Person Re-identification using Heterogeneous Local Graph Attention Networks

Zhong Zhang[1], Haijia Zhang[1], Shuang Liu[1]*
[1]Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission,
Tianjin Normal University, China
{zhong.zhang8848, haijia27zhang, shuangliu.tjnu}@gmai1.com

Abstract

Recently, some methods have focused on learning local relation among parts of pedestrian images for person re-identification (Re-ID), as it offers powerful representation capabilities. However, they only provide the intra-local relation among parts within single pedestrian image and ignore the inter-local relation among parts from different images, which results in incomplete local relation information. In this paper, we propose a novel deep graph model named Heterogeneous Local Graph Attention Networks (HLGAT) to model the inter-local relation and the intra-local relation in the completed local graph, simultaneously. Specifically, we first construct the completed local graph using local features, and we resort to the attention mechanism to aggregate the local features in the learning process of inter-local relation and intra-local relation so as to emphasize the importance of different local features. As for the inter-local relation, we propose the attention regularization loss to constrain the attention weights based on the identities of local features in order to describe the inter-local relation accurately. As for the intra-local relation, we propose to inject the contextual information into the attention weights to consider structure information. Extensive experiments on Market-1501, CUHK03, DukeMTMC-reID and MSMT17 demonstrate that the proposed HLGAT outperforms the state-of-the-art methods.

1. Introduction

Person re-identification (Re-ID) [5, 22, 43, 51] aims to match the given person of interest in different scenarios. It is a challenging task due to various complex factors, such as illumination, body poses and viewpoints.

Learning discriminative and robust features is the main issue for person Re-ID. With the prosperity of Convolutional Neural Network (CNN), researchers learn deep features to improve the performance of person Re-ID. Some methods [2, 4, 7, 53] design various deep models to extract global features from entire pedestrian images. In order to mine local information from different body parts, some part-based methods [13, 27, 34, 42] are proposed to extract local features. Among them, direct partition strategy [1, 6, 21, 36, 44] is usually adopted to split feature maps or pedestrian images into several horizontal grids. To locate accurate and meaningful body parts, some researchers [10,16,29,31,48] resort to human pose estimation or human paring techniques to process pedestrian images.

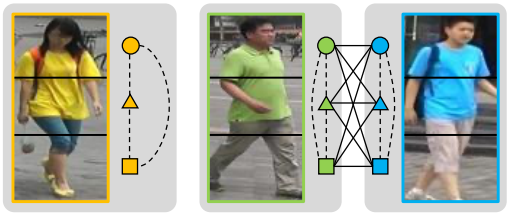

Figure 1. (a) The existing methods focus on learning the intra-local relation among parts within single pedestrian image. (b) The proposed HLGAT constructs the completed local graph to model the inter-local relation and the intra-local relation, simultaneously. The solid line indicates the inter-local edge, and the dotted line denotes the intra-local edge.

Recently, local relation information is exploited in several studies [12, 14, 26], where they compute relation maps based on the local similarity among feature maps, or aggregate local features to learn relation information from different body parts. However, they only focus on the intra-local relation among parts within single pedestrian image as shown in Fig. 1(a), and ignore the inter-local relation among parts from different pedestrian images. This results in a lack of local relation information from these parts and weakens the representation capabilities of local features. Hence, it is important to take the inter-local relation into consideration, and properly integrate with the intra-local relation for person Re-ID.

In this paper, we propose a novel deep graph model

*Corresponding author.

named Heterogeneous Local Graph Attention Networks (HLGAT) for person Re-ID, where we construct a completed local graph to model the inter-local relation and the intra-local relation, simultaneously. To this end, we regard the local features extracted from pedestrian images as the nodes of completed local graph, and design two types of edges to link the nodes, i.e., inter-local edge and intra-local edge as shown in Fig. 1(b). Concretely, we link the nodes belonging to different pedestrian images using the inter-local edges, and link the nodes belonging to the same pedestrian image using the intra-local edges. In order to jointly learn the inter-local relation and the intra-local relation, we utilize the attention mechanism to aggregate the local features linked by the inter-local edges and the intra-local edges in the completed local graph simultaneously so as to enhance the representation capabilities of local features.

As for the inter-local relation, if the local features belong to the same identity, they possess high correlation and therefore the attention weights should be large, otherwise small. Correspondingly, we propose the attention regularization loss to constrain the attention weights in order to describe the inter-local relation accurately. Furthermore, we link the local features from the corresponding and adjacent parts of different pedestrian images using the inter-local edges, and differentiate them via the attention weights in the aggregation process. As for the intra-local relation, we observe that the relation strength between two local features from the same pedestrian image enhances with the decrease of their spatial distance. Hence, we propose to inject the contextual information into the attention weights to consider structure information. Finally, we extract the aggregated local features from the completed local graph. In the test stage, we concatenate the aggregated local features to obtain discriminative features for person Re-ID.

