Cross-domain Person Re-identification using Heterogeneous Convolutional Network

Zhong Zhang, *Senior Member, IEEE,* Yanan Wang, Shuang Liu, *Senior Member, IEEE,* Baihua Xiao, Tariq S. Durrani, *Life Fellow, IEEE*

***Abstract*—Person re-identification (Re-ID) is a challenging task due to variations in pedestrian images, especially in cross-domain scenarios. The existing cross-domain person Re-ID approaches extract the feature from single pedestrian image, but they ignore the correlations among pedestrian images. In this paper, we propose Heterogeneous Convolutional Network (HCN) for cross-domain person Re-ID, which learns the appearance information of pedestrian images and the correlations among pedestrian images simultaneously. To this end, we first utilize Convolutional Neural Network (CNN) to extract the appearance features for pedestrian images. Then we construct a graph in the target dataset where the appearance features are treated as the nodes and the similarity represents the linkage between the nodes. Afterwards, we propose Dual Graph Convolution (DGConv) to explicitly learn the correlation information from the similar and dissimilar samples, which could avoid the over-smoothing caused by the fully connected graph. Furthermore, we design HCN as a multi-branch structure to mine the structural information of pedestrians. We conduct extensive evaluations for HCN on three datasets, i.e. Market-1501, DukeMTMC-reID and MSMT17, and the results demonstrate that HCN is superior to the state-of-the-art methods.**

***Index Terms*—cross-domain person re-identification, graph convolution network, dual graph convolution**

I. Introduction

Person re-identification (Re-ID) [1]–[4] aims to determine whether the same identity appears in different cameras, which could cooperate with other recognition technologies [5]–[7] to make up for the vision limitation of fixed cameras. Its practical application fields include criminal investigation, object localization, monitoring, security and so on [8]–[12]. Since the pedestrian appearance is easily affected by the changes of posture, occlusion and illumination, person Re-ID is significantly challenging.

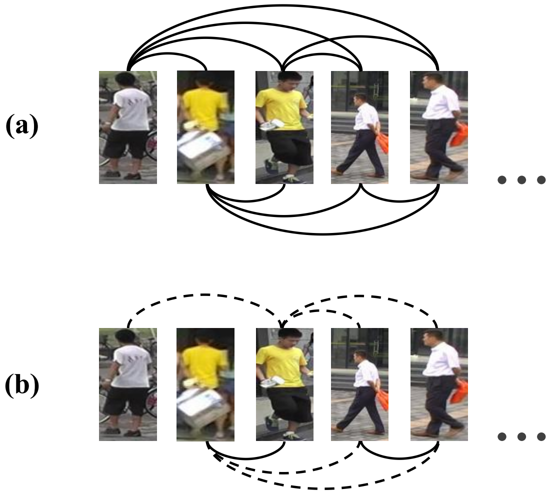


Fig. 1: (a) The linkages among samples for traditional graph convolution operations. (b) The linkages among samples for DGConv, where the solid line indicates that the connected sample pair has the same pseudo label, and the dotted line means the opposite.

Recently, person Re-ID has achieved promising performance when training and testing on the same dataset, but the performance is unsatisfactory when testing on a different dataset. The unsupervised domain adaptation (UDA) [13]–[15] trains a model with a labeled source dataset and an unlabeled target dataset, and it could perform well on the target dataset. Hence, researchers gradually resort to UDA to enhance the generalization ability of Re-ID model, and this kind of setting is known as cross-domain person Re-ID.

The UDA approaches are widely used for image classification and the assumption is that the categories in source and target datasets are same. But the categories (identities) in person Re-ID datasets are totally different, and therefore the UDA approaches can not be directly applied in the field of cross-domain person Re-ID. In order to overcome the above-mentioned limitation, some cross-domain person Re-ID approaches employ Generative Adversarial Network (GAN) [16] and its derivative version [17]–[21] to implement style transformation and information delivery for cross-domain scenarios. Specifically, some of them conduct the style transformation from the source dataset to the target dataset, while others focus on the camera style adaptation within the target dataset. Some other cross-domain person Re-ID approaches

This work was supported by National Natural Science Foundation of China under Grant No. 61711530240, Natural Science Foundation of Tianjin under Grant No. 20JCZDJC00180 and No. 19JCZDJC31500, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No. 202000002, and the Tianjin Higher Education Creative Team Funds Program. (Corresponding author: Shuang Liu.)

Zhong Zhang, Yanan Wang and Shuang Liu are with Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China (e-mail: zhong.zhang8848@gmail.com, yananwang585@gmail.com, shuangliu.tjnu@gmail.com).

Baihua Xiao is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: baihua.xiao@ia.ac.cn).

Tariq S. Durrani is with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow Scotland, UK (e-mail: t.durrani@strath.ac.uk).

Copyright © 20xx IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

are proposed to assign potential identity labels to the unlabeled target dataset using clustering algorithms [22]–[24]. As a result, the target dataset can be used to train the deep model in a supervised manner and the training data is augmented. However, the existing cross-domain person Re-ID approaches ignore the correlation information among pedestrian images, which could not learn useful information from other pedestrian images.

In this paper, we propose a novel Re-ID model named Heterogeneous Convolutional Network (HCN) for cross-domain person Re-ID, which applies Convolutional Neural Network (CNN) and Graph Convolution Network (GCN) to simultaneously learn the appearance features and the correlation information among pedestrian images. To this end, we first employ CNN to extract appearance features (CNN-based features) for the unlabeled target dataset. Then, we cluster the CNN-based features to generate pseudo labels for the unlabeled target dataset. In order to emphasize the correlations among the target samples, we establish the linkages between pedestrian images as shown in Fig. 1(a). Specifically, we treat the CNN-based feature of each pedestrian image as a node of graph. We further construct an adjacency matrix according to similarities between pedestrian images, which could reflect the strength of the linkage.

After obtaining the graph and the adjacency matrix, the node information can be transmitted in the graph using the graph convolution operation. Since the linkages are established among all the nodes, the graph is fully connected. When traditional graph convolution operations [25], [26] conduct on the fully connected graph, the trivial and redundant linkages could reduce the discrimination of samples and result in the model over-smoothing, especially when using multiple graph convolution layers. Hence, we propose the Dual Graph Convolution (DGConv) to explicitly learn the correlations from the similar and dissimilar samples by only selecting the high confidence linkages according to the adjacency matrix, where we regard that the samples with the same pseudo label are similar and the samples with different pseudo labels are dissimilar. As shown in Fig. 1(b), the high confidence linkages for similar samples and dissimilar samples are selected respectively, which could avoid the dissimilar linkages between similar samples and the similar linkages between dissimilar samples, so that the GCN-based features obtained by DGConv possess more discriminative correlation information.

Furthermore, we divide the pedestrian image into the global, upper and lower parts to capture the structural information and mine complete identity information. Correspondingly, we design a multi-branch structure for HCN where each branch handles the different parts of pedestrian image by the abovementioned process. In a nutshell, our main contributions are summarized as follows:

• We propose HCN to simultaneously consider the appearance features and the correlation information for cross-domain person Re-ID, which is the first one to consider the correlation information among pedestrian images in the target domain.

• We propose DGConv to explicitly propagate the important correlation information of similar and dissimilar samples, which effectively avoids the trivial and redundant linkages and improves the discrimination of pedestrian images, so that the similar samples are closer and the dissimilar samples are farther after the DGConv operation.

• We evaluate the performance of HCN on three person Re-ID datasets, i.e., Market-1501 [27], DukeMTMC-reID [28] and MSMT17 [29], and the results are significantly superior to the state-of-the-art methods.

