Masked Graph Attention Network for Person Re-identification

Liqiang Bao[1], Bingpeng Ma[1], Hong Chang[2], Xilin Chen[2,1]
[1]University of Chinese Academy of Sciences, Beijing 100049, China.
[2]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Bejing, 100190, China.
baoliqiang16@mails.ucas.ac.cn, bpma@ucas.ac.cn, {changhong, xlchen}@ict.ac.cn

Abstract

The mainstream methods for person re-identification (ReID) mainly focus on the correspondence between individual sample images and labels, while ignoring rich global mutual information resides in the whole sample set. We propose a method called Masked Graph Attention Network (MGAT) to address this problem. MGAT operates on the complete graph constructed with the extracted features, where nodes are able to directionally attend over other nodes' features under the guidance of label information in the form of mask matrix. By using MGAT module, the previously neglected global mutual information is exploited to generate an optimized feature space with more discriminant power. Meanwhile, we propose to feedback the optimization information learned by MGAT module to the feature-embedding network to enhance the mapping capability, thus avoiding the difficulty to handle large-scale graphs in testing phase. To evaluate our method, we conduct experiments on three commonly used ReID datasets. The results show that our method outperforms most mainstream methods, and is highly comparable to the state-of-the-art method.

1. Introduction

Person Re-identification (ReID) aims at matching pedestrians in different tracks from multiple non-overlapping cameras. This task has drawn increasing attention in recent years due to its importance in applications, such as surveillance [29], activity analysis [19] and tracking [38]. Important though it is, this task remains a challenging problem because of complex variations in camera viewpoints, human poses, lighting, occlusions, and background clutter.

However, as illustrated in Figure 1, most current mainstream methods learn the feature embedding network by independently estimating the class label for single feature with identification loss, while neglecting the rich mutual information resides in the whole graph constructed by all the features. In other words, they only pay attention to the classification characteristic of features which indicates to what extent the features correspond to their correct labels, while the clustering characteristic of features fails to receive as much attention, which indicates how greatly the features of the same class are clustered and the features of different classes are separated. Whereas the discriminant analysis shows that more discriminative features requires better clustering characteristic, current methods rarely take it into consideration.

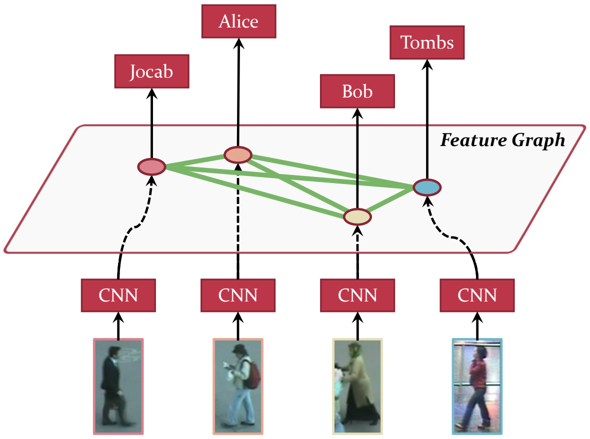


Figure 1: Illustration of node-focuesd ReID. Most existing approaches only used the vertical node classification pipeline (black arrows), but omitted the mutual information transfer (green lines) between the nodes in the graph formed by features.

There are several existing methods attempting to overcome this defect, such as manifold learning [3, 18] and re-ranking [44, 9, 35]. Both of them are capable of using mutual information to improve the clustering characteristic of feature space. But as Yantao *et al*. conclude in [23], they all have two major limitations: One is that most manifold learning and re-ranking approaches are weakly supervised or unsupervised, which could not fully exploit the provided training labels into the learning process. The other is that these two kinds of approaches could not benefit feature learning since they are not involved in training process.