Our contributions are summarized as follows: (1) We propose HLGAT to model the inter-local relation and the intra-local relation by constructing the completed local graph. (2) We propose the attention regularization loss to constrain the attention weights of inter-local edges to describe the inter-local relation accurately, and meanwhile we inject the contextual information into the attention weights of intra-local edges to provide structure information of pedestrian. (3) Experimental results on four large-scale person Re-ID datasets including Market-1501, CUHK03, DukeMTMC-reID and MSMT17 prove that the proposed HLGAT exceeds state-of-the-art methods.

2. Related Work

2.1. Part-based methods for person Re-ID

In order to mine local knowledge from different body parts, some researchers [10, 13, 34, 36, 39] design various part-based models. Direct partition is adopted in [24,34,44], where images or feature maps are divided into horizontal stripes. Yi *et al*. [44] propose Deep Metric Learning (DML) to partition pedestrian images into three parts, and then feed them into siamese convolutional neural network (SCNN) to learn local features. Sun *et al*. [34] utilize CNNs to extract feature maps, and then divide them into several fixed size grids so as to obtain local features from each grid independently. Fu *et al*. [9] design Horizontal Pyramid Matching (HPM) to partition feature maps into multi-scale horizontal regions, and then pool them with global average pooling and global max pooling to extract local features from multiple scales.

Different from direct partition strategy, some researchers resort to human parsing techniques or human pose estimation to locate accurate and meaningful body parts so as to alleviate parts misalignment. In EANet [13], pedestrian images are firstly processed by human pose detector so as to obtain body keypoints, and then divided into several parts with these keypoints. Dual Part-Aligned Block (DPB) [10] is composed by the human part branch which generates the human part masks to obtain the human part-aligned representation, and the latent part branch where the self-attention scheme is utilized to compute the latent part-aligned representation.

2.2. Relation learning for person Re-ID

It is important to exploit relation information among different body parts for the enhancement of representation ability, and therefore some studies [12, 26, 30, 47] are proposed to inject relation information into deep features. These methods mainly focus on learning pairwise relation [30], local relation [12, 26] and global relation [47]. Among them, Shen *et al*. propose Similarity-Guided Graph Neural Network (SGGNN) [30] to model the pairwise relation among different probe-gallery pairs so as to learn the probe-gallery relation features. Hou *et al*. [12] design the Spatial Interaction-and-Aggregation (SIA) module for person Re-ID, which aggregates the correlated spatial features to consider the local relation and the multi-scale appearance relations. In addition, Park and Ham [26] build the one-vs.-rest relation module in Relation-Net to exploit the relation between each part and the rest parts, and they incorporate the local information from other body parts to obtain the local relational features. To learn the global relation, Zhang *et al*. [47] propose Relation-Aware Global Attention (RGA) module to obtain the relation among feature vectors of different spatial positions in a feature map.

Different from the aforementioned methods, the proposed HLGAT simultaneously considers the inter-local relation and the intra-local relation in the completed local graph to learn discriminative features for person Re-ID.

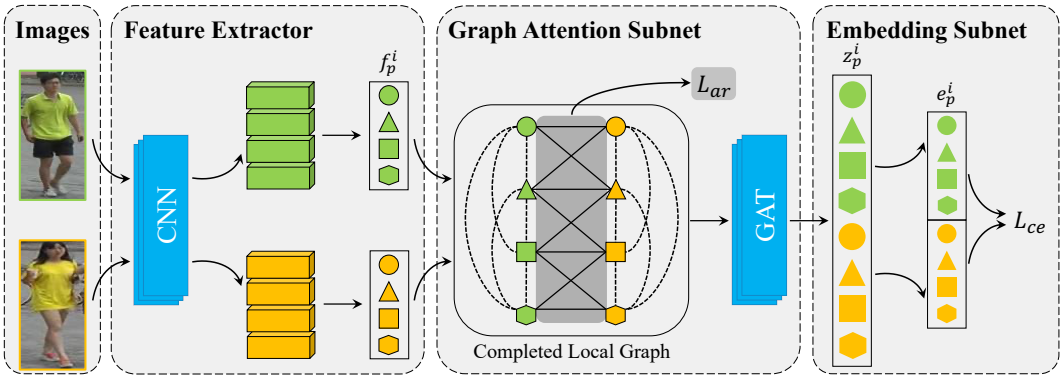

Figure 2. The pipeline of the proposed HLGAT. As for the completed local graph, the solid lines indicate the inter-local edges, and the dotted lines represent the intra-local edges.

2.3. Graph Attention Networks

Recently, Graph Convolutional Network (GCN) [17, 40] receives wide focus from both academia and industry due to its powerful capacity of reasoning and aggregating in coping with graph data. In order to focus on the important parts of data, Graph Attention Network (GAT) [18, 35] is proposed to improve GCN with attention mechanism. Some researchers [19, 25, 33, 37, 45] resort to GAT to address their tasks. Li *et al*. [19] propose Relation Aware Graph Attention Network (ReGAT) to model multi-type inter-object relations in order to represent geometric positions and semantic interactions between objects for visual question answering. Mi and Chen [25] design Hierarchical Graph Attention Network (HGAT) which constructs object-level graph and triplet-level graph to model interactions between objects and dependencies among relation triplets for visual relationship detection. However, these methods neglect the constraint on the attention mechanism of GAT in the optimization process. The proposed HLGAT injects the attention regularization loss and the contextual information into attention mechanism to accurately describe the inter-local relation and the intra-local relation for person Re-ID.