II. Related Work

A. Unsupervised Domain Adaptation

Our work is related to the UDA methods [13]–[15] which focus on reducing the domain discrepancy using a labeled source dataset and an unlabeled target dataset. Some researchers pay more attention to the alignment of different domain distributions [30]–[33]. For example, Ganin et al. [30] combine the optimized features and two discriminant classifiers to align the distribution difference between the source and target domains. Sun et al. [31] propose the CORrelation ALignment (CORAL) to reduce the domain shift by aligning the feature covariances of target domain and source domain. To simulate the target images, Hsu et al. [33] obtain an intermediate domain by translating the source images and utilize adversarial learning to conduct two adaptation subtasks.

The others attempt to enhance the model robustness through image or feature translation [13], [34], [35]. For example, Bousmalis et al. [13] present an unsupervised pixel-level domain adaptation method which maps the source images into the target style at the pixel level while keeping their original contents. Volpi et al. [34] perform data augmentation in the feature space using the Conditional GAN (CGAN) which learns the class distribution of training samples and generates sample features of desired classes.

B. Unsupervised Person Re-Identification

Unsupervised person Re-ID is closer to the real scene and it can be applied more flexibly. The predefined manual features [36]–[38] can be directly used for unsupervised person Re-ID, but they are ineffective for large-scale datasets. To solve this issue, some person Re-ID methods apply GAN to generate samples for data augmentation [39]–[42]. For instance, Deng et al. [39] utilize cycleGAN [18] to transform the style of pedestrian images from the source dataset to the target dataset, and apply SiaNet network to maintain more identity information for the generated images. In [40], starGAN [20] is employed to implement the camera-style translation in the unlabeled target dataset, and all generated samples are added into the training set in order to learn the camera and domain invariant model. Zhong et al. [41] propose the exemplar memory to learn three kinds of invariant issues from both intra-domain and inter-domain for cross-domain person Re-ID. Furthermore, they [42] present the Graph-based Positive Prediction (GPP) to promote the invariance learning. They optimize GCN in the source domain to infer the positive and negative neighbors in the target domain, while our work optimizes GCN using pedestrian images from the target domain in order to model the sample correlations in the target domain.

Some recent works realize unsupervised person Re-ID via clustering the unlabeled dataset [43]–[46]. Fan et al. [43] propose the Progressive Unsupervised Learning (PUL) which applies the K-means clustering to assign pseudo labels for the unlabeled samples. Fu et al. [44] propose the Self-similarity Grouping (SSG) to assign new labels to unlabeled target dataset after the self-similarity grouping. Wang et al. [45] present smoothing adversarial domain attack (SADA) to align the source and target images at the image level, and then the source-aligned images and the target images after clustering are utilized to optimize the Re-ID network where p-Memory Reconsolidation (pMR) is proposed to retain the source knowledge by a small probability p.

C. Graph Convolutional Networks

GCN [47] is proposed to handle the graph structure data, and it has shown special advantages in many fields, such as image generation, person search and multi-label image recognition [25], [26], [48]–[50]. There are two main ways to implement GCN where one is based on spectral domain and the other is based on spatial domain. The spectral based GCN [51]–[53] realizes the convolution operation of graph according to the graph Fourier transformation [54]. The spatial based GCN [55]–[58] relies on the neighborhood of nodes in the graph. Specifically, after determining node neighbors and the receptive field, the graph convolution layer is applied to propagate the information between nodes and their neighbors. Inspired by the spatial based GCN, we try to apply GCN to the field of cross-domain person Re-ID to fully learn the complex interaction of pedestrian images.

III. Approach

A. Overview

The training samples of HCN include the labeled source and unlabeled target datasets, and the proposed HCN is mainly composed of the CNN and GCN models. We first utilize the source dataset to pre-train the CNN model, so that the pre-trained CNN model can be employed to extract the appearance features of the samples in the unlabeled target dataset. Then the unsupervised clustering algorithm is applied to cluster these appearance features and assign pseudo labels for the target dataset. Afterwards, we apply GCN to explore the correlations among samples of the target dataset. Specifically, we exploit the similarity between the target samples to construct the graph, and then DGConv is conducted on the graph. Hence, the appearance features and the correlations among the target samples can be fully considered in the feature learning process. Furthermore, we extend the proposed HCN to the multi-branch version to capture the structural information and mine complete identity information. Next, we will introduce the above-mentioned process in detail.

B. Pre-training

Although the model trained on the specific domain performs poorly in another domain, the model pre-trained with the source dataset is helpful for the subsequent training in the target dataset. Hence, we pre-train the CNN model using the labeled source dataset in a supervised manner, as shown in Fig. 2. Specifically, we treat ResNet-50 [59] as CNN, and we replace the layers after the average pooling of ResNet-50 with two fully connected (FC) layers. One FC layer is FC-1024 with 1024 neurons, and the other is FC-ID where ID denotes the identity number of the source dataset. With the CNN model, the appearance features of samples can be extracted appropriately, and we name these appearance features as the CNN-based features. Furthermore, we apply the triplet loss and the cross-entropy loss to optimize the CNN model, where the triplet loss could learn the similarity measurement of positive and negative sample pairs and the cross-entropy loss is applied to train the classification ability of CNN. Finally, we utilize the pre-trained model to initialize the CNN model of HCN.

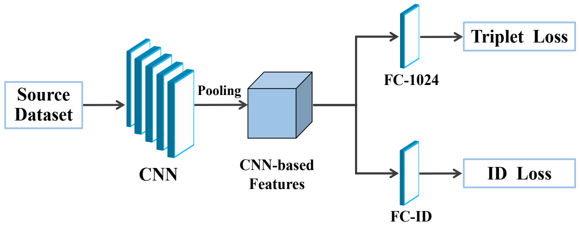


Fig. 2: The framework of pre-training stage, where the ID loss represents the cross-entropy loss.

C. Unsupervised Clustering for Training HCN

As shown in Fig. 3, we first apply the pre-trained CNN model to extract the CNN-based features for the unlabeled target dataset. Then we utilize the unsupervised clustering algorithm [60] to cluster these CNN-based features, in which the unlabeled target samples are divided into different regions in the feature space. For each region, the samples are assigned to the same identity label, so that a pseudo-labeled target dataset is obtained.

D. Correlation Modeling in the Target Dataset

Graph Construction. Considering the correlations among pedestrian images is beneficial for each sample to learn useful information from other samples. However, the existing cross-domain person Re-ID methods [40], [42], [61] tend to learn the appearance information from individual sample, thereby ignoring the correlations among pedestrian images. To overcome the limitation, we establish a graph based on the pseudo-labeled target dataset, and employ GCN to learn the correlations among the target samples.

As shown in Fig. 3, after clustering, we construct a graph where the CNN-based features of target samples are treated as the nodes and the linkages between the nodes are measured by the similarities. The feature matrix of the graph is composed by the nodes:

$$F = [f_1, \dots, f_n, \dots, f_N] \tag{1}$$

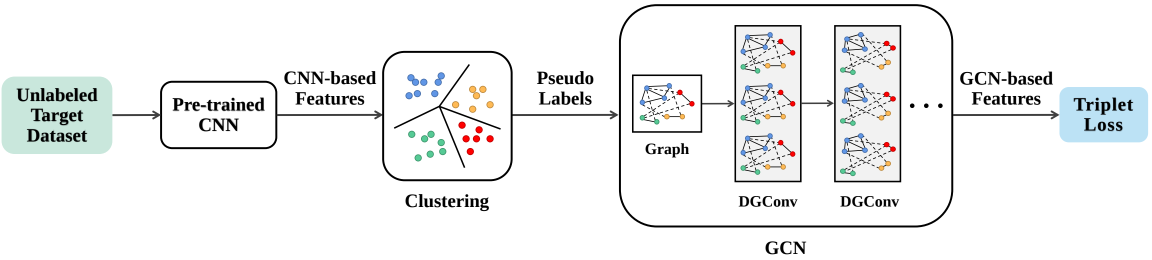


Fig. 3: The framework of HCN. We first utilize the pre-trained CNN model to extract the CNN-based features for the unlabeled target samples. Then we exploit the unsupervised clustering algorithm to cluster these CNN-based features and assign pseudo labels for the unlabeled target samples. Afterwards, the graph in the GCN model is constructed based on pseudo-labeled target dataset, where the different colors represent the samples with different pseudo labels, the solid line indicates the sample pair with the same pseudo label, the dotted line indicates the sample pair with different pseudo labels, and the length of each line represents the similarity between samples. The output of the GCN model is the GCN-based features. Finally, the model is optimized by the triplet loss with the pseudo labels of the target dataset.

