure 1: Sample images from cuhk01(a), cuhk03(b) and GRID(c) datasets, displaying parts of challenges to be resolved: (a)similar appearance across different identities with the same one in dashed rectangle; (b)viewpoint changes and illumination variations; (c)low quality of images and additional gallery images shown in red dashed rectangle.

MuVer'17, October 27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-5510-0/17/10...$15.00
DOI: https://doi.org/10.1145/3132384.3132385

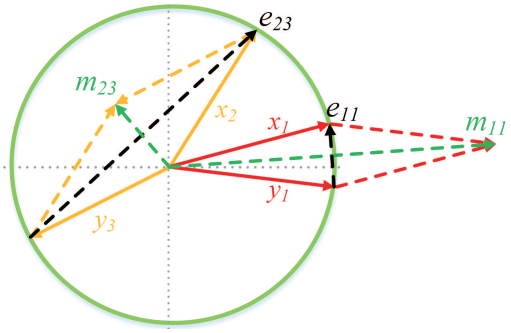


Figure 2: An illustration of *difference* and *commonness* in 2-dimensional Euclidean space. Each feature is l_2-normalized. (x_1,y_1) denotes a similar pair while (x_2,y_3) represents a dissimilar pair. The green circle is a unit one.

task [18] fully illustrates their powerful function in automatically extracting the advanced semantic information. Such end-to-end methods applied to person re-identification also boost the performance. The most popular frameworks used to date are siamese network, triplet network and their variants [10, 11, 25]. Given a pairwise input, a siamese network with two identical sub deep CNNs computes corresponding feature vectors. Then the distance between them is calculated according to a certain criterion. Some other works [1, 33] regard person verification as a classification problem, utilizing the cross correlation acquired from the middle layers to classify whether they are the same or not. Triplet network requires three images of two persons as input and shares the learning parameters across three branches, targeting at enlarging the distance of different persons and shortening that of the same. However, these frameworks only contain a single model. The ability of discrimination might be limited. Inspired by the cooperation mechanisms of human beings, collaboration of multiple networks is considered to perform better. In this paper, we regard each independent and complete network as an agent. Different agents collaborating with one another, either at the early stage by introducing gated functions or at the end by fusing the outputs, is a sort of learning strategy that helps the system extract more abundant and robust information. There are works [5, 7, 25, 33, 35] adopting collaborative learning strategy unconsciously in person verification with outstanding performance.

Following the above discussions, we propose a method of collaborative networks for person verification, which includes two agents. One of the agents improves the typical siamese network, where we introduce two novel components: *commonness* and *difference*. The other agent is called deep discriminative network, where we tell the difference between the input images without individual feature extraction. Then, these two networks are fused into a unified architecture under the strategy of collaborative learning. It is worth mentioning that while dealing with metric learning, Yang et al. [41] proposed the concept of *commonness* and *difference*, denoted as m and e respectively, showing the combination of both could better describe the similarity of two images, measured as follows:

$$r(x, y) = m^T Am - \lambda e^T Be, \tag{1}$$

where A and B are two matrices, parameterizing the similarity and dissimilarity scores, respectively, and the parameter λ is used to balance the effects between similarity and the dissimilarity scores. Fig. 2 better illustrates the concept of *commonness* and *difference*. Different from [41], we use deep framework to generate *commonness* and *difference* vectors and conclude them with a fully connected layer instead of the measurement of Eq. 1.

The main contributions of this paper are fourfold:

- First, we propose an improved siamese network for person verification, which is composed by *commonness* and *difference* components;
- Second, we propose a deep discriminative network for person verification, which learns how to discriminate the difference in the original difference space;
- Third, a collaborative learning strategy is put forward and different networks are fused into a unified architecture;
- Finally, experiments are conducted on several datasets and our method achieves competitive or superior performance compared to the state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, the proposed approach is described. Experimental results and the analyses are given in Section 4. Finally, we present the conclusion in Section 5.

2 RELATED WORK

For better describing a visual appearance of a person, many hand-crafted methods have been proposed, including color histograms [27, 39], local binary patterns (LBP) [19] and SIFT [44], *etc.* Yang et al. [40] proposed a novel salient color names based color descriptor (SCNCD) for semantic analysis of images. Farenzena et al. [12] proposed a symmetry-driven accumulation of local features (SDALF) which made use of symmetry property of human body to resort the viewpoint variation. Zhao et al. [45] proposed a mid-level filter learning method based on the selective patch clusters. In [22], the local maximal occurrence (LOMO) representation was proposed to deal with viewpoint changes by maximizing the horizontal occurrence of local features. Ma et al. [26] turned local features to a final global representation by Fisher vector encoding. Similarly, Matsukawa et al. [29] described a region area in an image via hierarchical Gaussian distribution.

