IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 27, 2020, accepted July 13, 2020, date of publication July 16, 2020, date of current version July 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009653

RRGCCAN: Re-Ranking via Graph Convolution Channel Attention Network for Person Re-Identification

XIAOQIANG CHEN[1,2], LING ZHENG[1], CHONG ZHAO[1,2], QICONG WANG[1,2], AND MAOZHEN LI[3], (Member, IEEE)

[1] School of Informatics, Xiamen University, Xiamen 361001, China
[2] Shenzhen Research Institute, Xiamen University, Shenzhen 518000, China
[3] Department of Electronics and Computer Engineering, Brunel University, London UB83PH, U.K.

Corresponding authors: Chong Zhao (zhc@xmu.edu.cn) and Qicong Wang (qcwang@xmu.edu.cn)

This work was supported by the Shenzhen Science and Technology Project under Grant JCYJ20180306173210774.

ABSTRACT The classical person re-identification methods are mostly focused on employing discriminative features amongst which the distance is measured on Euclidean space, while the effort of re-ranking is constrained as the lack of the utilization of quality context representation in embedding set. In this paper, we incorporate graph models on feature subsets resorting to the initial ranking by adopting the integration of the attention mechanism into graph convolution network. On the one hand, the context information regarding embedding pairs is considered to compute feature group similarity through the aggregation operation by using graph convolution networks. On the other hand, we adopt a channel attention mechanism to enhance the contribution of relevant feature channels, further strengthening the ability of similarity pulling and dissimilarity pushing of the overall network. Experimental study shows that the proposed network structure is superior to the state-of-the-art deep neural networks on three very challenging datasets that are popular in examining person re-identification techniques.

INDEX TERMS Person re-identification, graph convolution network, attention mechanism, context information.

I. INTRODUCTION

Given a set of probe images, the person re-identification (Re-ID) aims at searching for the identical person across multiple different cameras. It has attracted a great attention since it is an important task in analyzing video for security systems. Although deep neural networks have made remarkable successes in feature representation and thus the task of person Re-ID, large appearance variations in occlusions, viewpoints, illumination, and pose diversity remain challenging as a result of the problem of large inter-class similarity and small intra-class similarity [5]. This problem has been alleviated by learning additional feature discriminative features through the convolutional neural network that calculates feature similarities by either Euclidean distance or cosine distance [3]. In order to further improve the performance of the person Re-ID, Re-ranking is an important alternative which may consider the significance of a single feature and/or complex feature relationship, especially in the case of obvious appearance variations. The active re-ranking methods often use the original ranking with additional information to build distinguishable context information where the distance measurements are calculated for direct inference [18]. However, the predefined distance metrics may produce massive manual intervention. Moreover, the guidance of training data is neglected out of context information, which may result in insufficient reasoning ability [21]. Therefore, a re-ranking method incorporating similarity information obtained from trainable feature context is desired for weighting paired features.

Encouraged by the graph convolutional networks that have a powerful reasoning ability, a novel framework of graph convolution channel attention network is proposed with a learned re-ranking model integrated (RRGCCAN) for the person Re-ID task. Within our proposed neural network framework, the process of searching a probe image from gallery can be regarded as a task of predicting similarity. The flowchart drawn in Figure 1 presents the main principle of

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

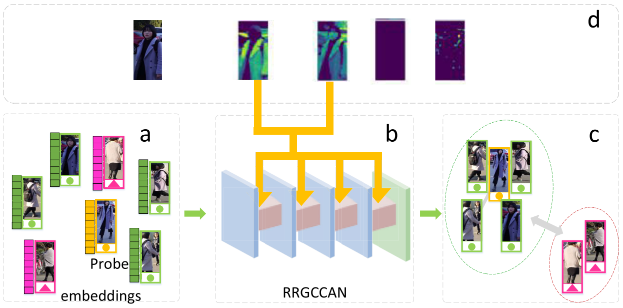


FIGURE 1. **The key idea of our method. (a) represents the original features distribution, (b) is our model (RRGCCAN), (c) is the features distribution after RRGCCAN, (d) is pedestrian image and visualization of some channels. RRGCCAN impose layer-by-layer aggregation to reasoning the similarity distribution of the pictures around the probe based on the group level in the embeding subset for re-ranking. It also utilize some salient channels to enhance the inference ability.**

