Person Re-identification with Deep Similarity-Guided Graph Neural Network

Yantao Shen[1], Hongsheng Li[1(✉)], Shuai Yi[2], Dapeng Chen[1], and Xiaogang Wang[1]

[1] CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong, Hong Kong, People's Republic of China
`{ytshen,hsli,dpchen}@ee.cuhk.edu.hk`
[2] SenseTime Research, Hong Kong, People's Republic of China
`yishuai@sensetime.com`

Abstract. The person re-identification task requires to robustly estimate visual similarities between person images. However, existing person re-identification models mostly estimate the similarities of different image pairs of probe and gallery images independently while ignores the relationship information between different probe-gallery pairs. As a result, the similarity estimation of some hard samples might not be accurate. In this paper, we propose a novel deep learning framework, named Similarity-Guided Graph Neural Network (SGGNN) to overcome such limitations. Given a probe image and several gallery images, SGGNN creates a graph to represent the pairwise relationships between probe-gallery pairs (nodes) and utilizes such relationships to update the probe-gallery relation features in an end-to-end manner. Accurate similarity estimation can be achieved by using such updated probe-gallery relation features for prediction. The input features for nodes on the graph are the relation features of different probe-gallery image pairs. The probe-gallery relation feature updating is then performed by the messages passing in SGGNN, which takes other nodes' information into account for similarity estimation. Different from conventional GNN approaches, SGGNN learns the edge weights with rich labels of gallery instance pairs directly, which provides relation fusion more precise information. The effectiveness of our proposed method is validated on three public person re-identification datasets.

Keywords: Deep learning · Person re-identification
Graph Neural Networks

1 Introduction

Person re-identification is a challenging problem, which aims at finding the person images of interest in a set of images across different cameras. It plays a significant role in the intelligent surveillance systems.
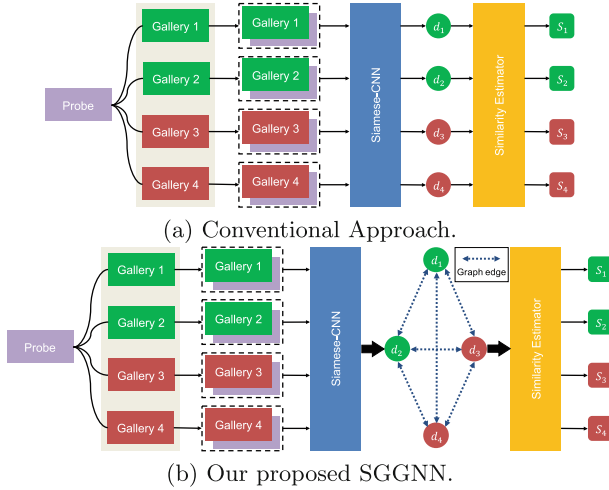
(a) Conventional Approach.

(b) Our proposed SGGNN.

Fig. 1. Illustration of our Proposed SGGNN method and conventional person re-identification approach. (a) The pipeline of conventional person re-identification approach, the pairwise relationships between different probe-gallery pairs are ignored. The similarity score of each probe-gallery pair d_i $(i = 1, 2, 3, 4)$ is estimated individually. (b) Our proposed SGGNN approach, pairwise relationships between different probe-gallery pairs are involved with deeply learned message passing on a graph for more accurate similarity estimation.

To enhance the re-identification performance, most existing approaches attempt to learn discriminative features or design various metric distances for better measuring the similarities between person image pairs. In recent years, witness the success of deep learning based approaches for various tasks of computer vision [12,17,20,25,31,39,51,59,62,63,67], a large number of deep learning methods were proposed for person re-identification [37,40,64,81]. Most of these deep learning based approaches utilized Convolutional Neural Network (CNN) to learn robust and discriminative features. In the mean time, metric learning methods were also proposed [3,4,72] to generate relatively small feature distances between images of same identity and large feature distances between those of different identities.

However, most of these approaches only consider the pairwise similarity while ignore the internal similarities among the images of the whole set. For instance, when we attempt to estimate the similarity score between a probe image and a gallery image, most feature learning and metric learning approaches only consider the pairwise relationship between this single probe-gallery image pair in both training and testing stages. Other relations among different pairs of images are ignored. As a result, some hard positive or hard negative pairs are difficult to obtain proper similarity scores since only limited relationship information among samples is utilized for similarity estimation.

To overcome such limitation, we need to discover the valuable internal similarities among the image set, especially for the similarities among the gallery set. One possible solution is utilizing manifold learning [2,42], which considers the similarities of each pair of images in the set. It maps images into a manifold with more smooth local geometry. Beyond the manifold learning methods, re-ranking approaches [16,70,78] were also utilized for refining the ranking result by integrating similarities between top-ranked gallery images. However, both manifold learning and re-ranking approaches have two major limitations: (1) most manifold learning and re-ranking approaches are unsupervised, which could not fully exploit the provided training data label into the learning process. (2) These two kinds of approaches could not benefit feature learning since they are not involved in training process.

Recently, *Graph Neural Network* (GNN) [6,18,23,45] draws increasing attention due to its ability of generalizing neural networks for data with graph structures. The GNN propagates messages on a graph structure. After message traversal on the graph, node's final representations are obtained from its own as well as other node's information, and are then utilized for node classification. GNN has achieved huge success in many research fields, such as text classification [13], image classification [6,46], and human action recognition [66]. Compared with manifold learning and re-ranking, GNN incorporates graph computation into the neural networks learning, which makes the training end-to-end and benefits learning the feature representation.