3. Approach

3.1. Overview

The framework of HLGAT is shown in Fig. 2 where it contains three key components including Feature Extractor, Graph Attention Subnet and Embedding Subnet.

- Feature Extractor. We feed pedestrian images into CNN to obtain feature maps, and then we adopt uniform partition strategy [34] to split these feature maps into several horizontal grids. Finally, we extract local features using the global max pooling operation on these grids.

- Graph Attention Subnet. We regard the local features as the nodes to construct the completed local graph to learn the inter-local relation and the intra-local relation, simultaneously. These nodes are linked by the inter-local edges and the intra-local edges. For each node in the graph, we weight its all neighbor nodes using the attention weights, and then aggregate them to obtain the two kinds of relations. Meanwhile, we constrain the attention weights of inter-local edges using the attention regularization loss to describe the inter-local relation accurately, and we inject the contextual information into the attention weights of intra-local edges to consider structure information.

- Embedding Subnet. In this subnet, we apply independent fully connected (FC) layers to reduce the dimension of the features extracted from Graph Attention Subnet. We utilize these dimension-reduced features as the final features, and make identity prediction on them.

3.2. Feature Extractor

In order to obtain local features from different body parts, we utilize the modified ResNet-50 [11] as the backbone of Feature Extractor due to its powerful capacity of feature representation. Specifically, we retain the architecture of ResNet-50 before the global average pooling layer to obtain high-level feature maps. Additionally, in order to extract the feature maps with larger spatial size, we modify the stride of $Conv5\_1$ from 2 to 1.

Given a set of pedestrian images $X = \{x_i\}$, where $i = 1, \cdots, N$ and N is the number of pedestrian images, we firstly resize each pedestrian image x_i into the fixed size of 384×128, and then feed them into the backbone. Afterwards, we obtain the feature maps with the size of $2048 \times 24 \times 8$, where 2048 denotes the channel number, and 24 and 8 are the height and the width of feature map, respectively. Finally, we split the feature maps into P uniform

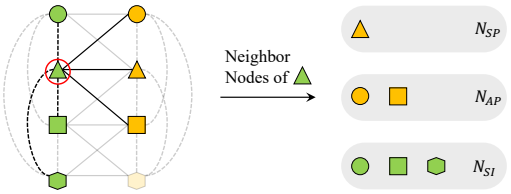

Figure 3. The neighbor nodes are divided into three types including N_{SP}, N_{AP} and N_{SI}. Concretely, the nodes from the corresponding parts of different pedestrian images are grouped into N_{SP}, the nodes from the adjacent parts of different pedestrian images are classified into N_{AP}, and the nodes from different parts of the same pedestrian image are grouped into N_{SI}.

horizontal grids and perform the global max pooling operation to obtain local features $F = \{f_p^i \in R^{2048}\}$, where $p = 1, \cdots, P$ and f_p^i denotes the local feature extracted from the p-th part of pedestrian image x_i.

3.3. Graph Attention Subnet

We treat the local features F as the nodes and construct the completed local graph $G = (F, E)$ where E denotes the edge set including the inter-local edges to link the nodes from corresponding parts and adjacent parts of different pedestrian images (see the solid lines in the completed local graph of Fig. 2), and the intra-local edges to link the nodes from the same pedestrian image (see the dotted lines in the completed local graph of Fig. 2). Hence, there are $N \times P$ nodes in the completed local graph which is not fully connected, and each node has several neighbor nodes with the linkage of the two types of edges. We divide these neighbor nodes into three types as shown in Fig. 3, i.e., N_{SP}, N_{AP} and N_{SI}. Specifically, N_{SP} consists of the neighbor nodes from the corresponding parts of different pedestrian images, N_{AP} is composed of the neighbor nodes from the adjacent parts of different pedestrian images, and N_{SI} includes the neighbor nodes from different parts of the same pedestrian image. It should be noticed that the adjacent parts of different pedestrian images possess higher correlation than the non-adjacent parts, and therefore we utilize the inter-local edges to link the adjacent parts.

In order to learn the inter-local relation and the intra-local relation in the completed local graph, we resort to the traditional GAT [35] to differentiate the importance of neighbor nodes and aggregate the information from them. The core equation of GAT is formulated as:

$$h_j^{'} = \sigma(\sum_{k \in N_j} \alpha_{jk} M h_k), \tag{1}$$

where h_k is the feature vector of node k, M is the transformation matrix, α_{jk} is the attention weight for node $k \in N_j$, and N_j is the neighbor node set of node j. σ denotes the non-linear activation function. However, it is impractical to directly utilize GAT to learn the inter-local relation and the intra-local relation in the completed local graph simultaneously. It is because for each node in the completed local graph, there are two types of edges and three types of neighbor nodes, but the traditional GAT equally treats all the nodes when computing the attention weights without discriminating them. Specifically, there are two limitations. (1) As for the intra-local relation, the difference between different relative parts of pedestrian images is neglected in the learning of weight attentions, which results in losing the structure information of pedestrian images; (2) as for the inter-local relation, the traditional GAT ignores the difference of attention weights from the same identity and different identities, which results in inaccurate attention weights learning.