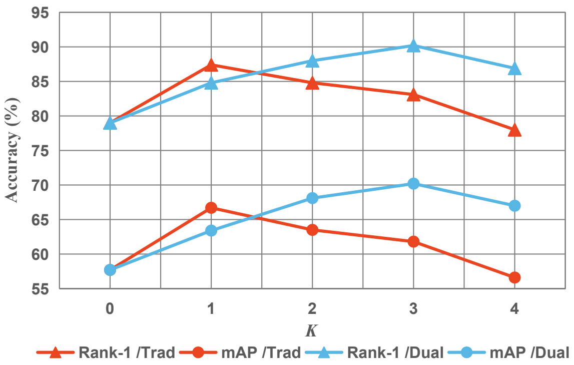


Fig. 4: The performance of HCN (for D-M) with the traditional GCN layers (red line) and the DGConv layers (blue line). "K=0" means that the GCN model is removed, "Trad" represents the traditional GCN operation and "Dual" is the DGConv operation.

where $F \in \mathbb{R}^{N \times d}$, N is the number of nodes, d is equal to 2048, and f_n represents the n-th node (i.e., the CNN-based feature of the n-th target sample) in the graph. To reflect the linkage strength among the target samples, we calculate the similarities between all node pair to obtain a fully connected graph. The adjacency matrix of the fully connected graph is based on the similarities:

$$A = [a_{ij}], 1 \leq (i,j) \leq N \tag{2}$$

where $A \in \mathbb{R}^{N \times N}$ and a_{ij} represents the linkage strength between the i-th node and the j-th node. It is formulated as:

$$a_{ij} = \frac{f_i \cdot {f_j}^T}{\sum_{u=1}^{N} \sum_{v=1}^{N} (f_u \cdot f_v^T)} \tag{3}$$

Dual Graph Convolution Operation. When utilizing the traditional graph convolution operations [25], [26], the feature matrix of the fully connected graph is updated by:

$$E^{k+1} = h(AE^k W^k), 1 \leq k \leq K \tag{4}$$

where W^k is the trainable parameters of the k-th graph convolution layer, E^k represents the input feature matrix of the k-th graph convolution layer, K is the total number of graph convolution layers, and h indicates the nonlinear transformation which is usually implemented by ReLU.

However, propagating information from all the nodes in the fully connected graph is prone to involving trivial and redundant linkages. When the node aggregates and absorbs the information from its connected nodes, excessive linkages may confuse the identity information of nodes. The nodes from different identities become similar especially when the number of graph convolution layers increases, which leads to the decline of the node discrimination and the model oversmoothing. As a result, the performance of the model becomes smooth or even worse as shown in the red curves of Fig. 4. To avoid this phenomenon, we propose DGConv to explicitly learn the important correlation information from the similar and dissimilar samples.

Since the target dataset is pseudo-labeled, we regard that the sample pair with the same pseudo label is similar and the sample pair with different pseudo labels is dissimilar. Then we only select the high confidence elements from the adjacency matrix, that is, we retain the linkages between high similar samples and the linkages between high dissimilar samples. The selection criterions are expressed as:

$$A_1 = [a_{ij}^1] \quad where \quad a_{ij}^1 = \begin{cases} a_{ij}, \ if \, a_{ij} > \theta(\tau_1) \ and \ l_i = l_j \\ 0, \quad otherwise \end{cases} \tag{5}$$

$$A_2 = [a_{ij}^2] \quad where \quad a_{ij}^2 = \begin{cases} a_{ij}, \ if \, a_{ij} < \theta(\tau_2) \ and \ l_i \neq l_j \\ 0, \quad otherwise \end{cases} \tag{6}$$

where $A_1 \in \mathbb{R}^{N \times N}$, $A_2 \in \mathbb{R}^{N \times N}$, l_i and l_j are the pseudo labels of the i-th and j-th nodes, and $\theta(\tau_1)$ and $\theta(\tau_2)$ are the thresholds which represent the element values at τ_1% and τ_2% of the adjacency matrix elements in the descending order. Hence, A_1 retains the linkages between the high similar nodes and A_2 keeps the linkages between the high dissimilar nodes.

Compared to the adjacency matrix A in Eq. (4), A_1 and A_2 not only reduce the trivial and redundant linkages among samples but also avoid the dissimilar linkages between similar samples and the similar linkages between dissimilar samples, so that the similar samples are closer and the dissimilar samples are farther after the DGConv operation. The proposed DGConv is formulated as:

$$E^{k+1} = h(W^k \sigma(A_1 E^k \parallel A_2 E^k)), 1 \leq k \leq K \tag{7}$$

where W^k, E^k and h are consistent with Eq. (4), $\parallel$ is the concatenation operation, and σ is the non-linear transformation to aggregate features for the similar and dissimilar samples. Here, σ is implemented by a FC layer which follows BN and ReLU, and the input feature matrix of the first DGConv layer E^1 is initialized by F.

The DGConv operation in Eq. (7) can be regarded as the graph convolution based on A_1 to make the similar samples close, and the graph convolution based on A_2 to push the dissimilar samples away, in order to improve the discrimination of sample features. In Fig. 4, as for the traditional graph convolution, the model achieves the best performance when the number of graph convolution layers is equal to 1, while as for the proposed DGConv, the best performance is obtained when the number of graph convolution layers is 3. Meanwhile, the proposed DGConv achieves better performance than the traditional graph convolution. Hence, the reduction of the number of linkages in the graph avoids the decline of the discrimination of nodes, so that the DGConv operation could relieve the model over-smoothing in a certain extent. Afterwards, we can obtain the GCN-based feature of each target sample which contains the appearance feature and the important correlation information from its similar and dissimilar samples.

E. Extension to Multi-branch HCN

To mine the structural information of pedestrians, we extend HCN to a four-branch structure where three branches are utilized to optimize the GCN-based features of global, upper and lower parts and one branch is applied to optimize the CNN-based features of global parts, and the framework of multi-branch HCN is shown in Fig. 5. Specifically, we first horizontally divide the pedestrian images into two uniform parts, and then extract and cluster the global, upper and lower CNN-based features. Thus, we can obtain three groups of pseudo labels corresponding to the global, upper and lower parts of pedestrian images for the target dataset. Assuming that the unlabeled target dataset is Z_t, we denote the i-th pedestrian image of the target dataset as z_t^i. Hence, the updated target dataset is represented as:

$$Z_t^{'} = [z_t^i : y_g^i, y_u^i, y_l^i], 1 \leq i \leq N_t, \tag{8}$$

where y_g^i, y_u^i and y_l^i denote the generated pseudo labels corresponding to the global, upper and lower parts of the i-th image z_t^i, and N_t is the total number of the target samples. Since the clustering process is carried out within each group of features, the three groups of labels are independent, i.e., $y_g^i \neq y_u^i \neq y_l^i$.

After labeling the global, upper and lower parts of pedestrian images, we further utilize DGConv to extract the GCN-based features for them. As a result, we employ the triplet loss to optimize the four-branch structure for HCN, in which the first three branches are based on the global, upper and lower GCN-based features, and the fourth branch learns the global CNN-based features with an additional FC layer, i.e., FC-1024.