In this paper, we propose a novel deep learning framework for person re-identification, named Similarity-Guided Graph Neural Network (SGGNN). SGGNN incorporates graph computation in both training and testing stages of deep networks for obtaining robust similarity estimations and discriminative feature representations. Given a mini-batch consisting of several probe images and gallery images, SGGNN will first learn initial visual features for each image (*e.g.*, global average pooled features from ResNet-50 [17].) with the pairwise relation supervisions. After that, each pair of probe-gallery images will be treated as a node on the graph, which is responsible for generating similarity score of this pair. To fully utilize pairwise relations between other pairs (nodes) of images, deeply learned messages are propagated among nodes to update and refine the pairwise relation features associated with each node. Unlike most previous GNNs' designs, in SGGNN, the weights for feature fusion are determined by similarity scores by gallery image pairs, which are directly supervised by training labels. With these similarity guided feature fusion weights, SGGNN will fully exploit the valuable label information to generate discriminative person image features and obtain robust similarity estimations for probe-gallery image pairs.

The main contribution of this paper is two-fold. (1) We propose a novel Similarity Guided Graph Neural Network (SGGNN) for person re-identification, which could be trained end-to-end. Unlike most existing methods, which utilize inter-gallery-image relations between samples in the post-processing stage, SGGNN incorporates the inter-gallery-image relations in the training stage to enhance feature learning process. As a result, more discriminative and accurate

person image feature representations could be learned. (2) Different from most Graph Neural Network (GNN) approaches, SGGNN exploits the training label supervision for learning more accurate feature fusion weights for updating the nodes' features. This similarity guided manner ensures the feature fusion weights to be more precise and conduct more reasonable feature fusion. The effectiveness of our proposed method is verified by extensive experiments on three large person re-identification datasets.

2 Related Work

2.1 Person Re-identification

Person re-identification is an active research topic, which gains increasing attention from both academia and industry in recent years. The mainstream approaches for person re-identification either try to obtain discriminative and robust feature [1,7,8,10,21,28,35,54–56,58,60,61,71] for representing person image or design a proper metric distance for measuring similarity between person images [3,4,41,47,72]. For feature learning, Yi *et al.* [71] introduced a Siamese-CNN for person re-identification. Li *et al.* [28] proposed a novel filter pairing neural network, which could jointly handle feature learning, misalignment, and classification in an end-to-end manner. Ahmed *et al.* [1] introduced a model called Cross-Input Neighbourhood Difference CNN model, which compares image features in each patch of one input image to the other image's patch. Su *et al.* [60] incorporated pose information into person re-identification. The pose estimation algorithm are utilized for part extraction. Then the original global image and the transformed part images are fed into a CNN simultaneously for prediction. Shen *et al.* [57] utilized kronecker-product matching for person feature maps alignment. For metric learning, Paisitkriangkrai *et al.* [47] introduced an approach aims at learning the weights of different metric distance functions by optimizing the relative distance among triplet samples and maximizing the averaged rank-k accuracies. Bak *et al.* [3] proposed to learn metrics for 2D patches of person image. Yu *et al.* [72] introduced an unsupervised person re-ID model, which aims at learning an asymmetric metric on cross-view person images.

Besides feature learning and metric learning, manifold learning [2,42] and re-rank approaches [16,69,70,78] are also utilized for enhancing the performance of person re-identification model, Bai *et al.* [2] introduced Supervised Smoothed Manifold, which aims to estimating the context of other pairs of person image thus the learned relationships with between samples are smooth on the manifold. Loy *et al.* [42] introduced manifold ranking for revealing manifold structure by plenty of gallery images. Zhong *et al.* [78] utilized k-reciprocal encoding to refine the ranking list result by exploiting relationships between top rank gallery instances for a probe sample. Kodirov *et al.* [24] introduced graph regularised dictionary learning for person re-identification. Most of these approaches are conducted in the post-process stage and the visual features of person images could not be benefited from these post-processing approaches.

2.2   Graph for Machine Learning

In several machine learning research areas, input data could be naturally represented as graph structure, such as natural language processing [38,44], human pose estimation [11,66,68], visual relationship detection [32], and image classification [48,50]. In [53], Scarselli *et al.* divided machine learning models into two classes due to different application objectives on graph data structure, named *node-focused* and *graph-focused* application. For *graph-focused* application, the mapping function takes the whole graph data G as the input. One simple example for *graph-focused* application is to classify the image [48], where the image is represented by a region adjacency graph. For *node-focused* application, the inputs of mapping function are the nodes on the graph. Each node on the graph will represent a sample in the dataset and the edge weights will be determined by the relationships between samples. After the message propagation among different nodes (samples), the mapping function will output the classification or regression results of each node. One typical example for *node-focused* application is graph based image segmentation [36,76], which takes pixels of image as nodes and try to minimize the total energy function for segmentation prediction of each pixel. Another example for *node-focused* application is object detection [5], the input nodes are features of the proposals in a input image.