In order to address the aforementioned limitations, we differentiate the attention weights in the learning process of the inter-local relation and the intra-local relation. In the completed local graph, we compute the attention weight $\alpha(f_p^i, f_q^j)$ between node f_p^i and its neighbor node f_q^j according to the cosine similarity. Specifically, we firstly utilize the transformation matrix $W \in R^{32 \times 2048}$ to transform f_p^i and f_q^j, and then we compute cosine similarity to measure the correlation between them. Finally, we apply the softmax function to obtain the attention weight $\alpha(f_p^i, f_q^j)$

$$\alpha(f_p^i, f_q^j) = \frac{exp(\phi(W f_p^i, W f_q^j))}{\sum_{f_q^k \in N(f_p^i)} exp(\phi(W f_p^i, W f_q^k))} \times K, \tag{2}$$

where $\phi(\cdot, \cdot)$ denotes the cosine similarity function, $N(f_p^i)$ is the neighbor node set of node f_p^i, and K is a regulation coefficient to consider different types of neighbor nodes. There are three types of neighbor nodes, and we define different K when node f_q^j belongs to different neighbor node sets.

- When $i \neq j$ and $p = q$, f_p^i and f_q^j are both from the corresponding parts of different images and f_q^j belongs to the neighbor node set N_{SP}, i.e., $N(f_p^i) = N_{SP}$. We experimentally set K to 1.

- When $i \neq j$ and $p = q + 1$ or $p = q - 1$, f_p^i and f_q^j are from the adjacent parts of different images and f_q^j belongs to the neighbor node set N_{AP}, i.e., $N(f_p^i) = N_{AP}$. We set K to a constant value which is smaller than that of N_{SP}, because the correlation decreases with the increase of the spatial distance.

- When $i = j$ and $p \neq q$, f_p^i and f_q^j are from different parts of the same image and f_q^j belongs to the neighbor node set N_{SI}, i.e., $N(f_p^i) = N_{SI}$. We observe that the correlation between two parts increases with the decrease of their spatial distance. Hence, we inject

the contextual information into the attention weights to consider structure information and K is formulated as

$$K = \frac{1}{|p-q| \times s}, \tag{3}$$

where s is the adjustment coefficient and $|p-q|$ is the relative spatial distance between the p-th part and the q-th part. For example, the relative spatial distance between the 3-th part and the 5-th part is $|3-5| = 2$. It should be noticed that we learn three independent transformation matrices W for three kinds of neighbor nodes in Eq. 2. Eq. 3 can distinguish the different relative parts of pedestrian images in the learning process of intra-local relation, which explicitly overcomes the first limitation.

In a word, we compute the attention weights according to different neighbor node sets. We learn the inter-local relation from the neighbor node sets N_{SP} and N_{AP}, and the intra-local relation from the neighbor node set N_{SI}. In order to integrate the inter-local relation and the intra-local relation, we aggregate these local features with the attention weights to obtain features z_p^i

$$z_p^i = \sigma(f_p^i + \sum_{f_q^j \in N(f_p^i)} \alpha(f_p^i, f_q^j) V f_q^j), \tag{4}$$

where σ denotes the non-linear activation function and $V \in R^{2048\times2048}$ is the transformation matrix. Like W, we learn different V for different neighbor nodes.

As for the inter-local relation, if the nodes belong to the same identity, then they possess high correlation and the attention weights should be large. In order to accurately describe the inter-local relation, we propose the attention regularization loss to constrain the attention weights

$$L_{ar} = \sum_{i=1}^{N} \sum_{p=1}^{P} \sum_{f_q^j \in (N_{SP} \cup N_{AP})} L_{bce}(\alpha(f_p^i, f_q^j), \tau), \tag{5}$$

where $L_{bce}(\cdot,\cdot)$ denotes the binary cross-entropy loss, and τ denotes the ground truth value. If node f_p^i and node f_q^j belong to the same identity, then $\tau = 1$, otherwise $\tau = 0$. After adding the attention regularization loss, the attention weights between the nodes from the same identity are enlarged, and meanwhile the attention weights between the nodes from different identities are reduced. As a result, the inter-local relation is accurately learned, which could overcome the second limitation.

3.4. Embedding Subnet

In this subnet, we utilize P independent FC layers to reduce the dimension of z_p^i from 2048 to 256 so as to obtain the final feature $e_p^i \in R^{256}$ for the p-th part of pedestrian image x_i, and we make identity prediction on them.