There are increasing researches on seeking an appropriate metric to compute similarity of visual features. The Mahalanobis distance has been the most common metric that has been adopted [9, 13, 37]. Köestinger et al. [15] considered a log likelihood ratio test of two Gaussian distributions and proposed a simple and straightforward algorithm (KISSME). Liao et al. [22] extended the method to cross-view quadratic discriminant analysis (XQDA), where the QDA metric was

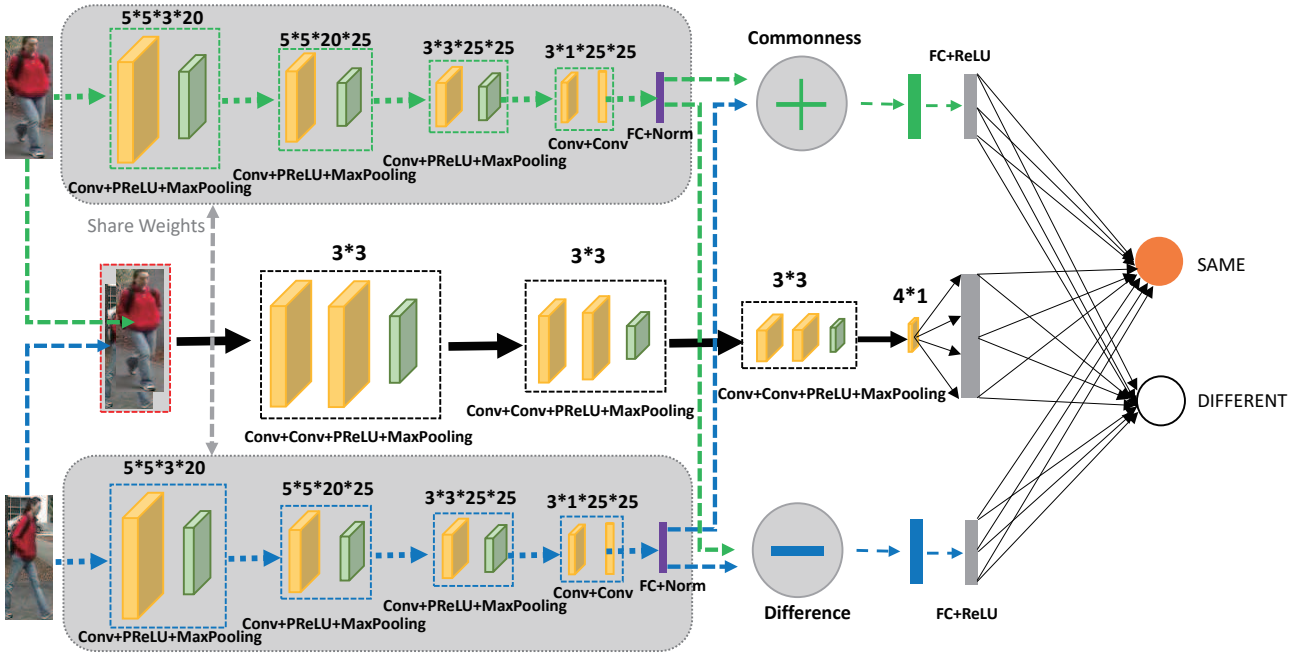


Figure 3: The architecture of our collaborative networks for person verification. The improved siamese network (iSN) has two branches with tied layers in grey background, sharing with each other the same parameters and extracting the input features. The components *commonness* and *difference* take the normalized outputs to calculate the correlation, followed by summary layers. The deep discriminative network (DDN) treats the input as a whole and processes it with a deep neural network. These two networks collaboratively produce the similarity of the input.

learned on the learned discriminant low dimensional subspace. In [41], Yang et al. based on the *commonness* and *difference* of an image pair, proposed a novel similarity measure as a combination of a bilinear similarity metric and a Mahalanobis metric.

Inspired by the great success of deep learning methods applied to other computer vision tasks, many works have achieved outstanding performance in person verification via deep methods. Yi et al. [11] proposed deep metric learning for person verification, where two person images are first separated into three overlapped parts and the image pairs are matched by three siamese convolutional neural networks. Li et al. [21] proposed a new filter pairing neural network, jointly optimizing the feature learning, photometric transforms, geometric transforms, misalignment, occlusions and classification under a unified deep architecture. Based on the siamese network, Ahmed et al. [1] designed a novel layer to compute cross-input neighborhood differences, after which a summary network is followed to classify whether the pairwise input is the same or not. Inspired by [1], Subramaniam et al. [33] improved the performance by normalizing the cross correlation under the inexact matching strategy. They also proposed a fused model of multiple deep networks, and gained a better result. The triplet loss deep convolutional network was proposed by Ding et al. [10] to learn with the relative distance comparison. Only treating person verification as a ranking problem or a classification problem has its insufficiency. Therefore, Chen et al. [5] proposed a multi-task deep network that trained both of the tasks together and gained better performance.

3 PROPOSED APPROACH

After analyzing some state-of-the-art methods [5, 7, 25, 33, 35] applied to person verification recently, we find a common point: they fused different networks in designed ways and achieved better results. Therefore, we summarize them as methods under the strategy of collaborative learning for person verification, referring to the proposal of collaborative networks in other fields [24, 42]. Our proposed approach is also a kind of collaborative method, with different types of networks training simultaneously and interacting with one another.