2.3   Graph Neural Network

Scarselli *et al.* [53] introduced Graph Neural Network (GNN), which is an extension for recursive neural networks and random walk models for graph structure data. It could be applied for both graph-focused or node-focused data without any pre or post-processing steps, which means that it can be trained end-to-end. In recent years, extending CNN to graph data structure received increased attention [6,13,18,23,33,45,66], Bruna *et al.* [6] proposed two constructions of deep convolutional networks on graphs (GCN), one is based on the spectrum of graph Laplacian, which is called spectral construction. Another is spatial construction, which extends properties of convolutional filters to general graphs. Yan *et al.* [66] exploited spatial construction GCN for human action recognition. Different from most existing GNN approaches, our proposed approach exploits the training data label supervision for generating more accurate feature fusion weights in the graph message passing.

3   Method

To evaluate the algorithms for person re-identification, the test dataset is usually divided into two parts: a probe set and a gallery set. Given an image pair of a probe and a gallery images, the person re-identification models aims at robustly determining visual similarities between probe-gallery image pairs. In the previous common settings, among a mini-batch, different image pairs of probe and gallery images are evaluated individually, *i.e.*, the estimated similarity between a pair of

images will not be influenced by other pairs. However, the similarities between different gallery images are valuable for refining similarity estimation between the probe and gallery. Our proposed approach is proposed to better utilize such information to improve feature learning and is illustrated in Fig. 1. It takes a probe and several gallery images as inputs to create a graph with each node modeling a probe-gallery image pair. It outputs the similarity score of each probe-gallery image pair. Deeply learned messages will be propagated among nodes to update the relation features associated with each node for more accurate similarity score estimation in the end-to-end training process.

In this section, the problem formulation and node features will be discussed in Sect. 3.1. The Similarity Guided GNN (SGGNN) and deep messages propagation for person re-identification will be presented in Sect. 3.2. Finally, we will discuss the advantage of similarity guided edge weight over the conventional GNN approaches in Sect. 3.3. The implementation details will be introduced in Sect. 3.4

3.1 Graph Formulation and Node Features

In our framework, we formulate person re-identification as a *node-focused* graph application introduced in Sect. 2.2. Given a probe image and N gallery images, we construct an undirected complete graph $G(V, E)$, where $V = \{v_1, v_2, ..., v_N\}$ denotes the set of nodes. Each node represents a pair of probe-gallery images. Our goal is to estimate the similarity score for each probe-gallery image pair and therefore treat the re-identification problem as a node classification problem. Generally, the input features for any node encodes the complex relations between its corresponding probe-gallery image pair.

In this work, we adopt a simple approach for obtaining input relation features to the graph nodes, which is shown in Fig. 2(a). Given a probe image and N gallery images, each input probe-gallery image pair will be fed into a Siamese-CNN for pairwise relation feature encoding. The Siamese-CNN's structure is based on the ResNet-50 [17]. To obtain the pairwise relation features, the last global average pooled features of two images from ResNet-50 are element-wise subtracted. The pairwise feature is processed by element-wise square operation and a Batch Normalization layer [19]. The processed difference features d_i ($i = 1, 2, ..., N$) encode the deep visual relations between the probe and the i-th gallery image, and are used as the input features of the i-th node on the graph. Since our task is node-wise classification, *i.e.*, estimating the similarity score of each probe-gallery pair, a naive approach would be simply feeding each node's input feature into a linear classifier to output the similarity score without considering the pairwise relationship between different nodes. For each probe-gallery image pair in the training mini-batch, a binary cross-entropy loss function could be utilized,

$$L = -\sum_{i=1}^{N} y_i \log(f(d_i)) + (1 - y_i) \log(1 - f(d_i)), \tag{1}$$

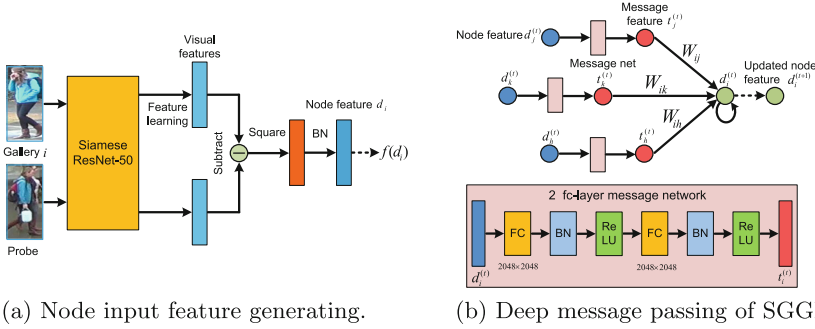(a) Node input feature generating.　　　(b) Deep message passing of SGGNN.

Fig. 2. The illustration of our base model and deep message passing of SGGNN. (a) Our base model is not only utilized for calculating the probe-gallery pairs' similarity scores, but also for obtaining the gallery-gallery similarity scores, which could be utilized for deep message passing to update the relation features of probe-gallery pairs. (b) For passing more effective information, probe-gallery relation features d_i are first fed into a 2 layer message network for feature encoding. With gallery-gallery similarity scores, the probe-gallery relation feature fusion could be deduced as a message passing and feature fusion schemes, which is defined as Eq. 4.

where $f()$ denotes a linear classifier followed by a sigmoid function. y_i denotes the ground-truth label of i-th probe-gallery image pair, with 1 representing the probe and the i-th gallery images belonging to the same identity while 0 for not.

