Learning Hybrid Relationships for Person Re-identification

Shuang Liu, Wenmin Huang, Zhong Zhang*
Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission,
Tianjin Normal University, Tianjin, China.
shuangliu.tjnu@gmail.com, huangwenmin2018@gmail.com, zhong.zhang8848@gmail.com

Abstract

Recently, the relationship among individual pedestrian images and the relationship among pairwise pedestrian images have become attractive for person re-identification (re-ID) as they effectively improve the ability of feature representation. In this paper, we propose a novel method named Hybrid Relationship Network (HRNet) to learn the two types of relationships in a unified framework that makes use of their own advantages. Specifically, for the relationship among individual pedestrian images, we take the features of pedestrian images as the nodes to construct a locally-connected graph, so as to improve the discriminative ability of nodes. Meanwhile, we propose the consistent node constraint to inject the identity information into the graph learning process and guide the information to propagate accurately. As for the relationship among pairwise pedestrian images, we treat the feature differences of pedestrian images as the nodes to construct a fully-connected graph so as to estimate robust similarity of nodes. Furthermore, we propose the inter-graph propagation to alleviate the information loss for the fully-connected graph. Extensive experiments on Market-1501, DukeMTMCreID, CUHK03 and MSMT17 demonstrate that the proposed HRNet outperforms the state-of-the-art methods.

1 Introduction

Person re-identification (re-ID) aims at matching the pedestrian across non-overlapping camera views, and it has a wide range of applications, such as picture grouping, activity analysis and multi-person tracking (Zheng et al. 2015; Fu et al. 2019; Zhou, Su, and Wu 2020). The challenges of person re-ID lie in complex variations in viewpoints, poses, resolutions, illumination, etc.

So far, Convolutional Neural Network (CNN) dominates the person re-ID field due to its promising performance (Ahmed, Jones, and Marks 2015; Park and Ham 2020). Most person re-ID methods utilize CNN to extract discriminative features from pedestrian images and then employ the classification set-up or metric learning to optimize the deep model. However, they only rely on individual pedestrian image in the feature learning process without providing the relationship among pedestrian images.

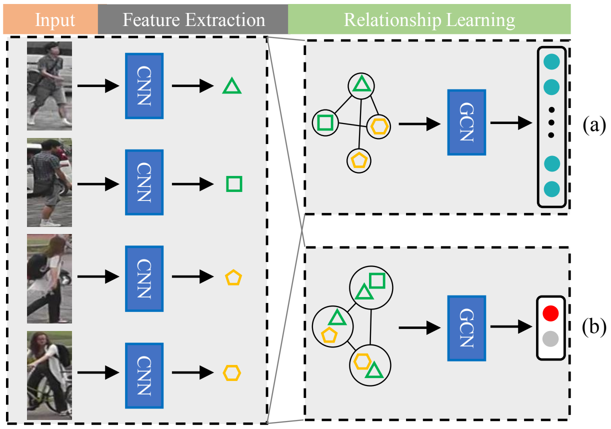


Figure 1: Two types of methods for learning the relationship among pedestrian images: (a) learning the relationship among individual pedestrian images, and (b) learning the relationship among pairwise pedestrian images.

The booming Graph Convolutional Network (GCN) has attracted extensive attention because of the powerful ability to learn the relationship of nodes in the graph (Kipf and Welling 2016; Zhou et al. 2018). Hence, several recent studies embed GCN into the person re-ID pipeline to establish the relationship among pedestrian images. These studies are roughly divided into two categories based on the type of relationship learning. The first category methods (Bao et al. 2019; Li et al. 2020) take single feature as the node of graph and utilize the classification loss to learn the relationship among individual pedestrian images, aiming to improve the discriminative ability of nodes, as shown in Figure 1 (a). As for the second category methods (Shen et al. 2018; Yan et al. 2019), they treat pairwise features as the node of graph and employ the verification loss to learn the relationship among pairwise pedestrian images, aiming to estimate robust similarity of nodes, as shown in Figure 1 (b). Since the two types of methods are different in node forms and loss functions, they could learn the relationship from different aspects. Hence, this motivates us to learn more comprehensive relationships among pedestrian images.

In this paper, we propose a novel method named Hybrid

[*] Corresponding Author.

Relationship Network (HRNet) to learn two types of relationships, i.e., relationship among individual pedestrian images and relationship among pairwise pedestrian images, for person re-ID. Specifically, in order to learn the relationship among individual pedestrian images, the features of pedestrian images are treated as the nodes to construct a locally-connected graph where each node is only connected to its nearest neighborhoods. In the process of information propagation, we introduce the attention mechanism to assign different edge weights to different neighborhoods, which could allow the nodes to decide where to focus by itself. Meanwhile, we propose the consistent node constraint between the input and output of graph convolutional layer to fully exploit the identity information of nodes. It encourages to bring the nodes from the same identity close and push away the nodes from different identities after the information propagation. This identity-guided manner ensures the information to be propagated in the graph accurately.