The burgeoning graph attention networks (GATs) [26] shows its potential to exploit the mutual information in nodes to improve the clustering characteristic, due to its intrinsic power to aggregate information from other nodes' features. The GATs successfully introduced the attention mechanism into graph neural networks (GNNs) [21], by which nodes are able to attend over their neighborhoods' features and specify different weights to different nodes in a neighborhood. More importantly, it requires no computational intensive matrix operation. Nevertheless, conventional GATs only utilize the relative importance of nodes without label information, which is capable of aggregating similar nodes, but is hard to directly separate nodes of different classes.

We propose a novel extension of GATs called Masked Graph Attention Network (MGAT) to exploit the rich mutual information between features in the present paper for ReID. The heart of MGAT lies in the innovative masked attention mechanism for node updating, which is different from the conventional GATs that only aggregate similar nodes by an attention matrix. Specifically, we first reform the features learned by the feature embedding network as a complete graph. Then our MGAT uses an *attention matrix* to provide the weights for updating, and a *mask matrix* guided with label information to decide in what direction to update (i.e. pull the nodes of the same class closer or push the nodes of different classes). Thus the features finally gain an improved clustering characteristic.

The optimized output features of MGAT are directly supervised by the identification loss to guarantee the classification characteristic. Besides, the optimization information learned from MGAT is further fed back to the original features using an *optimization feedback* (OF) loss. The purpose of which is to enhance the mapping capability of the feature embedding network, so as to avoid any post or non-end-to-end process like re-ranking.

2. Related Work

2.1. Person Re-identification

Existing mainstream methods [36, 42, 20, 46, 33, 17, 24, 39, 15] usually follow the routine that first extracts discriminative and robust features from person images with identification loss, and then adopts a metric distance to match between probe and gallery sets. For image-based ReID, Li *et al*. [16] 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 neighborhood difference CNN model, which compares image features in each patch of one input image to the other images' patch. For video-based ReID, McLaughlin *et al*. [20] first extracted features with CNN from images using identification loss and then used RNN and temporal pooling to aggregate those features. Chung *et al*. [7] presented a two stream method, of which each stream was a Siamese network and superivized by identification loss, then RNN and temporal pooling were used to aggregate features. Liu *et al*. [17] used a CNN model to learn the quality for each image with classification score, and then aggregated all the frame features weighted by the quality.

Besides feature representation learning with identification loss devoted to improve the classification characteristic, there were some preliminary attempts on incorporating affinities between gallery images into the ranking process [27, 34, 35, 44, 3, 18]. First, manifold learning [3, 18, 37] and re-rank approaches [44, 34, 35, 9] are utilized to enhance the performance of person re-identification model. Bai *et al*. [3] introduced Supervised Smoothed Manifold, which aimed at estimating the context of other pairs of person images thus the learned relationships between samples are smooth on the manifold. Loy *et al*. [18] introduced manifold ranking for revealing manifold structure by plenty of gallery images. Zhong *et al*. [44] utilized k-reciprocal encoding to optimize the ranking list result by exploiting relationships between top rank gallery instances for a probe sample. Note that all the methods mentioned above are conducted as post-processing procedure during testing, which is not end-to-end trainable to optimize the feature embedding network.

Shen *et al*. [22] proposed Group-Shuffling Random Walk Network to utilize the affinity information between gallery images in both training and testing stages. The approach tried to optimize the probe-gallery (P2G) affinities based on gallery-gallery (G2G) affinity information with a simple matrix operation, which can be integrated into deep neural networks. They [23] also proposed a deep learning framework named Similarity-Guided Graph Neural Network (SGGNN) to utilize relationships between different feature pairs. The input features as nodes to the graph are the relation features of different probe-gallery image pairs, and the node updating is performed by the messages passing, which takes other nodes' information into account for similarity estimation.

