Part-Guided Graph Convolution Networks for Person Re-identification

Zhong Zhang[a], Haijia Zhang[b], Shuang Liu[a,*], Yuan Xie[c], Tariq S. Durrani[d]

[a] *Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China*
[b] *Faculty of Psychology, Tianjin Normal University, Tianjin 300387, China*
[c] *The School of Computer Science and Software Engineering, East China Normal University, Shanghai, 200062, China*
[d] *Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow Scotland, UK*

Abstract

Recently, part-based deep models have achieved promising performance in person re-identification (Re-ID), yet these models ignore the inter-local relationship of the corresponding parts among pedestrian images and the intra-local relationship between adjacent parts in one pedestrian image. As a result, the feature representations are hard to learn the information from the same parts of other pedestrian images and are lack of the contextual information of pedestrian. In this paper, we propose a novel deep graph model named Part-Guided Graph Convolution Network (PGCN) for person Re-ID, which could simultaneously learn the inter-local relationship and the intra-local relationship for feature representations. Specifically, we construct the inter-local graph using the local features extracted from the same parts of pedestrian images and build the adjacency matrix using the similarity so as to mine the inter-local relationship. Meanwhile, we construct the intra-local graph using the local features extracted from different body parts in one pedestrian image, and propose the fractional dynamic mechanism (FDM) to accurately describe the correlations between adjacent parts in the optimization process. Finally, after the graph convolutional operation, the inter-local relationship and the intra-local relationship

*Corresponding author
Email address: shuangliu.tjnu@gmail.com (Shuang Liu)

are injected into the feature representations of pedestrian images. Extensive experiments are conducted on Market-1501, CUHK03, DukeMTMC-reID and MSMT17, and the experimental results show the proposed PGCN exceeds state-of-the-art methods by an overwhelming margin.

Keywords: Person re-identification, graph convolution network.

1. Introduction

In recent years, person re-identification (Re-ID) [1, 2, 3, 4] has received extensive attention from academia and industry due to wide applications in human activity analysis, multi-object tracking and other fields [5, 6, 7]. It originates from the basic demand in video surveillance for retrieving specific person of interest from a large gallery captured under various cameras or scenes. Since the appearance of pedestrian is easily affected by many factors including posture, viewpoint, illumination, occlusion, etc., person Re-ID is an extremely challenging task.

One main issue for person Re-ID is to develop completed and discriminative feature representations for pedestrian images. Some researchers [8, 9, 10] utilize Convolutional Neural Networks (CNNs) to extract global features from entire pedestrian images so as to discover appearance clues. Different from global features, part-based models [11, 12, 13, 14, 15, 16] are designed to learn local features. Among them, some methods [11, 15, 16] divide pedestrian images or feature maps into fixed-size parts according to human prior knowledge and then learn local features from each part. In order to overcome the misalignment, some methods [17, 18, 19] employ semantic segmentation or pose estimation to obtain more accurate human part location. Furthermore, the fusion strategies of global and local features [20, 21, 22] are usually adopted to fully exploit their advantages.

Although the part-based models could obtain discriminative representations for pedestrian images, the existing methods have two obvious drawbacks. (1) The inter-local relationship of the corresponding parts among pedestrian im-

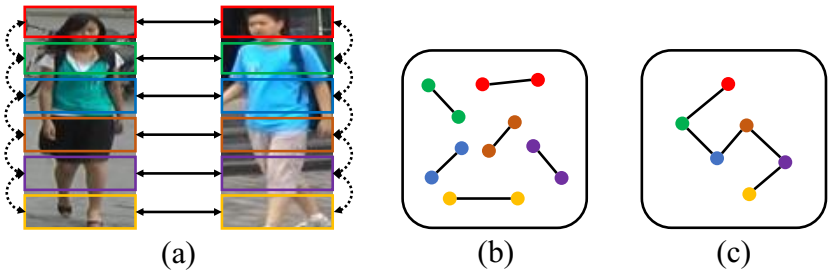


Figure 1: (a) The dotted line indicates the intra-local relationship, and the solid line denotes the inter-local relationship; (b) Inter-local graph. The local features learned from the same parts of pedestrian images are regarded as the nodes to build the inter-local graph; (c) Intra-local graph. The local features extracted from different body parts are treated as the nodes to construct the intra-local graph.

ages is ignored, which cannot learn the information from the same parts of other pedestrian images as shown in Fig. 1(a). (2) Local features are extracted independently without considering the intra-local relationship between adjacent parts as shown in Fig. 1(a). This results in the lack of contextual information of pedestrian. In a word, the absence of the two kinds of local relationships weakens the representation capacity of local features, and therefore it is significant for person Re-ID to fully consider the inter-local relationship and intra-local relationship for local feature representations.

The Graph Convolution Networks (GCNs) [23] possess powerful relationship reasoning and information aggregation capabilities, and they have promoted the performance of many fields [24, 25, 26]. The property of GCNs is to mine the relationship among the nodes by propagating the message in the graph. Hence, it is reasonable to resort to GCNs to build intra-local and inter-local relationships for person Re-ID.

In this paper, we propose a novel deep graph model named Part-Guided Graph Convolution Network (PGCN) for person Re-ID, which could simultaneously learn the inter-local relationship and the intra-local relationship for feature representations. To this end, the proposed PGCN is designed as a multi-stream network and correspondingly constructs two types of graphs as

shown in Fig. 1(b) and (c). In order to mine the inter-local relationship, we construct the inter-local graph where the local features extracted from the same parts of pedestrian images are regarded as the nodes. Then, we utilize the similarities between the same parts as the adjacency matrix where the elements of adjacency matrix are data-dependent.

As for the intra-local relationship, we build the intra-local graph where the local features extracted from different body parts are treated as the nodes. The adjacent parts of pedestrian have high correlations, and therefore the elements in the adjacency matrix corresponding to the adjacent parts are assigned to non-zero values, otherwise to zero. Furthermore, we propose the fractional dynamic mechanism (FDM) to optimize the adjacency matrix in the training stage, which is more flexible than the constant value setting. After constructing the two kinds of graphs, we apply the graph convolutional operation to aggregate the information from different local features. As a result, the inter-local and intra-local relationships are injected into the feature representations, which ensures to obtain the completed and discriminative features for pedestrian images.

Our contribution lies in three folds.

• We propose PGCN to learn the inter-local relationship and the intra-local relationship simultaneously by constructing the inter-local graph and the intra-local graph so that the representation capacities of local features are enhanced.

• The adjacency matrix of inter-local graph is totally data-driven, and meanwhile the adjacency matrix of intra-local graph is optimized by FDM. Hence, they could describe the relationship of local features accurately.