As for the relationship among pairwise pedestrian images, the feature differences of pedestrian images which reflect the similarities between pairwise pedestrian images are regarded as the nodes to construct a fully-connected graph. Similarly, we introduce the attention mechanism in the fully-connected graph to adaptively learn the edge weights between the nodes. As the information propagation in the graph, we could obtain robust similarity estimations for pairwise pedestrian images. It is worth noting that some discriminative information may be lost when the features are subtracted to form the nodes of the fully-connected graph, and therefore we propose the inter-graph propagation to relieve this loss. To this end, we add the edges between the corresponding nodes of the locally-connected graph and the fully-connected graph, and aggregate the node information of the full-connected graph from the locally-connected graph.

To summarize, we make the following contributions:

- We propose HRNet for person re-ID, which learns two types of relationships among pedestrian images to improve the ability of feature representation.
- As for the locally-connected graph, we propose the consistent node constraint to ensure the information to be propagated accurately, and as for the fully-connected graph, we propose the inter-graph propagation to alleviate the information loss.
- We evaluate the proposed HRNet on four large-scale person re-ID databases including Market-1501 (Zheng et al. 2015), DukeMTMC-reID (Ristani et al. 2016), CUHK03 (Zhao, Ouyang, and Wang 2014) and MSMT17 (Wei et al. 2018), and the results demonstrate the performance of HRNet outperforms the state-of-the-art methods.

2 Related Work

Person Re-ID. Person re-ID has made great progress in recent years and the mainstream methods either attempt to learn powerful feature representation or design appropriate metric learning to improve the performance of person re-ID. For feature representation, various CNN models (Li et al. 2014; Sun et al. 2018) have been proposed to perform robust and discriminative feature learning. Meanwhile, extra cues such as pose estimation and human parsing are also introduced to assist to mine discriminative visual information (Kalayeh et al. 2018).

As for the metric learning, some methods (Fu et al. 2019; Ro et al. 2019) treat person re-ID as a classification problem. They usually apply the softmax function to normalize the features, and then calculate the cross-entropy loss based on the ground-truth identity. Others (Ding et al. 2015; Zhou, Su, and Wu 2020) focus on learning an embedded space to push pedestrian images from the same identity close and those from different identities far away. The common used losses include triplet loss (Liu et al. 2017), hard triplet loss (Hermans, Beyer, and Leibe 2017), and quadruplet loss (Chen et al. 2017).

GCN and Its Applications in Person Re-ID. GCN (Kipf and Welling 2016; Zhou et al. 2018) has been proposed to process graph structure data. The principle of constructing GCN mainly follows two streams: spectral perspective and spatial perspective. The spectral-based GCN (Defferrard, Bresson, and Vandergheynst 2016) realizes convolution operation in the Fourier domain by calculating the eigenvectors and eigenvalues of graph Laplacian matrix. The spatial-based GCN (Chen et al. 2020) directly performs the convolution filters on the graph nodes and their neighborhoods.

Recently, some studies introduce GCN in the person re-ID field in order to learn the relationship among pedestrian images. For example, Shen et al. (Shen et al. 2018) propose the Similarity-Guided Graph Neural Network (SGGNN) for person re-ID, where they regard feature differences of pedestrian images as the nodes and the similarity among features as the edge weights. Yan et al. (Yan et al. 2019) design the Contextual Graph Model (CGM) to jointly consider the information among the target pair and the context pairs, where the target pair is linked to all the context pairs. Li et al. (Li et al. 2020) present the Spatial Preserved Graph Convolution (SPGC) network to model the relationship among individual pedestrian images, which employs the mask matrix with the identity information to determine the edges between the nodes and calculates the edge weights based on the similarity.

Attention Mechanism in Graph Neural Network. The attention mechanism allows the model to focus on the most task-relevant parts, which makes it attractive for deep learning communities (Wang et al. 2018; Mnih et al. 2014). Recent studies (Veličković et al. 2018) introduce the attention mechanism into the graph neural network to adaptively learn the edge weights between the nodes, thereby improving the ability of information aggregation.

The proposed HRNet follows the spatial perspective of GCN and applies the attention mechanism to learn two types of relationships among pedestrian images in a unified framework. Meanwhile, we propose the consistent node constraint and the inter-graph propagation to guide the information propagation and alleviate the information loss, respectively.

3 Approach

The structure of the proposed HRNet is shown in Figure 2. We first introduce the feature extraction of pedestrian

