PH-GCN: Person Retrieval with Part-based Hierarchical Graph Convolutional Network

Bo Jiang, Xixi Wang, Aihua Zheng, Jin Tang, and Bin Luo

***Abstract*—Compact feature representation of person image is important for person re-identification (Re-ID) task. Recently, part-based representation models have been widely studied for extracting the more compact and robust feature representation for person image to improve person Re-ID results. However, existing part-based representation models mostly extract the features of different parts independently which ignore the spatial relationship information among different parts. To address this issue, in this paper we propose a novel deep learning framework, named Part-based Hierarchical Graph Convolutional Network (PH-GCN) for person Re-ID problem. Given a person image, PH-GCN first constructs a hierarchical graph to represent the spatial relationships among different parts. Then, both local and global feature learning is achieved by the feature information passing in PH-GCN, which takes the information of other parts into account for part feature representation. Finally, a perceptron layer is adopted for the final person part label prediction and re-identification. The proposed framework provides a general solution that integrates *local*, *global* and *structural* feature learning simultaneously in a unified end-to-end network representation and learning. Extensive experiments on several widely used benchmark datasets demonstrate the effectiveness and benefits of the proposed PH-GCN approach for person Re-ID task.**

***Index Terms*—Person Re-identification, Graph Convolutional Network, Part-based Representation, Hierarchical Graph.**

I. Introduction

PERSON retrieval, also known as person re-identification (Re-ID) is an active research problem in computer vision, which aims to study how to re-identify a query person from a set of images taken by multiple cameras [1], [2], [3], [4], [5], [6]. With the development of deep learning, many of existing Re-ID methods adopt a person classification framework to determine the label of an input person image by using a classifier trained on the training samples [7], [8], [9], [10], [11], [12]. Although recent years have witnessed rapid advancements in person Re-ID, it is still a challenging task partly due to large changes of person appearance caused by variety of factors, such as pose, illumination, deformation and occlusion, etc.

One main issue for Re-ID is to develop a compact and robust feature representation for person image. Recently, part-based methods have been widely studied and verified beneficially to person Re-ID task [13], [2], [5], [14], [15], [16], [6]. These methods generally conduct feature representation on part-level and thus can extract both local and global representations for person image. In particular, deeply-learned features have been verified stronger discriminative ability, especially when aggregated from deeply-learned part features [2], [13], [5], [16], [14]. For example, Zhao *et al.* [13] develop a human part-aligned representation by detecting the human body regions (parts), computing and aggregating the similarities between the corresponding parts for Re-ID. Sun *et al.* [5] propose a Part-based Convolutional Baseline (PCB) and a refined part pooling (RPP) method for enhancing the consistency within each part. Wei *et al.* [15] propose Global-Local-Alignment descriptor (GLAD) to leverage both local and global cues in the human body. Zheng *et al.* [16] design a new coarse-fine pyramid model to conduct local and global representation.

However, the above existing part-based Re-ID models generally extract the feature of each person part independently which thus fails to consider the inherent spatial (or geometrical) relationships with other parts. We think this relationship information with other parts also provides an important discriminative feature for person image. Therefore, to overcome the limitation, our aim in this paper is trying generate the context-aware feature representation for each part of person image. Recently, Graph Convolutional Networks (GCNs) have drawn increasing attention in machine learning and computer vision area due to their abilities to generalize neural networks for graph data [17], [18], [19], [20], [21], [22], [23]. GCNs aim to propagate messages on a graph structure. After message passing on the graph, the final node representations are obtained from their own as well as the information of their neighboring nodes, which thus can naturally incorporate the contextual information for graph node.

Motivated by these, in this paper we propose a novel Part-based Hierarchical Graph Convolutional Network (PH-GCN) for person image representation and Re-ID task. PH-GCN aims to learn a **context-aware representation** for each person part that incorporates the geometrical structure information among parts while maintains the unary appearance feature of each part. PH-GCN exploits the inherent relationships of parts effectively, and thus performs robustly to part noises and/or corruptions.

Overall, the main contributions of this paper are summarized as follows.

- We propose a novel deeply-learned and context-aware part feature extraction and learning model for person Re-ID.

This research is supported in part by the Key Project of Research and Development of Anhui Province (201904b11020037), NSFC Key Projects of International (Regional) Cooperation and Exchanges under Grant (61860206004); the National Natural Science Foundation of China (62076004, 61976002).

B. Jiang, X. Wang, J. Tang and B. Luo are with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology of Anhui University, Hefei, 230601, China.

A. Zheng is with Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, AnHui University, Hefei, 230601, China.

Corresponding author: Aihua Zheng

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3095789, IEEE Transactions on Multimedia

- We propose a novel Part-based Hierarchical Graph Convolutional Network (PH-GCN) learning architecture for object representation. The proposed network provides a general solution that integrates *local*, *global* and *structural* feature representation and learning simultaneously in a unified network.

Extensive experiments on several benchmark datasets demonstrate the effectiveness and benefits of the proposed PH-GCN method. The remainder of this paper is organized as follows. In section II, we briefly review some related works on person re-identification and Graph Convolutional Networks (GCNs). We present the detail of PH-GCN in section III. In section IV, we implement PH-GCN on several benchmarks to demonstrate the effectiveness of the proposed model. In section V, we conclude our paper and future work.

II. Related Work

A. Part-based Person Re-identification

With the development of deep learning, many methods have been proposed for person Re-ID tasks [6], [3], [24], [25], [26], [27]. In this section, we briefly review some recent related works that are also devoted to generating deeply-learned *part* features for the Re-ID problem. For instance, Ustinova $et\ al.$ [27] propose a network architecture to learn a more effective embedding by performing bilinear pooling. Si $et\ al.$ [3] propose to develop Dual Attention Matching network (DuATM) to learn context-aware feature sequences. Su $et\ al.$ [2] propose Pose-driven Deep Convolutional (PDC) model, which aims to utilize the human part cues to alleviate the pose variations and thus learn robust features from both global image and local parts. Zhao $et\ al.$ [13] propose a human part-aligned representation by detecting the human body regions and computing between the corresponding parts. Sun $et\ al.$ [5] propose a strong convolution baseline method to further leverage a uniform partition strategy and learn a more compact representation for person Re-ID. To further incorporate the global information, Wei $et\ al.$ [15] propose Global-Local-Alignment Descriptor (GLAD) that explicitly leverages the local and global cues in the human body to generate a discriminative and robust representation. Wang $et\ al.$ [14] develop a feature learning strategy to integrate discriminative information by combining global and local information in different granularities. Zheng $et\ al.$ [16] propose a new coarse-fine pyramid model to conduct local and global representation simultaneously. Li $et\ al.$ [28] propose a tree branch network (TBN) to investigate joint learning global and local features.