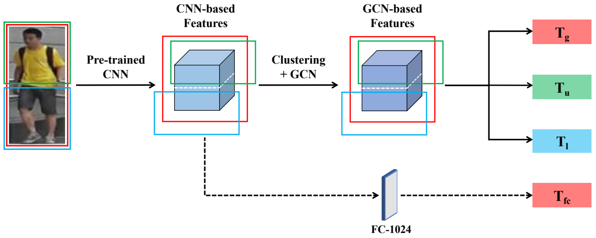


Fig. 5: Framework of multi-branch HCN, where the blocks in red, green and blue represent the operations on the global, upper and lower parts of samples, respectively. The black dashed line indicates that we add a FC (i.e., FC-1024) branch to optimize the global CNN-based features for pedestrian images, and T_g, T_u, T_l and T_{fc} represent the triplet losses of the four branches, respectively.

F. Optimization

In the unsupervised training stage, we conduct multiple iterations to update the pseudo labels and the network circularly. With the pseudo-labeled target dataset, the triplet loss is utilized as the optimization function for the multi-branch framework of HCN. Hence, the total loss of HCN is formulated as:

$$L_{HCN} = \omega_{T_g} T_g + \omega_{T_u} T_u + \omega_{T_l} T_l + \omega_{T_{fc}} T_{fc} \tag{9}$$

where T_g, T_u, T_l and T_{fc} denote the triplet losses of the four branches, and ω_{T_g}, ω_{T_u}, ω_{T_l} and $\omega_{T_{fc}}$ are weights of the four losses. Take the global part as an example, the loss of the first branch is expressed as:

$$T_g = \sum_{i=1}^{N_t} [m + \parallel f_{ig}^a - f_{ig}^p \parallel_2 - \parallel f_{ig}^a - f_{ig}^n \parallel_2]_+ \tag{10}$$

where m denotes the margin and $[x]_+ = max(x, 0)$. Here, f_{ig}^a is the global GCN-based feature of the i-th anchor image, f_{ig}^p is the global GCN-based feature of the positive sample farthest to the i-th anchor image, and f_{ig}^n is the global GCN-based feature of the negative sample nearest to the i-th anchor image. T_u, T_l and T_{fc} can be defined analogously to T_g.

In a word, HCN can not only considers the appearance features and the correlation information among pedestrian images, but also realizes the joint optimization of the global and partial features of pedestrian images.

IV. Experiments

A. Datasets

We evaluate the proposed HCN on three person Re-ID datasets: Market-1501 (Market) [27], DukeMTMC-reID (Duke) [28] and MSMT17 [29]. The Market dataset is captured

TABLE I: The time cost of HCN in pre-training, unsupervised training and test stages. M: Market, D: Duke and $M17$: MSMT17.

stage	M	D	M17
pre-training	45m	52m	110m
unsupervised training	14.4h	15.5h	27.8h
test	0.043s	0.059s	0.133s

by 6 cameras, in which the training set has 12936 images of 751 identities, the gallery set contains 19732 images of 750 identities, and the query set has 3368 images. The Duke dataset is captured by 8 cameras, and it possesses 36411 images of 1404 identities including 16522 training images, 17661 gallery images, and 2228 query images. The MSMT17 dataset is closer to the real scene and it is taken by 15 cameras. Specifically, the training set of MSMT17 contains 32621 images of 1041 identities. The test set of MSMT17 includes 93820 images of 3060 identities, where 11659 images are randomly selected from the test set as the query set, and the remaining 82161 images are regard as the gallery set. When employing the above datasets as the source dataset or the target dataset, only the training sets of them are used in the training stage, while the gallery and query sets of the target dataset are utilized to evaluate the performance of the model.

B. Implementation Details

1) Pre-training. We utilize the labeled source dataset to pre-train the CNN model using the framework in Fig. 2. Specifically, we resize the pedestrian images of the source dataset to 256×128 and then utilize the random cropping, flipping and erasing to conduct the data augmentation. We set the batch size to 128 which contains 16 identities and 8 pedestrian images for each identity. The margin of the triplet loss is set to 0.5 and the number of epochs is 70. The weight decay and momentum of the Adam optimizer [62] are set to 5×10^{-4} and 0.9 respectively, and the initial learning rate is 3×10^{-3} which is multiplied by 0.1 after 40 epochs.

2) Unsupervised Training. In the unsupervised training stage, the unlabeled target dataset is applied to train the multi-branch framework of HCN in Fig. 5, where the CNN model is initialized in the pre-training stage. Specifically, we resize the pedestrian images of the target dataset to 256×128 and conduct the same data augmentation with the pre-training stage. The margin of the triplet loss is set to 0.6 and the batch size is set to 128. The number of iterations is 20 and each iteration contains 60 epoches. After each iteration, the unsupervised clustering algorithm [60] is applied to cluster the CNN-based features of the target dataset, so that the pseudo labels of the target dataset can be updated in the optimization process. The initial learning rate is set to 8×10^{-4} and it is reduced to 0.1 times after 40 epochs. After training, we test the performance of HCN by the query and gallery sets of the target dataset, where the Cumulative Matching Characteristic (CMC) curve and the mean Average Precision (mAP) are employed as the evaluation criteria.

Our work is conducted on the Pytorch platform using two RTX 2080TI GPUs. The time cost of HCN in pre-training, unsupervised training and test stages is listed in Table I, where "pre-training" indicates the total pre-training time on the corresponding source dataset, "unsupervised training" is the total training time on the corresponding target dataset, and "test" represents the test time of each query pedestrian image.

TABLE II: The performance of the pre-trained CNN model, where A/B means the model is pre-trained on A and directly tested on B.

mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
M/M		M/D		M/M17	
81.0	92.7	17.8	32.9	2.8	8.6
D/D		D/M		D/M17	
72.4	84.0	26.6	54.6	6.1	18.3
M17/M17		M17/M		M17/D	
43.8	71.0	29.3	55.6	35.8	54.9

C. Ablation Study

1) The performance of the pre-trained CNN model. In Table II, we record the performance of the CNN model when training on the source domain and directly testing on the target domain. Obviously, when we train and test the CNN model on the same dataset in a supervised manner, the model obtains higher accuracies. But when the CNN model trained on the source dataset is directly tested on the target dataset, the performance degenerates greatly. For example, mAP and Rank-1 accuracy are 81.0% and 92.7% for M/M, but drop to 17.8% and 32.9% when directly tested on Duke. Similarly, the comparison of M/M and $M/M17$ shows that mAP and Rank-1 accuracy are dropped by 78.2% and 84.1%, respectively. There are similar performance degradations when we test the CNN model of D/D and $M17/M17$ on different target datasets. Therefore, the model trained on a single domain can not perform the task of person Re-ID well in different scenarios.

2) The effectiveness of the GCN model and DGConv. As we discuss above, we employ the GCN model to learn the correlation information among samples, and the GCN model can be implemented by the traditional graph convolution operation or the proposed DGConv operation. To verify the effectiveness of the GCN model and DGConv, we do experiments about "w/o GCN", "w/o DGConv" and "HCN", and list corresponding evaluation results in Table III. Here, "w/o GCN" indicates the GCN model is removed (i.e., only using the CNN model), "w/o DGConv" represents that we replace the DGConv operation in Eq. (7) with the traditional graph convolution operation in Eq. (4), and "HCN" utilizes the proposed DGConv operation.

Compared to "w/o GCN", "w/o DGConv" and "HCN" obtain the performance improvement, which fully proves the effectiveness of the GCN model. Specifically, for D-M, mAP of "w/o DGConv" is 68.0% which is 10.3% higher than "w/o GCN" and Rank-1 accuracy is 88.4% which is 9.4% higher than "w/o GCN". Similarly, mAP and Rank-1 accuracy of "HCN" are promoted by 12.5% and 11.2%, respectively. As for other cross-domain scenarios, the evaluation results with the GCN model are also superior to "w/o GCN". The

TABLE III: Ablation studies on the three datasets, where "w/o" means that we remove the component behind it.