To optimize the proposed HLGAT, we combine the cross-entropy loss with the proposed attention regularization loss

$$Loss = L_{ce} + \lambda L_{ar}, \tag{6}$$

where λ is the balance coefficient and L_{ce} denotes the cross-entropy loss. L_{ce} is defined as

$$L_{ce} = -\sum_{i=1}^{N} \sum_{p=1}^{P} y_i log Q(e_p^i), \tag{7}$$

where y_i is the ground truth label of e_p^i, and $Q(e_p^i) \in [0,1]$ denotes the prediction probability.

In the test stage, we compute the distance between query image x_q and gallery image x_g to measure their similarity, which is formulated as

$$d(x_q, x_g) = \sum_{p=1}^{P} \phi(e_p^q, e_p^g), \tag{8}$$

where e_p^q and e_p^g denote the final features extracted from the p-th parts of x_q and x_g, respectively.

4. Experiments

In order to validate the performance of the proposed HLGAT, we conduct amounts of experiments on four publicly available large-scale person Re-ID datasets including Market-1501 [50], CUHK03 [20], DukeMTMC-reID [28] and MSMT17 [38]. In this section, we firstly introduce these datasets, and then we present the implementation details. We conduct ablation experiments to analyze the effectiveness of different components of HLGAT. We compare the proposed HLGAT with the relation-based methods and the state-of-the-arts. Finally, we evaluate the effect of several important hyper-parameters.

4.1. Datasets and Evaluation Protocol

Market-1501 is composed of 32,668 pedestrian images (1,501 identities). These images are split into two subsets including training set (12,936 images, 751 identities) and test set (19,732 images, 750 identities). The test set contains 3,368 query images and 15,913 gallery images. **CUHK03** contains 14,097 pedestrian images of 1,467 identities. There are 7,365 training images of 767 identities, 1,400 query images and 5,332 gallery images of 700 identities. **DukeMTMC-reID** includes 16,522 training images of 702 identities, 2,228 query images of 702 identities and 17,661 gallery images of 1,100 identities (408 distractor identities). **MSMT17** contains more pedestrian images and identities than other three datasets. There are 32,621 training images of 1,041 identities, 11,659 query images and 82,161 gallery images of 3,060 identities.

Method	Market-1501		CUHK03		DukeMTMC-reID		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Ours (N_{SI}^{*})	92.3	96.4	79.8	81.4	85.3	91.7	72.6	86.0
Ours	**93.4**	**97.5**	**80.6**	**83.5**	**87.3**	**92.7**	**73.2**	**87.2**
Ours without N_{SI}	91.3	96.5	78.6	80.9	84.5	90.9	71.3	84.9
Ours without L_{ar}	91.4	96.7	78.3	81.9	85.3	91.1	71.0	84.7
Ours without N_{AP}	92.7	96.5	79.5	82.1	86.5	91.4	71.2	85.0
Ours without N_{SP}	84.0	95.5	59.6	63.7	75.6	88.5	54.6	82.8

Table 1. Ablation studies on four datasets.

We adopt the mean Average Precision (mAP) accuracy and the Cumulative Matching Characteristic (CMC) curve at Rank-1 accuracy and Rank-5 accuracy as the evaluation protocols to assess the performance of the proposed HLGAT.

4.2. Implementation Details

We take the pre-trained ResNet-50 as the backbone to extract local features. Before fed into the backbone, each pedestrian image is resized into the fixed size of 384×128, and is normalized on the RGB channels. We apply random horizontal flip operation for each image to perform data augmentation in the training phase. In order to optimize HLGAT, we utilize the stochastic gradient descent (SGD) algorithm as the optimizer. The learning rate is initialized to 0.001 for the backbone and multiplied by 0.1 after 30 epochs. In the training stage, we set the epoch number to 50 and the batch size to 16. As for the hyper-parameters, $P = 8$, and s and λ are set to 10 and 0.01, respectively.

4.3. Ablation Experiments

We conduct ablation experiments to analyze the effectiveness of different components of HLGAT. We remove each component from HLGAT respectively including the inter-local relation (denoted as N_{SP} and N_{AP}), the attention regularization loss (denoted as L_{ar}) and the intra-local relation (denoted as N_{SI}). Table 1 shows the experimental results where we can draw the following conclusions.

Firstly, we consider the inter-local relation in N_{SP} by aggregating local features from corresponding parts of different pedestrian images. From the last row of Table 1 we can see that there is a prominent decline on four datasets. Specifically, the performance of HLGAT is reduced from 93.4% to 84.0% in mAP accuracy and from 97.5% to 95.5% in Rank-1 accuracy on Market-1501. It demonstrates that the inter-local relation among corresponding parts of different pedestrian images is important to improve the performance of the proposed HLGAT.

Secondly, we further consider the inter-local relation in N_{AP} by integrating local information from adjacent parts of different pedestrian images. As shown in fifth row of Table 1, there is 1.1% in mAP accuracy and 1.4% in Rank-1 accuracy decline on CUHK03. Hence, the inter-local relation in N_{AP} where there is high relation between adjacent parts is beneficial to the performance.