2.2. Graph Attention Network

Many computer vision tasks involve data that can not be represented in a regularly used grid-like structure, like graph. GNNs were introduced in [21] as a generalization of recursive neural networks that can directly deal with a more general class of graphs. Then Bruna *et al*. [4] and Duvenaud *et al*. [8] started the research of Graph Convolu-

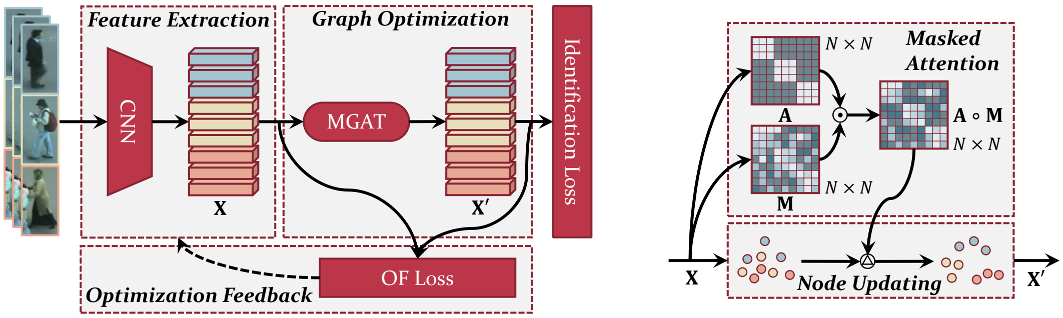


Figure 2: ***Left***: Overall depiction of MGAT integrated with CNN for ReID task. When a mini-batch of images is fed, CNN will first generate a feature set $\mathbf{X}$, which will be then optimized by MGAT to a new set $\mathbf{X}'$. OF loss will use the optimization information to further enhance the learning ability of CNN. ***Right***: Visualization of the inner mechanism of MGAT. The elementwise product of the attention matrix $\mathbf{A}$ and the mask matrix $\mathbf{M}$ represents the directional forces between nodes.

tional Networks (GCNs) in spectral and non-spectral manners respectively. Petar Velickovic *et al*. [26] introduced an attention-based architecture named Graph Attention Networks (GATs), which operate directly on graphs, leveraging masked self-attentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations. By stacking layers in which nodes are able to attend over their neighborhoods' features, thus enables implicitly specifying different weights to different nodes in a neighborhood, without requiring any kind of computationally intensive matrix operation or depending on knowing the graph structure upfront.

3. Method

In this section, we introduce our MGAT integrated with ResNet50 [10] baseline for ReID tasks. We first describe the overall network architecture and then elaborate on the design of MGAT module and the OF loss.

3.1. Overview

The pipeline is shown on the left in Figure 2. The architecture mainly consists of three components, the first is the extraction of features, what follows is the feature optimization by the proposed MGAT. The OF loss is applied to feed the leaned optimization information back to CNN (feature-embedding network).

Given a mini-batch of images, we first extract a set of features X with CNN, where each feature uniquely represents the visual information of the corresponding image. Considering the set of features as a set of nodes, we then construct a complete graph, on which each edge characterizes the similarity between connected nodes (including self-joins). Inspired by [30], the similarity function can be implemented in many ways. Then the constructed graph will be fed into the proposed MGAT to be optimized. Note that the output features X' of MGAT are directly supervised by identification loss to guarantee the classification characteristic.

We at the same time introduce the OF loss to constraint the difference between the output features and the original features. It is used to feed back the optimization information learned by MGAT to the feature-embedding network, so that the feature-embedding network is able to directly generate optimized features without applying MGAT or any post-processing methods in testing phase. The probe and gallery sets are always very large, it is inefficient or even impossible to directly process on them as graphs.

In general, the principle of the whole network architecture is to enhance the learning of feature-embedding network by utilizing the optimization information learned by the proposed MGAT, so that we can find a more discriminative feature space for ReID tasks.

3.2. Masked Graph Attention Network

MGAT is designed to address the person re-identification scenario where a lot of valuable mutual information is neglected to obtain the optimal clustering characteristic. Like the attentional structure in [26], the attention setup of MGAT also follows the work of Bahdanau *et al*. [2], but with different attention mechanism. We start with describing the input and output of MGAT, and then focus on building the interesting masked attention mechanism.