Methods	D-M		M17-M		M-D		M17-D		M-M17		D-M17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
w/o GCN	57.7	79.0	55.0	77.8	50.2	70.8	57.1	73.9	12.0	30.9	12.3	31.5
w/o global	60.0	81.9	55.7	79.1	51.9	71.6	57.8	74.4	14.9	35.4	16.1	38.3
w/o multi	61.3	83.2	56.4	80.1	52.7	72.5	58.2	75.4	19.7	43.8	23.8	50.7
w/o FC	66.3	86.5	68.6	87.4	54.0	74.0	62.7	80.6	21.7	47.8	24.4	52.0
w/o DGConv	68.0	88.4	69.3	88.5	55.5	76.7	63.1	81.3	25.5	53.5	28.3	56.6
HCN	**70.2**	**90.2**	**70.5**	**90.7**	**57.3**	**78.9**	**65.7**	**83.5**	**27.0**	**55.1**	**29.9**	**58.7**

TABLE IV: The statistical analysis for the traditional graph convolution operation and the DGConv operation (D-M).

	w/o DGConv K=1	w/o DGConv K=2
mean value	4.6×10^{-5}	9.0×10^{-9}
p-value	6.9×10^{-20}	
	w/o DGConv K=2	w/o DGConv K=3
mean value	9.0×10^{-9}	1.3×10^{-10}
p-value	9.6×10^{-9}	
	w/o DGConv K=1	HCN K=3
mean value	4.6×10^{-5}	9.1×10^{-4}
p-value	1.0×10^{-25}	

significant improvement demonstrates that the correlation information provided by the GCN model is vital for cross-domain person Re-ID. When removing the GCN model, the features of the target samples are independent and have no correlations. When using the GCN model, each target sample feature could aggregate and absorb the correlation information from its connected samples, so that the discrimination of target sample features can be improved.

Furthermore, the comparison results between "w/o DGConv" and "HCN" show that the proposed DGConv improves the performance, which could verify the effectiveness of DGConv. For example, for M17-D, mAP of "HCN" is increased from 63.1% to 65.7% and Rank-1 accuracy is improved from 81.3% to 83.5% compared to "w/o DGConv". It is because DGConv considers the linkages between high similar samples and the linkages between high dissimilar samples, which can effectively avoid the negative effects caused by the trivial and redundant linkages in traditional graph convolution operation. The similar samples are closer and the dissimilar samples are farther after the DGConv operation.

To further verify the contribution of DGConv, we conduct some statistical analysis to show that more traditional graph convolution layers lead to similar nodes for different identities and the proposed DGConv can improve this situation in a certain extent. Specifically, we randomly select 128 target samples and then extract the GCN-based features of them based on "w/o DGConv" and "HCN", respectively, so as to obtain two groups of GCN-based features. Note that "w/o DGConv" indicates the traditional graph convolution is employed. For each group of GCN-based features, we calculate the similarities between negative sample pairs to obtain a similarity vector and then calculate the mean value of similarity vector. Afterwards, we use t-test to obtain the significant difference between the two similarity vectors. The results of the statistical analysis for D-M are shown in Table IV, where K is the number of graph convolution layers. Here, a small mean value indicates the similar features (nodes) and "p-value<0.05" means the obvious difference between the two similarity vectors. As for "w/o DGConv", we can see that the mean value decreases with the increase of the number of traditional graph convolution layers and the p-value is much less than 0.05, so more traditional graph convolution layers lead to the similar nodes and the decline of the sample discrimination. Meanwhile, the mean value of "HCN" (K=3) is higher than "w/o DGConv" (K=1), and the p-value between them is much less than 0.05. Hence, the proposed DGConv could overcome the drawback of similar nodes from different identities in a certain extent, and learn more discriminative features than the traditional graph convolution. Note that the traditional graph convolution and the proposed DGConv achieve the best performance with the number of graph convolution layers 1 and 3, respectively, and therefore the comparison between them is convictive.

3) The effectiveness of the multi-branch structure. In our work, we design the proposed HCN as a multi-branch structure to capture the structural information of pedestrians. For comparison, we only utilize the global features to optimize the deep model and denote it as "w/o multi". From Table III, we can see that compared to "w/o multi", mAP of "HCN" is increased 8.9% and Rank-1 accuracy is improved 7.0% for D-M. Meanwhile, the comparison results on other cross-domain scenarios indicate that learning the structural information is beneficial to improve the performance of the Re-ID model.

4) The effectiveness of the FC branch. In Fig. 5, we add a FC branch to learn the global CNN-based features for the target samples. In this subsection, we remove the FC branch to verify its effectiveness, which can be denoted as "w/o FC" in Table III. The comparison of "w/o FC" and "HCN" confirms that the FC branch contributes a lot to the performance. For example, with the FC branch, mAP is promoted from 66.3% to 70.2% and Rank-1 accuracy is increased from 86.5% to 90.2% for D-M. Similar performance improvements are observed in other cross-domain scenarios. Thus, the FC branch is an indispensable component of HCN.

5) The effectiveness of global correlation information. If only learning correlation information from the upper and lower parts of pedestrian images, it may cause false positives because of similar clothes. As shown in Table III, "w/o global" indicates only considering correlation information from upper and lower parts of pedestrian images. Compared to "w/o

TABLE V: Comparison with state-of-the-art unsupervised person Re-ID methods, where "–" denotes the corresponding test results cannot be obtained.

Methods	D-M		M17-M		M-D		M17-D		M-M17		D-M17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
LOMO [37]	8.0	27.2	–	–	4.8	12.3	–	–	–	–	–	–
BOW [27]	14.8	35.8	–	–	8.3	17.1	–	–	–	–	–	–
UMDL [38]	12.4	34.5	–	–	7.3	18.5	–	–	–	–	–	–
PTGAN [29]	–	38.6	–	–	–	27.4	–	–	–	10.2	–	11.8
PUL [43]	20.5	45.5	–	–	16.4	30.0	–	–	–	–	–	–
SPGAN [39]	22.8	51.5	–	–	22.3	41.1	–	–	–	–	–	–
CAMEL [22]	26.3	54.5	–	–	–	–	–	–	–	–	–	–
SPGAN+LMP [39]	26.9	58.1	–	–	26.4	46.9	–	–	–	–	–	–
HHL [40]	31.4	62.2	30.9	62.8	27.2	46.9	28.1	47.7	5.3	17.2	6.1	19.4
CASCL [63]	35.6	64.7	35.5	65.4	30.5	51.5	37.8	59.3	–	–		
ECN [41]	43.0	75.1	43.8	77.1	40.4	63.3	42.0	65.8	8.5	25.3	10.2	30.2
CR-GAN [61]	54.0	77.7	–	–	48.6	68.9	–	–	–	–	–	–
PAUL [64]	40.1	68.5	–	–	53.2	72.0	–	–	–	–	–	–
SSG [44]	58.3	80.0	59.6	82.3	53.4	73.0	56.0	74.7	13.2	31.6	13.3	32.2
ACT [46]	60.6	80.5	–	–	54.5	72.4	–	–	–	–	–	–
pMR-SADA [45]	59.8	83.0	–	–	55.8	74.5	–	–	–	–	–	–
GPP [42]	63.8	84.1	–	–	54.4	74.1	–	–	15.2	40.4	16.0	42.5
HCN	**70.2**	**90.2**	**70.5**	**90.7**	**57.3**	**78.9**	**65.7**	**83.5**	**27.0**	**55.1**	**29.9**	**58.7**

TABLE VI: The performance of HCN with different number of graph convolution layers.

	mAP	Rank-1	mAP	Rank-1
K	D-M		M17-M	
1	63.4	84.8	63.8	84.7
2	68.1	88.0	69.7	88.8
3	**70.2**	**90.2**	**70.5**	**90.7**
4	67.0	86.9	68.8	87.0
K	M-D		M17-D	
1	53.3	72.5	57.8	75.6
2	56.0	76.9	62.2	80.8
3	**57.3**	**78.9**	**65.7**	**83.5**
4	54.5	75.1	60.8	78.9
K	M-M17		D-M17	
1	19.3	44.2	23.8	51.1
2	24.9	52.0	25.5	55.7
3	**27.0**	**55.1**	**29.9**	**58.7**
4	25.2	52.8	26.4	55.8

global", "HCN" obviously obtains better performance. It is because HCN simultaneously learns correlation information from the global, upper and lower parts of pedestrian images, so that the global, upper and lower parts of pedestrian images can complement each other in the optimization process.