the proposed neural network framework. In order to obtain more precise similarity measurements between pedestrian features, the adjacency subgraph and group prediction graph are introduced to mine more possible relationships for calculating the level of similarity. Specifically, a subgraph with the predefined edges is generated by taking each image embedding as the center and using the K nearest neighbor algorithm. The local context of the center contains all the node information in the subgraph and its size depends on the nearest neighbor rule. If any two subgraphs have neighborhood relationship between their centers, they are adjacency subgraphs. The similarity between them can be inferred step by step through GCN. For example, in Figure 1 (a), due to similarity, a subgraph centered on yellow marked image embedding is an adjacency subgraph of the one centered on the green marked image embedding, but not with the red mark. In Figure 1(c), each color ellipse represents a group prediction map of different pedestrians. Group prediction graph is the result of GCN inference for all input adjacency subgraphs. Therefore, the model does not use single image embedding to measure the similarity between pedestrians, but uses the group features obtained by GCN aggregation of adjacency subgraphs to calculate it. In order to obtain the feature similarities between all central nodes, an inference is performed on these subgraphs via the graph convolutional network. As a result, the subgraphs are learned with edges re-weighted by new feature similarities. The merger of all the subgraphs is organized into a new predictive graph, which can be used to obtain relations between a probe and gallery images. This resultant predictive graph is then a re-ranking output of the original graph that is ranked by graph convolutional networks. Furthermore, we explore the different contribution of each feature channel of nodes and locate some channels may help to reduce intra-class distance and expand inter-class distance. To this end, a channel attention subnet is designed for graph convolution framework, where the importance of each channel is adjusted adaptively during training. The subnet model can strengthen the relevant channels (for example channels related to contour) while suppressing irrelevant channels (for example channels related to some appearance changes). This character of the subnet may further improve its reasoning ability.

Our main contributions are summarized as follows. (1) We leverage the graph model and graph convolution network to explore complex features-to-features relationship, which used to optimize the ranking results for person Re-ID. The constructed subgraphs model the local context information of samples in the embedding set which consists of all probes and gallery, and the inferred group similarity graph through graph convolutional network describes their intra/inter-class relations. (2) A channel attention subnet is incorporated into the graph convolution framework, which can adjust channel responses adaptively. This not only enhances the importance of relevant channels but also suppresses the impact of irrelevant channels.

II. RELATED WORK

A. RE-RANKING FOR PERSON RE-ID

To improve the performance of Person Re-ID, the prevalent methods in most literatures are used either via a modification of the feature extraction algorithm [3] or an alternative of feature similarity metrics [6]. There are very few re-ranking schemes being optimized in person Re-ID, most of which are unsupervised. A ranking method with k-reciprocal encoding which imposes Jaccard distance on calculating similarity is proposed in [18] where the image matching is effective when either of two images should be in the ranged ranking list of the other while, in [19], the expanded cross neighborhood distance is calculated for measuring the distance between paired images via an integrating across between the probe neighbors and images in gallery. Some re-ranking methods exploit the potential of training data, for example, [22] exploits training data to optimize each initial ranking and obtains the information of Discriminant Context to improve the ranking. Considering that the guidance of labels has a positive impact on similarity reasoning and the reuse of training data has no impact on the test stage, that is, the computational cost is acceptable during training. We make use of training data to enhance the reasoning ability of the model to context information. Finally, unlike its idea of image pair and node classification [21], our framework can make use of the context information composed of initial ranking and it is equivalent to a reasoning network of edge prediction.

B. LINK PREDICTION WITH GCN AND ATTENTION MECHANISM

As a universal structure used to describe the complex relationship between things, graph models have been successfully applied in machine learning problems, for example, molecular structure prediction, software code plagiarism detection, etc [7]. A series of advanced convolutional neural networks are designed to deal with graph structure data. By the definition of GCN, it can be fell into either the class of spectral

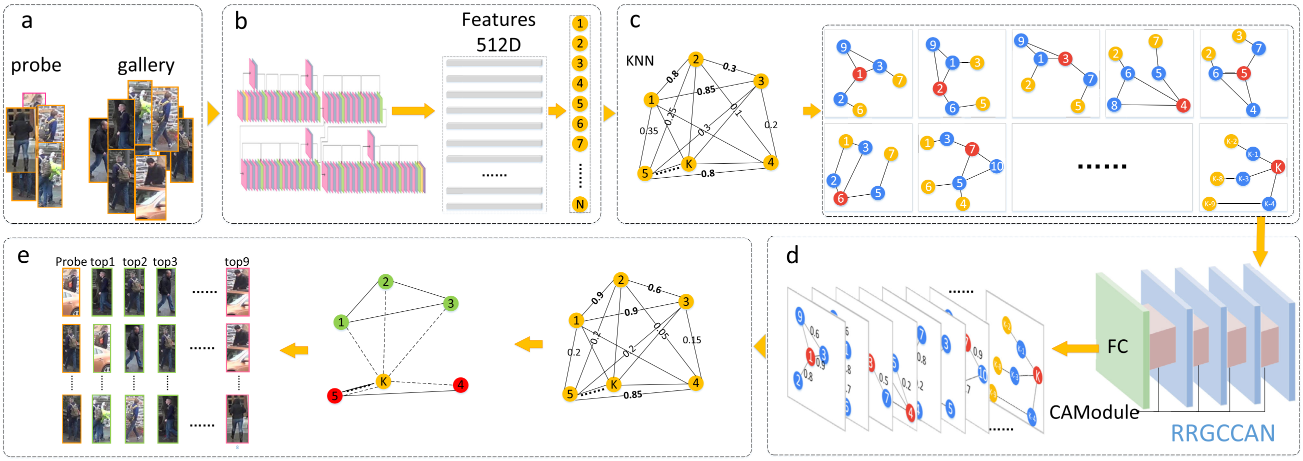


FIGURE 2. Illustration of the proposed approach. (a) and (b) represent the feature extraction process of pedestrian pictures. (c) Using the nearest neighbor rule to create subgraphs based on (b). (c) and (d) stand for obtaining group similarity graphs by feeding the subgraphs into our model. (e) shows the process from local graphs to the final graph to realize the evaluation. *K* represents the *k* nearest neighbors of the probe selected from *N* nodes.