3.2   Similarity-Guided Graph Neural Network

Obviously, the naive node classification model (Eq. (1)) ignores the valuable information among different probe-gallery pairs. For exploiting such vital information, we need to establish edges E on the graph G. In our formulation, G is fully-connected and E represents the set of relationships between different probe-gallery pairs, where W_{ij} is a scalar edge weight. It represents the relation importance between node i and node j and can be calculated as,

$$W_{ij} = \begin{cases} \frac{\exp(S(g_i,g_j))}{\sum_j \exp(S(g_i,g_j))}, & i \neq j \\ 0, & i = j \end{cases}, \tag{2}$$

where g_i and g_j are the i-th and j-th gallery images. $S()$ is a pairwise similarity estimation function, that estimates the similarity score between g_i and g_j and can be modeled in the same way as the naive node (probe-gallery image pair) classification model discussed above. Note that in SGGNN, the similarity score $S(g_i, g_j)$ of gallery-gallery pair is also learned in a supervised way with person identity labels. The purpose of setting W_{ii} to 0 is to avoid self-enhancing. To enhance the initial pairwise relation features of a node with other nodes' information, we propose to propagate deeply learned messages between all connecting nodes. The node features are then updated as a weighted addition fusion of

all input messages and the node's original features. The proposed relation feature fusion and updating is intuitive: using gallery-gallery similarity scores to guide the refinement of the probe-gallery relation features will make the relation features more discriminative and accurate, since the rich relation information among different pairs are involved. For instance, given one probe sample p and two gallery samples g_i, g_j. Suppose that (p, g_i) is a hard positive pair (node) while both (p, g_j) and (g_i, g_j) are relative easy positive pairs. Without any message passing among the nodes (p, g_i) and (p, g_j), the similarity score of (p, g_i) is unlikely to be high. However, if we utilize the similarity of pair (g_i, g_j) to guide the refinement of the relation features of the hard positive pair (p, g_i), the refined features of (p, g_i) will lead to a more proper similarity score. This relation feature fusion could be deduced as a message passing and feature fusion scheme.

Before message passing begins, each node first encodes a deep message for sending to other nodes that are connected to it. The nodes' input relation features d_i are fed into a message network with 2 fully-connected layers with BN and ReLU to generate deep message t_i, which is illustrated in Fig. 2(b). This process learns more suitable messages for node relation feature updating,

$$t_i = F(d_i) \quad \text{for } i = 1, 2, ..., N, \tag{3}$$

where F denotes the 2 FC-layer subnetwork for learning deep messages for propagation.

After obtaining the edge weights W_{ij} and deep message t_i from each node, the updating scheme of node relation feature d_i could be formulated as

$$d_i^{(1)} = (1 - \alpha)d_i^{(0)} + \alpha \sum_{j=1}^{N} W_{ij} t_j^{(0)} \quad \text{for } i = 1, 2, ..., N, \tag{4}$$

where $d_i^{(1)}$ denotes the i-th refined relation feature, $d_i^{(0)}$ denotes the i-th input relation feature and $t_j^{(0)}$ denotes the deep message from node j. α represents the weighting parameter that balances fusion feature and original feature.

Noted that such relation feature weighted fusion could be performed iteratively as follows,

$$d_i^{(t)} = (1 - \alpha)d_i^{(t-1)} + \alpha \sum_{j=1}^{N} W_{ij} t_j^{(t-1)} \quad \text{for } i = 1, 2, ..., N, \tag{5}$$

where t is the iteration number. The refined relation feature $d_i^{(t)}$ could substitute then relation feature d_i in Eq. (1) for loss computation and training the SGGNN. For training, Eq. (5) can be unrolled via back propagation through structure.

In practice, we found that the performance gap between iterative feature updating of multiple iterations and updating for one iteration is negligible. So we adopt Eq. (4) as our relation feature fusion in both training and testing stages. After relation feature updating, we feed the relation features of probe-gallery image pairs to a linear classifier with sigmoid function for obtaining the similarity score and trained with the same binary cross-entropy loss (Eq. (1)).

3.3   Relations to Conventional GNN

In our proposed SGGNN model, the similarities among gallery images are served as fusion weights on the graph for nodes' feature fusion and updating. These similarities are vital for refining the probe-gallery relation features. In conventional GNN [45,66] models, the feature fusion weights are usually modeled as a nonlinear function $h(d_i, d_j)$ that measures compatibility between two nodes d_i and d_j. The feature updating will be

$$d_i^{(t)} = (1 - \alpha)d_i^{(t-1)} + \alpha \sum_{j=1}^{N} h(d_i, d_j)t_j^{(t-1)} \quad \text{for } i = 1, 2, ..., N. \tag{6}$$

They lack directly label supervision and are only indirectly learned via back-propagation errors. However, in our case, such a strategy does not fully utilize the similarity ground-truth between gallery images. To overcome such limitation, we propose to use similarity scores $S(g_i, g_j)$ between gallery images g_i and g_j with directly training label supervision to serve as the node feature fusion weights in Eq. (4). Compared with conventional setting of GNN Eq. (6), these direct and rich supervisions of gallery-gallery similarity could provide feature fusion with more accurate information.