3.1 Overall Architecture

The overall architecture is illustrated in Fig. 3, including two different types of deep convolutional neural networks. One is an improved siamese network (iSN) composed by *commonness* and *difference* components and the other is deep discriminative network (DDN). The input of our networks is pairwise images. At the end it outcomes the probabilities

Table 1: The performance (%) of iSN with and without summary layer after the *commonness* and *difference* components on QMUL GRID dataset.

Type	rank 1	rank 5	rank 10	rank 20
without summary	0.80	4.00	8.00	16.00
with summary	**16.00**	**36.00**	**51.20**	**69.60**

indicating whether they belong to the same person or not by the logsoftmax function. Note that the binary cross entropy loss might be more popular to be used here for person verification, but in our case the performance of using logsoftmax layer seems to be better. The output of the last logsoftmax layer is a binary variable p defined as:

$$p(y^i = j|x^i; \theta, b) = \log \frac{e_j^T x^i + b_j}{\sum_{l=1}^{k} e_l^T x^i + b_l}. \quad (2)$$

Let y^i=1 if two input images are the same, y^i=0 otherwise. θ and b are the weights and bias terms to be learned. We adopt the negative log-likelihood as criterion. Given the true labels of m training sample pairs, the cost could be written as:

$$J_{cost} = -\frac{1}{m} \sum_{i}^{m} y^i p(y^i = 1|x^i; \theta, b) + (1-y^i) p(y^i = 0|x^i; \theta, b), \quad (3)$$

where m represents the number of training pairs in a batch. Our target is to minimize the J_{cost} by adjusting the parameters of the networks.

3.2 Improved Siamese Network

The iSN in Fig. 3 covers the top and bottom components. The parts in grey background are tied layers with an identical structure and share with each other the same parameters, *e.g.* weights and grad weights. This constraint guarantees the features extracted from two different images with the same filters.

Tied layers: In order to compare the similarity of two person images, distinctive features should be exploited. Convolutional features are proven to be effective for many computer vision tasks. Therefore, we use five convolutional layers to extract the advanced semantic features of the input images. The filters of the first three convolutional layers are symmetry, 5×5, 5×5 and 3×3 respectively. The last two convolutional layers' filters are asymmetry, 3×1 for both. Such strategy only pools the local features along the same row, which also helps reduce the number of parameters compared to symmetric filters. Each convolutional layer, except for the last two, is followed by PReLU (Parametric Rectified Linear Unit) [14] clipping and a max pooling layer which reduces the output dimension by a factor of 2. PReLU is learnable, whose performance is shown to be better than ReLU and the number of its parameters would not intensify the overfitting issue. We further reduce the number by setting the same parameters across different channels. The generated feature maps of the last convolutional layer is then reshaped to a vector before being fed to the fully connected layer with normalization. It finally produces a 500 dimensional vector, ready for the subsequent operations.

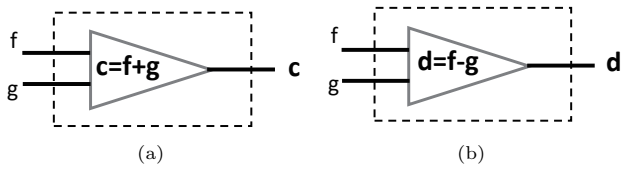


Figure 4: (a) *Commonness* component of additive operation. With the input feature vectors of f and g, the output vector c is the sum of them. (b) *Difference* component of subtractive operation. With the input feature vectors of f and g, the output vector d is the difference of them.

***Commonness* & *difference* components:** In deep learning, there are seldom works exploiting the *commonness* property of pairwise image features for person verification. Our designed *commonness* component takes the two normalized outputs of tied layers as input, and summarizes them with an additive operation (Fig. 4(a)). On the contrary, the *difference* component takes the same two normalized outputs but summarizes them with a subtractive operation (Fig. 4(b)). Then we get two corresponding vectors of *commonness* and *difference*. The 2-norm value of the *commonness* vector is supposed to be large if the input images belong to the same person, and to be small otherwise. Since we attempt to exploit more information from the raw vector, not only with the 2-norm value, the *commonness* vector is preserved for the subsequent processing. The same way is conducted on the *difference* component. However, directly feeding the *commonness* and *difference* vectors into the last deterministic layer will hinder the performance of iSN. The vectors of *commonness* and *difference* contain much mutual information between input images. Without refinement, they might not work. That's why we need to add another fully connected layer to summarize them. To simplify our model, we add only one linear layer followed by ReLU clipping. The performance with and without the summarized layer are compared in Table 1.

We can also see iSN as a combination of two sub networks, one is a siamese network with *commonness* and the other is a siamese network with *difference*. Through collaborating with each other, they could perform better.