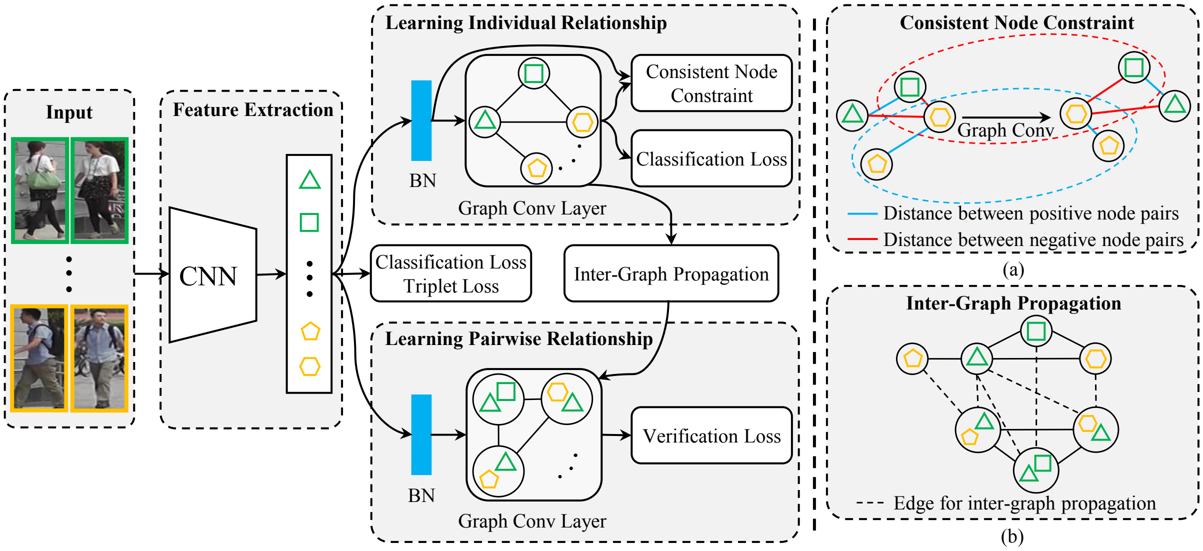

Input
Feature Extraction
CNN
Learning Individual Relationship
BN
Graph Conv Layer
Consistent Node Constraint
Classification Loss
Classification Loss
Triplet Loss
Inter-Graph Propagation
Learning Pairwise Relationship
BN
Graph Conv Layer
Verification Loss
Consistent Node Constraint
Graph Conv
Distance between positive node pairs
Distance between negative node pairs
(a)
Inter-Graph Propagation
Edge for inter-graph propagation
(b)

Figure 2: The structure of the proposed HRNet.

images. Then, we explain the relationship learning in detail including relationship among individual pedestrian images, consistent node constraint, relationship among pairwise pedestrian images and inter-graph propagation. Finally, we present the optimization strategy of HRNet.

3.1 Feature Extraction

We utilize the ResNet-50 (He et al. 2016) pre-trained on ImageNet as the CNN model of HRNet to extract the features of pedestrian images. Specifically, following (Luo et al. 2019), the fully connected layer FC-1000 of ResNet-50 is removed and the stride of Conv5_1 is set to 1. Each pedestrian image is resized to 256 × 128 and then fed into the ResNet-50 to extract the convolutional activation maps with the size of $2048 \times 16 \times 8$, where 2048 is the number of convolutional activation maps, and 16 and 8 are the height and width of convolutional activation map, respectively. Finally, we apply the global average pooling to the convolutional activation maps and obtain the features of pedestrian images $\mathcal{F} = \{f_1, f_2, \cdots, f_N\}$, where $f_n \in \mathbb{R}^d$, and N is the number of pedestrian images. Here, d is the feature dimension and it is equal to 2048.

3.2 Relationship Learning

Motivation. There are rich relationships among pedestrian images, which is beneficial to improving the ability of feature representation. Hence, some recent studies are working in this direction, and learn the relationship among individual pedestrian images or the relationship among pairwise pedestrian images. The methods for learning the relationship among individual pedestrian images could improve the discriminative ability of features by allowing them to aggregate the information from other features. The methods for learning the relationship among pairwise pedestrian images could make full use of sample pairs to generate robust similarity estimation. Since the two types of methods have different advantages, this motivates us to learn more comprehensive relationships among pedestrian images to further improve the performance of person re-ID.

Learning the Relationship among Individual Pedestrian Images. We treat each feature of pedestrian image as one node to learn the relationship among individual pedestrian images. Furthermore, the relationship between the node and its neighborhoods has a great impact on this kind of relationship, and therefore we only construct the links with the neighborhood nodes. Specifically, we take the feature f_n as the node v_n to construct a locally-connected graph, where each node is only connected to the k nearest neighborhoods. For the node set $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$, where $v_n \in \mathbb{R}^d$, we utilize the adjacency matrix $C = [c_{i,j}] \in \mathbb{R}^{N \times N}$ to describe the connection relationship among the nodes, and $c_{i,j}$ represents the edge between v_i and v_j. It is defined as:

$$c_{i,j} = \begin{cases} 1, \; if \; r(m(v_i, v_j)) \leq k \\ 0, \; otherwise \end{cases} \qquad (1)$$

where $m(v_i, v_j)$ indicates the Euclidean distance between v_i and v_j, $r(m(v_i, v_j))$ represents the ranking of $m(v_i, v_j)$ in the distances between v_i and all nodes in the ascending sort, and k is the hyperparameter. Eq. 1 indicates that we build the edges between the node and its k nearest neighborhoods.

From Eq. 1 we can see that C only defines the connection relationship between the nodes using 0 or 1, however it can not capture the importance among different nodes. In order to model the accurate relationship among the nodes, we introduce the attention matrix $A = [a_{i,j}] \in \mathbb{R}^{N \times N}$ to assign different edge weights to different nodes. Here, $a_{i,j}$ is the attention coefficient and represents the edge weight between v_i and v_j:

$$a_{i,j} = \frac{e^{\mu([v_i||v_j]^\top H_s)}}{\sum_{n=1}^{N} e^{\mu([v_i||v_n]^\top H_s)}} \qquad (2)$$

where $||$ represents the vector concatenation, $\top$ indicates the transposition operation, $H_s \in \mathbb{R}^{2d}$ is the learnable parameter vector, and $\mu(\cdot)$ is a non-linear function implemented by the $LeakyReLU$ function (the negative input slope is set to 0.2).