methods or that of spatial methods. Graph neural network plays a significant role in Person Re-ID [21], [23]. In [23], to obtain part feature representation based on hierarchical context, the author exploits the inherent relationships of parts through the convolutional network of parts-based hierarchical graphs. In addition, for link prediction task, it is towards to predicting if two member nodes in a complex network should be connected or not. The approach like in [17] was developed in order to evaluate the likelihood of links. Inspired by the above works, a predictive model for the similarity inference on pedestrian graph data is being integrated into the proposed framework. Attention is a distribution mechanism that targets giving more computing resources to the region of the richest information. SENet proposed the Squeeze-and-Excitation subnetwork which pays attention to the lightweight module and the relations between channels [8]. A residual attention network was recommended to obtain mixed attention that are a result of multiple different modules [9]. The Convolutional Block Attention Module (CBAM) uses the channel and spatial modules to improve network architecture, and generates more robust feature representation [24]. The direct use of the idea of channel adaptive response from SENet to GCN does not apply to GCN performance. To this end, a modification of the channel attention module based on graph convolution is then conducted in order to estimate similarities adaptively.

III. MODEL DESIGN

We design RRGCCAN for person Re-ID to learn the similarity relations between probe images and gallery. The whole Re-ID procedure and method proposed will be outlined (Section 3.1). The method includes two parts. (1) A re-ranking via graph convolution network (RRGCN) for person Re-ID will be presented, with the purpose of obtaining the similarity relationship between the probe and its 1-hop neighbors (Section 3.2). (2) The channel attention module (CAModule) will be introduced to recalibrate the importance of channels (Section 3.3). Finally, we will introduce the loss function and discuss the model (Section 3.4).

A. RRGCCAN

The Re-ID system based on our proposed method is shown in Figure 2. Due to its powerful ability of extracting data features and alleviating gradient disappearance effectively, ResNet [1] serves as the backbone of most Re-ID models. Compared with ResNet, the parameters of DenseNet [2] are less and it can extract more discriminative features on our current used devices.

Firstly, we exploit ResNet-18, ResNet-34, ResNet-50, DenseNet-121, DenseNet-161 and DenseNet-201 respectively, to extract the pedestrian features, rather than a method only effective under a particular model. To compare our method with the state-of-the-art methods, we also use an advanced deep neural network called DG-Net [5] to extract features. The output feature dimension takes 512, that is, each pedestrian image is represented as a 512-dimensional feature (Figure 2.b).

Secondly, the construction of subgraphs occurs in the component of Figure 2.c. The general neighbor rule is used to construct the subgraphs. In this paper, the dot product between vectors is adopted as the standard for initializing edges:

$$S_{i,j} = F_i \cdot F_j \quad i, j \in \{1, \ldots, n\} \tag{1}$$

where symbol dot represents dot product, $S_{i,j}$ is the general similarity score of node i and node j, F_i is the features of the i-th node while F_j is the features of the j-th node, and n is the number of nodes in a subgraph. The larger $S_{i,j}$ is, the more similar nodes i and node j are. This means the possibility of node i and node j being the same person is higher.

In a subgraph, it is worth noting that nodes need to subtract the feature of the central node to form the final node feature, which can intuitively express the difference between the central node feature and other nodes. Such difference is calculated as follows:

$$F_{i,j} = F_{i,j} - F_{z,j} \quad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, c\} \tag{2}$$

where $F_{i,j}$ represents the j-th channel of the i-th node, $F_{z,j}$ represents the j-th channel of the central node and c is the number of channels. This equation can highlight the differences between the central node and neighbor nodes, which may allow neighbor nodes to obtain relevant information from the central node.

Moreover, the size n of the subgraph is determined by the predefined number of 1-hop neighbors (directly-linked neighbors) and 2-hop neighbors (neighbors of its individual neighbors):

$$n = h_1 * h_2 + 1 \tag{3}$$

where h_1 denotes 1-hop neighbors of the central node and h_2 denotes 2-hop neighbors of the central node. We set a unified size for subgraphs to enable data-parallel processing. The number of initialized edges for each node is set to 5. In other words, according to similarity scores, we choose the top five 1-hop neighbors to initialize edges.

The proposed framework is trained by CNN features of the pedestrian training set, so that it facilitates to minimize the distances between positive pairs while maximizing the distances between negative pairs. During the model (Figure 2.d) testing phase, we use features of all possible images that are from both gallery and probes to infer the global similarity scores (Figure 2.e), *i.e.*, the link prediction of each probe node for each pedestrian node in the gallery.