3.3 Deep Discriminative Network

Convolutional neural network (CNN) is well known for its powerful ability of feature extraction. Most works benefit a lot from it and compare the difference from the extracted feature space. However, the discriminative power of CNN might be constrained because there may lose something informative in the feature space. In order to tap the potential of CNN to discriminate the difference of pairwise images, we define the concatenation of two images along the color dimension as

Table 2: The CMC performance comparison (%) with the state-of-the-art methods on CUHK01 dataset. The bold indicates the best performance.

Method	CUHK01(test100)					Method	CUHK01(test486)				
	Rank 1	Rank 5	Rank 10	Rank 20	Reference		Rank 1	Rank 5	Rank 10	Rank 20	Reference
eSDC [44]	22.84	43.89	57.67	49.84	2013	mFilters [45]	34.30	55.00	65.00	74.90	2014
KISSME [15]	29.40	57.67	62.43	86.07	2012	Mirror-KFMA [6]	40.40	64.6	75.30	84.10	2015
FPNN [21]	27.87	58.20	73.46	86.31	2014	IDLA [1]	47.50	71.00	80.00	87.44	2015
IDLA [1]	65.00	89.50	93.12	97.20	2015	DeepRanking [4]	50.40	75.90	84.00	91.30	2016
SIRCIR [35]	72.50	91.00	95.50	–	2016	Ensembles [30]	51.90	75.10	83.00	89.40	2015
PersonNet [38]	71.14	90.07	95.00	98.06	2016	ImpTrpLoss [7]	**53.70**	**84.30**	**91.00**	**96.30**	2016
Ours	**77.33**	**93.67**	**97.33**	**99.67**	–	Ours	51.64	74.00	82.30	89.78	–

original difference space. Different images that concatenated into one may bring about noises, but the noises may also conform to a certain distribution which can reflect the images belonging to the same one or not. And our DDN learns to discriminate the difference from the original difference space.

As shown in the middle of Fig. 3, the DDN concatenates the images along the color dimension at the beginning. The original difference space is supposed to contain more complicated but meaningful information for discrimination, thus we push the depth of the network to 8 weight layers, including 7 convolutional layers and 1 fully connected layer. According to VGG-Net [31], compared with larger filters, small filters help make the decision function more discriminative while decreasing the number of parameters. Therefore, we design a conv-block that consists of 2 tightly connected convolutional layers with small filters sized 3×3, followed by PReLU clipping and finally a max pooling layer that reduces the dimension by a factor of 2. The conv-block is continuously repeated for three times, with feature maps of size $64 \times 16 \times 4$ generated. The last convolutional layer uses asymmetry filters of size 4×1, pooling column values within the same row and producing $64 \times 16 \times 1$ sized feature maps. Then the fully connected layer further summarizes the features by generating a 1024 dimensional feature vector.

Note that the architecture of our DDN is simple but it tries to deal with the verification issue from a different angle, which is later proved to be effective.

3.4 Collaborative Learning

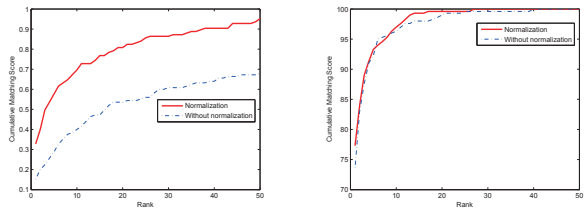
Collaborative learning is a machine learning strategy we proposed, in which two or more agents learn something together. Unlike individual learning, agents engaged in collaborative learning share different experiences and take on asymmetry roles for a common goal. A similar effort can be seen in [36]. For person verification, we design two kinds of agents based on the collaborative learning strategy, one is the iSN and the other is DDN. The collaborative learning of our model reflects in two aspects. For the first one, it happens inside the iSN. Taking a close look at the iSN, it is in fact a fusion model of siamese network with *commonness* and siamese network with *difference*, both of which share the same architecture mostly. Since *commonness* and *difference* components extract different mutual information from the previous tied layers, they will impact each other during the back propagation phase, together instructing how the parameters of tied layers should update. Such kind of collaborative learning makes two networks interact with each other by sharing different experience. For the other, it happens at the end of decision making stage. The outputs of *commonness*, *difference* and DDN are finally fed to a 2-node fully connected layer with 2 logsoftmax units, which determines the probabilities of the same identity. Therefore the loss that propagates back to each network contains the error from another networks, which we show in experiments are essential for improving the performance of person verification. Such kind of collaborative learning helps different agents improve the robustness of themselves by learning the mistakes of others. Using these two kinds of collaborative learning methods, we fuse different networks into a unified architecture and the performance is expected to be promising.

4 EXPERIMENT

4.1 Datasets and Evaluation Protocol

We evaluate our method on four datasets: CUHK01, CUHK03, PRID2011 and QMUL GRID, each divided into training sets and testing sets randomly. Our goal is to match the probe images (captured by one camera) with every gallery image (captured by another camera) in the same dataset. The matching results are recorded as similarity values, and then be ranked from high to low. The percentage of true matches founded among the first m ranked samples is computed and denoted as *rank(m)*. The adopted evaluation metric is called the cumulative match characteristic (CMC), generally used for person verification. Note that the experiments are all conducted in the single shot setting.