The input of MGAT is a set of features that is extracted by CNN (i.e. ResNet50 in our implementation), $X = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}, \vec{x}_i \in \mathbb{R}^d$, where N is the number of features, and d is the dimension of single feature. The proposed MGAT generates a new set of optimized features $X' = \{\vec{x}'_1, \vec{x}'_2, \cdots, \vec{x}'_N\}, \vec{x}'_i \in \mathbb{R}^{d'}$ as its output. In order to use the optimized features to further supervise the original features, we specify that the dimension of the output is the same as that of the input, meaning $d' = d$.

MGAT works on data with graph structure. Considering the set of input features X as a set of nodes, and the distances between any two nodes as a set of edges E, we construct a complete graph $G(X, E)$. In our implementation, we use the Euclidean distance to determine the edge e_{ij} between two nodes $\vec{x}_i$ and $\vec{x}_j$,

$$e_{ij} = \|\vec{x}_i - \vec{x}_j\|_2. \tag{1}$$

The core of MGAT lies in its innovative masked attention mechanism. It is specifically designed on edges to reach the goal of improving the clustering characteristic, by aggregating nodes of the same class and separating nodes of different classes based in an attentional manner. To be specific, our masked attention mechanism includes two main components, namely the *attention matrix* $\mathbf{A}$ and the *mask matrix* $\mathbf{M}$, see the visualization on the right in Figure 2.

3.2.1 Attention Matrix

The attention mechanism is usually used to reveal the relative importance between two features. In our graph, since the relation between two nodes is uniquely decided by the edge, we can simply define an attention function $f : \mathbb{R} \to \mathbb{R}$ to map edge to attention. In practice, we define the attention as:

$$\alpha_{ij} = \exp(-e_{ij}^2/\gamma). \tag{2}$$

In the above α_{ij} specifies the relative importance of the j-th node to the i-th node, and γ is a hyper parameter that helps to map the attention in a small range near zero. We can observe that the shorter the edge, the higher the attention. Note that in many implementations of GCNs, to integrate the graph structure, a node usually considers the influence of the nodes in the first-order domain adjacent to it. Yet since the graph we construct is a complete graph involving only a mini-batch of features, we can calculate the attention for each node with all other nodes to capture the global information without worrying about computational complexity. In order to make the attention with different nodes comparable, we then perform L1 normalization:

$$\alpha_{ij} = \frac{\exp(-e_{ij}^2/\gamma)}{\sum_{k \in N} \exp(-e_{ik}^2/\gamma)}. \tag{3}$$

For a mini-batch containing N images, we can get a row normalized $N \times N$ *attention matrix* $\mathbf{A}$, in which the attention values of the i-th node to all the nodes serves as the i-th row.

3.2.2 Mask Matrix

The *attention matrix* represents the mutual importance information of the nodes in the graph, conventional GCNs and GATs utilize this information to update the nodes due to the assumption that connected nodes in the graph are likely to share the same label [13]. However, this assumption might restrict modeling capacity, because it only considers similarity, but neglects the dissimilarity. And it fails to handle hard samples.

To address this problem, alone with the *attention matrix*, we introduce a *mask matrix* to decide in which direction we aggregate the nodes (to shorten or lengthen the edges). For example, we shorten the edges between nodes with the same labels, and lengthen the edge otherwise, in an attentional manner. More specifically, the N-size mini-batch we use contains M person identities with each identity has K images, in which the label distribution has the following structure

$$\Big\{\underbrace{y_1, \cdots, y_1}_{K}\Big\}, \Big\{\underbrace{y_2, \cdots, y_2}_{K}\Big\}, \cdots, \Big\{\underbrace{y_M, \cdots, y_M}_{K}\Big\}$$

where y_i is the label for i-th person identity. Element of the *mask matrix* is then formulated as:

$$\mu_{ij} = \begin{cases} 1 & \text{if } \lfloor \frac{i}{K} \rfloor = \lfloor \frac{j}{K} \rfloor, \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

In the above, $\lfloor\ \rfloor$ is the floor function, making $\mathbf{M}$ a matrix filled with M $\mathbf{1}^{(K \times K)}$ matrices aligning on the diagonal while all other elements is -1.