After obtaining C and A, we perform the graph convolutional operation to learn the relationship among the nodes. Specifically, the graph convolutional layer is formulated as:

$$X' = \sigma((A \circ C + I)XW_s) \tag{3}$$

where $X \in \mathbb{R}^{N \times d}$ is the node feature matrix, $X' \in \mathbb{R}^{N \times d'}$ is the updated node feature matrix, $W_s \in \mathbb{R}^{d \times d'}$ is the parameter matrix of graph convolutional layer, $\circ$ represents the element-wise product, and $\sigma(\cdot)$ is an activation function implemented by the $ReLU$ function in our experiments. Here, $I \in \mathbb{R}^{N \times N}$ is an identity matrix which is utilized to increase the edge weight of the node itself in order to prevent the over-smoothing. For the first graph convolutional layer, $X = [v_1, v_2, \cdots, v_N]$.

Consistent Node Constraint. The traditional graph convolutional operations (Shen et al. 2018; Chen et al. 2020) usually ignore the identity information of nodes as expressed in Eq. 3. In order to fully exploit the identity information of nodes and supervise the graph learning process, some methods (Bao et al. 2019; Li et al. 2020) utilize the identity information to construct the graph via connecting the nodes with the same identity so as to ensure the rationality of information aggregation. A main problem for this manner is that the graph construction requires the identity information, and therefore the graph cannot be built when the identity information is unavailable, such as in the test stage. As a result, these methods can not directly extract the features from GCN in the test stage, which weakens the ability of relationship representation.

To overcome the above-mentioned drawback, we propose the consistent node constraint which injects the identity information in the graph learning process and does not require the identity information when constructing the graph. Specifically, we utilize the identity information to constrain the distance between the input node and the output node of graph convolutional layer. Assuming that the graph convolutional layer updates the nodes v_i and v_j to v_i' and v_j', the consistent node constraint for the node pairs with the same identity is defined as:

$$\mathcal{L}_{c+} = (m(v_i', v_j') - z(m(v_i, v_j)) + \omega)_+ \tag{4}$$

where $(\cdot)_+ = \max(\cdot, 0)$, ω is the pre-defined margin, and $z(\cdot)$ represents the zero gradient function which treats the variable as the constant when calculating gradients, and stops the back propagation during training. The consistent node constraint for the node pairs with different identities is defined as:

$$\mathcal{L}_{c-} = (z(m(v_i, v_j)) - m(v_i', v_j') + \omega)_+ \tag{5}$$

It is worth noting that for each node, we only select the hardest positive node and the hardest negative node to calculate the loss. As shown in Figure 2 (a), the consistent node constraint brings the node pairs with the same identity (blue ellipse) close and push away the node pairs with different identities (red ellipse) after graph convolution, so that the identity information can be injected into the graph learning process to guide the information to be propagated accurately.

Learning the Relationship among Pairwise Pedestrian Images. The relationship among pairwise pedestrian images is very important for estimating the similarity of image pair. Hence, we take the feature differences of pedestrian images as the nodes to construct a fully-connected graph. Specifically, for each feature of pedestrian image f_n, we first obtain the node set $\mathcal{U}_n = \{u_{n,1}, u_{n,2}, \cdots, u_{n,N}\}$ of the fully-connected graph by performing the subtraction operation between f_n and each feature in $\mathcal{F}$, i.e., $u_{n,i} = f_n - f_i$. Note that one feature of pedestrian image corresponds to a fully-connected graph, and there are N fully-connected graphs for $\mathcal{F}$. Then, we treat the attention coefficient $g_{i,j}^n$ as the edge weight between $u_{n,i}$ and $u_{n,j}$:

$$g_{i,j}^n = \frac{e^{\mu([u_{n,i}||u_{n,j}]^\top H_p)}}{\sum_{b=1}^{N} e^{\mu([u_{n,i}||u_{n,b}]^\top H_p)}} \tag{6}$$

where $H_p \in \mathbb{R}^{2d}$ is the learnable parameter vector.