CUHK01 Dataset [20]: The CUHK01 dataset is a mid-sized collection of 3884 images belonging to 971 pedestrians, each with 4 images captured from two disjoint cameras in a campus environment. Under one camera view, images show the front and back of a person while under the other camera view, images show the side of the person. There are 2 images for every pedestrian under each camera view and the quality is relatively good, 160×60 uniformly. We conducted two experiments on CUHK01, with one randomly selecting 871 identities for training and the rest 100 identities for testing,



(a) (b)

Figure 5: The comparison of performance with image preprocessing and without image preprocessing. (a)The performance on QMUL GRID dataset. (b)The performance on CUHK01 dataset.

and the other randomly selecting 486 identities for training and the rest 485 identities for testing.

CUHK03 Dataset [21]: The CUHK03 dataset is the first person re-identification dataset that is large enough for deep learning. It provides the bounding boxes detected from deformable part models (DPM) and manually labelled. There are totally 1360 pedestrians of 13164 images captured from 6 different surveillance cameras over months. The misalignment, occlusions and body part missing problems are considered to be more realistic in the dataset. In our experiments, 1260 identities are randomly selected for training and the other 100 identities are for testing, conducted only on the labeled version.

PRID2011 Dataset [16]: The Person Re-ID (PRID) 2011 dataset consists of images captured by two static surveillance cameras, with viewpoint change and a stark difference in illumination, background and camera characteristics. Camera view A and B contain 385 and 749 persons, respectively, with 200 persons appearing in both views. Camera view A is regarded as the probe set and B the gallery set. We randomly select 100 persons for training. For testing, the other 100 persons in camera view A are used as probe set and all the remaining persons in camera view B are used as gallery set, including the irrelevant 549 persons.

QMUL GRID Dataset [2]: The QMUL underGround verification (GRID) dataset contains 250 pedestrian image pairs, each containing two images of the same individual seen from different camera views in a busy underground station. It is a challenging person verification benchmark due to the variations of poses, colors, lighting changes as well as poor image quality caused by the low spatial resolution. We randomly select half of these 250 identities for training and the rest half plus 900 irrelevant persons for testing.

4.2 Implementation Details

We implement the experiments on a machine with Tesla K80 GPU and use the Torch [8] framework. All images are resized to 160×60 to train our model before being fed to the networks. Then we randomly divide them into mini-batches of 128. After performing the forward propagation for each min-batch data, we adopt stochastic gradient descent (SGD) method to update weights for optimization. The parameters of momentum, starting learning rate, learning rate decay and weight decay during training are set to be 0.9, 0.05, 1×10^{-4}, and 5×10^{-4} respectively for all our experiments. Although PRID2011 and QMUL GRID datasets are relatively small, we do not use cross training strategy to fine-tune the model. The training of our model for different datasets are all from scratch. During the training phase, we calculate the mean class accuracy every epoch and stop the training when it converges.

Data Augmentation: Since the number of positive pairs are far less than that of negative pairs in training datasets, it will cause a biased issue if all of them are directly used for training. Thus we control the proportion of positive and negative pairs to be 1:2. Furthermore, we need more data for better training so that the augmentation of the existing data is adopted. For an original image of size $W \times H$, we sample 5 images (2 images for CUHK03) around the image center, with 2D translation drawn from a uniform distribution in the range $[-0.05H, 0.05H] \times [-0.05W, 0.05W]$. We also augment the data by retaining the horizontal flipping version of the original images.

Image Preprocessing: It is mentioned that all images will be resized uniformly, and we also try to normalize them to see whether it could help to improve the performance. Images with normalization and without normalization are both used respectively to train the models for comparison. Meanwhile we found something interesting that data normalization is beneficial for person verification, but could be more useful for small datasets than that for the larger datasets. Comparison can be seen in Fig. 5, which shows the data normalization of images for GRID helps boost the performance greatly. However, for CUHK01 dataset, it has little difference. As a result, the image preprocessing is operated on all the datasets.

4.3 Analysis of the Proposed Model

We have introduced the architecture of our model in Section 3. Since the model takes effect by fusing different components, we would like to probe into the individual performance of different parts and their assembly. Therefore, five cases of three hierarchies are considered in our ablation experiments on QMUL GRID dataset, which are *MC*, *MD*, *MS*, *M6* and *MF* respectively. *MC* represents the case that only uses the *commonness* component in our model. *MD* represents the case that only uses the *difference* component. These two cases are treated as the first hierarchy. The second hierarchy includes *MS* that only uses the iSN in our model and *M6* that only uses the DDN. *MF* indicates the case that uses the complete model, which is also the third hierarchy. With the hierarchy from low to high, the complexity of the cases are increasing. Fig. 6 shows the CMC curves of different cases from rank 1 to rank 50 and we summarize as follows:

The effect of collaborative learning: The performance of different cases shown in Fig. 6 are regular, turning better with the hierarchy from low to high. For example, *MD* and