• Extensive experiments are conducted on four large-scale person Re-ID datasets, i.e., Market-1501, CUHK03, DukeMTMC-reID and MSMT17, and the results prove that PGCN surpasses state-of-the-art methods by an overwhelming margin. For example, on MSMT17, PGCN outperforms many state-of-the-art methods by at least 3.9% in Rank-1 accuracy and 11.9% in mAP accuracy.

2. Related Work

2.1. Part-based Models for Person Re-ID

With the rapid development of deep learning, many CNN-based models have been proposed for person Re-ID, and they hold dominant position in this community [27, 28, 29]. The part-based model for person Re-ID is closely related to our work, and we briefly introduce them in this subsection.

Several researchers extract the structure information of pedestrian via partitioning the pedestrian image or feature maps into uniform parts [11, 15, 16]. Sun et al. [11] divide feature maps into several uniform stripes, and then pool them to extract local features. Quan et al. [15] design a part-aware module to learn the body structure information, which first splits feature maps into several parts and then adopts the self-attention mechanism to learn more specific part-based information.

However, these direct partition strategies are prone to causing part misalignment due to posture changes, and therefore many approaches [12, 13, 14] are proposed to address this problem. Suh et al. [12] develop a two-stream network to separately generate appearance and part maps, and then perform bilinear pooling operation to compute the final part-aligned feature representations. Wei et al. [13] resort to the pose detection technology to estimate four human body keypoints, and then divide the pedestrian image into three regions for local part alignment. Guo et al. [14] utilize the human parsing model to obtain the semantic human part masks, and then extract human part-aligned features using the human part branch.

In order to fully exploit the advantages of global and local features, many researchers fuse them to represent pedestrian images [20, 21, 22]. Li et al. [20] design the Joint Learning Multi-Loss (JLML) CNN model to extract global and local features concurrently, and utilize the joint learning scheme to learn correlated complementary information between local and global feature selections. In [16], Zheng et al. propose the coarse-to-fine pyramidal model to capture the discriminative information of different scales, and employ identification and

triplet losses to learn global and multi-scale local features.

2.2. Graph Convolutional Networks

Although CNNs have been successful in handling data in the Euclidean space [30], they encounter obstacles in processing data in the non-Euclidean space. The non-Euclidean data is usually represented as a graph with complex relationships and interdependency among objects. Inspired by CNNs, GCNs are designed for handling complex graph data [23]. According to the manner of convolutional operation, there are mainly two types of GCNs, i.e., spectral-based and non-spectral (spatial).

Spectral-based GCNs [31, 32, 33] depend on the eigen decomposition of the Laplacian matrix, and they perform the convolutional operation based on the Graph Fourier Transform. While non-spectral approaches [24, 25] utilize the first-order approximation of spectral filters to simplify GCNs, and they implement the convolutional operation by manually defining convolution on nodes as well as neighbor nodes of graph.

Recently, some researchers utilize GCNs to handle person Re-ID task [34, 35, 36]. Shen et al. [34] develop Similarity-Guided Graph Neural Network (SGGNN) to learn similarity estimation by constructing graph with different probe-gallery pairs, and obtain pairwise relationships among them. Chen et al. [35] propose to model multi-scale local similarity between image pairs and group similarity between probe-gallery pairs via CRF in order to learn robust similarity metric. Ye et al. [37] propose a dynamic dual-attentive aggregation (DDAG) learning method to mine both intra-modality part-level and cross-modality graph-level contextual cues for visible-infrared person Re-ID. In short, there are two differences between DDAG and PGCN. Firstly, DDAG focuses on visible-infrared person Re-ID, while the proposed PGCN studies on visible person Re-ID. Secondly, DDAG considers the relation in the inter-modality and the intra-local modality, while the proposed PGCN models the inter-local relationship and the intra-local relationship by constructing the inter-local graph and the intra-local graph.

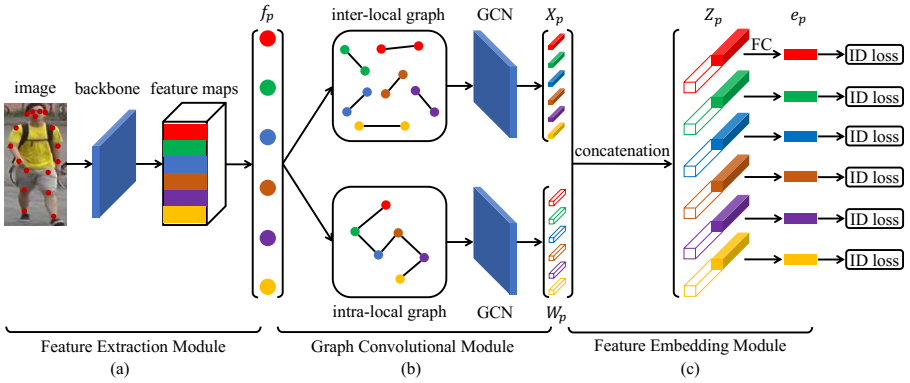


Figure 2: The overview architecture of PGCN. (a) In the Feature Extraction Module, we first resize the pedestrian images, and then feed them into backbone to extract feature maps with the size of $2048 \times 24 \times 8$. Afterwards we divide the pedestrian image into P regions by pooling these feature maps with body keypoints to obtain local features $f_p \in R^{2048}$ for the *p-th* region. (b) In the Graph Convolutional Module, we utilize these local features to construct the inter-local and intra-local graphs respectively, and then perform the graph convolutional operation on the two types of graphs to obtain GCN features X_p and W_p respectively. (c) In the Feature Embedding Module, we concatenate X_p and W_p extracted from the *p-th* regions to obtain Z_p, and then we apply the independent FC layer to reduce the dimensionality of Z_p. Finally, we learn the final feature representation e_p for the *p-th* region of pedestrian image.

3. Approach

In this section, we first give an overview of the proposed PGCN, and then detail the key components of PGCN including Feature Extraction Module, Graph Convolutional Module and Feature Embedding Module. Finally, we formulate the loss function of PGCN.

3.1. Overview

The proposed PGCN is mainly composed of the following three key components as shown in Fig. 2.

Feature Extraction Module. We utilize CNN to obtain feature maps, and meanwhile locate the keypoints of pedestrian using the pose estimation model [38]. Then, we extract local features from different body parts with these keypoints.

Graph Convolutional Module. We regard the local features extracted from different body parts as the nodes of graph, and then construct the inter-local graph and the intra-local graph. Afterwards, we perform the graph convolutional operation on the two types of graphs to learn the local relationship.

Feature Embedding Module. After graph convolutional operation, we obtain GCN features from the inter-local and intra-local graphs respectively, and then we concatenate these features from the same parts. Finally, we apply the independent fully connected (FC) layers to reduce the dimensionality of them.