B. RRGCN

The basic framework of the proposed RRGCCAN is shown in Figure 3. The traditional fully connected layer is used to predict the link likelihood between the central node and its 1-hop neighbors and then determine whether two nodes in the subgraph should be linked or not. Finally, all the output subgraphs are combined together to obtain the similarity scores of all nodes.

Our model employs spatial graph convolution as the backbone for re-ranking. Output features V^i of the i-th layer are obtained by the following the convolutional aggregation:

$$V^i = \sigma(M^{-1}\widetilde{A}V^{i-1}W) \tag{4}$$

where $V^i \in R^{n \times c^i}$ and $i \in \{0, 1, 2, 3, 4\}$, V^0 is the original node, which represents feature information of all nodes in the subgraph, c^0 denotes the initial feature channel number of each node in the network, σ is a ReLU non-linear activation function, $W \in R^{c^{i-1} \times c^i}$ is a trainable parameter in the network, c^i represents the number of channels in the i-th layer, $\widetilde{A}$ is any adjacent matrix of a graph with self-connection, and $\widetilde{A} = A + I$, and $M = \sum_i \widetilde{A}_{ij}$ is the diagonal matrix of a graph.

Further explanation regarding equation (4) is given as follows, the matrix multiplication between V^{i-1} and W is the result of reconstructing context information of V^{i-1} while $\widetilde{A}V^{i-1}W$ calculates similarities between V^{i-1} and its directly-linked neighbor nodes. M^{-1} is a regular matrix which is used for regularizing the matrix $\widetilde{A}V^{i-1}W$ in order to maintain a fixed scale of V^{i-1}.

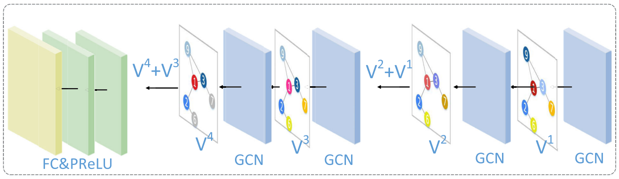


FIGURE 3. Illustration of re-ranking with graph convolution network (RRGCN). From right to left, the network consists of two steps: (1) Graph convolution networks based on residual idea; (2) fully connected layers and non-linear activation function.

Four graph convolutional layers are employed in our backbone neural network, each layer takes a value from the list [512,512,256,256] in turn as its dimensional size. The following equation is applicable where the idea of ResNet is used:

$$V^i = V^{i-1} + V^i \tag{5}$$

where $i \in \{2, 4\}$ since the ResNet is only applied at 2nd and 4th layers, V^i indicates the output of the i-th layer, V^{i-1} signifies the input of the i-th layer, and V^i is the input of the next layer of the i-th.

Then the traditional fully connected layer (FC) outputs the link probabilities of between the central node and its 1-hop neighbor nodes. The neurons number of FC layer is 256 and 2 respectively. Each 1-hop node corresponding to outputs indicate the probability $output \in R^{h_1 \times 2}$ of connecting to the central node as follows:

$$output = W_2\sigma_1(W_1X) \tag{6}$$

where W_1 and W_2 represent the trainable weights of the first and second fully connected layer respectively, σ_1 is the PReLU activation function. $X \in R^{h_1 \times c^4}$ is a partial output after the graph convolutional layers.

C. CHANNEL ATTENTION MODULE (CAModule)

The channel attention mechanism is suitable to each graph convolutional layer in Section 3.2. The feature map of each layer are normalized, and then the global information $Y^i \in R^{1 \times c^i}$ of feature map at the i-th layer is obtained by the global average pooling (GAP):

$$Y^i_z = \frac{1}{N}\sum_{j=1}^{n} V^i_{z,j} \quad z \in \{1, \ldots, c^i\} \tag{7}$$

where Y^i_z represents information of the z-th channel at the i-th layer and $V^i_{z,j}$ is the z-th channel of the j-th node feature map in the i-th layer.

Interacted between two fully connected (FC) layers, the network outputs the adaptive response information $S^i \in R^{1 \times c^i}$ of the i-th layer at channel level, which is calculated as follows:

$$S^i = \sigma_2(W_2\sigma_1(W_1Y^i)) \tag{8}$$

where $W1$ and $W2$ represents the trainable parameters, σ_1 represents the non-linear activation function RReLU, and σ_2 represents the sigmoid activation function which limits the output between 0 and 1.

Finally the output $\widetilde{V}^i \in R^{n\times c^i}$ of each graph convolutional layer is computed by integrating S^i and its corresponding channel information as follows:

$$\widetilde{V}^i = S^i \cdot V^i \quad i \in \{1, \ldots, h\} \tag{9}$$

where V^i represents the output feature map of the i-th graph convolutional layer, and the symbol dot is the dot product of the vector. By using these equations, the importance of original channels is then further enhanced or suppressed.