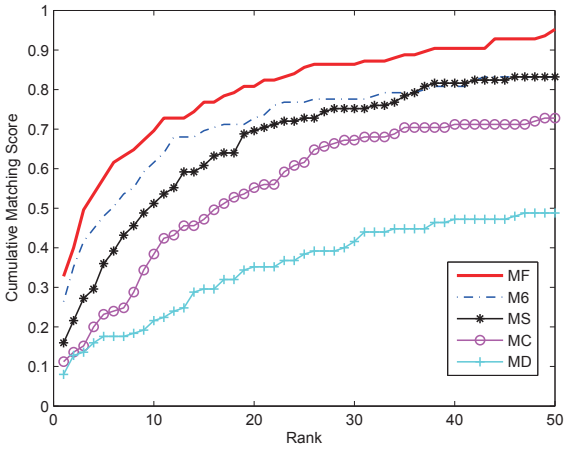


Figure 6: The results of our ablation experiments on QMUL GRID dataset. MF, M6, MS, MC and MD represent the model that is complete, uses DDN only, uses iSN only, uses *commonness* component only and uses *difference* component only respectively.

MC only gained 8% and 11.2% rank 1 accuracy but cooperate with each other, *MS* gained 16.0% rank 1 accuracy. And the collaborative learning with siamese network and 6-channal network helps the complete case *MF* perform the best, with 32.8% rank 1 accuracy. It demonstrates the collaborative learning is not only effective but also significant for the improvement of multiple agents.

The effect of *commonness*: Cases *MD* and *MC* are both very simple compared to [1] and [33], thus we do not attempt to analyze the individual performance here, which will be discussed later. Apparently, *MC* performs better than *MD*. Therefore, we might think that *commonness* plays an more important role than *difference*, which verifies the effectiveness of *commonness* used in deep networks.

The effect of DDN: *M6* performs the second in Fig. 6, fully proving the effectiveness of our DDN. To our knowledge, existing deep methods for person verification have not explored the potential by discriminating the difference of input images from the original difference space. Our work might provide a new direction for study of person verification.

4.4 Comparison with State-of-the-Art Methods

Comparison on CUHK01: The comparison on CUHK01 dataset is recorded in Table 2. The left shows the performance on CUHK01 100 Test Ids. Except for eSDC [44] and KISSME [15], the other four methods all use DCNN models to learn features. Compared with them, our model achieves the highest performance, with 4.83% better than the second one of rank 1 accuracy. The right shows the performance on CUHK01 486 Test Ids. Although the result of our proposed method is not the best compared to other methods listed, it is only about 2% worse than [7] of rank 1 accuracy. The performance is also comparative, validating the effectiveness of our method.

Table 3: The CMC performance comparison (%) with the state-of-the-art methods on CUHK03 (labeled) dataset.

Method	Rank 1	Rank 5	Rank 10	Rank 20
eSDC [44]	8.76	24.07	38.28	53.44
LDML [13]	13.51	40.73	52.13	70.81
KISSME [15]	14.17	48.54	52.57	70.53
FPNN [21]	20.65	51.50	66.50	80.00
LOMO+XQDA [22]	52.20	82.23	92.14	96.25
IDLA [1]	54.74	86.50	93.88	98.10
LOMO+MLAPG [23]	57.96	87.09	94.74	98.00
ensembles [30]	62.10	89.10	94.30	97.80
PersonNet [38]	64.80	89.40	94.92	98.20
NullReid [43]	62.55	90.05	94.80	98.10
MTDnet [5]	**74.68**	**95.99**	**97.47**	–
Ours	62.63	91.63	97.00	**99.65**

Table 4: The CMC performance comparison (%) with the state-of-the-art methods on PRID2011 dataset.

Method	Rank 1	Rank 5	Rank 10	Rank 20
ITML [9]	12.00	–	36.00	47.00
KISSME [15]	15.00	–	39.00	52.00
kLFDA [39]	22.40	46.60	58.10	–
DML [11]	17.90	37.50	45.90	–
NullReid [43]	29.80	52.90	66.00	76.50
Ensembles [30]	17.90	40.00	50.00	62.00
ImpTrpLoss [7]	22.00	–	47.00	57.00
MTDnet [5]	32.00	51.00	62.00	–
Ours	**43.00**	**76.00**	**85.00**	**95.00**

Comparison on CUHK03: Table 3 summarizes the results of the experiments on the CUHK03 Labeled dataset. Compared with most works listed in the table, especially classic deep methods like Li et al. [21] and Ahmed et al. [1], our proposed model achieves better performance. In addition, compared with recent popular works like [30], [38] and [43], the result we got is also competitive. Moreover, we notice the latest result reported by [5], of which the performance is very promising. There might still be a certain gap between the result of our method and the best one, but the difference narrows with the ranking increasing. Our experiments can also validate that it works with the proposed new networks and their combination.