3.2. Feature Extraction Module

Backbone. The target of Feature Extraction Module is to extract local features from pedestrian images. Concretely, we utilize ResNet-50 [39] as backbone due to its promising performance in many tasks.

In order to obtain high-level feature maps with large spatial size, we conduct some modifications on the original ResNet-50. Specifically, the global average pooling and the following layers are removed, and the stride of Conv5_1 layer is changed from 2 to 1 so as to enlarge the spatial size of feature maps.

Feature Extraction. We first resize pedestrian images into 384×128, and then feed them into backbone. Afterwards, we obtain the high-level feature maps with the size of $2048 \times 24 \times 8$ where the channel number of feature maps is 2048, and the height and width of each feature map are 24 and 8 respectively. Beside, in order to alleviate the part misalignment, we utilize the human pose estimation model to predict 17 keypoints of the pedestrian image as shown in Fig. 2(a). Similar to [27], we partition the pedestrian image into P regions, and then obtain the local feature $f_p \in R^{2048}$ $(p = 1, 2, \cdots, P)$ by pooling feature maps. In addition, we denote $v_p \in \{0, 1\}$ as the visibility of the *p-th* region. If the confidence of keypoints in the *p-th* region is low, v_p is set to 0 as [27], and meanwhile the local feature f_p is replaced by a zero vector.

3.3. Graph Convolutional Module

The inter-local relationship of the corresponding parts among pedestrian images and the intra-local relationship between adjacent parts are beneficial to enhance the representation capacity of local features, and therefore we design Graph Convolutional Module to simultaneously learn them for feature representations.

Graph Construction. We utilize the local features learned from the Feature Extraction Module as the nodes of inter-local graph and intra-local graph. Specifically, in order to mine the inter-local relationship, we treat the similarity between the same parts of pedestrian images as the element of adjacency matrix. Hence, each part of pedestrian image corresponds to one inter-local graph. Given a set of pedestrian images $\Omega = \{I_1, I_2, \cdots, I_N\}$ where N is the pedestrian image number, the similarity between the p-th region of I_a and I_b is formulated as

$$s_p^{ab} = \frac{f_p^a \odot f_p^b}{||f_p^a||_2 \cdot ||f_p^b||_2 + \xi}, \ a, b \in \{1, 2, \cdots, N\} \tag{1}$$

where $\odot$ denotes the dot product, $||\cdot||_2$ is l_2 norm, and f_p^a and f_p^b are the local features extracted from the p-th region of I_a and I_b, respectively. Here, ξ is set a small positive constant in order to avoid zero vectors caused by region invisibility.

We formulate the adjacency matrix of inter-local graph for the p-th region as

$$S_p = \left[s_p^{ab}\right] \tag{2}$$

where $S_p \in R^{N \times N}$.

The intra-local graph aims to build the relationship among local parts in one pedestrian image. As shown in Fig. 3, we can see that the adjacent parts possess high correlations, and vice versa. For example, R_3 and R_4 have similar appearances, while R_1 and R_3 are significantly different. Based on this observation, the elements in the adjacency matrix corresponding to adjacent parts are set to non-zero, otherwise to zero. Hence, we formulate the adjacency matrix

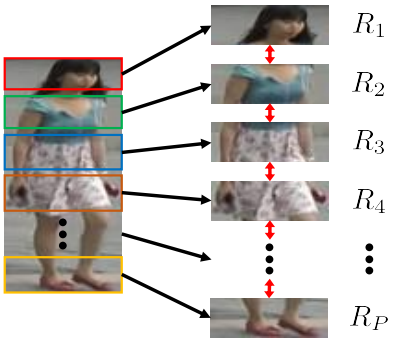


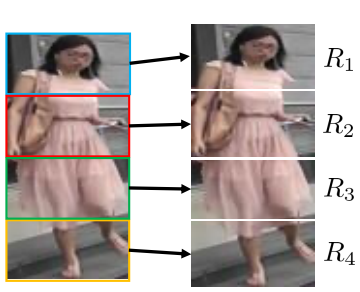
Figure 3: The pedestrian image is divided into P regions, from R_1 to R_P.

of intra-local graph as

$$T = [t_{ij}] \tag{3}$$

where $T \in R^{P \times P}$. If i and j are adjacent, t_{ij} is set to non-zero, otherwise zero.

We have qualitatively analyzed the correlations between adjacent parts, but how to quantify them? It is naive to set all non-zero elements to identical constants while neglecting the difference between adjacent parts. From Fig. 3 we can see that the correlation between R_3 and R_4 is higher than that of R_2 and R_3. The elements of traditional adjacency matrix are pre-defined, which is so rigid to accurately describe the local relation among different parts. Hence, we propose FDM to optimize the adjacency matrix of the intra-local graph. Specifically, we design the adjacency matrix as the optimizable parameters of PGCN where the elements corresponding to adjacent parts are learnable and independent, and the remaining elements corresponding to non-adjacent parts are fixed to zero values. In a word, in the training phase, the proposed FDM only optimizes the elements corresponding to adjacent parts in the adjacency matrix, and therefore we named this optimization strategy as FDM. In this way, we could mine the local information and accurately describe the latent intra-local relationship between these adjacent parts. Moreover, we get rid of the interference information from non-adjacent parts by fixing their elements of the adjacency matrix to zero values. In order to intuitively understand the proposed FDM, we take an example to show how FDM works. In Fig. 4(a), we



(a)

	R_1	R_2	R_3	R_4
R_1	L	L	0	0
R_2	L	L	L	0
R_3	0	L	L	L
R_4	0	0	L	L

(b)

Figure 4: (a) The pedestrian image is divided into four regions, from R_1 to R_4. (b) The adjacency matrix of intra-local graph for four regions. L denotes the element is learnable.

divide the pedestrian image into four parts and construct the intra-local graph using these parts. Fig. 4(b) shows the adjacency matrix of intra-local graph. In the training phase, the proposed FDM only optimize the elements marked by L (corresponding to adjacent parts) in the adjacency matrix and keep the elements marked by 0 (corresponding to non-adjacent parts). In short, there is no restriction for these non-zero elements, and therefore it is more flexible than the constant value setting. In this way, we can learn more accurate intra-local relationship from adjacent parts. It should be noticed that these non-zero elements are initialized to 1 in the training stage.

Graph Convolution. After constructing the inter-local and intra-local graphs, we perform the graph convolutional operation on them. Specifically, the graph convolutional operation on the inter-local graph is formulated as

$$X_p^l = \sigma(S_p X_p^{l-1} U^{l-1}) \tag{4}$$

where X_p^{l-1} is the activation matrix of the $l-1$-th graph convolution layer, U^{l-1} denotes the parameter matrix in the $l-1$-th graph convolution layer to be learned, and $\sigma(\cdot)$ denotes the ReLU activation function. The activation matrix $X_p^0 \in R^{N\times 2048}$ is initialized by the local features f_p from N pedestrian images. After the graph convolutional operation, we obtain the GCN features of inter-local graph $X_p \in R^{N\times d_1}$ (d_1 is the dimensionality of output features).