3.4   Implementation Details

Our proposed SGGNN is based on ResNet-50 [17] pretrained on ImageNet [14]. The input images are all resized to 256×128. Random flipping and random erasing [79] are utilized for data augmentation. We will first pretrain the base Siamese CNN model, we adopt an initial learning rate of 0.01 on all three datasets and reduce the learning rate by 10 times after 50 epochs. The learning rate is then fixed for another 50 training epochs. The weights of linear classifier for obtaining the gallery-gallery similarities is initialized with the weights of linear classifier we trained in the base model pretraining stage. To construct each mini-batch as a combination of a probe set and a gallery set, we randomly sample images according to their identities. First we randomly choose M identities in each mini-batch. For each identity, we randomly choose K images belonging to this identity. Among these K images of one person, we randomly choose one of them as the probe image and leave the rest of them as gallery images. As a result, a $K \times M$ sized mini-batch consists of a size K probe set and a size $K \times (M - 1)$ gallery set. In the training stage, K is set to 4 and M is set to 48, which results in a mini-batch size of 192. In the testing stage, for each probe image, we first utilize $l2$ distance between probe image feature and gallery image features by the trained ResNet-50 in our SGGNN to obtain the top-100 gallery images, then we use SGGNN for obtaining the final similarity scores. We will go though all the identities in each training epoch and Adam algorithm [22] is utilized for optimization.

We then finetune the overall SGGNN model end-to-end, the input node features for overall model are the subtracted features of base model. Note that for

gallery-gallery similarity estimation $S(g_i, g_j)$, the rich labels of gallery images are also used as training supervision. we train the overall network with a learning rate of 10^{-4} for another 50 epochs and the balancing weight α is set to 0.9.

4 Experiments

4.1 Datasets and Evaluation Metrics

To validate the effectiveness of our proposed approach for person re-identification. The experiments and ablation study are conducted on three large public datasets.

CUHK03 [28] is a person re-identification dataset, which contains 14,097 images of 1,467 person captured by two cameras from the campus. We utilize its manually annotated images in this work.

Market-1501 [75] is a large-scale dataset, which contains multi-view person images for each identity. It consists of 12,936 images for training and 19,732 images for testing. The test set is divided into a gallery set that contains 16,483 images and a probe set that contains 3,249 images. There are totally 1501 identities in this dataset and all the person images are obtained by DPM detector [15].

DukeMTMC [52] is collected from campus with 8 cameras, it originally contains more than 2,000,000 manually annotated frames. There are some extensions for DukeMTMC dataset for person re-identification task. In this paper, we follow the setting of [77]. It utilizes 1404 identities, which appear in more than two cameras. The training set consists of 16,522 images with 702 identities and test set contains 19,989 images with 702 identities.

We adopt mean average precision (mAP) and CMC top-1, top-5, and top-10 accuracies as evaluation metrics. For each dataset, we just adopt the original evaluation protocol that the dataset provides. In the experiments, the query type is single query.

4.2 Comparison with State-of-the-art Methods

Results on CUHK03 Dataset. The results of our proposed method and other state-of-the-art methods are represented in Table 1. The mAP and top-1 accuracy of our proposed method are 94.3% and 95.3%, respectively. Our proposed method outperforms all the compared methods.

Quadruplet Loss [9] is modified based on triplet loss. It aims at obtaining correct orders for input pairs and pushing away negative pairs from positive pairs. Our proposed method outperforms quadruplet loss 19.8% in terms of top-1 accuracy. OIM Loss [65] maintains a look-up table. It compares distances between mini-batch samples and all the entries in the table. to learn features of person image. Our approach improves OIM Loss by 21.8% and 17.8% in terms of mAP and CMC top-1 accuracy. SpindleNet [73] considers body structure information for person re-identification. It incorporates body region features and features from different semantic levels for person re-identification. Compared with SpindleNet, our proposed method increases 6.8% for top-1 accuracy.

Table 1. mAP, top-1, top-5, and top-10 accuracies by compared methods on the CUHK03 dataset [28].

Methods	Conference	CUHK03 [28]			
		mAP	top-1	top-5	top-10
Quadruplet Loss [9]	CVPR 2017	-	75.5	95.2	99.2
OIM Loss [65]	CVPR 2017	72.5	77.5	-	-
SpindleNet [73]	CVPR 2017	-	88.5	97.8	98.6
MSCAN [26]	CVPR 2017	-	74.2	94.3	97.5
SSM [2]	CVPR 2017	-	76.6	94.6	98.0
k-reciprocal [78]	CVPR 2017	67.6	61.6	-	-
VI+LSRO [77]	ICCV 2017	87.4	84.6	97.6	98.9
SVDNet [61]	ICCV 2017	84.8	81.8	95.2	97.2
OL-MANS [80]	ICCV 2017	-	61.7	88.4	95.2
Pose Driven [60]	ICCV 2017	-	88.7	98.6	**99.6**
Part Aligned [74]	ICCV 2017	-	85.4	97.6	99.4
HydraPlus-Net [39]	ICCV 2017	-	91.8	98.4	99.1
MuDeep [49]	ICCV 2017	-	76.3	96.0	98.4
JLML [29]	IJCAI 2017	-	83.2	98.0	99.4
MC-PPMN [43]	AAAI 2018	-	86.4	98.5	**99.6**
Proposed SGGNN		**94.3**	**95.3**	**99.1**	**99.6**