However, the above existing part-based Re-ID models mainly extract the features of different person parts independently or concatenate parts to encode some relations. The main differences between former methods and our proposed method are two points: First, we propose to construct a hierarchical graph model to encode the relationships among different parts. Comparing with the concatenation model, the proposed hierarchical graph explicitly encodes both spatial and hierarchical relationships together via graph (weighted) edges. Second, based on the constructed graph, we can thus employ a powerful learning tool (Graph Convolutional Networks [18], [21], [20], [19]) to learn the context-aware representation of each part to obtain more robust representation.

B. Graph Convolutional Network

Recently, Graph Convolutional Networks (GCNs) have been demonstrated effectively for graph structure data representation and learning in machine learning area [18], [19], [20], [21]. For example, Bruna $et\ al.$ [29] propose a CNN-like neural architecture on graphs in Fourier domain. Kipf and Welling [18] propose a simple Graph Convolutional Network (GCN) based on the first-order approximation of spectral filters. Atwood and Towsley [30] propose Diffusion-Convolutional Neural Networks (DCNNs). Monti $et\ al.$ [31] present mixture model CNNs (MoNet) to generalize CNN architecture on graphs. Veličković $et\ al.$ [32] present Graph Attention Networks (GAT) by further designing an edge attention layer. Jiang $et\ al.$ [33] propose Graph Learning-Convolutional Networks (GLCNs) for graph representation and semi-supervised learning.

In addition, GCNs (or GNNs) have also been employed in computer vision tasks in recent years [34], [35], [36], [21], [37], [38], [39], [40], [8]. For example, Qi $et\ al.$ [34] propose a 3D graph neural network model for RGB-D semantic segmentation. Knyazev $et\ al.$ [38] propose to explore multi-edge GCN for image classification. Gao $et\ al.$ [39] propose a novel Graph Convolution Tracking (GCT) method to jointly learn the structured representation of historical target exemplars and target localization for visual tracking. Michelle $et\ al.$ [36] develop neural graph matching networks for few-shot 3D action recognition. Yan $et\ al.$ [21] propose Spaptial Temporal Graph Convolutional Network (ST-GCN) for skeleton-based action recognition. Recently, some works [7], [8] use graph convolution network models for person re-identification. For example, Shen $et\ al.$ [7] propose Deep Similarity-Guided Graph Neural Network (SGGNN) to model relationships between probe-gallery image pairs. Yan $et\ al.$ [8] propose to employ the GCN model for person search of the complex scene. Yang $et\ al.$ [41] propose Spatial-Temporal Graph Convolutional Network (STGCN) to model the temporal relations from adjacent frames and the spatial relations in each frame for video person re-identification.

III. PH-GCN Model

In this section, we propose our Part-based Hierarchical Graph Convolutional Network (PH-GCN) for person image representation and re-identification.

A. Overview

Fig. 1 shows the overall framework of our PH-GCN which mainly contains four modules, i.e., 1) CNN based part feature extraction, 2) part-based hierarchical graph construction, 3) graph convolutional module and 4) perceptron layer.

- **CNN based part feature extraction:** We utilize a deep CNN network module to extract the appearance feature for each part of a person image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3095789, IEEE Transactions on Multimedia

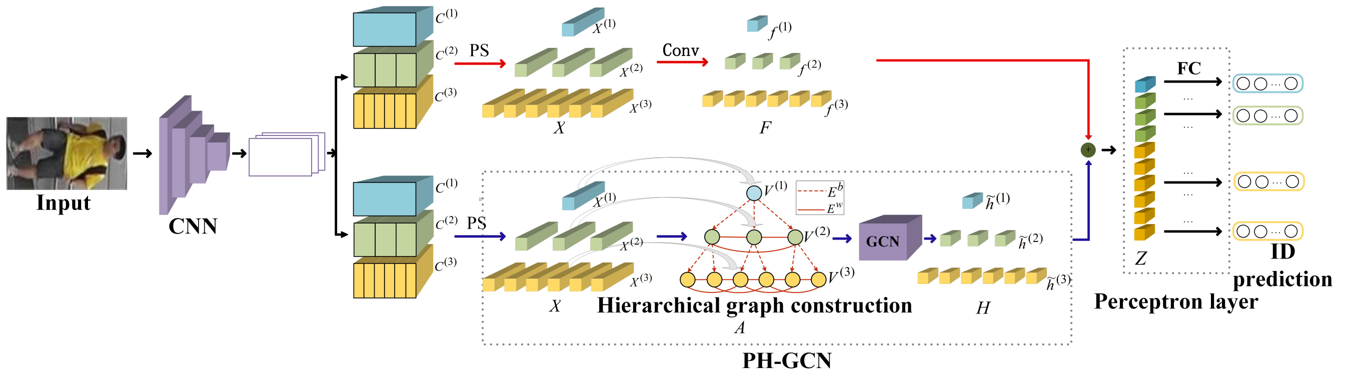


Fig. 1: Architecture of the proposed PH-GCN network for person Re-ID. For input image I, we first use ResNet50 [42] network slightly modified by removing global average pooling (GAP) layer and fully-connected (FC) layer and then set the stride of conv4_1 to 1. Then we employ pooling strategy (PS) to spatially down-sample the feature descriptor to obtain column vectors (parts) X. Next, we construct a hierarchical part graph and use multi-layer GCN (purple cube) to learn the structural feature H, one-layer orthodox convolutional (Conv) to learn visual feature F. Finally, we make ID prediction through the perceptron layer. More detail can see Section III. Best viewed in color.

- **Part-based hierarchical graph construction:** A hierarchical structural graph is constructed to encode/represent the spatial relationships among different person parts.
- **Graph convolutional module:** We employ a graph convolutional network (GCN) architecture to extract the context-aware representations for person parts.
- **Perceptron layer:** A perceptron layer is employed for the final person ID prediction.

In the following, we present the details of each module in our PH-GCN network, respectively.

B. CNN Based Part Feature Extraction

For each person image I, we first extract convolutional feature descriptor through slightly modifying ResNet50 [42] network pre-trained on ImageNet [43]. Specifically, we remove global average pooling (GAP) layer and fully-connected (FC) layer and set the stride of conv4_1 to 1. Then, we copy the feature descriptor three times to conduct three uniform partitions (layers), respectively. Taking the p-th partition $\mathcal{C}^{(p)}$ as an example, we adopt pooling strategy to spatially down-sample the feature descriptor of the p-th partition into n_p pieces of column vectors (parts), where $X^{(p)} = (x_1^p, x_2^p, \cdots, x_i^p, \cdots, x_{n_p}^p)$ denotes the feature collection for different parts of the p-th partition of image I in the following sections. Finally, as shown in the red solid line of Fig. 1, we leverage one-layer orthodox convolutional (Conv) to reduce the dimension to obtain the visual feature of each part in the p-th partition level, which can be formulated as

$$f_i^p = Conv(x_i^p), \quad i = 1, 2, ..., n_p \tag{1}$$

where $f_i^p \in \mathbb{R}^{d\times 1}$, x_i^p denotes the feature descriptor of the i-th part in p-th partition level. d is the feature dimension of each part and n_p is the number of parts in the p-th partition level. Thus, the ultimate visual feature presentation of the p-th partition level can be denoted as $F^{(p)} = (f_1^p, f_2^p, \cdots, f_{n_p}^p) \in \mathbb{R}^{d\times n_p}$. We empirically set p = 1, 2, 3 (the superscripts will not repeat the description unless necessary).