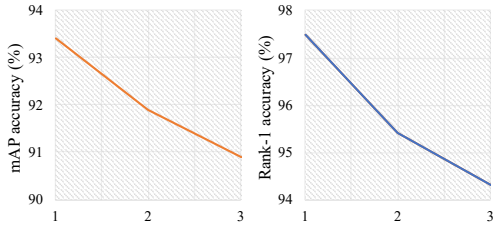

Figure 4. The influence of non-corresponding part number.

We conduct extra experiments on Market-1501 to exploit the effect when considering more non-corresponding parts from different pedestrian images for inter-local relation learning. As shown in Fig. 4, we can see that the performance drops when more non-corresponding parts are considered. It is because with the increase of the spatial distance between parts, the correlation between them decreases and distant parts are not similar in appearance, which may bring in some interference information. Hence, we only utilize the corresponding and adjacent parts to learn inter-local relation.

Thirdly, from the fourth row of the Table 1, we can see that after removing the attention regularization loss L_{ar}, there is 2.0% decline in mAP accuracy, and 1.6% decline in Rank-1 accuracy on DukeMTMC-reID. It proves the attention regularization loss could force the deep network to focus on the important parts of different pedestrian images so as to describe the inter-local relation accurately.

Finally, removing the intra-local relation in N_{SI} decreases mAP accuracy and Rank-1 accuracy from 73.2% to 71.3% and 87.2% to 84.9% on MSMT17, respectively. It is because we not only lose the intra-local relation but also ignore the structure information, which weakens representation capacities of local features. Moreover, in order to verify the effectiveness of the contextual information, we set K to the constant value in Eq. 3 (denoted as N_{SI}^{*}). From the

Method	Market-1501			CUHK03			DukeMTMC-reID			MSMT17		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
PCB [34]	77.4	92.3	97.2	54.2	61.3	77.7	66.1	81.8	89.7	-	-	-
PABR [32]	79.6	91.7	96.9	-	-	-	69.3	84.4	92.2	-	-	-
PABR + reranking [32]	89.9	93.4	96.4	-	-	-	83.9	88.3	93.1	-	-	-
CAMA [41]	84.5	94.7	98.1	64.2	66.6	82.7	72.9	85.8	93.1	-	-	-
OSNet [54]	84.9	94.8	-	67.8	72.3	-	73.5	88.6	-	52.9	78.7	-
Auto-ReID [27]	85.1	94.5	-	69.3	73.3	-	-	-	-	52.5	78.2	88.2
DG-Net [52]	86.0	94.8	-	-	-	-	74.8	86.6	-	52.3	77.2	87.4
BAT-net [8]	87.4	95.1	98.2	73.2	76.2	-	77.3	87.7	94.7	56.8	79.5	89.1
SFT [23]	87.5	94.1	97.5	71.7	74.3	85.6	79.6	90.0	94.0	58.3	79.0	85.8
DSA-reID [46]	87.6	95.7	98.4	73.1	78.2	-	74.3	86.2	-	-	-	-
SAN [15]	88.0	96.1	-	74.6	78.4	-	75.5	87.9	-	55.7	79.2	-
Pyramid [49]	88.2	95.7	98.4	74.8	78.9	-	79.0	89.0	-	-	-	-
ABD-Net [3]	88.3	95.6	-	-	-	-	78.6	89.0	-	60.8	82.3	90.6
VA-reID [55]	91.7	96.2	98.7	-	-	-	84.5	91.6	96.2	-	-	-
VA-reID + reranking [55]	95.4	96.8	98.3	-	-	-	91.8	93.9	96.5	-	-	-
HLGAT	93.4	97.5	98.9	80.6	83.5	92.6	87.3	92.7	96.5	73.2	87.2	93.0
HLGAT + reranking	**97.5**	**98.0**	**99.0**	**89.9**	**90.6**	**94.4**	**94.4**	**94.7**	**96.7**	**87.1**	**91.9**	**94.6**

Table 2. Comparison with state-of-the-art methods on four datasets.

Method	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
SGGNN [30]	82.8	92.3	68.2	81.1
IANet [12]	83.1	94.4	73.4	87.1
RGA-SC [47]	88.4	96.1	-	-
Relation-Net [26]	88.9	95.2	78.6	89.7
HLGAT	**93.4**	**97.5**	**87.3**	**92.7**

Table 3. Comparison with relation-based methods on Market-1501 and DukeMTMC-reID.

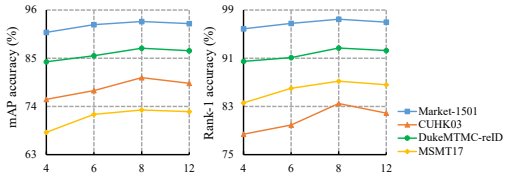

Figure 5. The influence of the number of uniform horizontal grids P.

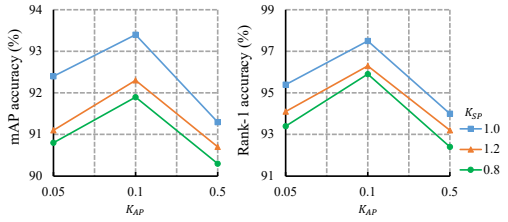

Figure 6. The influence of regulation coefficient K in the inter-local relation learning.

first row of Table 1, we can see that the performance drops, which proves the importance of the contextual information when learning the intra-local relation.