Inter-Graph Propagation. Since the nodes of the fully-connected graph are the feature differences of pedestrian images, some discriminative information may be lost. Correspondingly, we propose the inter-graph propagation to alleviate the information loss. Specifically, we add the edges between the corresponding nodes of the locally-connected graph and the fully-connected graph to allow the fully-connected graph to aggregate the information from the locally-connected graph. The dotted lines in Figure 2 (b) indicate this kind of edges. As a result, for the graph convolutional layer with the inter-graph propagation, the update of $u_{n,i}$ is expressed as:

$$\begin{aligned} u_{n,i}' =& \alpha u_{n,i} + (1-\alpha)\sigma(\beta \sum_{j=1}^{N} g_{i,j}^n u_{n,j}^\top W_p + \\ & (1-\beta)(q_{i,n}^n v_n^\top W_l + q_{i,i}^n v_i^\top W_r)) \end{aligned} \tag{7}$$

where W_p, W_l and $W_r \in \mathbb{R}^{d \times d'}$ are the parameter matrices of graph convolutional layer, α is the hyperparameter to balance the proportion of original information and aggregated information, and β is the hyperparameter to balance the proportion of aggregated information of fully-connected graph and locally-connected graph. Here, $q_{i,n}^n$ is the attention coefficient between $u_{n,i}$ and v_n, and $q_{i,i}^n$ is the attention coefficient between $u_{n,i}$ and v_i. They are defined as:

$$q_{i,n}^n = \frac{e^{\mu([u_{n,i}||v_n]^\top H_l)}}{e^{\mu([u_{n,i}||v_n]^\top H_l)} + e^{\mu([u_{n,i}||v_i]^\top H_r)}} \tag{8}$$

$$q_{i,i}^n = \frac{e^{\mu([u_{n,i}||v_i]^\top H_r)}}{e^{\mu([u_{n,i}||v_n]^\top H_l)} + e^{\mu([u_{n,i}||v_i]^\top H_r)}} \tag{9}$$

where H_l and $H_r \in \mathbb{R}^{2d}$ are the learnable parameter vectors.

3.3 Optimization

We apply the classification loss $\mathcal{L}_{cl}$ with the label smoothing and the hard triplet loss $\mathcal{L}_{tr}$ to jointly supervise the training of CNN model of HRNet (Luo et al. 2019). Specifically, $\mathcal{L}_{cl}$ is defined as:

$$\mathcal{L}_{cl} = \sum_{s=1}^{S} -p_s log(\hat{y}_s) \quad (10)$$

where S is the number of pedestrian identities, and $\hat{y}_s$ indicates the predicted probability of the s-th identity. p_s is the smoothed identity label and it is defined as:

$$p_s = \begin{cases} 1 - \frac{S-1}{S}\epsilon, & s = y \\ \frac{\epsilon}{S}, & s \neq y \end{cases} \quad (11)$$

where y is the ground-truth identity label, and ϵ is the s-moothed coefficient and it is set to 0.1 in our experiments. Meanwhile, $\mathcal{L}_{tr}$ is defined as:

$$\mathcal{L}_{tr} = \sum_{f_a, f_p, f_n \in \mathcal{F}} (m(f_a, f_p) - m(f_a, f_n) + \theta)_+ \quad (12)$$

where θ is the pre-defined margin, and f_a, f_p and f_n denote the anchor, the hardest positive and the hardest negative sample features, respectively.

As for the module of learning individual relationship, our task is to classify the nodes to improve the discriminative ability of features, so we apply the classification loss with the label smoothing in Eq. 10 as the loss function, denoted as $\mathcal{L}_{in}$. As for the module of learning pairwise relationship, we employ the verification loss to determine whether the feature pairs of nodes belong to the same identity:

$$\mathcal{L}_{pa} = -tlog\hat{t} - (1-t)log(1-\hat{t}) \quad (13)$$

where t is the ground-truth of feature pairs of nodes, $t = 1$ if the feature pair from the same identity, otherwise $t = 0$, and $\hat{t}$ indicates the predicted probability that the feature pair belongs to the same identity.

In summary, the total loss of HRNet is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{cl} + \mathcal{L}_{tr} + \mathcal{L}_{in} + \mathcal{L}_{pa} + \lambda(\mathcal{L}_{c+} + \mathcal{L}_{c-}) \quad (14)$$

where λ is the hyperparameter and it is experimentally set to 0.1.

4 Experiments

4.1 Databases

Market-1501 (Zheng et al. 2015) is shot by six disjoint cameras at the Tsinghua University campus, and it consists of 32,668 images of 1,501 identities. According to the database setting, the training set contains 12,936 images of 751 identities while the test set comprises of 3,368 query images and 16,364 gallery images from the other 750 identities.

DukeMTMC-reID (Ristani et al. 2016) consists of 36,411 images of 1,404 identities, among which 16,522 images of 702 identities are utilized as the training set. And 19,889 images of 702 non-overlapping identities are treated as the test set with 2,228 query images as well as 17,661 gallery images. Furthermore, DukeMTMC-reID is collected by eight high-resolution cameras.

CUHK03 (Zhao, Ouyang, and Wang 2014) is composed of 14,097 images of 1,467 identities, and each identity is captured by two of ten cameras at the CUHK campus. According to the database setting, the training set consists of 7,365 images of 767 identities and the test set includes 1,400 query images and 5,332 gallery images of 700 identities. CUHK03 provides two types of annotations for all images, i.e., manually labeled bounding-boxes and DPM-detected bounding-boxes. In this work, we evaluate the proposed HRNet on DPM-detected bounding-boxes which are more challenging.

MSMT17 (Wei et al. 2018) comprehends 126,441 images of 4,101 identities from 15 cameras and it is divided into the training set including 32,621 images of 1,041 identities and the test set including 93,820 images of 3,060 i-dentities. MSMT17 is more challenging than other person re-ID databases because of more complex changes in poses, viewpoints, illumination and scenes.