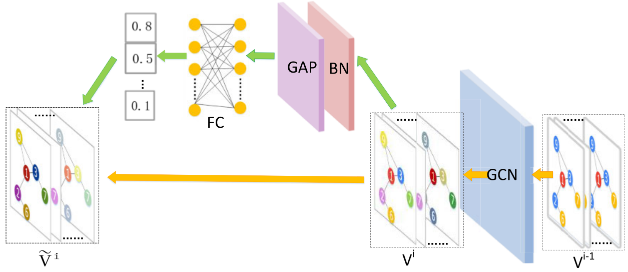


FIGURE 4. Channel attention module (CAModule). The output V^i is obtained by the input V^{i-1} through graph convolution. Then through the batch normalization (BN), the global average pooling (GAP) and the fully connected (FC) layer we obtain the channel response (0.8, 0.5, . . ., 0.1). Finally, we weight V^i with channel response to generate an adaptive adjustment result $\widetilde{V}^i$.

As shown in Figure 4, the channel attention module can generate a response related to the importance of each channel of pedestrian feature map, and then weights the original channel with it. These regards can highlight the relevant channels of intra-class while weakening the interference of the irrelevant channels. The color changes of nodes from a layer to the other in the figure presentation indicate that the weights of node channels are adjusted.

D. LOSS FUNCTION AND MODEL DISCUSSION

1) LOSS FUNCTION

A sample is in line with only one label in the classical cross entropy function. However, in the training process of RRGCCAN, a sample may correspond to multiple labels. an alternative cross entropy with logsoftmax is then adopted as the loss function, which is formulated at below:

$$loss(X, L) = \frac{1}{d}\sum_{i=1}^{d}(-X_{i,L_i} + log(\sum_j X_{i,j})) \tag{10}$$

where $X \in R^{d\times 2}$ and $L \in R^d$ represent prediction results and labels respectively, $d = b * h_1$ with b been well known as a mini-batch number, and $j \in \{0, 1\}$. The result of loss function indicates the link likelihoods of multiple edges are predicted.

2) DenseNet [2] AND ResNet [1]

Why don't we mainly use the most advanced network just published to extract features or based on this evaluation? Since the two networks, DenseNet and ResNet, are the basic networks of most models at present, we pay more attention to the universality and effectiveness of similarity learning methods, rather than the effective method only under a particular model. We choose DenseNet and ResNet as feature extraction models in the paper. Although most of the newly published feature extraction models do not have open source code, which is a problem for us to extract training set features, we also use one of the most advanced networks for feature extraction in the experimental part.

3) RESIDUAL NETWORK AND CHANNEL ATTENTION MECHANISM

On the one hand, the purpose of adding the residual network in layer 2,4 is the stability of the RRGCN, because we lose gradient in training. Secondly, it is one of our future research directions to study the effectiveness of the deep network of graph convolution in feature representation. On the other hand, for the channel attention mechanism, we use the idea of channel adaptive response of SENet for reference. The effect of using SENet module to convolute graph directly is not rational, so we redesigned a channel attention module based on the graph convolution module, and proved its effectiveness in the experiment. The accuracy potential of attention mechanism for graph network and person Re-ID is worth further study. It is worth noting that the re-ranking algorithm is divided into manual design [18] and learnable parameters. The former relies on experience and manual design parameters. Although it has some limitations, it does not need training process. Therefore, the time complexity is lower than the algorithm which needs to use the training set to optimize the parameters (ours).

IV. EXPERIMENTS

A. DATASET

To validate the effectiveness of our proposed methods, we conduct a series of experiments on three large-scale person Re-ID datasets, including DukeMTMC-reID [11], Market-1501 [10], and MSMT17 [12]. These datasets have diverse changes, for example, viewpoint variations, occlusion, illumination fluctuations, pose changes, and background clutters. DukeMTMC dataset used for the person Re-ID task contains 1,404 identities and their corresponding 34,183 image boxes, of which 702 identities are assigned as the training set and the rest of 702 as the test set. The test set is comprised of 2228 probes and 17661 gallery images. Market-1501 has 1,501 identities, 12,936 training images and 19,732 gallery images (of which 2,793 are distractors). It is split into 751 identities for training and 750 identities for testing. MSMT17 is similar to the setting of Market-1501, but the pedestrian situation is far more complicated. An overview of the datasets is shown in Table 1.

TABLE 1. The details of datasets.

Datasets	ID	Cam	Images	Train	Test
DukeMTMC-reID	1404	8	36411	702	702
Market-1501	1501	6	32668	751	750
MSMT17	4101	15	126441	1041	3060

B. EXPERIMENT SETUP

In the test process of extracting CNN features, we set the image size to (256, 128) and the batch to 32. The CNN's classification layer is removed to obtain feature maps with a fixed dimension. For the RRGCCAN at the test phase, we combine the feature maps of probes with gallery as a set to construct adjacency subgraphs. We follow the same training method to CNN with [5]. In the training process of RRGCCAN, 1-hop and 2-hop are set to 200 and 5 respectively for building subgraphs. The trainable parameters of our network are initialized by Xavier. We use Stochastic Gradient Descent (SGD) as the optimizer, in which the initial classifier learning rate is configured as 0.01, weighting decay is set to 0.0001 (that is, avoid over-fitting), and momentum is assigned a value of 0.9. The network is trained using four epochs.