D. Comparison with State-of-the-arts

In this subsection, we compare the proposed HCN with the state-of-the-art methods. As shown in Table V, most state-of-the-art methods are evaluated in two cross-domain scenarios, i.e., M-D and D-M, while we evaluate HCN in six cross-domain scenarios including M-D, M-M17, M17-M, M17-D, D-M and D-M17. From Table V, we can see that the proposed HCN outperforms the state-of-the-art methods in all cross-domain scenarios.

Firstly, the methods based on the hand-crafted features, such as BOW [27], LOMO [37] and UMDL [38] evidently show low accuracy and poor generalization ability. It is because they are predefined and can not adapt different datasets. Secondly, some methods apply GAN for the style translation of pedestrian images to reduce the domain gap, including PTGAN [29], SPGAN [39], HHL [40], CR-GAN [61] and GPP [42]. Comparing with the hand-crafted methods, the GAN-based methods perform better. Finally, the clustering-based methods, such as PUL [43], CAMEL [22], SSG [44], ACT [46] and pMR-SADA [45] have shown advantages in the field of unsupervised person Re-ID. The proposed HCN is the clustering-based method and its performance outperforms the above-mentioned methods by a large margin. It is because the proposed HCN learns the appearance, structural and correlation information of pedestrian images in a unified framework. For example, as for D-M, mAP of "HCN" is 10.4% higher than pMR-SADA, and its Rank-1 accuracy is 7.2% higher than pMR-SADA.

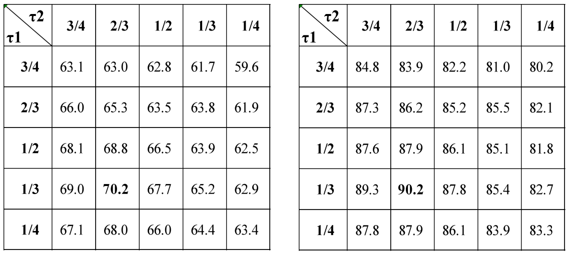
At present, only a few Re-ID methods, i.e., PTGAN [29], ECN [41], SSG [44] and GPP [42] conduct experiments on MSMT17. From Table V, we can see that the evaluation results of HCN still surpass these methods, which demonstrates the strong generalization ability of HCN. Specifically, for M-M17, Rank-1 accuracy of "HCN" is 23.5% higher than SSG and 14.7% higher than GPP. For D-M17, Rank-1 accuracy of "HCN" is promoted by 26.5% comparing with SSG and 16.2% comparing with GPP.

E. Parameter Analysis

1) The influences of the number of graph convolution layers. We set the number of graph convolution layers K from 1 to 4, and the evaluation results of HCN are presented in Table VI. From this table, we can see that when the number of graph convolution layers increases the performance of HCN is improved, and HCN obtains the best results when K reaches to 3. On the contrary, with the traditional graph convolution, the performance of the model decreases as the number of layers, as shown in the red curves of Fig. 4. The comparison

TABLE VII: The performance of HCN with different dimensions of GCN-based features.

	mAP	Rank-1	mAP	Rank-1
dimension	D-M		M17-M	
256	67.7	87.6	68.5	88.1
512	**70.2**	**90.2**	**70.5**	**90.7**
1024	68.4	88.5	67.7	88.9
dimension	M-D		M17-D	
256	54.5	75.7	60.4	79.8
512	**57.3**	**78.9**	**65.7**	**83.5**
1024	56.1	76.6	63.2	81.7
dimension	M-M17		D-M17	
256	23.9	49.0	25.2	54.9
512	**27.0**	**55.1**	**29.9**	**58.7**
1024	25.0	52.1	27.2	55.6



τ1 \ τ2	3/4	2/3	1/2	1/3	1/4
3/4	63.1	63.0	62.8	61.7	59.6
2/3	66.0	65.3	63.5	63.8	61.9
1/2	68.1	68.8	66.5	63.9	62.5
1/3	69.0	**70.2**	67.7	65.2	62.9
1/4	67.1	68.0	66.0	64.4	63.4

(a)

τ1 \ τ2	3/4	2/3	1/2	1/3	1/4
3/4	84.8	83.9	82.2	81.0	80.2
2/3	87.3	86.2	85.2	85.5	82.1
1/2	87.6	87.9	86.1	85.1	81.8
1/3	89.3	**90.2**	87.8	85.4	82.7
1/4	87.8	87.9	86.1	83.9	83.3

(b)

Fig. 6: (a) mAP and (b) Rank-1 accuracy of HCN with different combinations of τ_1 and τ_2 for D-M.

results prove that the proposed DGConv could relieve the over-smoothing in a certain extent.

2) The influences of GCN-based feature dimension. We list the results of different dimensions of GCN-based features in Table VII. From the table, we can see that when the dimension is set to 512, we obtain the best results in all cross-domain scenarios.

3) The influences of τ_1 and τ_2. The thresholds τ_1 and τ_2 in Eq. (5) and Eq. (6) are utilized to select the high confidence elements from the adjacency matrix for the similar and dissimilar images. We conduct experiments with different combinations of τ_1 and τ_2 for D-M and the results are listed in Fig. 6, where we can see that the results achieve the best when τ_1=$1/3$ and τ_2=$2/3$. Note that our experiments have shown that the conclusions can be generalized to other cross-domain scenarios as well.

4) The influences of the batch size. The batch size of the triplet loss is $P \times Q$, where P is the number of pedestrian identities and Q is the number of pedestrians for each identity. The results are listed in Table VIII where we can see that the performance varies with different batch sizes. It is because the batch size determines the size of the graph, and further determines the number of linkages in the graph. When the batch size is less than 128, the performance of HCN is relatively poor, because the linkages between samples are not enough to learn discriminative features. When the batch size is greater than 128, HCN shows slight performance degradation. Hence, proper batch size is key for the performance of our method. When the batch size is set to 128 (P=16, Q=8), mAP and Rank-1 accuracy of HCN achieve the best performance in all cross-domain scenarios.