Comparison on PRID2011: Generally, PRID2011 is more challenging than CUHK01 and CUHK03 datasets, due to its variants and the small size with additional gallery images. Therefore, the performance on this dataset has been unsatisfactory. We compare our methods with state-of-the-art methods on PRID2011 in Table 4. We can see that our result achieves the best performance, which is greatly better than the others. The rank 1, rank 5 and rank 10 accuracy beat the second one by 11%, 25% and 23% respectively. We did ablation experiments on PRID2011 and found that *MS* and *M6* contribute almost the same. The result shows that even in

Figure 7: The matching results of several probe images on PRID2011 dataset (from rank 1 to rank 10, from left to right). The left-most column indicates the probe images, and the others on the right in the same row indicate the re-identified gallery images. The images surrounded by dashed rectangle are the true matches. The images surrounded by rectangle are the additional gallery images.

the small dataset, without a fine-tune strategy, deep methods could also be able to shine. We also exhibit some matching results in Fig. 7, which shows the difficulties to re-identify the true match and the effectiveness of our method.

Comparison on QMUL GRID: QMUL GRID is another challenging dataset for person verification due to its small size and low quality, *etc.*, with 775 irrelevant person images in the gallery. The performance of previous state-of-the-art methods is low, which could be seen in Table 5. Note that the best performance of rank 1 accuracy of the existing methods is NLML [17], combination of non-deep and deep methods. We compare our proposed model with these methods, and our model performs the best. Likewise, no fine-tune strategy is taken on this dataset. Our gain in rank 1, rank 5, rank 10 and rank 20 accuracy are 8.26%, 13.04%, 15.52% and 14.40% respectively compared to the second best. We believe such improvement is significant for the challenging dataset of small size. It proves that it is possible for deep methods to achieve better performance than non-deep methods even in small datasets.

4.5 Discussions

In this part, we provide some discussions about the proposed method. For the first, the iSN we design benefits much from the *commonness* and *difference* components. However, compared to [1] and [33], the components are relatively simple in form. We believe there still leaves much to be explored.

Table 5: The CMC performance comparison (%) with the state-of-the-art methods on QMUL GRID dataset.

Method	Rank 1	Rank 5	Rank 10	Rank 20
LOMO+XQDA [22]	16.56	33.84	41.84	47.68
KEPLER [28]	18.40	39.12	50.24	57.04
Norm X-Corr [33]	19.20	38.40	53.60	66.40
NLML [17]	24.54	35.86	43.53	–
SSDAL+XQDA [32]	22.40	39.20	48.00	–
DR-KISS [34]	20.60	39.30	51.40	–
SCSP [3]	24.24	44.56	54.08	59.68
Ours	**32.80**	**57.60**	**69.60**	**80.80**

For the second, the effectiveness of our DDN shows that for the verification task, individual feature extracting might be important but is not the must. More information for discrimination from the original difference space can also be detected by CNN. Finally, we get the trade-off between the complexity of our model architecture and the generalization ability. That's why we could achieve good performance on small datasets without fine-tune strategy. As for the larger datasets, we believe more complicated models could help.

5 CONCLUSION

In this paper, we propose a novel architecture of collaborative networks for person verification, which is composed by two networks. One is the iSN with *commonness* and *difference* components, and the other is the DDN with pairwise input mergence. These two networks play asymmetry roles and are fused into a unified architecture under a collaborative learning strategy. They extract the correlation between two images and interact with each other at different stages, which we show are beneficial for enhancing the robustness of the model. Extensive experiments on four benchmark datasets show promising performance and fully validate the effectiveness of our method. In the future works, we will try to design a better component to extract *commonness* information between pairwise images, which we believe to be helpful for further improving the performance of person verification.

ACKNOWLEDGMENTS

This work was supported by Shenzhen Peacock Plan (20130408-183003656), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405467) and Guangdong Science and Technology Project (2014B010117007).

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K. Marks. 2015. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3908–3916.

[2] C. L. Chen, T. Xiang, and S. Gong. 2009. Multi-camera activity correlation analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1988–1995.

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. 2016. Similarity Learning with Spatial Constraints for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1268–1277.

[4] S. Z. Chen, C. C. Guo, and J. Lai. 2016. Deep Ranking for Person Re-identification via Joint Representation Learning. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 25, 5 (2016), 2353–2367.

[5] W. Chen, X. Chen, J. Zhang, and K. Huang. 2017. A Multi-task Deep Network for Person Re-identification. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[6] Y. C. Chen, W. S. Zheng, and J. Lai. 2015. Mirror representation for modeling view-specific transform in person re-identification. In *International Conference on Artificial Intelligence*. 3402–3408.

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. 2016. Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1335–1344.

[8] Ronan Collobert, Koray Kavukcuoglu, and Clment Farabet. 2011. Torch7: A Matlab-like Environment for Machine Learning. In *BigLearn, NIPS Workshop*.

[9] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. 2007. Information-theoretic metric learning. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference*. 209–216.

[10] S. Ding, L. Lin, G. Wang, and H. Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* 48, 10 (2015), 2993–3003.