Similar to the inter-local graph, we formulate the graph convolutional operation on the intra-local graph as

$$Y^l = \sigma(TY^{l-1}V^{l-1}) \tag{5}$$

where Y^{l-1} is the activation matrix of the $l-1$-th graph convolution layer, V^{l-1} denotes the parameter matrix in the $l-1$-th graph convolution layer to be learned. The activation matrix $Y^0 \in R^{P\times 2048}$ is initialized by local features extracted from all the P regions, i.e., $[f_1, f_2, \cdots, f_P]$. After the graph convolutional operation, we obtain the GCN features of intra-local graph $Y \in R^{P\times d_2}$ (d_2 is the dimensionality of output features). It should be noticed that Y is learned from one pedestrian image. We change Y of all N pedestrian images to $W_p \in R^{N\times d_2}$ for the p-th region using a dimensionality transformation.

3.4. Feature Embedding Module

After the graph convolutional operation on the inter-local and intra-local graphs, we can obtain the GCN features $X_p \in R^{N\times d_1}$ and $W_p \in R^{N\times d_2}$, respectively. In this module, we concatenate them which are both extracted from the p-th region

$$Z_p = X_p \oplus W_p \tag{6}$$

where $\oplus$ denotes the concatenation operation, and $Z_p \in R^{N\times(d_1+d_2)}$.

Afterwards, we employ the independent FC layers on Z_p to reduce the dimensionality from d_1+d_2 to 256. Finally, we obtain the final feature representation $e_p \in R^{256}$ for the p-th region of pedestrian image. Note that the Feature Embedding Module can also be implemented by other methods, such as [40]. In order to intuitively understand Eq. 4, Eq. 5 and Eq. 6, we show the learning procedure of GCN features in Fig. 5.

3.5. Loss Function

In order to effectively supervise the feature learning process, we apply the independent classifiers and cross-entropy losses in PGCN. Specifically, each classifier is composed of the FC layer and the softmax function. In addition, the part

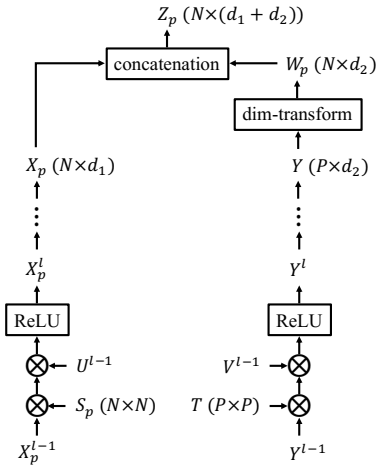


Figure 5: The learning procedure of GCN features. $\otimes$ denotes the matrix multiplication.

absence is inevitably caused by occlusion or imperfect detection, and therefore we inject the part visibility into the loss function. The loss function of PGCN is formulated as

$$Loss = \sum_{p=1}^{P} L_p \cdot v_p \tag{7}$$

where L_p is the cross-entropy loss of the p-th region

$$L_p = -\sum_{c=1}^{C} p_c(e_p) \log q_c(e_p) \tag{8}$$

where C is the number of identities, $q_c(e_p) \in [0, 1]$ is the prediction value of e_p for the c-th identity. $p_c(e_p)$ is true identity of e_p, and $p_c(e_p)$ is equal to 1 if e_p belongs to the c-th identity, otherwise 0.

In the test stage, we measure the similarity between the query image I_q and the gallery image I_g

$$Dist = \frac{\sum_{p=1}^{P} cosine(e_p^q, e_p^g) \cdot v_p^q}{\sum_{k=1}^{P} v_k^q} \tag{9}$$

where v_p^q is the visibility of the p-th region of I_q and $cosine$ denotes the cosine distance between the p-th regions of I_q and I_g.

4. Experiments

In this section, we evaluate the proposed PGCN on four large-scale person Re-ID datasets, i.e., Market-1501 [41], CUHK03 [42], DukeMTMC-reID [43] and MSMT17 [44]. We briefly introduce the four datasets, and then provide the implementation details. Afterwards, we perform ablation studies to analyze the contribution of inter-local graph, intra-local graph and FDM, and compare the proposed PGCN with GCN-based methods and state-of-the-art methods. Finally, we conduct extensive experiments to study the impact of hyper-parameters.

4.1. Datasets

Market-1501 contains 32,668 images from 1,501 pedestrians. There are 12,936 images of 751 pedestrians in the training set, and 3,368 query images and 15,913 gallery images from the other 750 pedestrians. **CUHK03** includes 14,097 images from 1,467 pedestrians. Among them, 7,365 images of 767 pedestrians constitute the training set, and the remaining images are divided into 1,400 query images and 5,332 gallery images from 700 pedestrians. **DukeMTMC-reID** consists of 16,522 training images of 702 pedestrians, 2,228 query images of 702 pedestrians and 17,661 gallery images of 1,110 pedestrians (702 pedestrians and 408 distractor pedestrians). **MSMT17** is the largest dataset including 32,621 training images (1,041 pedestrians), 11,659 query images and 82,161 gallery images (3,060 pedestrians).

We list the statistical information of these datasets in Table 1, and show some pedestrian images in Fig. 6. In order to comprehensively assess the effectiveness of PGCN, we regard the Cumulative Matching Characteristic (CMC) curve at Rank-1 accuracy and the mean Average Precision (mAP) accuracy as the evaluation metrics.

4.2. Implementation Details

Model. We take ResNet-50 pre-trained on ImageNet [45] as backbone, and utilize the human pose estimation model trained on COCO [46] to locate the keypoints of pedestrian. Following EANet [27], we first extract 17 keypoints

Table 1: The statistical information of the four datasets. A / B indicates pedestrians / images.

datasets	training set	query	gallery
Market-1501	751 / 12,936	750 / 3,368	750 / 15,913
CUHK03	767 / 7,365	700 / 1,400	700 / 5,332
DukeMTMC-reID	702 / 16,522	702 / 2,228	1,110 / 17,661
MSMT17	1,041 / 32,621	3,060 / 11,659	3,060 / 82,161

Figure 6: Several pedestrian image samples from the four datasets.

using the human pose estimation model, and partition each pedestrian image into 6 parts (head, upper torso, lower torso, upper leg, lower leg, foot) based on these keypoints. Afterwards, we extract the local feature from each part via max pooling. Details can be found in [27].

Pre-processing. We first resize the input image into 384×128, and then perform data augmentation using the random horizontal flip with the probability of 0.5 and the image normalization on RGB channels.