C. Hierarchical Part Graph Construction

Based on the above hierarchical partitions (layers), we then construct a hierarchical graph $G = (V, E)$ to define the spatial and appearance relationships among different parts. In particular, we construct a three-layer graph whose nodes and edges are introduced below.

Nodes. A three-layer hierarchical graph $G = (V, E)$ is constructed, where $V = \{V^{(1)}, V^{(2)}, V^{(3)}\}$ with each $V^{(p)}$ corresponding to the partition level $\mathcal{C}^{(p)}$. Each node $v_i^{(p)} \in V^{(p)}$ corresponds to a specific part which is assigned with a CNN based feature vector x_i^p, and there exist more nodes on the higher layers. Obviously, in multi-layer hierarchical graph representation, the higher layer contains more local information while the lower layer encodes more global representation for input person image I. We empirically set $\{|V_3|, |V_2|, |V_1|\}$ to $\{6, 3, 1\}$ respectively in all the experiments, as shown in Fig. 1 in detail.

Edges. Let $E = \{E^w, E^b\}$ be the edge set, where E^w denotes the edges within each layer and E^b denotes the edges existing between different layers in our hierarchical graph $G = (V, E)$. Specifically, in each intra-layer, an edge $e_{ij} \in E^w$ exists between node v_i^p and v_j^p if they are either neighbor or they have common neighboring nodes. For different layers, an edge $e_{ij}^{pq}, p \leq q$ exists between node v_i^p and v_j^q if the i-th part in p-layer involves the j-th part in q-layer. Finally, we compute the edge weight A_{ij}^{pq} for each edge as

$$A_{ij}^{pq} = \exp\big(-\frac{\|x_i^p - x_j^q\|_2}{\delta}\big) \tag{2}$$

where δ is a hyper-parameter.

D. Graph Convolutional Representation

As an extension of CNNs from the regular grid to irregular graph, Graph Convolutional Networks (GCNs) [20], [18] have been widely studied for graph data representation and learning. Our GCN representation aims to extract a contextual and compact representation for each person part by exploring the representation information of its neighboring parts, which thus can exploit the more discriminative structure information for person Re-ID. Our GCN module contains several convolutional hidden layers that take a feature map matrix $H^{(t)} \in \mathbb{R}^{N\times d_t}$ as the input and output a feature map $H^{(t+1)} \in \mathbb{R}^{N\times d_{t+1}}$ by using a graph convolution operator. In general, we set $d_{t+1} \leq d_t$, and thus the convolution operation also provides a kind of low-dimensional representation for each graph node.

Formally, let $X = [X^{(1)} \| X^{(2)} \| X^{(3)}]$ be the concatenation of $X^{(p)}$, where $X^{(p)} = (x^p_1, x^p_2 \cdots x^p_{n_p})$ denotes the extracted CNN feature vector collection of all parts. Let A be the whole adjacency matrix of the above hierarchical graph $G(X, A)$, i.e., A has a form as

$$A = \begin{pmatrix} A^{11} & A^{12} & A^{13} \\ A^{21} & A^{22} & A^{23} \\ A^{31} & A^{32} & A^{33} \end{pmatrix} \tag{3}$$

where A^{pq} is defined in Eq.(1). Formally, given an input feature matrix $X = H^{(0)} \in \mathbb{R}^{N\times d_0}$ and hierarchical graph $A \in \mathbb{R}^{N\times N}$. Similar to GCN [18], we propose to conduct the following layer-wise propagation as

$$H^{(t+1)} = \sigma\big[(\epsilon \tilde{A} H^{(t)} + (1-\epsilon) H^{(t)})\Theta^{(t)}\big] \tag{4}$$

where $t = 0, 1 \cdots T-1$ and $\sigma(\cdot)$ denotes an activation function. In this paper, we set $T = 2$. We define it as $\sigma(\cdot) = \mathrm{ReLU}(\cdot) = \max(0, \cdot)$. Parameters $\Theta = \{\Theta^{(0)}, \Theta^{(2)} \cdots \Theta^{(T-1)}\}$ denote the trainable weight matrices and $A = AD^{-1}$ (D is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$) denotes the row-normalization of adjacency matrix A [18]. Parameter $\epsilon \in (0, 1)$ denotes the fraction of feature information that nodes receive from their neighbors.

E. Perceptron Layer

In the final perceptron layer, we combine the visual appearance information and the structure information together and then adopt a full connection layer (FC) to predict part ID. For simplicity, we denote the final output feature map as $\tilde{H} = H^{(T)} = \{\tilde{h}^{(1)}_1, \tilde{h}^{(2)}_1, \tilde{h}^{(2)}_2, \tilde{h}^{(2)}_3, \tilde{h}^{(3)}_1 \cdots \tilde{h}^{(3)}_6\}$. Let $F = \{f^{(1)}_1, f^{(2)}_1, f^{(2)}_2, f^{(2)}_3, f^{(3)}_1 \cdots f^{(3)}_6\}$ be the appearance feature extracted by CNN. First, we combine H and F together into Z as,

$$z^{(p)}_i = f^{(p)}_i + \beta \tilde{h}^{(p)}_i \tag{5}$$

where β is a balancing hyper-parameter and $z^{(p)}_i$ is a part component of Z. Then, for each part $z^{(p)}_i$, we adopt a Fully connected (FC) layer to predict the ID label of the corresponding person.

Loss Function. For each region (part) of the person image, we train a specific classifier by using the cross-entropy loss function $\mathcal{L}^{(p)}_i$[61]. The final overall loss function is designed as the aggregation of them,

$$\mathcal{L} = -\frac{1}{N} \sum_{p=1}^{3} \sum_{i=1}^{n_p} \mathcal{L}^{(p)}_i \tag{6}$$

where N is the total number of parts, p and i denote the p-th partition level and the i-th part, respectively. n_p indicates the number of parts in the p-th partition level.

IV. Experiments

To verify the effectiveness of the proposed PH-GCN Re-ID method, we conduct experiments on three benchmarks including Market-1501 [44], DukeMTMC-reID [62] and CUHK03 [65], [66]. We compare our PH-GCN with some recent related state-of-art methods, including attention-based methods (HA-CNN [25], MGCAM-Siamese [56]), part-based methods (VPM [6], PCB [5], TBN [28], AANet [64], MGN [14]) and graph-based methods (SGGNN [7], M-GAT [58]). Finally, we implement our method with two versions, i.e., PH-GCN and PH-GCN+RR. PH-GCN+RR further uses the re-ranking [65] approach to improve the Re-ID results.