Evaluation criteria are delegated as follows. To compare our method with existing methods, we take advantage of Cumulative Matching Characteristics (CMC) at rank-1 and mean Average Precision (mAP) on the datasets. Rank-k is the accuracy that the matched image in the gallery is contained in the top-k answers according to the similarity scores. Note that all our results are obtained based on single-query setup.

C. COMPARISON WITH GENERAL SIMILARITY METRIC WITHOUT RE-RANKING

Most of the existing works always compute the distance between two images in Euclidean space. We compare with it to validate the effectiveness of our proposed re-ranking method. The extracted CNN features need to be normalized, so Euclidean distance can be obtained by the product of vectors.

1) THE RESULT OF DukeMTMC-reID AND MARKET-1501

From Table 2 and Table 3, it can be seen that compared with the general similarity metric (Euclidean distance), our proposed method shows superiority. On the one hand, with an increase of ResNet model depth, although its effects are also improved, our method improves the recognition results at large scale. It can be concluded that our model can generalize features extracted from different depth networks. On the other hand, from the comparison of values, it can be observed intuitively that our method has a significant improvement. The specific mAP increase is between 9% and 13%, while the rank-1 increase is around 5%. The test results using the DenseNet features are similar to those using the ResNet features. Due to few parameters in single layer, DenseNet can use more layers under the same memory conditions. The increase of mAP is between 8% and 13%, while the increase of rank-1 is between 3% and 6%, which further demonstrates our method has excellent ability to pull the intra-class distance and push the inter-class distance at the same time.

TABLE 2. Comparison results of original method (l_2 distance) and the proposed method on DukeMTMC-reID. "original" means no re-ranking. "+" means the degree to which our method is superior to the original method.

Methods	DukeMTMC-reID				-	
	original		RRGCCAN		-	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet-18	58.3	76.7	71.2	82.5	+12.9	+5.8
ResNet-34	61.5	78.3	73.5	83.2	+12.0	+4.9
ResNet-50	65.9	81.6	77.7	86.0	+11.8	+4.4
DenseNet-121	67.7	82.7	80.0	88.4	+12.3	+5.7
DenseNet-161	70.0	84.0	80.5	88.3	+10.5	+4.3
DenseNet-201	69.4	84.2	80.1	87.9	+10.7	+3.7

TABLE 3. Comparison results of original method and the proposed method on Market-1501.

Methods	Market-1501				-	
	original		RRGCCAN		-	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet-18	67.2	85.5	79.8	88.9	+12.6	+3.4
ResNet-34	70.3	87.2	81.8	90.3	+11.5	+3.1
ResNet-50	75.2	89.9	84.8	91.2	+9.6	+1.3
DenseNet-121	77.0	91.3	86.9	93.6	+9.9	+2.3
DenseNet-161	79.2	91.8	87.8	93.1	+8.6	+1.3
DenseNet-201	78.7	91.6	87.6	93.7	+8.9	+2.1

TABLE 4. Ablation experiments and comparison results on MSMT17.

Methods	MSMT17					
	original		RRGCN		-	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet-18	30.4	57.3	40.3	64.3	+9.9	+7.0
ResNet-34	33.7	60.3	43.6	66.5	+9.9	+6.2
ResNet-50	41.0	67.6	52.1	73.8	+11.1	+6.2
DenseNet-121	44.5	71.7	56.3	76.9	+11.8	+5.2
DenseNet-161	48.6	74.3	60.1	79.4	+11.5	+5.1
DenseNet-201	47.5	73.2	59.2	79.2	+11.7	+6.0

TABLE 5. Ablation experiments and comparison results on MSMT17.

Methods	MSMT17					
	RRGCN		RRGCCAN		-	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet-18	40.3	64.3	41.3	65.6	+1.0	+1.3
ResNet-34	43.6	66.5	43.9	66.7	+0.3	+0.2
ResNet-50	52.1	73.8	53.0	74.8	+0.9	+1.0
DenseNet-121	56.3	76.9	57.6	78.2	+1.3	+1.3
DenseNet-161	60.1	79.4	61.1	80.4	+1.0	+1.0
DenseNet-201	59.2	79.2	59.8	79.8	+0.6	+0.6

2) THE RESULT OF MSMT17

As far as we know, MSMT17 is currently one of the most challenging person Re-ID datasets, which including more pedestrian situations. From the MSMT17 experimental results as shown in Table 4 and Table 5, compared to the

Euclidean distance, our proposed approach has great advantages. In addition, the overall increase of mAP is around 12%, and the increase of rank-1 is between 5% and 8%. It is worth noting that the 1-hop and 2-hop settings have a great impact on the effectiveness of the whole network when building the subgraphs, because they directly determine the size of the graph and the number of edges. If the subgraph is too small, too few positive and negative samples will lead to poor learning ability. Furthermore, the subgraph is not the bigger the better. (1) The number of pedestrian samples in the same category is relatively small. When there are too many negative samples, there may be many interferences in the training phase. (2) The pressure on GPU memory and computing costs is multiplied, so it needs to be adjusted according to the dataset. A comparative experiment on the 1-hop node can be seen in Table 8.