4.4. Comparison with the State-of-the-Art

Table 2 shows the comparison of the proposed HLGAT with the state-of-the-art methods on four datasets. We compare the proposed HLGAT with the second best methods in mAP accuracy on four datasets. Specifically, the proposed HLGAT surpasses VA-reID [55] with $+1.7\%$ and $+2.8\%$ in mAP accuracy on Market-1501 and DukeMTMC-reID, respectively. Meanwhile, the proposed HLGAT exceeds Pyramid [49] and ABD-Net [3] with $+5.8\%$ and $+12.4\%$ in mAP accuracy on CUHK03 and MSMT17, respectively. The comparative results validate the superiority of the proposed HLGAT.

4.5. Comparison with the Relation-based Methods

We conduct several experiments on Market-1501 and DukeMTMC-reID so as to compare the performance of the proposed HLGAT with other relation-based methods. The experimental results are shown in Table 3 where we can see that the proposed HLGAT exceeds Relation-Net [26] with $+4.5\%$ and $+8.7\%$ in mAP accuracy on Market-1501 and DukeMTMC-reID, respectively. Similar to the proposed HLGAT, Relation-Net learns local features and the intra-local relation among different parts within single pedestrian image. The proposed HLGAT also take the inter-local relation into consideration with attention regularization loss to constrain attention weights. In a word, the proposed HLGAT learns two types of local relation in the completed local graph so as to obtain discriminative features for person Re-ID. It should be noticed that these compared methods do not conduct experiments on CUHK03 and MSMT17, but the proposed HLGAT obtain excellent performance on these two datasets.

4.6. Parameter Analysis

There are some key hyper-parameters in the proposed HLGAT including the number of uniform horizontal grids

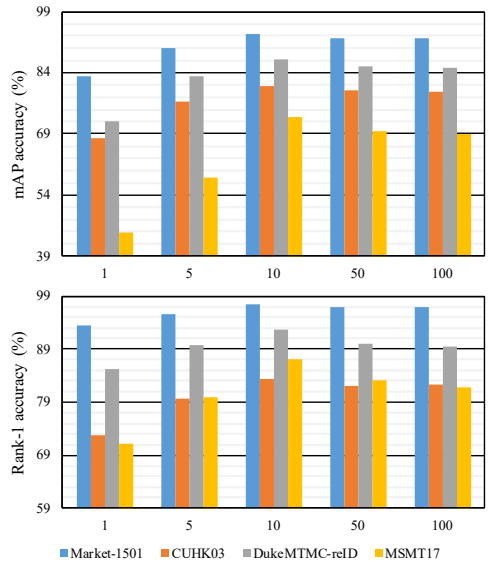

Figure 7. The influence of adjustment coefficient s.

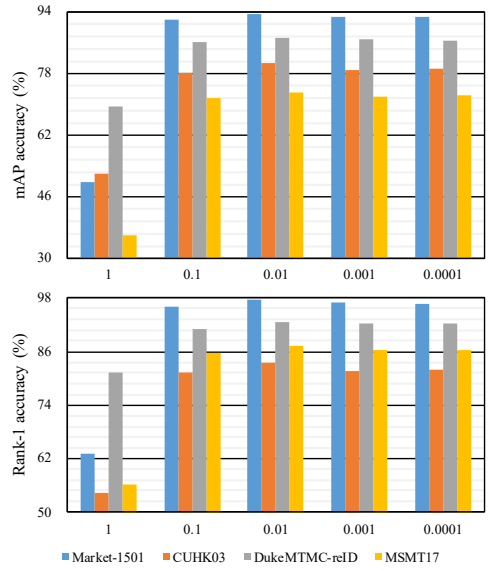

Figure 8. The influence of balance coefficient λ.

P, the regulation coefficient K and the balance coefficient λ. In order to comprehensively analyze the effect of these hyper-parameters, we conduct a series of experiments on four datasets.

The number of uniform horizontal grids P. We test different number of uniform horizontal grids P on four datasets, which is one issue to determine the size of the completed local graph. The experimental results are shown in Fig. 5 where we can see that when P is larger than 8, the performance drops. We argue that when P increases too large, the parts are more similar to each other and some parts may be partitioned into empty parts (e.g. background) where the proposed HLGAT can not learn useful local relation information from them. The proposed HLGAT obtains highest results when $P = 8$.

The influence of regulation coefficient K. In Eq. 2, we define different K for different neighbor node sets. As for the inter-local relation learning, we denote K_{SP} when nodes belong to the neighbor node set N_{SP}, and K_{AP} when nodes belong to the neighbor node set N_{AP}. As shown in Fig. 6, we can see that the proposed HLGAT achieves the best results when $K_{SP} = 1$ and $K_{AP} = 0.1$. It should be noticed that we conduct experiments on Market-1501, and the results can be generalized to other datasets.