A. Datasets and Settings

Market1501 [44] dataset consists of 1501 persons obtained from six camera viewpoints including five high-resolution cameras and one low-resolution camera. It contains 19,732 gallery images and 12,936 training images which are all detected by DPM [70].

DukeMTMC-reID [62] dataset is a subset of DukeMTMC dataset [62], which contains 1812 identities observed from 8 different camera viewpoints, where 1404 identities appear in more than two cameras with more than 500 occluded identities. It mainly contains 16522 training images, 2228 queries and 17661 gallery images.

CUHK03 [65], [66] dataset contains 13164 images with 1,467 identities. Each identity is observed from two cameras. It contains two kinds of bounding boxes (hand-labeled, DPM-detected) and we use both ways to validate our method in experiments. We adopt the new training/testing protocol proposed in work [65] on this dataset.

Evaluation Metrics. Following many previous works [5], [65], we use the measurement Cumulative Matching Characteristic (CMC), eg. rank-1, rank-5 and rank-10, and mean Average Precision (mAP) for evaluation, where mAP denotes the mean value of average precision across all queries.

B. Implementation Details

As shown in Fig. 1, we use ResNet50 [42] pre-trained on ImageNet [43] as our backbone network to extract a convolutional feature map for two-stream network, respectively. Then, we use the one-layer orthodox convolution and multi-layer graph convolutional network respectively to process the deeply learned part features. Thus, the output feature dimension in the perceptron layer is set to 256. We implement our model with PyTorch and training the network on two NVIDIA TITAN XP GPUs 12G in an end-to-end manner. All input images are adjusted to a resolution of 384×128, which is the same as in PCB [5]. Data augmentation is also adopted for training with horizontal flip, normalization and random erase as suggested in work [5]. We set the total number of epochs to 90 and the batch size is set as 80 in general for all datasets. We initialize the learning rate of backbone network to 0.1 and set the learning rate of GCN to 0.01 and other layers of network to 1.0. The learning rate is not fixed and begins to decay after 40 epochs until the convergence of learning. We train the whole

TABLE I: Comparisons results(%) on Market-1501[44] dataset on metrics mAP, Rank-1, Rank-5 and Rank-10. ”RR” denotes re-ranking [45] operation for refining person Re-ID performance.

Methods	Reference	mAP	Rank-1	Rank-5	Rank-10
SVDNet[46]	ICCV2017	62.1	82.3	92.3	95.2
HydraPlus[47]	ICCV2017	-	76.9	91.3	94.5
PAR[13]	ICCV2017	63.4	81.0	92.0	94.7
PDC* [2]	ICCV2017	63.4	84.1	92.7	94.9
MultiLoss[48]	IJCAI2017	64.4	83.9	-	-
SSM[49]	CVPR2017	68.8	82.2	-	-
DPFL$^{(2+)}$ [50]	CHI2017	73.1	88.9	-	-
GLAD(*)[15]	MM2017	73.9	89.9	-	-
BraidNet-CS+SRL[51]	CVPR2018	69.4	83.7	-	-
Pose-transfer[52]	CVPR2018	57.9	79.7	-	-
IDE+CamStyle[53]	CVPR2018	68.7	88.1	95.1	97.0
PSE [54]	CVPR2018	69.0	87.7	-	-
SafeNet[55]	IJCAI2018	72.7	90.2	-	-
MGCAM-Siamese[56]	CVPR2018	74.3	83.7	-	-
HA-CNN[25]	CVPR2018	75.7	91.2	-	-
PCB[5]	ECCV2018	77.3	92.4	97.0	97.9
MGN[14]	MM2018	86.9	95.7	-	-
SGGNN[7]	ECCV2018	82.8	92.3	96.1	97.4
PAAN[57]	IEEE2019	64.4	81.7	-	-
MGAT[58]	CVPRW2019	76.5	91.5	97.2	98.0
APDR[59]	PR2019	80.1	93.1	97.2	98.2
TBN[28]	ICME2019	79.9	92.9	-	-
VPM[6]	CVPR2019	80.8	93.0	97.8	98.8
PH-GCN	**Ours**	**81.4**	**93.7**	**97.8**	**98.6**
Rerank[60]	CVPR2017	63.6	77.1	-	-
IDE+CamStyle(RR)[53]	CVPR2018	86.0	91.5	95.0	96.3
PSE(RR) [54]	CVPR2018	84.0	90.3	-	-
MGN(RR) [14]	MM2018	94.2	96.6	-	-
APDR(RR)[59]	PR2019	90.2	94.4	97.0	97.9
PH-GCN(RR)	**Ours**	**92.1**	**94.6**	**96.9**	**97.7**

network by using stochastic gradient descent (SGD) [71] in each mini-batch. During testing, we concatenate all the part features for each query image to generate its final feature representation. The inference time is 0.043s for each image.

C. Comparison with the Related Works

Table I-III summarize the comparison results on Market-1501 [44], DukeMTMC-reID [62] and CUHK03 [65], [66] datasets, respectively. The result of all comparison methods have been reported in their papers. Here, we use them directly. Most of the comparison methods use ResNet50 [42] for fair comparison including PCB [5] , SGGNN [7], MGAT [58], TBN[28], VPM [6] and so on. Overall, PH-GCN generally obtains competitive results on these benchmarks. More concretely, we can observe the following,

Results on Market1501 dataset: In Table I, we can observe that the performance of our proposed PH-GCN is improved by +4.1% in mAP and +1.3% in rank-1 than the baseline method Part-based Convolution Baseline (PCB) [5]. Compared to other part-based methods, such as TBN [28] and VPM [6], we also have the better accuracy. PH-GCN works slightly overshadowed than MGN with more triplet losses. Compared to other graph-based methods, PH-GCN outperforms SGGNN [7] by +1.4% in rank-1 and +1.7% rank-5 and MGAT[58] by +2.2% in rank-1 and +0.6% in rank-5, respectively. This clearly demonstrates the effectiveness of PH-GCN by further exploiting the structural information of parts via GCN learning. In addition, PH-GCN performs better than some recent Re-ID approaches, which demonstrates the effectiveness of the proposed Re-ID approach. The comparison results are listed in Table I.