The *mask matrix* functions as an attention mask, when elementwisely multiplied to the *attention matrix*, it ensures that the attention values between nodes in the same class is positive, and the attention values between nodes from different classes are negative. In this way, the similarity between nodes in the same class increases (shorter edge) after node updating, while that of different classes reduces (longer edge). In brief, the *mask matrix* converts the information carried by node labels into the supervision for attention, thus to reach optimized clustering characteristic.

There is a doubt that the negative masks will damage the normalization results, but in fact the role of normalization here is to make the attention values comparable, and such operations also have the effect of weight decay.

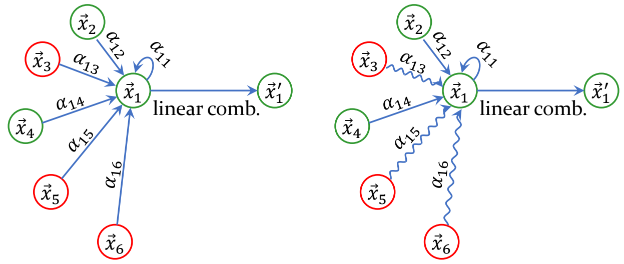
3.2.3 Node Updating

Let's represent the feature sets X and X' as $\mathbf{X}$ and $\mathbf{X}'$ in the form of matrix. While updating, to review that the conventional GATs only utilize the *attention matrix* $\mathbf{A}$, and obtain the output features of nodes by linear combination:

$$\mathbf{X}' = \mathbf{A}\mathbf{X}. \tag{5}$$

And the update for a single node $\vec{x}_i$ is:

$$\vec{x}_i' = \sum_{k \in N} \alpha_{ik} \vec{x}_k. \tag{6}$$



(a) Node updating of conventional GATs.

(b) Node updating of MGAT.

Figure 3: An comparison of the node updating process between the conventional GATs and the proposed MGAT. Circles of different colors denote different classes, and different arrow styles denote different attention processing (straight arrows for aggregation, and wavy arrows for separation). The nodes are linearly combined based on the directional forces to obtain $\vec{x}_1'$.

In our work, with the introduction of an extra *mask matrix* $\mathbf{M}$, we have got the label supervised directional information to tackle with the clustering characteristic of node features. The output has the following formulation:

$$\mathbf{X}' = (\mathbf{A} \circ \mathbf{M})\mathbf{X}, \tag{7}$$

where $\mathbf{A} \circ \mathbf{M}$ is the element-wise product of $\mathbf{A}$ and $\mathbf{M}$, called the *masked attention matrix*. Similarly, the updating for a single node is:

$$\begin{aligned} \vec{x}_i' &= \sum_{k \in N} \operatorname{sgn}(y_i, y_k)\alpha_{ik}\vec{x}_k && (8) \\ &= \sum_{k \in N_p} \alpha_{ik}\vec{x}_k - \sum_{k \in N_n} \alpha_{ik}\vec{x}_k && (9) \end{aligned}$$

where N_p is the number of nodes with the same class label as $\vec{x}_i$, N_n is the number of nodes with different class label from $\vec{x}_i$, and sgn is the sign function:

$$\operatorname{sgn}(y_i, y_j) = \begin{cases} 1, & \text{if } y_i = y_j \\ -1, & \text{otherwise} \end{cases}. \tag{10}$$

The node updating process of MGAT compared to the conventional GATs is illustrated by Figure 3. Note that the convolutional GATs use attention values to compute a linear combination of the node features corresponding to them to serve as the final output node features, while no label supervision is involved to directly separate nodes of different classes. As a comparison, given the mask information form the *mask matrix*, our proposed MGAT applies different attention processing to nodes from different classes. Intuitively, the first term of Formula 9 depicts the aggregation of the i-th node to nodes from the same class, while the second term depicts the separation of the i-th node to nodes from different classes. After such an updating process, each node is subjected to the information transmission from the surrounding environment. Thus it is expected to obtain an improved clustering characteristic, and lead to a considerable promotion on node-focused recognition task like ReID.