TABLE VIII: The performance of HCN with different batch sizes.

	mAP	Rank-1	mAP	Rank-1
$P \times Q$	D-M		M17-M	
16×4	27.0	53.9	30.1	54.3
8×8	29.8	55.1	31.2	56.0
32×4	67.8	88.3	68.3	89.0
16×8	**70.2**	**90.2**	**70.5**	**90.7**
32×8	64.0	84.0	64.8	85.1
16×16	66.6	85.2	65.2	86.6
$P \times Q$	M-D		M17-D	
16×4	18.1	33.7	20.7	36.1
8×8	19.9	34.4	22.7	37.6
32×4	55.8	76.7	62.2	80.9
16×8	**57.3**	**78.9**	**65.7**	**83.5**
32×8	53.7	74.0	58.0	78.1
16×16	54.6	75.5	60.6	80.5
$P \times Q$	M-M17		D-M17	
16×4	8.9	14.7	9.4	15.5
8×8	9.1	15.0	10.7	17.0
32×4	24.9	52.1	27.2	56.4
16×8	**27.0**	**55.1**	**29.9**	**58.7**
32×8	24.2	51.8	25.1	54.0
16×16	25.3	53.7	26.7	55.4

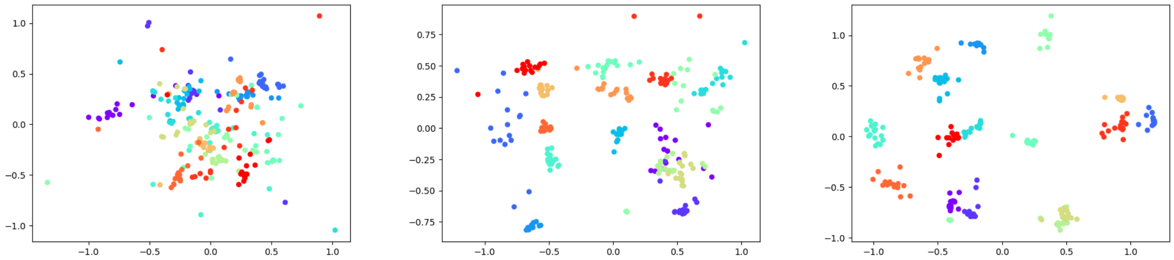
TABLE IX: The performance of HCN (for D-M) with different loss weights.

w_{T_g}	$w_{T_{fc}}$	w_{T_u}	w_{T_l}	mAP	Rank-1
0.6	0.6	1	1	66.8	87.1
0.8	0.8	1	1	67.5	88.3
1	1	0.6	0.6	67.8	87.5
1	1	0.8	0.8	68.9	89.3
1	1	1	1	**70.2**	**90.2**
1.2	1.2	1	1	68.6	89.1
1.4	1.4	1	1	67.1	88.5
1	1	1.2	1.2	68.2	88.0
1	1	1.4	1.4	66.4	86.5

5) The influences of the loss weight. We do experiments for HCN with different weights of the four losses in Eq. (9). As an example, we present the experimental results for D-M in Table IX, where HCN achieves the best performance with the same loss weight and shows a slight performance degradation with different loss weights. Note that our experiments have shown that the conclusions can be generalized to other scenarios as well.

F. Visualization Results

1) Visualization of clustering results. In Fig. 7, we show the clustering results of different features including CNN-based features, GCN-based features of traditional graph convolution, and GCN-based features of DGConv. As shown in Fig. 7(a), when clustering with the CNN-based features, there are obvious crossings among different identities, and the distribution of the CNN-based features is relatively scattered. Fig. 7(b) and Fig. 7(c) are the clustering results of GCN-based features of traditional graph convolution and DGConv, in which the crossings between different identities are greatly reduced. It is because GCN-based features contain the correlations among pedestrian images. The clustering result based on DGConv is



(a) (b) (c)

Fig. 7: Visualization of clustering results for different features: (a) CNN-based features, (b) GCN-based features of traditional graph convolution, and (c) GCN-based features of DGConv.

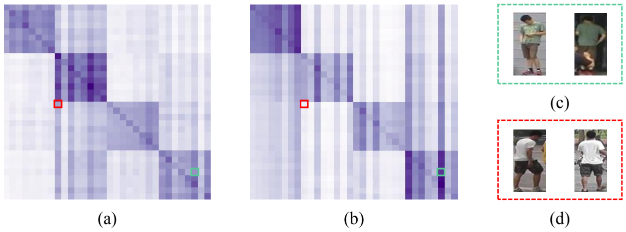


Fig. 8: Visualization of the adjacency matrices for (a) traditional graph convolution and (b) DGConv. (c) Pedestrian image pair with the same identity corresponding to the green box; (d) Pedestrian image pair with different identities corresponding to the red box. The deeper color indicates the larger similarity between pedestrian images.

better than that of the traditional graph convolution, because DGConv can explicitly learn the correlation information from the similar and dissimilar pedestrian images. Thus, the proposed DGConv can help the clustering algorithm to obtain more accurate identity labels for unlabeled target dataset.

2) Visualization of the adjacency matrices. We visualize the adjacency matrices of the traditional graph convolution and DGConv in Fig. 8. The green box indicates the similarity between pedestrian images with the same identity, and the corresponding image pair is shown in Fig. 8(c). From the figure, we can see that the proposed DGConv gives a large similarity value than the traditional graph convolution, which is closer to the ground truth. As for the red box, it represents the similarity between the pedestrian images with different identities, and the corresponding image pair is presented in Fig. 8(d). From the figure, we can see that the proposed DGConv gives more reasonable similarity value than the traditional graph convolution. It is because DGConv selects the high confidence elements of the adjacency matrix for similar and dissimilar pedestrian images to optimize the model.

V. Conclusion

In this paper, we have proposed HCN for cross-domain person Re-ID, which applies CNN and GCN to learn the appearance and correlation information for pedestrian images. As for the graph, we exploit the similarity between pedestrian images to establish the linkage. Then we propose DGConv to explicitly learn the correlation information from the similar and dissimilar pedestrian images, which could avoid to propagate the trivial information in the fully connected graph. Moreover, we design HCN as a multi-branch structure to discover the structural information of pedestrian images. The evaluation results on Market, Duke and MSMT17 indicate that HCN is superior to the start-of-the-art methods and it possesses the remarkable adaptability for the cross-domain scenarios.

References

[1] J. Wang, Z. Wang, C. Gao, N. Sang and R. Huang, "Deeplist: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 513-524, 2017.

[2] L. Wu, R. Hong, Y. Wang and M. Wang, "Cross-entropy adversarial view adaptation for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2081-2092, 2020.

[3] H. X. Yu and W. S. Zheng, "Weakly supervised discriminative feature learning with state information for person identification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5527-5537, 2020.

[4] L. Qi, L. Wang, J. Huo, Y. Shi and Y. Gao, "Progressive cross-camera soft-label learning for semi-supervised person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2815-2829, 2020.

[5] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan and J. Feng, "Towards pose invariant face recognition in the wild," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2207-2216, 2018.

[6] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," *ACM International Conference on Multimedia*, pp. 792-800, 2018.

[7] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, H. Liu, S. Yan and J. Feng, "Multi-prototype networks for unconstrained set-based face recognition," *arXiv preprint arXiv:1902.04755*, 2019.

[8] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270-286, 2014.

[9] L. Zheng, Y. Yang and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[10] F. Zhao, J. Li, J. Zhao and J. Feng, "Weakly supervised phrase localization with multi-scale anchored transformer network," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5696-5705, 2018.

[11] K. Kansal, A. V. Subramanyam, Z. Wang and S. Satoh, "SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3422-3432, 2020.

[12] Z. Wang, J. Zhao, C. Lu, F. Yang, H. Huang, L. Li and Y. Guo, "Learning to detect head movement in unconstrained remote gaze estimation in the wild," *IEEE Winter Conference on Applications of Computer Vision*, pp. 3443-3452, 2020.

[13] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3722-3731, 2017.

[14] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulo, N. Sebe and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9471-9480, 2019.

[15] L. Luo, L. Chen, S. Hu, Y. Lu and X. Wang, "Discriminative and geometry-aware unsupervised domain adaptation," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3914-3927, 2020.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversial nets," *Advances in Neural Information Processing Systems*, pp. 2672-2680, 2014.

[17] J. Zhao, L. Xiong, K. Jayashree, J. Li, F. Zhao, Z. Wang, S. Pranata, S. Shen, S. Yan and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," *Neural Information Processing Systems*, vol. 30, pp. 66-76, 2017.