4.2 Implementation Details

Training. The batch size is set to 66 where we randomly select 11 identities and 6 images for each identity, and the epoch number is set to 200. We adopt the random cropping and the horizontal flipping for data augmentation. We utilize Adam as the optimizer and set the weight decay to 5×10^{-4}. The learning rate is initialized to 3.5×10^{-4} and it is decreased by the factor of 0.1 at the 40-th and 120-th epochs. The pre-defined margins ω in Eq. 4 and θ in Eq. 12 are both set to 0.3, and the hyperparameters α and β in Eq. 7 are set to 0.9 and 0.8, respectively.

Testing. We first use the CNN model of HRNet to extract the CNN-based features of all pedestrian images. Then, for each query image, we obtain the top-120 gallery images based on the cosine similarity between the CNN-based features of query image and gallery images. Afterwards, each query image and its top-120 gallery images are fed into the two relationship learning modules, respectively. Finally, the similarity scores calculated from the two relationship learning modules are added to obtain the final similarity scores between the query image and the gallery images.

4.3 Ablation Study

In this subsection, we perform comprehensive ablation s-tudies to demonstrate the effectiveness of different components in HRNet. The experimental results are shown in Table 1, where CNN is the baseline implemented by ResNet-50, IR represents the relationship among individual pedestrian images, IR^* indicates that we use the fully-connected graph to replace the locally-connected graph for the individual relationship learning, CNC denotes the consistent node constraint, PR represents the relationship among pairwise pedestrian images, and IGP denotes the inter-graph propagation.

Effectiveness of Different Components in HRNet. From Table 1 we draw the following conclusions. Firstly, both CNN+IR and CNN+PR significantly improve CNN due to considering the relationship among pedestrian images. Specifically, on the four databases, the improvements of CNN+IR in mAP are +4.8%, +5.4%, +16.0% and +8.8%,

Methods	Market-1501		DukeMTMC-reID		CUHK03		MSMT17	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
CNN	84.2	92.8	74.7	85.1	57.7	59.4	48.3	72.2
CNN+IR	89.0	95.1	80.1	89.3	73.7	77.1	57.1	79.6
CNN+IR^*	86.9	93.7	76.9	87.5	70.3	72.8	52.1	77.0
CNN+IR+CNC	90.6	96.0	82.8	90.7	77.1	79.6	58.9	81.3
CNN+PR	87.4	94.7	78.1	88.9	71.9	76.1	52.3	76.6
CNN+PR+IGP	89.7	95.2	80.8	90.2	75.9	77.5	57.4	79.7
HRNet	**91.3**	**96.7**	**83.7**	**91.9**	**78.4**	**81.1**	**60.6**	**82.0**

Table 1: Ablation study on Market-1501, DukeMTMC-reID, CUHK03 and MSMT17.

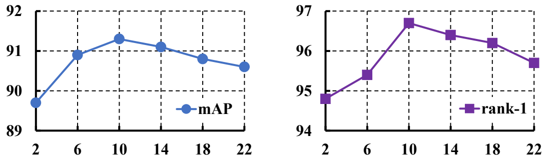


Figure 3: The effect of k for HRNet on Market-1501.

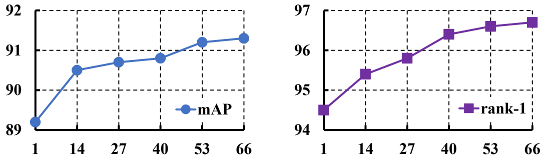


Figure 4: Performance of different edge number for each node when learning the relationship among pairwise pedestrian images on Market-1501.

respectively; the improvements of CNN+PR in mAP are +3.2%, +3.4%, +14.2% and +4.0%, respectively. Secondly, compared with CNN+IR, CNN+IR+CNC gains higher mAP and rank-1 on the four databases. It is because the consistent node constraint injects the identity information in the graph learning process, and effectively guides the information to propagate accurately. Thirdly, CNN+PR+IGP significantly exceeds CNN+PR in mAP by +2.3%, +2.7%, +4.0% and +5.1% on the four databases, respectively. It shows that the inter-graph propagation could effectively alleviate the information loss for robust similarity estimation learning. Note that when without IR, $v_n = f_n$, and $v_i = f_i$ in Eq. 7. Finally, HRNet achieves better performance than CNN+IR+CNC and CNN+PR+IGP. It demonstrates that fusion of the two types of graphs could learn the relationship among pedestrian images from different aspects so as to improve the representation ability.

Discussion about Locally-Connected Graph and Fully-Connected Graph. For the relationship among individual pedestrian images, CNN+IR (locally-connected graph) improves CNN+IR^* (fully-connected graph) in mAP from 86.9%, 76.9%, 70.3% and 52.1% to 89.0% (+2.1%), 80.1% (+3.2%), 73.7% (+3.4%) and 57.1% (+5.0%) on the four databases, respectively. Hence, we adopt the locally-connected graph to learn the relationship among individual pedestrian images. Furthermore, we analyze the hyper-parameter k in Eq. 1 which controls the number of nearest neighborhoods for each node in the graph. From Figure 3, we can see that the proposed HRNet achieves the best performance on Market-1501 when k is equal to 10. Note that the experiments have shown that the conclusions can be generalized to the other three databases as well.