Results on DukeMTMC-reID dataset: Comparing with PCB [5], our proposed PH-GCN has +7.2% and +3.3% improvements on mAP and rank-1, respectively. PH-GCN performance better TBN [28] by 1.0% in mAP and has the same mAP as AANet [64] that is assisted by person attributes. Compared with graph-based method SGGNN [7], our result improves +4.3% and +4.1% on mAP and rank-1, respectively. It further demonstrates the effectiveness of our PH-GCN based representation and learning. In addition, PH-GCN outperforms some recent Re-ID approaches, which demonstrates the benefit of the proposed Re-ID approach. The comparison results are listed in Table II.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3095789, IEEE Transactions on Multimedia

TABLE II: Comparisons results(%) on DukeMCMT-reID [62] dataset on metrics mAP, Rank-1, Rank-5 and Rank-10. ”RR” denotes re-ranking [45] operation for refining person Re-ID performance.

Methods	Reference	mAP	Rank-1	Rank-5	Rank-10
SVDNet[46]	ICCV2017	56.8	76.7	-	-
GAN[62]	ICCV2017	47.1	67.7	-	-
PSE [54]	CVPR2018	62.0	79.8	-	-
IDE+CamStyle[53]	CVPR2018	53.4	75.2	84.6	87.9
Pose-transfer[52]	CVPR2018	56.4	78.5	-	-
SafeNet[55]	IJCAI2018	57.0	82.7	-	-
BraidNet-CS+SRL[51]	CVPR2018	59.4	76.4	-	-
Inception-V1+OpenPose[63]	ECCV2018	64.2	82.1	90.2	92.7
HA-CNN[25]	CVPR2018	63.8	80.5	-	-
PCB [5]	ECCV2018	65.3	81.9	89.4	91.6
SGGNN[7]	ECCV2018	68.2	81.1	88.4	91.2
MGN[14]	MM2018	78.4	88.7	-	-
AANet[64]	CVPR2019	72.5	86.4	-	-
TBN[28]	ICME2019	71.5	85.2	-	-
VPM[6]	CVPR2019	72.6	83.6	91.7	94.2
APDR[59]	PR2019	69.7	84.3	92.4	94.7
PH-GCN	**Ours**	**72.5**	**85.2**	**92.5**	**94.9**
PSE(RR) [54]	CVPR2018	79.8	85.2	-	-
IDE+CamStyle(RR)[53]	CVPR2018	73.3	81.4	88.1	90.8
Inception-V1+OpenPose(RR)[63]	ECCV2018	83.9	88.3	93.1	95.0
AANet-152(RR)[64]	CVPR2019	86.8	90.3	-	-
APDR(RR)[59]	PR2019	83.2	87.3	93.0	95.2
PH-GCN(RR)	**Ours**	**86.8**	**89.3**	**93.9**	**96.1**

TABLE III: Comparisons results(%) on CUHK03 [65], [66] dataset used the new protocol. ”RR” denotes re-ranking [45] operation for refining person Re-ID performance.

Method	Reference	Labeled		Detected	
		Rank-1	mAP	Rank-1	mAP
TriNet[67]	arXiv2017	-	-	55.5	50.7
MultiScale[50]	ICCVW2017	-	-	40.7	37.0
PAN[68]	TCSVT2018	36.9	35.0	36.3	34
MLFN[69]	CVPR2018	54.7	49.2	52.8	47.8
MGCAM-Siamese[56]	CVPR2018	50.1	50.2	46.7	46.9
HA-CNN[25]	CVPR2018	44.4	41.0	41.7	38.6
PCB[5]	ECCV2018	57.6	56.5	61.3	54.2
MGN[14]	MM2018	68.0	67.4	66.8	66.0
TBN[28]	ICME2019	-	-	59.1	54.5
PH-GCN	**Ours**	**65.6**	**62.0**	**61.8**	**58.5**
PH-GCN(RR)	**Ours**	**71.7**	**74.1**	**66.5**	**70.3**

TABLE IV: Result(%) of PH-GCN with different parts number on three datasets. ”1+2” indicates hierarchical graph with two layer, which contains 1 part and 2 parts respectively. The rest can be explained in a similar way.

	Market-1501		DukeMTMC-reID		CUHK03	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
1+2	64.2	85.7	56.9	77.0	34.5	36.7
1+3	73.3	90.9	66.2	82.7	47.8	49.4
1+4	77.5	91.9	68.8	83.8	54.2	56.9
1+3+6	81.4	93.7	72.5	85.2	62.0	65.6
1+4+8	78.5	91.6	70.9	84.1	48.2	47.5
1+3+6+8	76.0	90.9	66.5	79.7	50.3	51.1
1+3+6+12	71.8	89.0	62.1	77.4	43.2	43.4

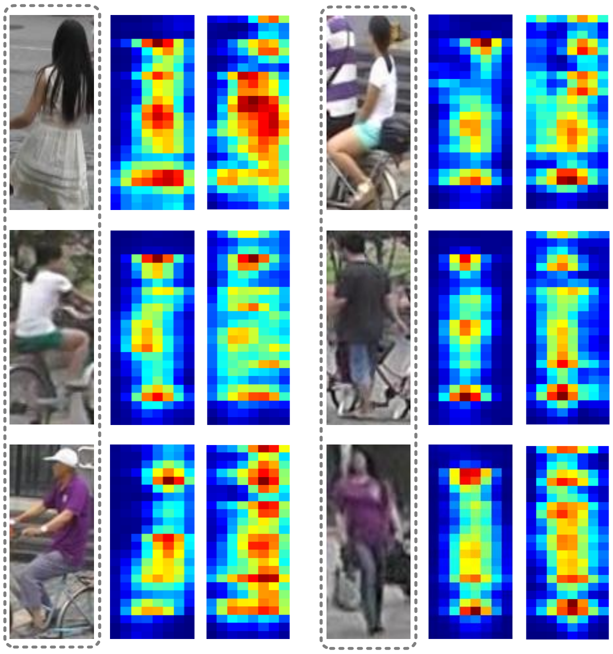


Fig. 2: Illustration of the example feature maps on Market-1501 [44]. The first column of each image shows the feature map learned by baseline. The second column shows the feature map learned by our proposed PH-GCN.

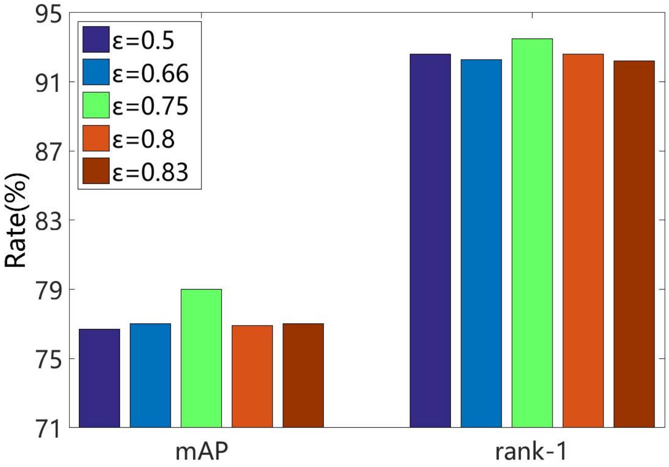


Fig. 3: Results of PH-GCN ϵ on Market-1501 [44] dataset.