Optimization. In order to optimize the parameters of PGCN, we utilize

Table 2: Ablation studies on the four datasets. "w/o FDM" indicates without FDM.

Method	Market-1501		CUHK03		DukeMTMC-reID		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
baseline	94.0	83.0	66.1	60.9	84.3	70.0	76.3	48.1
$+S_p$	96.1	92.7	78.4	76.3	86.5	79.1	82.6	68.3
$+T$ (w/o FDM)	97.0	94.0	80.0	79.1	88.4	81.8	85.1	70.7
+FDM	**98.0**	**94.8**	**86.7**	**83.6**	**91.1**	**85.2**	**87.7**	**72.7**
+FDM (all learnable)	97.2	94.1	83.0	81.1	89.4	82.8	86.1	70.9
−keypoints	96.9	93.8	81.3	77.8	90.1	84.4	82.7	65.4

the stochastic gradient descent (SGD) algorithm as the optimizer. Specifically, the momentum and the weight decay are set to 0.9 and 0.0005, respectively. The learning rate is initialized to 0.01 for the backbone, and 0.02 for the graph convolution layers and FC layers. The learning rate is multiplied by 0.1 after 50 epochs. Meanwhile, we initialize the non-zero elements in the adjacency matrix T of intra-local graph to 1. In the training stage, we set 80 epochs and 16 batch sizes to optimize PGCN. In addition, d_1 and d_2 (the dimensionality of GCN features) are both set to 2048. It should be noticed that we maintain the same parameter settings on the four datasets.

4.3. Ablation Studies

In order to evaluate the effectiveness of the inter-local graph, the intra-local graph and FDM, we conduct ablation experiments on the four datasets. Specifically, we treat the method which learns local features without considering the relationship among them as the *baseline*, and we implement the baseline by combining the Feature Extraction Module and the Feature Embedding Module. Based on this, we add the inter-local graph, the intra-local graph (without FDM), and FDM to *baseline*, one by one, which are denoted as $+S_p$, $+T$ (w/o FDM) and +FDM. Besides, we replace keypoints with the uniform partition to evaluate the effectiveness of keypoints (denoted as −keypoints). The experimental results are shown in Table 2.

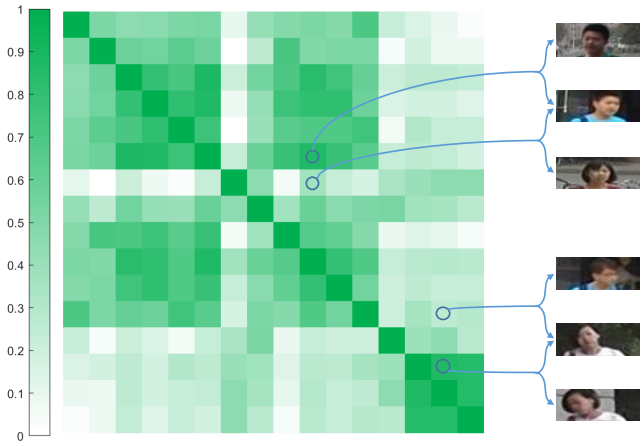


Figure 7: Visualization of the adjacency matrix of inter-local graph. The deeper color indicates larger value.

The effectiveness of the inter-local graph. We first add the inter-local graph ($+S_p$) to *baseline*. In Table 2, we can see that there is an obvious improvement on Rank-1 accuracy and mAP accuracy for the four datasets. Specifically, it increases mAP accuracy from 83.0% to 92.7%, and Rank-1 accuracy from 94.0% to 96.1% on Market-1501. As for the largest dataset MSMT17, it obtains prominent improvement with +6.3% and +20.2% in Rank-1 accuracy and mAP accuracy, respectively. It is because the inter-local relationship is taken into consideration by aggregating the information from the same parts of other pedestrian images. In the process of aggregation, the parts which have high correlation are emphasized, and the local features can learn discriminative information from these parts to enhance the representation capacities.

In order to intuitively understand the effectiveness of the inter-local graph, we visualize its adjacency matrix. Specifically, we take the first regions of pedestrian images as an example, and show the adjacency matrix in Fig. 7 where we normalize the elements in the adjacency matrix to $[0, 1]$. From this figure, we can see that the regions with high scores have similar appearances and vice versa.

The effectiveness of the intra-local graph. As shown in Table 2, both

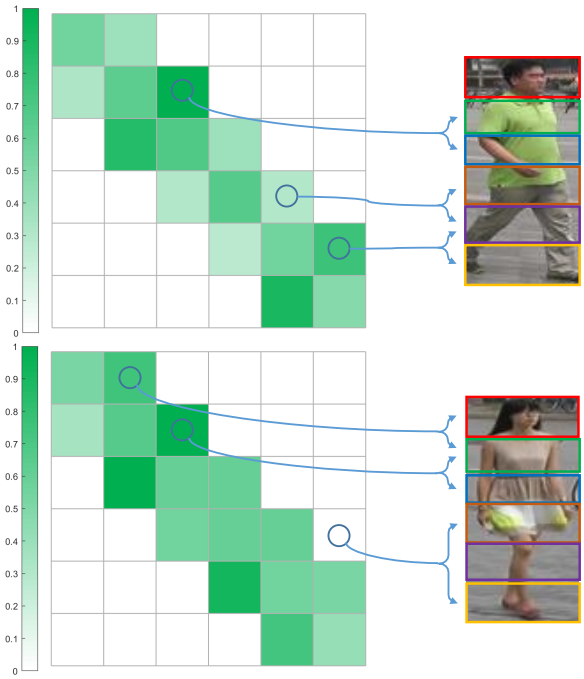


Figure 8: Visualization of the adjacency matrix of intra-local graph. The deeper color indicates larger value.

Rank-1 accuracy and mAP accuracy are further improved after adding the intra-local graph (without FDM). It is because the contextual information of pedestrian is learned using the intra-local relationship between adjacent parts.

The effectiveness of FDM. There is consistently an improvement on both Rank-1 and mAP accuracy after adding FDM which is utilized to optimize the adjacency matrix of intra-local graph. From Table 2 we can see that "+FDM" achieves 91.1% (+2.7%) in Rank-1 accuracy and 85.2% (+3.4%) in mAP accuracy on DukeMTMC-reID. It is because the proposed FDM parameterizes the non-zero elements in the adjacency matrix of intra-local graph to differentiate the correlations between adjacent parts. We visualize the adjacency matrix of intra-local graph to intuitively evaluate the effectiveness of the intra-local graph and FDM. In Fig. 8, we show the adjacency matrices of two pedestrian images where we normalize the elements in the two adjacency matrices to $[0, 1]$. From

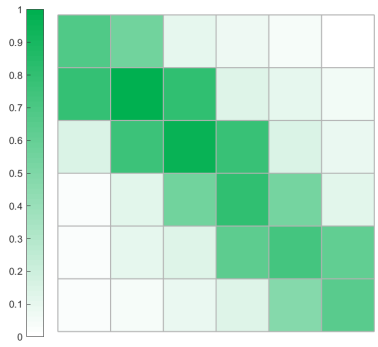


Figure 9: Visualization of the adjacency matrix learned by FDM (all learnable). The deeper color indicates larger value.

the figure we can see that adjacent parts possess different correlations while non-adjacent parts possess no correlation.