For the relationship among pairwise pedestrian images, mAP and rank-1 improves when the edge number for each node increases as shown in Figure 4. It demonstrates that global similarity information (fully-connected graph) is more robust than the local similarity information (locally-connected graph) for estimating the similarity of feature pairs of nodes. Therefore, we adopt the fully-connected graph to learn the relationship among pairwise pedestrian images in all the experiments.

4.4 Comparison with State-of-the-Art

Table 2 presents the comparison results of HRNet with the state-of-the-art methods on Market-1501, DukeMTMC-reID, CUHK03 and MSMT17. The first part of Table 2 indicates the methods extract features from individual pedestrian images, and the second part of Table 2 represents the methods learn the relationship among pedestrian images.

From Table 2, we have the following observations. Firstly, on the four databases, the proposed HRNet gains 91.3%, 83.7%, 78.4% and 60.6% on mAP, and 96.7%, 91.9%, 81.1% and 82.0% on rank-1, respectively, which outperforms all the compared methods. It fully demonstrates the effectiveness of HRNet. Secondly, Our method substantially outperforms SGGNN (Shen et al. 2018) by +8.5% and +15.5% on mAP on Market-1501 and DukeMTMC-reID, respectively. It is because SGCNN only considers the relationship among pairwise pedestrian images, while our method learns two types of relationships among pedestrian images in a unified framework. Thirdly, the proposed HRNet improves SPGC+PPE (Li et al. 2020) in terms of mAP by +4.8%, +4.7% and 3.0% on Market-1501, DukeMTMC-reID and CUHK03, respectively. SPGC+PPE provides one kind of relationship during training, and it relies on the features extracted from CNN to conduct the pedestrian matching in the test stage. Our method not only provides two kinds of relationship, but also incorporates the relationship into both during training and testing. Finally, HRNet also achieves the better performance compared with other hybrid method-

Methods	Market-1501		DukeMTMC-reID		CUHK03		MSMT17	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
V-F (Zheng, Zheng, and Yang 2017)	59.9	79.6	49.3	68.9	-	-	-	-
PCB (Sun et al. 2018)	77.3	92.4	65.3	81.9	54.2	61.3	40.4	68.2
Mancs (Wang et al. 2018)	82.3	93.1	71.8	84.9	60.5	65.5	-	-
HPM (Fu et al. 2019)	82.7	94.2	74.3	86.6	57.5	63.9	-	-
Auto-ReID (Quan et al. 2019)	85.1	94.5	-	-	69.3	73.3	52.5	78.2
DG-Net (Zheng et al. 2019)	86.0	94.8	74.8	86.6	-	-	52.3	77.2
PISNet (Zhao et al. 2020)	87.1	95.6	78.7	88.8	-	-	-	-
SAN (Jin et al. 2020)	88.0	96.1	75.5	87.9	74.6	79.6	55.7	79.2
RGA-SC (Zhang et al. 2020)	88.4	96.1	-	-	74.5	79.6	57.5	80.3
SGGNN (Shen et al. 2018)	82.8	92.3	68.2	81.1	-	-	-	-
RNet-S (Park and Ham 2020)	88.0	94.8	77.1	89.3	69.5	72.5	-	-
SPGC+PPE (Li et al. 2020)	86.5	95.4	79.0	89.3	75.4	79.0	-	-
HRNet	**91.3**	**96.7**	**83.7**	**91.9**	**78.4**	**81.1**	**60.6**	**82.0**

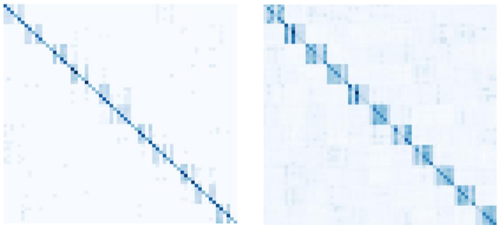
Table 2: Performance (%) comparisons with the state-of-the-art methods on Market-1501, DukeMTMC-reID, CUHK03 and MSMT17.

s. For example, V-F (Zheng, Zheng, and Yang 2017) combines the verification and identification models to learns a discriminative embedding. HRNet significantly surpasses V-F. Mancs (Wang et al. 2018) performs ranking and classification tasks in a unified framework. Compared with Mancs, we improve 9%, 11.9% and 17.9% on mAP on Market-1501, DukeMTMC-reID and CUHK03. It proofs that learning the hybrid relationship among images is more powerful.

4.5 Visualization

In order to understand the learning process of relationship among individual pedestrian images intuitively, we visualize the matrix $A \circ C$ in Eq. 3 as shown in Figure 5 (a) where the deeper color indicates the larger value. We sample 11 identities and 6 pedestrian images for each identity to form a batch, and distribute the identity labels in a batch as $\{y_1, y_1, y_1, y_1, y_1, y_1\}, \cdots, \{y_{11}, y_{11}, y_{11}, y_{11}, y_{11}, y_{11}\}$, where y_i represents the i-th identity label. Hence, the elements around the matrix diagonal indicate the edge weights between pedestrian images from the same identity. As can be seen from Figure 5 (a), the large values are mainly concentrated around the diagonal of the matrix. It shows that there is more information propagation between the nodes from the same identity than the nodes from different identities.