Results on CUHK03 dataset: This dataset is a challenging dataset under the new protocol [65]. Here we use pedestrian boxes annotated by two methods, which are denoted as cuhk03-labeled dataset and cuhk03-detected dataset respectively. On the cuhk03-labeled dataset, PH-GCN obtains +5.5% and +8.0% improvements on mAP and rank-1 respectively when comparing with related method PCB [5]. On the cuhk03-detected dataset, Compared to TBN [28], PH-GCN improves +4.0% in mAP and +2.7% in rank-1. This further demonstrates the robustness and effectiveness of PH-GCN on person image representation and thus recognition.

Qualitative Visualization: Fig. 2 shows the example feature maps learned by baseline and PH-GCN on Market-1501 [44], respectively. Intuitively, we can observe that our proposed PH-GCN can learn more detail information and enhance the discriminative abilities for person images.

D. Parameter Analysis

Balances Parameters: The proposed PH-GCN model has two main parameters, i.e., ϵ in Eq.(4) and β in Eq.(5). ϵ balances the feature representation of node itself and that received from its neighbors. β balances the visual appearance information and the structure information. We empirically set $\epsilon = 0.75$ and $\beta = 0.3$. The proposed model is relatively insensitive to these parameters when we slightly adjust the parameters. The final Re-ID results only change a little on Market-1501 [44] as shown in Fig. 3 and Fig. 4. For example, in Fig. 4 when β changes from 0.1 to 0.5, the final performance of our method changes slightly. It can obtain the best performance when $\beta = 0.3$. This phenomenon demonstrates the insensitivity of the proposed model w.r.t parameter β in range (0.1, 0.5). However, if β continuously increases, we find that the performance of our method begins to decline significantly when β exceeds 0.5, as shown in Fig. 4.

Number of Parts: In order to analyze the effect of different part numbers in our proposed PH-GCN, we conduct experiments across different part numbers on three datasets, as shown in Table IV. We can observe that the performance is improving as the part number increasing. However, when the number of parts is larger than 8, the performance declines dramatically because over partition may lose some discriminative information for each part.

E. Ablation Study

Effectiveness of PH-GCN Structure: To further understand and verify the core components (GCN module, hierarchical graph construction) of our PH-GCN model, we conduct ablation analysis experiments on three datasets. On CUHK03, we select cuhk03-labeled dataset for evaluation. First, to verify the effectiveness of GCN module, we implement a special variant of our model, i.e., Ours-NoGCN that removes GCN module in our PH-GCN network. Second, to demonstrate the benefit of hierarchical graph, we implement a special variant of our model with a single layer graph (denoted as P-GCN). As a baseline, we also report the results of PCB [5]. Fig. 5 summarizes the comparison results. Here, one can observe that (1) Both PH-GCN and P-GCN obtain better performance than PCB, which demonstrates the effectiveness of the proposed PH-GCN (or P-GCN) by incorporating the inherent spatial relationship information among different parts. (2) PH-GCN performs better than P-GCN, which demonstrates the benefit of hierarchical graph by capturing the structural information of both local and global cues. (3) PH-GCN obviously outperforms Ours-NoGCN, which further shows the desired advantage of the proposed deeply learned context-aware part representation in the Re-ID task.

Number of Layers in Hierarchical Graph: Fig. 7 further shows the performance of PH-GCN across different layers in our hierarchical graph on three datasets. We implement two special variants of our model with single layer graph

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3095789, IEEE Transactions on Multimedia

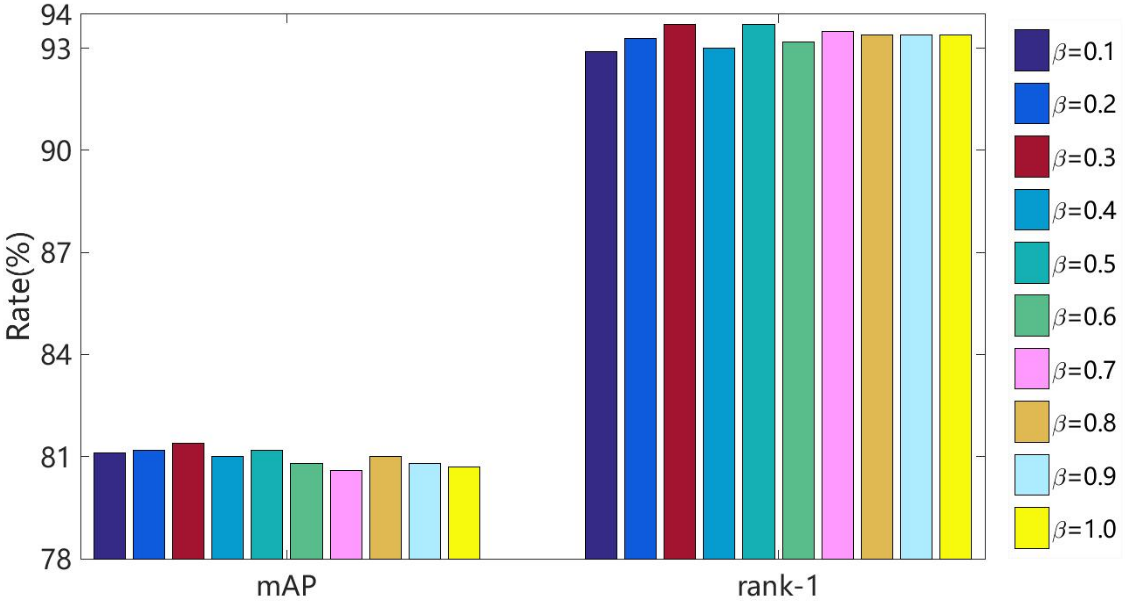


Fig. 4: Results of PH-GCN with β on Market-1501 [44] dataset.