In addition, we optimize all the elements in the adjacency matrix of the intra-local graph and then optimize the elements corresponding to the non-adjacent parts to near 0, which is denoted as +FDM (all learnable). Specifically, as for the implementation of FDM (all learnable), we first optimize all elements before 40 epochs, and then we minimize the L_2 loss to constrain the weight learning of non-adjacent parts. In this way, we can not ensure that the elements corresponding to non-adjacent parts are zero, but near zero, as shown in Fig. 9. There are some non-zero elements of non-adjacent parts, because in the optimization process the cross-entropy loss and the L_2 loss should be balanced. Compared with FDM (all learnable), FDM is simple and more effective with less parameters. Experimental results in Table 2 show that +FDM (all learnable) weakens the performance of PGCN. It may introduce some interference and noise information from non-adjacent parts, and therefore we directly fix these elements into zero values.

The effectiveness of keypoints. After replacing keypoints with the uniform partition, the performance on both Rank-1 and mAP accuracy is weakened on the four datasets as shown in Table 2. It is because the uniform partition is easily affected by posture changes which are prone to causing part misalignment.

Table 3: Comparison with GCN-based methods on Market-1501, CUHK03 and DukeMTMC-reID.

Method	Market-1501		CUHK03		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
DGDNet [47]	83.6	63.3	-	-	-	-
SGGNN [34]	92.3	82.8	-	-	81.1	68.2
Deep CRF [35]	93.5	81.6	-	-	84.9	69.5
PH-GCN [48]	93.5	79.0	64.9	61.5	85.0	70.7
PGCN	**98.0**	**94.8**	**86.7**	**83.6**	**91.1**	**85.2**

4.4. Comparison with GCN-based Methods

In order to comprehensively evaluate the performance of PGCN, we conduct experiments on Market-1501, CUHK03 and DukeMTMC-reID to compare PGCN with other GCN-based methods. The experimental results are listed in Table 3 where we can see that PGCN achieves the best results on these datasets. Specifically, PGCN obtains the highest accuracy on DukeMTMC-reID with 91.1% in Rank-1 accuracy and 85.2% in mAP accuracy, which exceeds the best compared PH-GCN [48] with +6.1% in Rank-1 accuracy and +14.5% in mAP accuracy, respectively. It is because PH-GCN only considers the intra-local relationship between different parts within single pedestrian image, while PGCN takes two types of local relationship, i.e., inter-local and intra-local relationship, into consideration so that the learned features contain the information from the same parts of other pedestrian images and the contextual information of pedestrian. It should be noticed that these compared GCN-based methods do not conduct experiments on the largest dataset MSMT17 while the proposed PGCN obtains competitive performance on MSMT17 as shown in Table 4.

Table 4: Comparison with state-of-the-art methods on the four datasets. Both Rank-1 and mAP accuracy are listed.

Method	Market-1501		CUHK03		DukeMTMC-reID		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
PABR [12]	91.7	79.6	-	-	84.4	69.3	-	-
PCB [11]	92.3	77.4	61.3	54.2	81.8	66.1	-	-
PCB+RPP [11]	93.8	81.6	63.7	57.5	83.3	69.3	-	-
AlignedReID++ [49]	91.8	79.1	61.5	59.6	82.1	69.7	69.8	43.7
SFT [50]	94.1	87.5	74.3	71.7	90.0	79.6	79.0	58.3
Auto-ReID [15]	94.5	85.1	73.3	69.3	-	-	78.2	52.5
DG-Net [51]	94.8	86.0	-	-	86.6	74.8	77.2	52.3
ABD-Net [52]	95.6	88.3	-	-	89.0	78.6	82.3	60.8
Pyramid [16]	95.7	88.2	78.9	74.8	89.0	79.0	-	-
Deep-Person [4]	92.3	79.6	-	-	80.9	64.8	-	-
SCSN [53]	92.4	88.3	84.7	81.0	91.0	79.0	83.8	58.5
RGA-SC [54]	96.4	88.4	79.6	74.5	-	-	80.3	57.5
PGCN	**98.0**	**94.8**	**86.7**	**83.6**	**91.1**	**85.2**	**87.7**	**72.7**
$PGCN^*$	**98.2**	**95.1**	**86.9**	**83.9**	**91.6**	**85.7**	**88.1**	**73.2**

4.5. Comparison with State-of-the-art Methods

We perform extensive experiments on the four datasets to compare PGCN with state-of-the-art methods. The experimental results are shown in Table 4 where we can see that PGCN obtains the highest Rank-1 accuracy and mAP accuracy on all four datasets, and surpasses state-of-the-art methods by an overwhelming margin. Specifically, PGCN exceeds the second best methods RGA-SC [54], SCSN [53], SFT [50] and ABD-Net [15] by +6.4%, +2.6%, +5.6% and +11.9% in mAP accuracy on Market-1501, CUHK03, DukeMTMC-reID and MSMT17, respectively. We argue that the proposed PGCN not only utilizes CNN to extract appearance features, but also uses GCN to learn the relationship between local features from two different angles for information transfer.

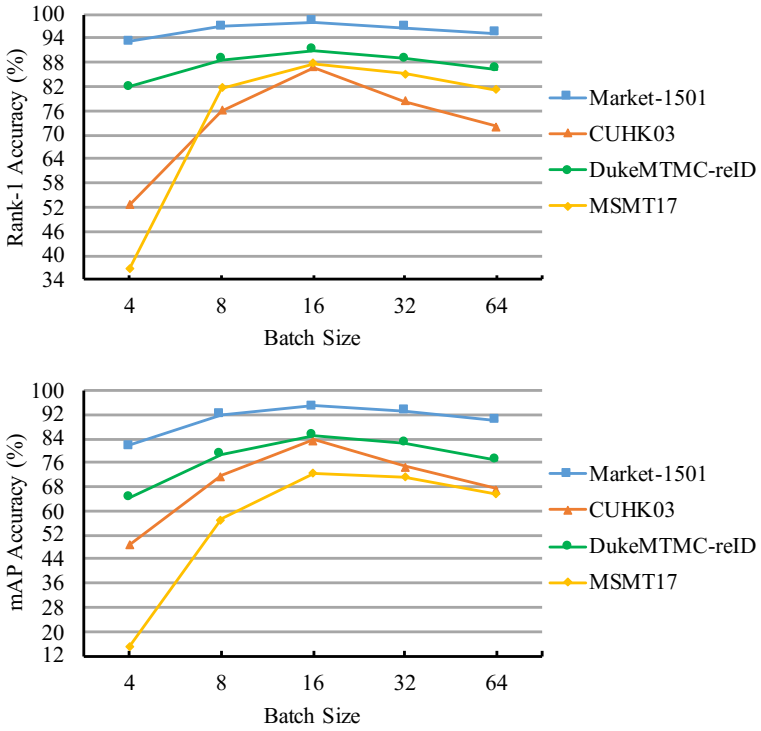


Figure 10: The influence of the batch size on the four datasets. Both Rank-1 (%) accuracy and mAP (%) accuracy are listed.