For learning the relationship among pairwise pedestrian images, we visualize the matrix $G_n = [g^n_{i,j}]$ in Eq. 6. Since we build a fully-connected graph for each feature of pedestrian image f_n, we add the fully-connected graphs from different identities in a batch to visualize the relationship among pairwise pedestrian images as shown in Figure 5 (b). Obviously, like $A \circ C$, high values mainly distribute in the area near the diagonal of the matrix. Therefore, we have the similar conclusion to $A \circ C$. In addition, an intuitive advantage of having large edge weights between feature pairs (nodes) from the same identity is that the relative easy positive feature pairs can be used to guide the similarity update of the hard positive feature pairs, so as to improve the similarity scores of the hard positive feature pairs.



(a) Matrix $A \circ C$

(b) Matrix G_n

Figure 5: Visualization of the matrices $A \circ C$ and G_n.

5 Conclusion

In this paper, we have proposed HRNet to learn two types of relationships among pedestrian images simultaneously. Specifically, we construct the locally-connected graph and the fully-connected graph to learn the relationship among individual pedestrian images and the relationship among pairwise pedestrian images, respectively. Meanwhile, for the locally-connected graph, we propose the consistent node constraint to ensure the information to be propagated accurately. As for the fully-connected graph, we propose the inter-graph propagation to alleviate the information loss. We have verified HRNet on four large-scale person re-ID databases, and the experimental results show the proposed method surpasses the state-of-the-art methods.

Acknowledgments

This work was supported 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

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*, 3908–3916.

Bao, L.; Ma, B.; Chang, H.; and Chen, X. 2019. Masked Graph Attention Network for Person Re-Identification. In *CVPRW*, 1496–1505.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 403–412.

Chen, X.; Zheng, L.; Zhao, C.; Wang, Q.; and Li, M. 2020. RRGCCAN: Re-Ranking via Graph Convolution Channel Attention Network for Person Re-Identification. *IEEE Access* 8: 131352–131360.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 3844–3852.

Ding, S.; Lin, L.; Wang, G.; and Chao, H. 2015. Deep feature learning with relative distance comparison for person re-identification. *PR* 48(10): 2993–3003.

Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; and Huang, T. 2019. Horizontal pyramid matching for person re-identification. In *AAAI*, 8295–8302.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

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

Jin, X.; Lan, C.; Zeng, W.; Wei, G.; and Chen, Z. 2020. Semantics-Aligned Representation Learning for Person Re-Identification. In *AAAI*, 11173–11180.

Kalayeh, M. M.; Basaran, E.; Gökmen, M.; Kamasak, M. E.; and Shah, M. 2018. Human semantic parsing for person re-identification. In *CVPR*, 1062–1071.

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

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 152–159.

Li, Z.; Zhou, Z.; Jiang, N.; Han, Z.; Xing, J.; and Jiao, J. 2020. Spatial Preserved Graph Convolution Networks for Person Re-identification. *ACM TOMM* 16(1): 1–14.

Liu, H.; Feng, J.; Qi, M.; Jiang, J.; and Yan, S. 2017. End-to-end comparative attention networks for person re-identification. *IEEE TIP* 26(7): 3492–3506.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *CVPRW*.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *NIPS*, 2204–2212.

Park, H.; and Ham, B. 2020. Relation Network for Person Re-identification. In *AAAI*.

Quan, R.; Dong, X.; Wu, Y.; Zhu, L.; and Yang, Y. 2019. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*, 3750–3759.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 17–35.

Ro, Y.; Choi, J.; Jo, D. U.; Heo, B.; Lim, J.; and Choi, J. Y. 2019. Backbone cannot be trained at once: Rolling back to pre-trained network for person re-identification. In *AAAI*, volume 33, 8859–8867.

Shen, Y.; Li, H.; Yi, S.; Chen, D.; and Wang, X. 2018. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 486–504.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 480–496.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. *ICLR* .

Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2018. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 365–381.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 79–88.

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning context graph for person search. In *CVPR*, 2158–2167.

Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; and Chen, Z. 2020. Relation-Aware Global Attention for Person Re-identification. In *CVPR*, 3186–3195.

Zhao, R.; Ouyang, W.; and Wang, X. 2014. Learning mid-level filters for person re-identification. In *CVPR*, 144–151.

Zhao, S.; Gao, C.; Zhang, J.; Cheng, H.; Han, C.; and Jiang, X. 2020. Do Not Disturb Me: Person Re-identification Under the Interference of Other Pedestrians. In *ECCV*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*, 1116–1124.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2138–2147.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM TOMM* 14(1): 1–20.

Zhou, J.; Cui, G.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; and Li, C. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* .

Zhou, J.; Su, B.; and Wu, Y. 2020. Online Joint Multi-Metric Adaptation from Frequent Sharing-Subset Mining for Person Re-Identification. In *CVPR*, 2909–2918.