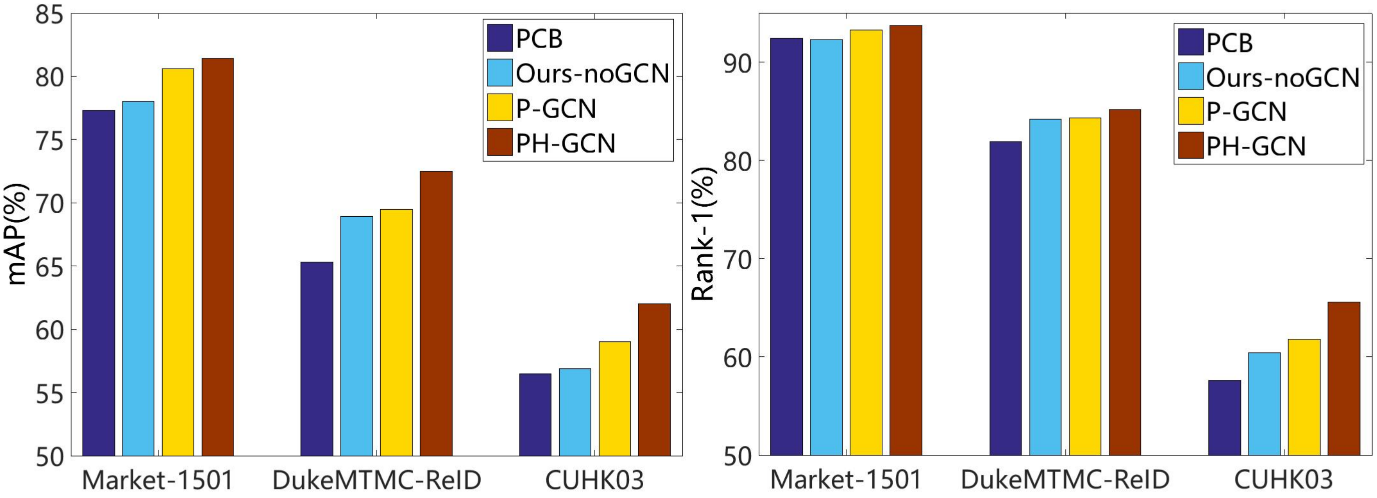


Fig. 5: Ablation study on three datasets, where $'Ours\text{-}NoGCN'$ denotes our PH-GCN network without GCN module and $'P\text{-}GCN'$ represents single layer graph with six-part.

(denoted as 1-layer) that contains six parts and two-layer graph (denoted as 2-layer) that contains six parts and three parts, respectively. Here, we can note that PH-GCN with three-layer graph performs better than 1-layer and 2-layer graphs, which indicates the effectiveness of the proposed hierarchical graph construction by integrating both local and global information together for person image representation.

Analysis of Different Baseline: We implement a new part-based model which uses a hierarchical partition for part-based ReID as used in HPM [72]. We implement it by using the similar model and parameter setting as our PH-GCN and denote it as HPM*. We leverage it to replace PCB [5] as our new baseline. Then we conduct experiments on three public datasets, as shown in Fig 6. We can find that our proposed PH-GCN can also improve the performance well for the new baseline HPM*.

V. Conclusion

Compact feature representation of person image is important for person Re-ID task. This paper proposes a novel Part-based Hierarchical Graph Convolutional Network (PH-GCN) which aims to learn a hierarchical context-aware part feature representation for person image representation and Re-ID problem. PH-GCN also provides a general solution for object (e.g., person) representation and recognition that integrates *local*, *global* and *structural* feature learning simultaneously in a unified end-to-end network. Extensive experiments on three commonly used datasets demonstrate the effectiveness and benefits of the proposed PH-GCN method. In our future, we will employ some more optimal graph convolutional network, such as Graph Attention Networks (GATs), for part-based person representation and Re-ID problem.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3095789, IEEE Transactions on Multimedia

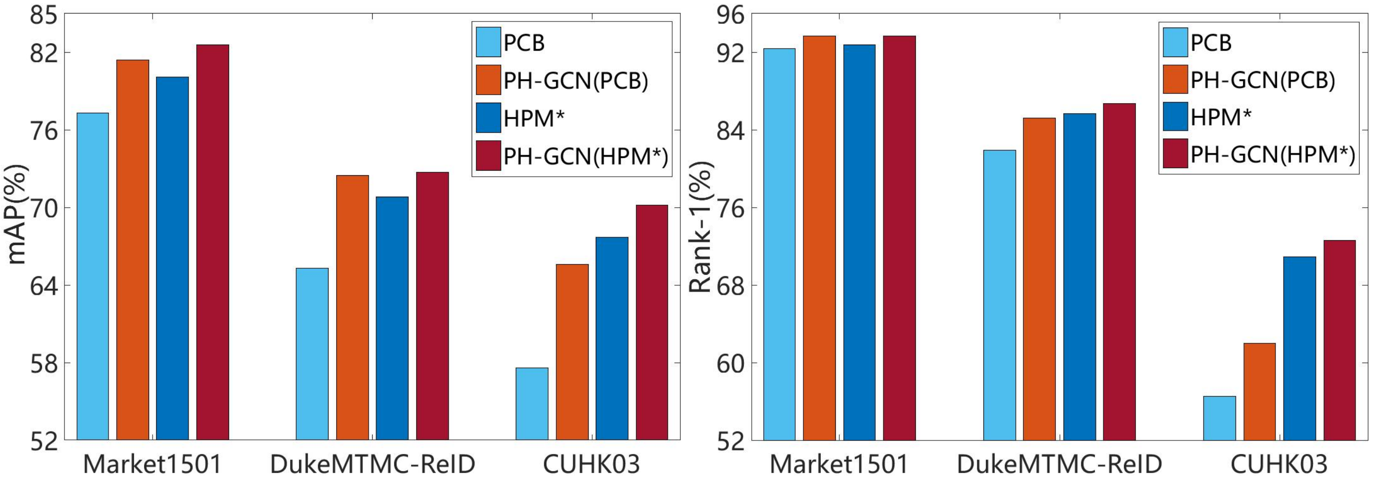


Fig. 6: Ablation study with different baseline on three datasets, where $'PH\text{-}GCN(PCB)'$ denotes the baseline is PCB [5] and $'PH\text{-}GCN(HPM)'$ represents HPM* as the baseline.

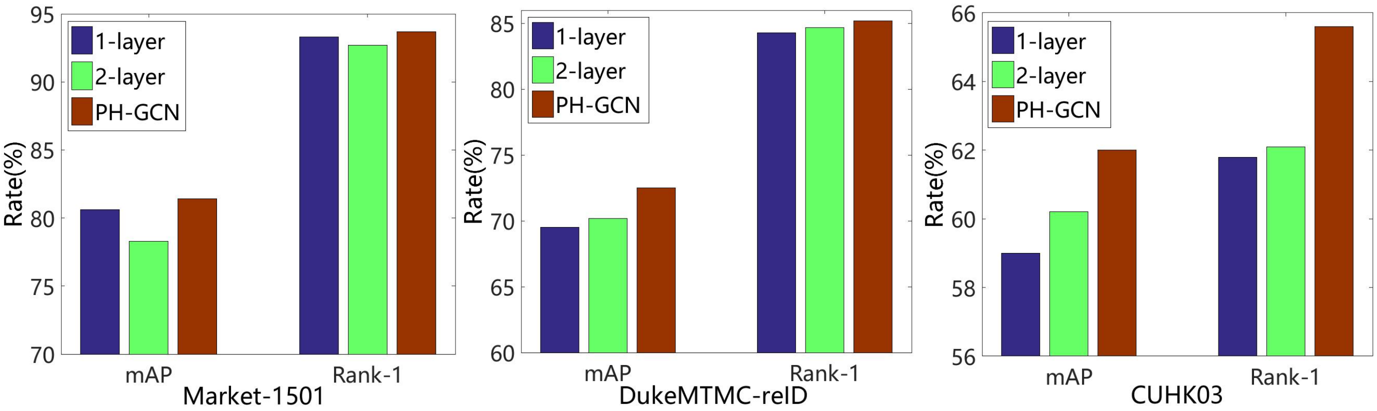


Fig. 7: Ablation study for the Part-based hierarchical graph construction on Market-1501 [44], DukeMTMC-reID [62] and CUHK03 [65], [66] datasets. $'1\text{-}layer'$ denotes single layer hierarchical graph with six part. $'2\text{-}layer'$ is hierarchical graph with two layer, which include three part and six part in our PH-GCN module, respectively.