MSCAN [27] stands for Multi-Scale ContextAware Network. It adopts multiple convolution kernels with different receptive fields to obtain multiple feature maps. The dilated convolution is utilized for decreasing the correlations among convolution kernels. Our proposed method gains 21.1% in terms of top-1 accuracy. SSM stands for Smoothed Supervised Manifold [2]. This approach tries to obtain the underlying manifold structure by estimating the similarity between two images in the context of other pairs of images in the post-processing stage, while the proposed SGGNN utilizes instance relation information in both training and testing stages. SGGNN outperforms SSM approach by 18.7% in terms of top-1 accuracy. k-reciprocal [78] utilized gallery-gallery similarities in the testing stage and uses a smoothed Jaccard distance for refining the ranking results. In contrast, SGGNN exploits the gallery-gallery information in the training stage for feature learning. As a result, SGGNN gains 26.7% and 33.7% increase in terms of mAP and top-1 accuracy.

Results on Market-1501 Dataset. On Market-1501 dataset, our proposed methods outperforms significantly state-of-the-art methods. SGGNN achieves mAP of 82.8% and top-1 accuracy of 92.3% on Market-1501 dataset. The results are shown in Table 2.

Table 2. mAP, top-1, top-5, and top-10 accuracies of compared methods on the Market-1501 dataset [75].

Methods	Reference	Market-1501 [75]			
		mAP	top-1	top-5	top-10
OIM Loss [65]	CVPR 2017	60.9	82.1	-	-
SpindleNet [73]	CVPR 2017	-	76.9	91.5	94.6
MSCAN [26]	CVPR 2017	53.1	76.3	-	-
SSM [2]	CVPR 2017	68.8	82.2	-	-
k-reciprocal [78]	CVPR 2017	63.6	77.1	-	-
Point 2 Set [81]	CVPR 2017	44.3	70.7	-	-
CADL [35]	CVPR 2017	47.1	73.8	-	-
VI+LSRO [77]	ICCV 2017	66.1	84.0	-	-
SVDNet [61]	ICCV 2017	62.1	82.3	92.3	95.2
OL-MANS [80]	ICCV 2017	-	60.7	-	-
Pose Driven [60]	ICCV 2017	63.4	84.1	92.7	94.9
Part Aligned [74]	ICCV 2017	63.4	81.0	92.0	94.7
HydraPlus-Net [39]	ICCV 2017	-	76.9	91.3	94.5
JLML [29]	IJCAI 2017	65.5	85.1	-	-
HA-CNN [30]	CVPR 2018	75.7	91.2	-	-
Proposed SGGNN		**82.8**	**92.3**	**96.1**	**97.4**

HydraPlus-Net [39] is proposed for better exploiting the global and local contents with multi-level feature fusion of a person image. Our proposed method outperforms HydraPlus-Net by 15.4 for top-1 accuracy. JLML [29] stands for Joint Learning of Multi-Loss. JLML learns both global and local discriminative features in different context and exploits complementary advantages jointly. Compared with JLML, our proposed method gains 17.3 and 7.2 in terms of mAP and top-1 accuracy. HA-CNN [30] attempts to learn hard region-level and soft pixel-level attention simultaneously with arbitrary person bounding boxes and person image features. The proposed SGGNN outperforms HA-CNN by 7.1% and 1.1% with respect to mAP and top-1 accuracy.

Results on DukeMTMC Dataset. In Table 3, we illustrate the performance of our proposed SGGNN and other state-of-the-art methods on DukeMTMC [52]. Our method outperforms all compared approaches. Besides approaches such as OIM Loss and SVDNet, which have been introduced previously, our method also outperforms Basel+LSRO, which integrates GAN generated data and ACRN that incorporates person of attributes for person re-identification significantly. These results illustrate the effectiveness of our proposed approach.

Table 3. mAP, top-1, top-5, and top-10 accuracies by compared methods on the DukeMTMC dataset [52].

Methods	Reference	DukeMTMC [52]			
		mAP	top-1	top-5	top-10
BoW+KISSME [75]	ICCV 2015	12.2	25.1	-	-
LOMO+XQDA [34]	CVPR 2015	17.0	30.8	-	-
ACRN [54]	CVPRW 2017	52.0	72.6	84.8	88.9
OIM Loss [65]	CVPR 2017	47.4	68.1	-	-
Basel.+LSRO [77]	ICCV 2017	47.1	67.7	-	-
SVDNet [61]	ICCV 2017	56.8	76.7	86.4	89.9
Proposed SGGNN		**68.2**	**81.1**	**88.4**	**91.2**

4.3   Ablation Study

To further investigate the validity of SGGNN, we also conduct a series of ablation studies on all three datasets. Results are shown in Table 4.