D. ABLATION EXPERIMENTS

To investigate whether our basic graph convolutional framework (RRGCN) can improve the quality of group similarity, we compare the effect of original distance with RRGCN as shown in Table 4. The results show that RRGCN can promote the performance of specific networks with different scales. Specifically, the mAP based on ResNet can increase by around 10%, while rank-1 increased by around 6%. The results based on DenseNet shows that the mAP rose by 11% to 12%, while rank-1 can grow by around 5%. The results show that using the images around the probe to optimize its features is advantageous and the reasoning ability of the graph convolution network can promote the person Re-ID. From the experimental results, it can be concluded that with the deepening of the network structure, the performance can be gradually improved to a certain extent due to the gradual enhancement of feature robustness.

In order to assess the essentials of using the channel attention module (CAModule) in re-ranking, we compare the basic framework (RRGCN) with the complete network (RRGCCAN) with the attention mechanism. The gap between the results of RRGCN and RRGCCAN as shown in Table 5 indicates that the attention mechanism can boost our basic framework effectively. For the challenging MSMT17 dataset, although different CNN models may affect the similarity and the quality of methods, Re-ID performance by using RRGCCAN improves nearly 1% in mAP and 1%-2% in rank-1, which is common for all CNN models used. The proposed channel attention module boosts the performance of the model significantly in the experiment. It can not only promote GCN, but also has stable performance in multiple backbone networks.

E. COMPARISON WITH STATE-OF-THE-ART APPROACHES

1) COMPARISON ON MARKET-1501 AND DukeMTMC-reID

TABLE 6. Comparison with state-of-the-art methods on the Market-1501 and DukeMTMC-reID datasets. rank-1 (%) and mAP (%) are reported. (rr) represents the re-ranking method. Bold texts represent the best result in the same evaluation standard.

Methods	Market-1501		DukeMTMC	
	mAP	rank-1	mAP	rank-1
k-reciprocal(rr) [18]	63.6	77.1	-	-
PSE+ECN(rr) [19]	84.0	90.3	**79.8**	85.2
SGGNN*(rr) [21]	76.7	**91.5**	64.6	79.1
expanded k-rnn(rr) [20]	84.0	90.7	79.2	85.5
ResNet-50+RRGCCAN	**84.8**	91.2	77.7	**86.0**
GCSL+DCRF [6]	81.6	93.5	69.5	84.9
OSNet [14]	84.9	94.8	73.5	88.6
DG-Net [5]	86.0	94.8	74.8	86.6
FPR [15]	86.6	95.4	78.4	88.6
DFLCAR [16]	84.7	**96.1**	73.1	86.3
k-reciprocal*(rr)	90.7	93.3	**85.6**	88.6
DG-Net+RRGCCAN	**91.7**	95.6	83.5	**90.1**

As shown in Table 6 we compare the proposed method with the latest methods on Market-1501 and DukeMTMC-reID datasets, including the latest published Re-ID precision, the latest re-ranking methods (rr), and the re-ranking approach (*) based on the same conditions. Since a large number of Re-ID methods have been reported, comparing all methods is impractical. We mainly compare the latest results, which range from 2017 to 2019, where k-reciprocal, PSE+ECN, expanded k-rnn, SGGNN* are re-ranking approaches as ours. The k-reciprocal* method is reproduced with publicly available codes using the same baselines to ensure a fair comparison. Since many re-ranking methods employ ResNet-50 or variants as the baseline architecture, we also compare the results of our proposed RRGCCAN based on ResNet-50. It is noteworthy that SGGNN* is essentially a similarity learning framework. We compare it with our approach as a re-ranking method, and quote the experimental results under the same conditions. In summary, it can be seen clearly that the proposed approach is more advantageous when compared with state-of-the-art methods due to the consideration of the group similarity during the testing phase. In practice, compared with the similarity learning framework [21], we find that the re-ranking module with learning ability can be integrated with other advanced models (e.g. DG-Net) easily. And it can use the original ranking information to learn group-level similarity. Compared with the traditional methods, our method can use the training set to train the network. The GCN can learn more accurate similarity relationship without relying on customized parameters to evaluate the similarity between images.

2) ACHIEVING THE STATE-OF-THE-ART IN MSMT17

Although we pay more attention to the generalization ability of the method, we are still surprised to find that on the MSMT17 dataset, our method can achieve 64.3% mAP and 82.3% rank-1 accuracy, which is very competitive with the state-of-the-art approaches as shown in Table 7. Because the compared methods are more focused on the performance improvement of feature extraction networks, we mainly compare the accuracy of the evaluation results rather than the feature extraction model improvement. In MSMT17, our

TABLE 7. Comparison with state-of-the-arts on MSMT17. -: not available.