REFERENCES

[1] W. Zheng, H. Ruimin, L. Chao, Y. Yi, J. Junjun, Y. Mang, C. Jun, and L. Qingming, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, pp. 260–272, 2016.

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

[3] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5363–5372.

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

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

[6] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 393–402.

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

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

[9] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia*, vol. 21, pp. 2376–2386, 2019.

[10] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, vol. 21, pp. 1412–1424, 2018.

[11] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Feature affinity-based pseudo labeling for semi-supervised person re-identification," *IEEE Transactions on Multimedia*, vol. 21, pp. 2891–2902, Nov 2019.

[12] C. Wan, Y. Wu, X. Tian, J. Huang, and X. Hua, "Concentrated local part discovery with fine-grained part representation for person re-identification," *IEEE Transactions on Multimedia*, vol. 22, pp. 1605–1618, 2020.

[13] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3219–3228.

[14] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM International Conference on Multimedia*,

2018, pp. 274–282.

[15] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 420–428.

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

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

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

[19] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 2014–2023.

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

[21] Y. X. Sijie Yan and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 7444–7452.

[22] F. Hu, Y. Zhu, S. Wu, L. Wang, and T. Tan, "Hierarchical graph convolutional networks for semi-supervised node classification," *ArXiv preprint arXiv:1902.06667*, 2019.

[23] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," *ArXiv preprint arXiv:1908.07325*, 2019.

[24] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2265–2274.

[25] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.

[26] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1169–1178.

[27] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2017, pp. 1–6.

[28] H. Li, M. Yang, Z. Lai, W. Zhen, and Z. Yu, "Pedestrian re-identification based on tree branch network with local and global learning," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2019, pp. 694–699.

[29] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *Proceedings of the International Conference on Learning Representations*, 2014.

[30] J. Atwood and D. Towsley, "Diffusion-convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 1993–2001.

[31] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5423–5434.

[32] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *ArXiv preprint arXiv:1710.10903*, 2017.

[33] B. Jiang, Z. Zhang, D. Lin, J. Tang, and B. Luo, "Semi-supervised learning with graph learning-convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 313–11 320.

[34] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5199–5208.

[35] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *ArXiv preprint arXiv:1711.04043*, 2017.

[36] M. Guo, E. Chou, D.-A. Huang, S. Song, S. Yeung, and L. Fei-Fei, "Neural graph matching networks for fewshot 3d action recognition," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 653–669.

[37] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 401–417.

[38] B. Knyazev, X. Lin, M. R. Amer, and G. W. Taylor, "Image classification with hierarchical multigraph networks," *ArXiv preprint arXiv:1907.09000*, 2019.

[39] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4649–4659.

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

[41] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3286–3296.

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

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. Mar, pp. 211–252, 2015.

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

[45] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3652–3661.

[46] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3800–3808.

[47] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[48] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *ArXiv preprint arXiv:1705.04724*, 2017.

[49] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2530–2539.

[50] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2590–2600.

[51] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1470–1478.

[52] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4099–4108.

[53] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157–5166.

[54] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 420–429.

[55] K. Yuan, Q. Zhang, C. Huang, S. Xiang, C. Pan, and H. Robotics, "Safenet: Scale-normalization and anchor-based feature extraction network for person re-identification." in *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2018, pp. 1121–1127.

[56] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[57] Y. Zhang, X. Gu, J. Tang, K. Cheng, and S. Tan, "Part-based attribute-aware network for person re-identification," *IEEE Access*, vol. PP, Apr.

[58] L. Bao, B. Ma, H. Chang, and X. Chen, "Masked graph attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[59] S. Li, H. Yu, W. Huang, and J. Zhang, "Attributes-aided part detection and refinement for person re-identification," *ArXiv preprint arXiv:1902.10528*, 2019.

[60] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[61] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, no. Jul, pp. 19–67, 2005.

[62] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[63] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 418–437.

[64] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7134–7143.

[65] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.

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

[67] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv preprint arXiv:1703.07737*, 2017.

[68] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. Jul, pp. 1–1, 2018.

[69] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2109–2118.

[70] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. Feb, pp. 1627–1645, 2010.

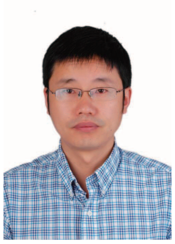
[71] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, 2010, pp. 177–186.

[72] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. S. Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Feb 2019, pp. 8295–8302.

Bo Jiang received the B.S. degrees in mathematics and applied mathematics and the M.Eng. and Ph.D. degree in computer science from AnHui University of China in 2009, 2012, and 2015, respectively. He is currently an associate professor in computer science at Anhui University. His current research interests include image feature extraction and matching, data representation and learning.

Xixi Wang is currently a Ph.D student in computer science at AnHui University. Her current research interests include person re-identification and multi-model learning.

AiHua Zheng received her B.Eng. degrees and finished her Master-Doctor combined program in computer science and technology from AnHui University of China in 2006 and 2008 respectively. And received her Ph.D degree in computer science from the University of Greenwich of UK in 2012. She is currently an associated professor in computer science at AnHui University. Her main research interests include computer vision and artificial inteligent, especially on person/vehicle re-identification, audio-visual learning and multi-modal and cross-modal learning.



Jin Tang received the B.Eng. degree in automation in 1999, and the Ph.D. degree in computer science in 2007 from AnHui University, Hefei, China. Since 2012, he has been a professor at the School of Computer Science and Technology at the AnHui University. His research interests include image processing, pattern recognition and computer vision.

Bin Luo received his B.Eng. degree in electronics and M.Eng. degree in computer science from AnHui University of China in 1984 and 1991, respectively. In 2002, he was awarded the Ph.D. degree in Computer Science from the University of York, the United Kingdom. He has published more than 200 papers in journal and refereed conferences. He is a professor at AnHui University of China. At present, he chairs the IEEE Hefei Subsection. He served as a peer reviewer of international academic journals such as IEEE Trans. on PAMI, Pattern Recognition, Pattern Recognition Letters, etc. His current research interests include random graph based pattern recognition, image and graph matching, spectral analysis.