[18] J. Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE International Conference on Computer Vision*, pp. 2223-2232, 2017.

[19] F. Zhao, J. Zhao, S. Yan and J. Feng, "Dynamic conditional networks for few-shot learning," *European Conference on Computer Vision*, pp. 19-35, 2018.

[20] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim and J. Choo, "StarGAN: unified generative adversarial networks for multi-domain image-to-image translation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8789-8797, 2018.

[21] H. Zhang, V. Sindagi and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3943-3956, 2020.

[22] H. X. Yu, A. Wu and W. S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," *IEEE International Conference on Computer Vision*, pp. 994-1002, 2017.

[23] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang and N. Zheng, "Semi-supervised person re-identification using multi-view clustering," *Pattern Recognition*, vol. 88, pp. 285-297, 2019.

[24] Y. Lin, X. Dong, L. Zheng, Y. Yan and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8738-8745, 2019.

[25] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu and X. Yang, "Learning context graph for person search," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2158-2167, 2019.

[26] Z. M. Chen, X. S. Wang, P. Wang and Y. Guo, "Multi-label image recognition with graph convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5177-5186, 2019.

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

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *European Conference on Computer Vision*, pp. 17-35, 2016.

[29] L. Wei, S. Zhang, W. Gao and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 79-88, 2018.

[30] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[31] B. Sun, J. Feng and K. Saenko, "Return of frustratingly easy domain adaptation," *AAAI Conference on Artificial Intelligence*, pp. 2058-2065, 2016.

[32] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu and J. Huang, "Progressive feature alignment for unsupervised domain adaptation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 627-636, 2019.

[33] H. K. Hsu, C. H. Yao, Y. H. Tsai, W. C. Hung, H. Y. Tseng, M. Singh, and M. H. Yang, "Progressive domain adaptation for object detection," *IEEE Winter Conference on Applications of Computer Vision*, pp. 749-757, 2020.

[34] R. Volpi, P. Morerio, S. Savarese and V. Murino, "Adversarial feature augmentation for unsupervised domain adaptation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5495-5504, 2018.

[35] F. Lv, K. Zhu, G. Yang and L. Duan, "TarGAN: Generating target data with class labels for unsupervised domain adaptation," *Knowledge-Based Systems*, vol. 172, pp. 123-129, 2019.

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

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

[38] P. Peng, T. xiang, Y. Wang, M. Pontil, S. Gong, T. Huang and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1306-1315, 2016.

[39] W. Deng, L. Zheng, Q. Ye, G. Yang and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 994-1003, 2018.

[40] Z. Zhong, L. Zheng, S. Li and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," *European Conference on Computer Vision*, pp. 176-192, 2018.

[41] Z. Zhong, L. Zheng, Z. Luo, S. Li and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 598-607, 2019.

[42] Z. Zhong, L. Zheng, Z. Luo, S. Li and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DoI: 10.1109/TPAMI.2020.2976933, 2020.

[43] H. Fan, L. Zheng, C. Yan and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 14, no. 4, pp. 1-18, 2018.

[44] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," *IEEE International Conference on Computer Vision*, pp. 6112-6121, 2019.

[45] G. Wang, J. Lai, W. Liang and G. Wang, "Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10565-10574, 2020.

[46] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji and S. Li, "Asymmetric co-teaching for unsupervised cross domain person re-identification," *AAAI Conference on Artificial Intelligence*, pp. 12597-12604, 2020.

[47] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[48] J. Johnson, A. Gupta and L. Fei-Fei, "Image generation from scene graphs," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1219-1228, 2018.

[49] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong and P. H.S. Torr, "Dual graph convolutional network for semantic segmentation," *British Machine Vision Conference*, 2019.

[50] A. Pareja, G. Domeniconi, J. Chen, T. Ma, T. Suzumura, H. Kanezashi, T. Kaler, T. B. Schardl and C. E. Leiserson, "Evolvegcn: Evolving graph convolutional networks for dynamic graphs," *AAAI Conference on Artificial Intelligence*, pp. 5363-5370, 2020.

[51] J. Bruna, W. Zaremba, A. Szlam and Y. LeCun, "Spectral networks and locally connected networks on graphs," *arXiv preprint arXiv:1312.6203*, 2013.

[52] M. Henaff, J. Bruna and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[53] M. Defferrard, X. Bresson and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in Neural Information Processing Systems*, pp. 3844-3852, 2016.

[54] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83-98, 2013.

[55] W. Hamilton, Z. Ying and J. Leskovec, "Inductive representation learning on large graphs," *Advances in Neural Information Processing Systems*, pp. 1024-1034, 2017.

[56] C. Zhuang and Q. Ma, "Dual graph convolutional networks for graph-based semi-supervised classification," *The World Wide Web Conference*, pp. 499-508, 2018.

[57] L. Shi, Y. Zhang, J. Cheng and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12026-12035, 2019.

[58] Z. Qiu, K. Qiu, J. Fu and D. fu, “DGCN: Dynamic graph convolutional network for efficient multi-person pose estimation,” *AAAI Conference on Artificial Intelligence*, pp. 11924-11931, 2020.

[59] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[60] M. Ester, H. P. Kriegel, J. Sander and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” *Knowledge Discovery and Data Mining*, vol. 96, no. 34, pp. 226-231, 1996.

[61] Y. Chen, X. Zhu and S. Gong, “Instance-guided context rendering for cross-domain person re-identification,” *IEEE International Conference on Computer Vision*, pp. 232-242, 2019.

[62] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[63] A. Wu, W. S. Zheng and J. H. Lai, “Unsupervised person re-identification by camera-aware similarity consistency learning,” *IEEE International Conference on Computer Vision*, pp. 6922-6931, 2019.

[64] Q. Yang, H. X. Yu, A. Wu and W. S. Zheng, “Patch-based discriminative feature learning for unsupervised person re-identification,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3633-3642, 2019.

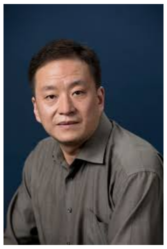
Zhong Zhang (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He is a Professor at Tianjin Normal University, Tianjin, China. He has published about 110 papers in international journals and conferences such as IEEE Transactions on Fuzzy Systems, Pattern Recognition, IEEE Transactions on Circuits Systems Video Technology, IEEE Transactions on Information Forensics and Security, Signal Processing (Elsevier), CVPR, ICPR and ICIP. His research interests include computer vision, pattern recognition, and deep learning.

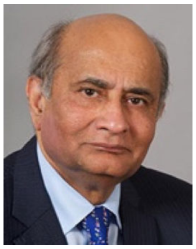
Yanan Wang is a master student at Tianjin Normal University, Tianjin, China. Her research interests include the cross-domain person re-identification and deep learning.

Shuang Liu (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

She is an Associate Professor at Tianjin Normal University, Tianjin, China. She has published over 50 papers in major international journals and conferences. Her research interests include computer vision, pattern recognition, and deep learning.



Baihua Xiao received the B.S. degree in Electronic Engineering from Northwestern Polytechnical University, Xian, China and the Ph.D. degree in Pattern Recognition and Artificial Intelligence from the Institute of Automation, Chinese Academy of Science, Beijing, China, in 1995 and 2000, respectively. From 2005, he has been a Professor at the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Science, Beijing, China. His research interests include pattern recognition, computer vision, image processing and machine learning.



Tariq S. Durrani is Research Professor at University of Strathclyde, Glasgow Scotland. His research covers AI, Signal Processing and Technology Management. He has authored 350 publications; supervised 45 PhDs. He is a Fellow of the: IEEE, UK Royal Academy of Engineering, Royal Society of Edinburgh, IET, and the Third World Academy of Sciences. In 2018 he was elected Foreign Member of the US National Academy of Engineering.