Methods	MSMT17			
	mAP	rank-1	rank-5	rank-10
Glad [13]	34.0	61.4	76.8	81.6
IANet [4]	46.8	75.5	85.5	88.7
DG-Net [5]	52.3	77.2	87.4	90.5
OSNet [14]	52.9	78.7	-	-
DG-Net+RRGCCAN	**64.3**	**82.3**	**89.0**	**91.0**

model infers similarity based on the information around the probe, to improve the pedestrian features. The improvement effect is significant. It is worth noting that to explore the impact of increasing the number of GCN layers [25] on person Re-ID tasks, we conducted experiments on networks with layers 4, 6, and 8 based on DenseNet-161 backbone network and Market-1501 dataset. It can be seen clearly in Table 8. The network doesn't add other modules (such as attention, residual model). Available from experimental data, without considering the overall performance of the network structure, blindly increasing the number of layers will cause performance degradation.

TABLE 8. The influence of subgraphs and network layers.

	subgraphs (1-hop)				layers		
	50	100	150	200	4	6	8
rank-1	93.2	93.0	93.0	93.1	90.9	87.9	86.3
mAP	84.0	86.7	87.4	87.8	86.1	84.0	82.4

V. CONCLUSION

We propose a competitive re-ranking method based on graph convolution channel attention network for person re-identification. Our strategies are based on graph model and attention mechanism. We construct the local subgraphs, which represent the context similarity for the selected subsets by the nearest neighbor rule. Then, the group intra/inter-class similarity relations are predicted by graph convolutional network. Moreover, our channel attention mechanism alleviates the influences of irrelevant channels and enhances the corresponding weights to relevant channels for more robust similarity estimation. Extensive experiments on the three public datasets demonstrated the generalization ability of our method, and it can achieve more competitive Re-ID accuracy in comparison with the state-of-the-art approaches. A series of ablation studies proved the effectiveness of each component of our approach.

REFERENCES

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

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[3] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7127–7136.

[4] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9309–9318.

[5] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2133–2142.

[6] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8649–8658.

[7] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 17, 2020, doi: 10.1109/TKDE.2020.2981333.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[9] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian, "Person re-identification meets image search," 2015, *arXiv:1502.02171*. [Online]. Available: http://arxiv.org/abs/1502.02171

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

[12] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

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

[14] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3701–3711.

[15] H. Lingxiao, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8449–8458.

[16] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8039–8048.

[17] Z. Wang, L. Zheng, Y. Li, and S. Wang, "Linkage based face clustering via graph convolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1117–1125.

[18] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3652–3661.

[19] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.

[20] Y. Chen, J. Yuan, Z. Li, Y. Wu, M. Nouioua, and G. Xie, "Person re-identification based on re-ranking with expanded k-reciprocal nearest neighbors," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 486–494, Jan. 2019.

[21] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 486–504.

[22] J. C. S. Jacques, X. Baró, and S. Escalera, "Exploiting feature representations through similarity learning, post-ranking and ranking aggregation for person re-identification," *Image Vis. Comput.*, vol. 79, pp. 76–85, Nov. 2018.

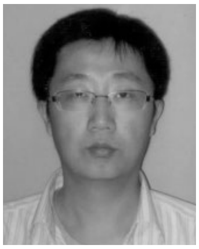
[23] B. Jiang, X. Wang, and B. Luo, "PH-GCN: Person re-identification with part-based hierarchical graph convolutional network," 2019, *arXiv:1907.08822*. [Online]. Available: http://arxiv.org/abs/1907.08822

[24] S. Woo, J. Park, J. Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[25] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9267–9276.

XIAOQIANG CHEN received the B.E. degree from the College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fujian, China, in 2018. He is currently pursuing the master's degree with the School of Information Science and Engineering, Xiamen University, Fujian. His research interests include robot navigation, machine vision, and reinforcement learning.

LING ZHENG received the B.S. degree from the Faculty of Software Engineering, Fujian Normal University, and the M.S. and Ph.D. degrees from the Department of Computer Science, Aberystwyth University. He currently holds a postdoctoral position at the Department of Computer Science, Xiamen University. His research interests include image processing, feature selection, and rule-based advanced reasoning.



CHONG ZHAO received the B.S. degree from the Department of Computer Science, Jilin University, the M.S. degree from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong. He is currently an Assistant Professor with the Department of Computer Science, Xiamen University. His research interests include computer graphics, geometry processing, and wavelet analysis.

QICONG WANG received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2007. He is currently an Associate Professor with the Department of Computer Science and the Shenzhen Research Institute, Xiamen University, China. His research interests include machine vision, robot navigation, and machine learning.

MAOZHEN LI (Member, IEEE) received the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences, in 1997. He is a Professor with the Department of Electronics and Computer Engineering, Brunel University London, U.K. He has over 180 research publications in his research areas, including four books. His main research interests include high-performance computing, big data analytics, and intelligent systems with applications to smart grid, smart manufacturing, and smart cities. He has served in over 30 IEEE conferences and serves on the editorial board of a number of journals. He is a Fellow of the British Computer Society and the IET.