[11] Y. Dong, L. Zhen, S. Liao, and Stan Z. Li. 2014. Deep Metric Learning for Person Re-identification. In *International Conference on Pattern Recognition*. 34–39.

[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR)*. 2360–2367.

[13] M Guillaumin, J Verbeek, and C Schmid. 2009. Is that you? Metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*. 498–505.

[14] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision*. 1026–1034.

[15] Martin Hirzer. 2012. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition*. 2288–2295.

[16] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. 2011. Person Re-Identification by Descriptive and Discriminative Classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*.

[17] S. Huang, J. Lu, J. Zhou, and Anil K. Jain. 2015. Nonlinear Local Metric Learning for Person Re-identification. *Computer Science* (2015).

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*. 1097–1105.

[19] W. Li and X. Wang. 2013. Locally Aligned Feature Transforms across Views. In *Computer Vision and Pattern Recognition*. 3594–3601.

[20] W. Li, R. Zhao, and X. Wang. 2012. Human Reidentification with Transferred Metric Learning. In *ACCV*.

[21] W. Li, R. Zhao, T. Xiao, and X. Wang. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.

[22] S. Liao, Y. Hu, X. Zhu, and Stan Z. Li. 2015. Person re-identification by Local Maximal Occurrence representation and metric learning. In *Computer Vision and Pattern Recognition (CVPR)*. 2197–2206.

[23] S. Liao and Stan Z. Li. 2015. Efficient PSD Constrained Asymmetric Metric Learning for Person Re-Identification. In *IEEE International Conference on Computer Vision*. 3685–3693.

[24] K. Lin, S. Wang, and J. Zhou. 2017. Collaborative Deep Reinforcement Learning. *arXiv preprint arXiv:1702.05796* (2017).

[25] J. Liu, Z. J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. 2016. Multi-Scale Triplet CNN for Person Re-Identification. In *ACM on Multimedia Conference*. 192–196.

[26] B. Ma, Y. Su, and Frederic Jurie. 2012. Local Descriptors Encoded by Fisher Vectors for Person Re-identification. In *European Conference on Computer Vision (ECCV)*. 413–422.

[27] L. Ma, X. Yang, and D. Tao. 2014. Person Re-Identification Over Camera Networks Using Multi-Task Distance Metric Learning. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 23, 8 (2014), 3656–70.

[28] Niki Martinel, Christian Micheloni, and Gian Luca Foresti. 2015. Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning. *IEEE Transactions on Image Processing* 24, 12 (2015), 5645–5658.

[29] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. 2016. Hierarchical Gaussian Descriptor for Person Re-identification. In *Computer Vision and Pattern Recognition (CVPR)*. 1363–1372.

[30] Sakrapee Paisitkriangkrai, Chunhua Shen, and Van Den Hengel Anton. 2015. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1846–1855.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2014).

[32] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. 2016. Deep Attributes Driven Multi-Camera Person Re-identification. *arXiv preprint arXiv:1605.03259* (2016).

[33] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. 2016. Deep Neural Networks with Inexact Matching for Person Re-Identification. In *Advances in Neural Information Processing Systems 29*. 2667–2675.

[34] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang. 2016. Person Re-Identification by Dual-Regularized KISS Metric Learning. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 25, 6 (2016), 2726–2738.

[35] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. 2016. Joint Learning of Single-Image and Cross-Image Representations for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1288–1296.

[36] Jinzhuo Wang, Wenmin Wang, Ronggang Wang, Wen Gao, and others. 2016. Deep Alternative Neural Network: Exploring Contexts as Early as Possible for Action Recognition. In *Advances in Neural Information Processing Systems*. 811–819.

[37] Kilian Q Weinberger and Lawrence K Saul. 2006. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *NIPS*. 207–244.

[38] L. Wu, Shen C., and Anton van den Hengel. 2016. PersonNet: Person Re-identification with Deep Convolutional Neural Networks. *arXiv preprint arXiv:1601.07255* (2016).

[39] F. Xiong, M. Gou, Octavia Camps, and Mario Sznaier. 2014. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*. Springer, 1–16.

[40] Y. Yang. 2014. Salient color names for person re-identification. In *European Conference on Computer Vision*.

[41] Y. Yang, S. Liao, Z. Lei, and Stan Z Li. 2016. Large scale similarity learning using similar pairs for person verification. In *AAAI*. 3655–3661.

[42] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, H. Zhou, and X. Wang. 2016. Crafting GBD-Net for Object Detection. *arXiv preprint arXiv:1610.02579* (2016).

[43] L. Zhang, T. Xiang, and S. Gong. 2016. Learning a Discriminative Null Space for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1239–1248.

[44] R. Zhao, W. Ouyang, and X. Wang. 2013. Unsupervised Salience Learning for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3586–3593.

[45] R. Zhao, W. Ouyang, and X. Wang. 2014. Learning Mid-level Filters for Person Re-identification. In *Computer Vision and Pattern Recognition (CVPR)*. 144–151.