In addition, our method is independent to the *baseline*, and we combine the proposed method with the AGW baseline [55] (denoted as $PGCN^*$). From the table, we can see that the experimental results are improved, which proves that AGW is a strong baseline.

4.6. Parameter Analysis

There are three key hyper-parameters in PGCN, i.e., the batch size, the number of graph convolution layers, and the neuron number of FC layers. We evaluate the influence of these parameters on the four datasets.

Batch size. The batch size affect the performance of PGCN, because it determines the size of adjacency matrix S_p in the inter-local graph. Fig. 10 shows the performance of PGCN under different batch size. From Fig. 10 we

Table 5: The influence of the number of graph convolution layers on Market-1501, CUHK03, DukeMTMC-reID and MSMT17, respectively. n_1 and n_2 denote the number of graph convolution layers on the inter-local graph and the intra-local graph, respectively. Both Rank-1 and mAP accuracy are listed.

n_1	n_2	Market-1501		CUHK03		DukeMTMC-reID		MSMT17	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1	1	**98.0**	**94.8**	**86.7**	**83.6**	**91.1**	**85.2**	**87.7**	**72.7**
1	2	94.2	86.8	66.0	61.6	86.1	75.8	72.8	49.1
1	3	68.9	49.9	33.6	30.5	52.6	31.6	60.8	36.2
2	1	93.5	87.4	71.7	68.9	80.4	72.1	77.3	60.3
2	2	80.4	65.2	49.4	45.9	62.5	51.6	57.0	36.7
2	3	66.1	53.4	38.1	35.1	46.7	35.8	38.4	19.4
3	1	86.3	77.1	64.1	61.9	66.6	56.3	55.5	38.6
3	2	67.8	51.9	38.5	36.2	36.2	27.4	27.2	14.9
3	3	53.5	38.9	29.6	28.1	28.8	20.2	25.2	11.4

can see that as the batch size increases, the performance of PGCN gradually improves until the batch size is equal to 16. It is because that the model cannot aggregate enough local information from other parts when setting the small batch size. When batch size is larger than 16, the performance of PGCN gradually drops. It may absorb in noise and inference information when setting the large batch size. Hence, we set the batch size to 16 on the four datasets.

The number of graph convolution layers. It is usually beneficial for performance improvement to stack more convolutional layers in CNNs, but is it also the deeper the better for GCNs in PGCN? We apply different number of graph convolution layers on the inter-local graph and the intra-local graph to test the performance. From Table 5 we can see that with the increase of graph convolution layers, the performance of PGCN becomes worse. PGCN achieves the best results when using one graph convolution layer on both the inter-local graph and the intra-local graph. After processed by each graph convolution layer, each node feature aggregates and absorbs the information from other

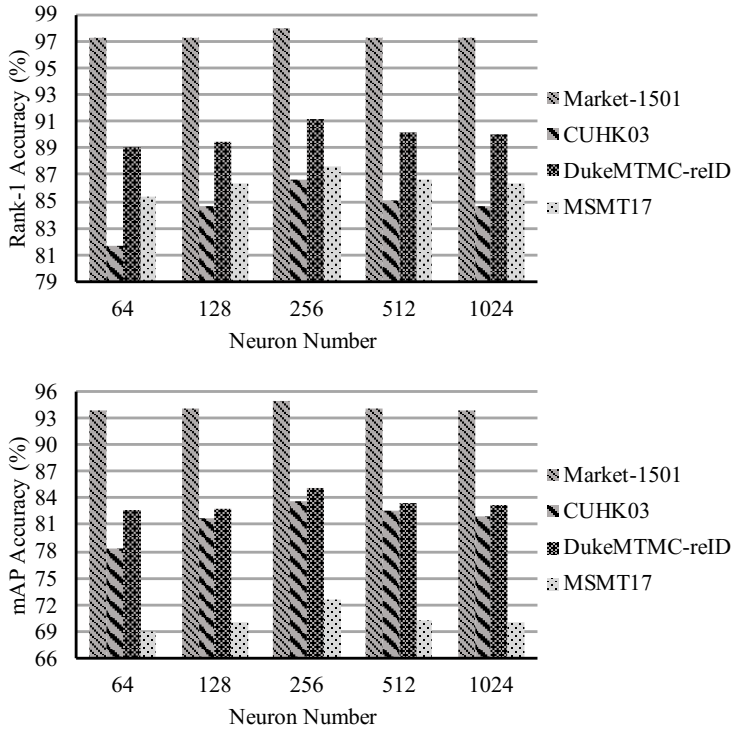


Figure 11: The influence of neuron number of FC layers. We conduct experiments on the four datasets. Both Rank-1 (%) accuracy and mAP (%) accuracy are listed.

node features. Hence, when stacking multiple graph convolution layers, the features of the nodes may converge to similar values, which is liable to result in model over-smoothing. And, the over-smoothing hurts the performance of person Re-ID.

The neuron number of FC layers. After the concatenation operation, we apply the independent FC layers to reduce the dimensionality of features. We vary the neuron number of FC layers from 64 to 1024. The detailed results are shown in Fig. 11 where PGCN achieves the best performance when the neuron number is set to 256.

5. Conclusion

In this paper, we have proposed PGCN to model the inter-local relationship and the intra-local relationship for person Re-ID. Meanwhile, we propose FDM to accurately describe the intra-local relationship by learning the correlations among adjacent parts of pedestrian. In this way, we could aggregate local information from corresponding parts of different pedestrian images and adjacent parts of single pedestrian image to enhance the representation capacity of local features. Experimental results on the four datasets have proved the excellent performance of PGCN and the importance of learning local relationships. Although the proposed PGCN achieves the excellent performance, it can not learn completed relationship among local features. Hence, in the future work, we will study how to inject the relationship among non-corresponding parts of different pedestrian images into the inter-local relationship and how to model the structure information among the parts of single pedestrian image when learning the intra-local relationship.