In Eq. 3, the adjustment coefficient s could control the regulation coefficient K in the intra-local relation learning. Hence, we test different s in Fig. 7 where we we can see that HLGAT obtains the best results when $s = 10$.

The influence of balance coefficient λ. The balance coefficient λ in Eq. 6 controls the importance of the cross-entropy loss and the attention regularization loss. As shown in Fig. 8, we set different values for λ to evaluate the performance of the proposed HLGAT, and we find $\lambda = 0.01$ is the optimal value.

5. Conclusion

In this paper, we have proposed HLGAT to model the inter-local relation and the intra-local relation in a unified framework for person Re-ID. Specifically, we regard the local features as the nodes to construct the completed local graph where we simultaneously learn the inter-local relation among corresponding and adjacent parts from different pedestrian images, and the intra-local relation among different parts from the same pedestrian image. In order to model the local relation accurately, we propose the attention regularization loss to constrain the attention weights for the inter-local relation, and propose to inject the contextual information into the attention weights for the intra-local relation. The experimental results on four publicly available person Re-ID datasets have proved that the proposed HLGAT surpasses the state-of-the-art methods by an overwhelming margin.

Acknowledgements

This work was supported in part by Natural Science Foundation of Tianjin under Grant No. 20JCZDJC00180 and 19JCZDJC31500, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No. 202000002, and the Tianjin Higher Education Creative Team Funds Program.

References

[1] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognit*, 98:107036, 2020. 1

[2] Haoran Chen, Yaowei Wang, Yemin Shi, Ke Yan, Mengyue Geng, Yonghong Tian, and Tao Xiang. Deep transfer learning for person re-identification. In *BigMM*, pages 1–5, 2018. 1

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019. 7

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *AAAI*, page 3988–3994, 2017. 1

[5] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *CVPR*, pages 3300–3310, 2020. 1

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016. 1

[7] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognit*, 48(10):2993–3003, 2015. 1

[8] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *ICCV*, pages 8030–8039, 2019. 7

[9] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, pages 8295–8302, 2019. 2

[10] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, pages 3642–3651, 2019. 1, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[12] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, pages 9317–9326, 2019. 1, 2, 7

[13] Houjing Huang, Wenjie Yang, Xiaotang Chen, Xin Zhao, Kaiqi Huang, Jinbin Lin, Guan Huang, and Dalong Du. Eanet: Enhancing alignment for cross-domain person re-identification. *arXiv preprint arXiv:1812.11369*, 2018. 1, 2

[14] Bo Jiang, Xixi Wang, and Bin Luo. Ph-gcn: Person re-identification with part-based hierarchical graph convolutional network. *arXiv preprint arXiv:1907.08822*, 2019. 1

[15] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, pages 11173–11180, 2020. 7

[16] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071, 2018. 1

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3

[18] John Boaz Lee, Ryan A. Rossi, Sungchul Kim, Nesreen K. Ahmed, and Eunyee Koh. Attention models in graphs: A survey. *ACM Trans. Knowl. Discov. Data*, 13(6), 2019. 3

[19] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, pages 10313–10322, 2019. 3

[20] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 5

[21] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, pages 2194–2200, 2017. 1

[22] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 1

[23] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *ICCV*, pages 4976–4985, 2019. 7

[24] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognit*, 94:53–61, 2019. 2

[25] Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *CVPR*, pages 13886–13895, 2020. 3

[26] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *AAAI*, pages 11839–11847, 2020. 1, 2, 7

[27] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*, pages 3750–3759, 2019. 1, 7

[28] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016. 5

[29] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, pages 420–429, 2018. 1

[30] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, pages 486–504, 2018. 2, 7

[31] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, pages 3960–3969, 2017. 1

[32] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, pages 402–419, 2018. 7

[33] Mohammed Suhail and Leonid Sigal. Mixture-kernel graph attention network for situation recognition. In *ICCV*, pages 10363–10372, 2019. 3

[34] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 1, 2, 3, 7

[35] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 3, 4

[36] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018. 1, 2

[37] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *CVPR*, pages 10296–10305, 2019. 3

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 5

[39] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, pages 420–428, 2017. 2

[40] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst*, 32(1):4–24, 2021. 3

[41] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, pages 1389–1398, 2019. 7

[42] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process*, 28(6):2860–2871, 2019. 1

[43] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell*, 2021. doi: 10.1109/TPAMI.2021.3054775. 1

[44] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In *ICPR*, pages 34–39, 2014. 1, 2

[45] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *ICCV*, pages 9587–9595, 2019. 3

[46] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, pages 667–676, 2019. 7

[47] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 2, 7

[48] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, pages 1077–1085, 2017. 1

[49] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, pages 8514–8522, 2019. 7

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 5

[51] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[52] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 7

[53] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimedia Comput. Commun. Appl*, 14(1), 2017. 1

[54] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 7

[55] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Weishi Zheng, and Xing Sun. Viewpoint-aware loss with angular regularization for person re-identification. In *AAAI*, pages 13114–13121, 2020. 7