We treat the siamese CNN model that directly estimates pairwise similarities from initial node features introduced in Sect. 3.1 as the base model. We utilize the same base model and compare with other approaches that also take inter-gallery image relations in the testing stage for comparison. We conduct k-reciprocal re-ranking [78] with the image visual features learned by our base model. Compared with SGGNN approach, The mAP of k-reciprocal approach drops by 4.3%, 4.4%, 3.5% for Market-1501, CUHK03, and DukeMTMC datasets. The top-1 accuracy also drops by 0.8%, 3.1%, 1.2% respectively. Except for the visual features, base model could also provides us raw similarity scores of probe-gallery pairs and gallery-gallery pairs. A random walk [2] operation could be conducted to refine the probe-gallery similarity scores with gallery-gallery similarity scores with a closed-form equation. Compared with our method, The performance of random walk drops by 3.6%, 4.1%, and 2.2% in terms of mAP, 0.8%, 3.0%, and 0.8% in terms of top-1 accuracy. Such results illustrate the effectiveness of end-to-end training with deeply learned message passing within SGGNN. We also validate the importance of learning visual feature fusion weight with gallery-gallery similarities guidance. In Sect. 3.3, we have introduced that in the conventional GNN, the compatibility between two nodes d_i and d_j, $h(d_i, d_j)$ is calculated by a non-linear function, inner product function without direct gallery-gallery supervision. We therefore remove the directly gallery-gallery supervisions and train the model with weight fusion approach in Eq. (6), denoted by *Base Model + SGGNN w/o SG*. The performance drops by 1.6%, 1.6%, and 0.9% in terms of mAP. The top-1 accuracies drops 1.7%, 2.6%, and 0.6% compared with our SGGNN approach, which illustrates the importance of involving rich gallery-gallery labels in the training stage.

To demonstrate that our proposed model SGGNN also learns better visual features by considering all probe-gallery relations, we evaluate the re-

identification performance by directly calculating the l_2 distance between different images' visual feature vectors outputted by our trained ResNet-50 model on three datasets. The results by visual features learned with base model and the conventional GNN approach are illustrated in Table 5. Visual features by our proposed SGGNN outperforms the compared base model and conventional GNN setting significantly, which demonstrates that SGGNN also learns more discriminative and robust features.

Table 4. Ablation studies on the Market-1501 [75], CUHK03 [28] and DukeMTMC [52] datasets.

Methods	Market-1501 [75]		CUHK03 [28]		DukeMTMC [52]	
	mAP	top-1	mAP	top-1	mAP	top-1
Base Model	76.4	91.2	88.9	91.1	61.8	78.8
Base Model + k-reciprocal [78]	78.5	91.5	89.9	92.2	64.7	79.9
Base Model + random walk [2]	79.2	91.5	90.2	92.3	66.0	80.3
Base Model + SGGNN w/o SG	81.2	90.6	92.7	93.6	67.3	80.5
Base Model + SGGNN	**82.8**	**92.3**	**94.3**	**95.3**	**68.2**	**81.1**

Table 5. Performances of estimating probe-gallery similarities by l_2 feature distance on the Market-1501 [75], CUHK03 [28] and DukeMTMC [52] datasets.

Model	Market-1501 [75]		CUHK03 [28]		DukeMTMC [52]	
	mAP	top-1	mAP	top-1	mAP	top-1
Base Model	74.6	90.4	87.6	91.0	60.3	77.6
Base Model + SGGNN w/o SG	75.4	90.4	87.7	91.5	61.7	78.1
Base Model + SGGNN	**76.7**	**91.5**	**88.1**	**93.6**	**64.6**	**79.1**

5   Conclusion

In this paper, we propose Similarity-Guided Graph Neural Neural to incorporate the rich gallery-gallery similarity information into training process of person re-identification. Compared with our method, most previous attempts conduct the updating of probe-gallery similarity in the post-process stage, which could not benefit the learning of visual features. For conventional Graph Neural Network setting, the rich gallery-gallery similarity labels are ignored while our approach utilized all valuable labels to ensure the weighted deep message fusion is more effective. The overall performance of our approach and ablation study illustrate the effectiveness of our proposed method.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
2. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. arXiv preprint arXiv:1703.08359 (2017)
3. Bak, S., Carr, P.: Person re-identification using deformable patch metric learning. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. IEEE (2016)
4. Bak, S., Carr, P.: One-shot metric learning for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
5. Bianchini, M., Maggini, M., Sarti, L., Scarselli, F.: Recursive neural networks learn to localize faces. Pattern Recogn. Lett. **26**(12), 1885–1895 (2005)
6. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013)
7. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: 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, pp. 1169–1178 (2018)
8. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep CRF for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8649–8658 (2018)
9. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
10. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)
11. Chu, X., Ouyang, W., Wang, X., et al.: CRF-CNN: modeling structured information in human pose estimation. In: Advances in Neural Information Processing Systems, pp. 316–324 (2016)
12. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. arXiv preprint arXiv:1702.07432, 1(2) (2017)
13. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems, pp. 3844–3852 (2016)
14. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 248–255. IEEE (2009)

15. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
16. Garcia, J., Martinel, N., Micheloni, C., Gardel, A.: Person re-identification ranking optimisation by discriminant context information analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1305–1313 (2015)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Henaff, M., Bruna, J., LeCun, Y.: Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163 (2015)
19. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
20. Kang, K., et al.: Object detection in videos with tubelet proposal networks. arXiv preprint arXiv:1702.06355 (2017)
21. Karaman, S., Lisanti, G., Bagdanov, A.D., Del Bimbo, A.: Leveraging local neighborhood topology for large scale person re-identification. Pattern Recogn. **47**(12), 3767–3778 (2014)
22. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
24. Kodirov, Elyor, Xiang, Tao, Fu, Zhenyong, Gong, Shaogang: Person re-identification by unsupervised ℓ_1 graph learning. In: Leibe, Bastian, Matas, Jiri, Sebe, Nicu, Welling, Max (eds.) ECCV 2016. LNCS, vol. 9905, pp. 178–195. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_11
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
26. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
27. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 384–393 (2017)
28. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
29. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. arXiv preprint arXiv:1705.04724 (2017)
30. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. arXiv preprint arXiv:1802.08122 (2018)
31. Li, Y., Ouyang, W., Wang, X., Tang, X.: ViP-CNN: visual phrase guided convolutional neural network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7244–7253. IEEE (2017)
32. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017)

33. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph LSTM. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 125–143. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_8

34. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)

35. Lin, J., Ren, L., Lu, J., Feng, J., Zhou, J.: Consistent-aware deep learning for person re-identification in a camera network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017

36. Liu, F., Lin, G., Shen, C.: CRF learning with CNN features for image segmentation. Pattern Recogn. **48**(10), 2983–2992 (2015)

37. Liu, J., et al.: Multi-scale triplet CNN for person re-identification. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 192–196. ACM (2016)

38. Liu, X., Li, H., Shao, J., Chen, D., Wang, X.: Show, tell and discriminate: image captioning by self-retrieval with partially labeled data. arXiv preprint arXiv:1803.08314 (2018)

39. Liu, X., et al.: Hydraplus-net: attentive deep features for pedestrian analysis. In: The IEEE International Conference on Computer Vision (ICCV), Oct 2017

40. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR, vol. 2, p. 8 (2017)

41. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), Oct 2017

42. Loy, C.C., Liu, C., Gong, S.: Person re-identification by manifold ranking. In: 2013 20th IEEE International Conference on Image Processing (ICIP), pp. 3567–3571. IEEE (2013)

43. Mao, C., Li, Y., Zhang, Y., Zhang, Z., Li, X.: Multi-channel pyramid person matching network for person re-identification. arXiv preprint arXiv:1803.02558 (2018)

44. Mills, M.T., Bourbakis, N.G.: Graph-based methods for natural language processing and understandinga survey and analysis. IEEE Trans. Syst. Man Cybern. Syst. **44**(1), 59–71 (2014)

45. Niepert, M., Ahmed, M., Kutzkov, K.: Learning convolutional neural networks for graphs. In: International Conference on Machine Learning, pp. 2014–2023 (2016)

46. van der Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759 (2016)

47. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1855 (2015)

48. Pavlidis, T.: Structural Pattern Recognition, vol. 1. Springer (2013)

49. Qian, X., Fu, Y., Jiang, Y.-G., Xiang, T., Xue, X.: Multi-scale deep learning architectures for person re-identification. arXiv preprint arXiv:1709.05165 (2017)

50. Quek, A., Wang, Z., Zhang, J., Feng, D.: Structural image classification with graph neural networks. In: 2011 International Conference on Digital Image Computing Techniques and Applications (DICTA), pp. 416–421. IEEE (2011)

51. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)

52. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking (2016)

53. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Netw. **20**(1), 61–80 (2009)
54. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1435–1443. IEEE (2017)
55. Shen, Y., Li, H., Xiao, T., Yi, S., Chen, D., Wang, X.: Deep group-shuffling random walk for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2265–2274 (2018)
56. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1918–1927. IEEE (2017)
57. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: End-to-end deep Kronecker-product matching for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6886–6895 (2018)
58. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. arXiv preprint arXiv:1711.08766 (2017)
59. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
60. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), Oct 2017
61. Sun, Y., Zheng, L., Deng, W., Wang, S.: SVDNet for pedestrian retrieval. In: The IEEE International Conference on Computer Vision (ICCV), Oct 2017
62. Wu, F., Li, S., Zhao, T., Ngan, K.N.: Model-based face reconstruction using sift flow registration and spherical harmonics. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 1774–1779. IEEE (2016)
63. Wu, F., Li, S., Zhao, T., Ngan, K.N.: 3D facial expression reconstruction using cascaded regression. arXiv preprint arXiv:1712.03491 (2017)
64. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1249–1258 (2016)
65. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the CVPR (2017)
66. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. arXiv preprint arXiv:1801.07455 (2018)
67. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: The IEEE International Conference on Computer Vision (ICCV), vol. 2 (2017)
68. Yang, W., Ouyang, W., Li, H., Wang, X.: End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3073–3082 (2016)
69. Ye, M., Liang, C., Wang, Z., Leng, Q., Chen, J.: Ranking optimization for person re-identification via similarity and dissimilarity. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1239–1242. ACM (2015)
70. Ye, M., et al.: Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Trans. Multimed. **18**(12), 2553–2566 (2016)

71. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 34–39. IEEE (2014)

72. Yu, H.-X., Wu, A., Zheng, W.-S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), Oct 2017

73. Zhao, H., et al.: Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1077–1085 (2017)

74. Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. arXiv preprint arXiv:1707.07256 (2017)

75. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)

76. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1537 (2015)

77. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision (2017)

78. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017

79. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)

80. Zhou, J., Yu, P., Tang, W., Wu, Y.: Efficient online local metric adaptation via negative samples for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), Oct 2017

81. Zhou, S., Wang, J., Wang, J., Gong, Y., Zheng, N.: Point to set similarity based deep feature learning for person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017