Acknowledgment

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.

[1] X. Y. Jing, X. Zhu, F. Wu, R. Hu, X. You, Y. Wang, H. Feng, J. Y. Yang, Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning, IEEE Transactions on Image Processing 26 (3) (2017) 13631378.

[2] L. He, J. Liang, H. Li, Z. Sun, Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, p. 70737082.

[3] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, N. Zheng, Semi-supervised person re-identification using multi-view clustering, Pattern Recognition 88 (2019) 285–297.

[4] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-person: Learning discriminative deep features for person re-identification, Pattern Recognition 98 (2020) 107036.

[5] Z. He, S. Yi, Y. Cheung, X. You, Y. Tang, Robust object tracking via key patch sparse representation, IEEE Transactions on Cybernetics 47 (2) (2017) 354–364.

[6] M. Cai, F. Lu, Y. Gao, Desktop action recognition from first-person point-of-view, IEEE Transactions on Cybernetics 49 (5) (2019) 1616–1628.

[7] C. Wang, B. Ma, H. Chang, S. Shan, X. Chen, Tcts: A task-consistent two-stage framework for person search, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 11952–11961.

[8] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian, T. Xiang, Deep transfer learning for person re-identification, in: IEEE International Conference on Multimedia Big Data, 2018, pp. 1–5.

[9] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, L. Lin, Scan: Self-and-collaborative attention network for video person re-identification, IEEE Transactions on Image Processing 28 (10) (2019) 4870–4882.

[10] Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Transactions on Circuits and Systems for Video Technology 29 (10) (2019) 3037–3045.

[11] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: European Conference on Computer Vision, 2018, pp. 480–496.

[12] Y. Suh, J. Wang, S. Tang, T. Mei, K. Mu Lee, Part-aligned bilinear representations for person re-identification, in: European Conference on Computer Vision, 2018, pp. 402–419.

[13] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, Glad: Global-local-alignment descriptor for pedestrian retrieval, in: ACM International Conference on Multimedia, 2017, pp. 420–428.

[14] J. Guo, Y. Yuan, L. Huang, C. Zhang, J. Yao, K. Han, Beyond human parts: Dual part-aligned representations for person re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 3641–3650.

[15] R. Quan, X. Dong, Y. Wu, L. Zhu, Y. Yang, Auto-reid: Searching for a part-aware convnet for person re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 3750–3759.

[16] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, R. Ji, Pyramidal person re-identification via multi-loss dynamic training, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8514–8522.

[17] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: IEEE International Conference on Computer Vision, 2017, pp. 3960–3969.

[18] M. M. Kalayeh, E. Basaran, M. Gkmen, M. E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1062–1071.

[19] M. S. Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 420–429.

[20] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: International Joint Conference on Artificial Intelligence, 2017, pp. 2194–2200.

[21] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.

[22] Z. Zhang, H. Zhang, S. Liu, Coarse-fine convolutional neural network for person re-identification in camera sensor networks, IEEE Access 7 (2019) 65186–65194.

[23] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks (2016). arXiv:1609.02907.

[24] Z. Wang, L. Zheng, Y. Li, S. Wang, Linkage based face clustering via graph convolution network, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1117–1125.

[25] Z. Chen, X. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5177–5186.

[26] K. Liu, L. Gao, N. M. Khan, L. Qi, L. Guan, A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action recognition, IEEE Transactions on Multimediadoi:10.1109/TMM.2020.2974323.

[27] H. Huang, W. Yang, X. Chen, X. Zhao, K. Huang, J. Lin, G. Huang, D. Du, Eanet: Enhancing alignment for cross-domain person re-identification (2018). arXiv:1812.11369.

[28] M. Ye, P. C. Yuen, Purifynet: A robust person re-identification model with noisy labels, IEEE Transactions on Information Forensics and Security 15 (2020) 2655–2666.

[29] M. Ye, X. Lan, Q. Leng, J. Shen, Cross-modality person re-identification via modality-aware collaborative ensemble learning, IEEE Transactions on Image Processing 29 (2020) 9387–9399.

[30] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems (2020) 1–21.

[31] R. Li, S. Wang, F. Zhu, J. Huang, Adaptive graph convolutional neural networks, in: AAAI Conference on Artificial Intelligence, 2018, pp. 3546–3553.

[32] Y. C. Ng, N. Colombo, R. Silva, Bayesian semi-supervised learning with graph gaussian processes, in: Neural Information Processing Systems, 2018, pp. 1683–1694.

[33] S. Kumar, J. Ying, J. V. de Miranda Cardoso, D. Palomar, Structured graph learning via laplacian spectral constraints, in: Neural Information Processing Systems, 2019, pp. 11647–11658.

[34] Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, in: European Conference on Computer Vision, 2018, pp. 486–504.

[35] D. Chen, D. Xu, H. Li, N. Sebe, X. Wang, Group consistent similarity learning via deep crf for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8649–8658.

[36] J. Yang, W. Zheng, Q. Yang, Y. Chen, Q. Tian, Spatial-temporal graph convolutional network for video-based person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3289–3299.

[37] M. Ye, J. Shen, D. J. Crandall, L. Shao, J. Luo, Dynamic dual-attentive aggregation learning for visible-infrared person re-identification, in: European Conference on Computer Vision, 2020, pp. 229–247.

[38] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: European Conference on Computer Vision, 2018, pp. 466–481.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[40] M. Ye, J. Shen, X. Zhang, P. C. Yuen, S. F. Chang, Augmentation invariant and instance spreading feature for softmax embedding, IEEE Transactions on Pattern Analysis and Machine Intelligencedoi:10.1109/TPAMI.2020.3013379.

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

[42] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

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

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

[45] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[46] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, 2014, pp. 740–755.

[47] D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann, N. Zheng, Deep feature learning via structured graph laplacian embedding for person re-identification, Pattern Recognition 82 (2018) 94–104.

[48] B. Jiang, X. Wang, B. Luo, Ph-gcn: Person re-identification with part-based hierarchical graph convolutional network (2019). arXiv:1907.08822.

[49] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang, Alignedreid++: Dynamically matching local information for person re-identification, Pattern Recognition 94 (2019) 53–61.

[50] C. Luo, Y. Chen, N. Wang, Z. Zhang, Spectral feature transformation for person re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 4976–4985.

[51] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2138–2147.

[52] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, Z. Wang, Abd-net: Attentive but diverse person re-identification, in: IEEE International Conference on Computer Vision, 2019, pp. 8351–8361.

[53] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, Y. Yang, Salience-guided cascaded suppression network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3300–3310.

[54] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3186–3195.

[55] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. H. Hoi, Deep learning for person re-identification: A survey and outlook, arXiv preprint arXiv:2001.04193.