3.3. OF loss

As is mentioned in Section 3.1 that it is always not a good idea to process probe set and gallery set as graphs in the test phase. We propose to make it possible for CNN to directly generate optimized features by applying the OF loss. We adopt the simplest implementation by using the *mean squared error* (MSE) loss to constraint the difference between the output features of MGAT and the original features:

$$\begin{aligned} \mathcal{L}_{OF} &= \sum_{i \in N} \|\vec{x}_i' - \vec{x}_i\|_2 && (11) \\ &= \sum_{i \in N} \| \sum_{k \in N_p} (\alpha_{ik}\vec{x}_k - \vec{x}_i) - \sum_{k \in N_n} \alpha_{ik}\vec{x}_k\|_2 && (12) \end{aligned}$$

Note that the OF loss is an auxiliary component for MGAT to enhance the learning of CNN, so as to avoid the massive graph construction work in test phase, thus we don't independently study how it contribute to the final performances.

4. Experiments

To validate the effectiveness of our proposed approach for ReID, we conduct extensive experiments and ablation study on three popular video-based ReID datasets namely iLIDS-VID [28], PRID2011 [12] and MARS [42], all of them have multi-shot images for a person identity that make it possible to optimize features space. Besides, to justify that our method also works on image-based dataset, We also evaluate our method on Market1501 [43].

4.1. Datasets and metric

iLIDS-VID. iLIDS-VID dataset collected 600 trajectories for 300 identities, based on the assumption that the real ReID system should have the trajectory for each identity. The problem of extremely heavy occlusion makes it a very challenging dataset for ReID task.

PRID2011. PRID2011 dataset has 385 trajectories from camera A and 749 trajectories from camera B. Among them, only 200 persons are commonly used in ReID tasks, in that they appear in both cameras. Although some trajectories are not well-synchronized, this dataset is much easier for simple and clean backgrounds.

MARS. MARS is the first large scale video based ReID dataset. Since all bounding boxes and tracklets are generated automatically, it contains distractors and each identity may have more than one tracklets.

Market1501. Market1501 is a classic image-based ReID dataset, which contains a large number of identities and each identity has several images from six dis-joint cameras. This dataset also includes 2793 false alarms from DPM as distractors to mimic the real scenario.

Evaluation metrics. We adopt the *Cumulative Matching Characteristics* (CMC) top-1, top-5, top-10, top-20 accuracies and *Mean Average Precision* (mAP) as evaluation metrics and strictly adopt the original evaluation protocol provided by the dataset. The final results are reported as the average of "10-fold cross validation".

4.2. Implementation details

As is mentioned in Section 3, we use ResNet50-bn as our baseline. The original ResNet50 is pretrained on the ImageNet dataset and then used to initialize the modules in ResNet50-bn except for the BN layer and the customized classifier module. All the input images are resized to 256×128. For data augmentation, only random horizontal flipping is adopted.

The most important part is our training procedure, which consists of two stages. In the first stage, we train the feature embedding network (the baseline). For all datasets, we set 0.01 as the initial learning rate for the BN layer and the classifier module, and 0.001 for other pretrained modules. We reduce the learning rate by 10 times every 3 epoches, and train a total of 9 epoches. In the second stage, we enhance the learned feature embedding network by MGATs. We adopt the previously trained classifier in the first stage to classify the optimized features. For initial learning rate for all the modules and the hyper parameter γ, we empirically set them to 1e-6 and 250. For every 10 epoches, we again reduce the learning rate by 10 times. Note that we use different mini-batch settings for the two stages mentioned above. A mini-batch for the training of feature embedding network includes 8 randomly selected persons, and only one image is in turn selected for every person from the cross-camera image set. But for the training of MGAT, a mini-batch contains M persons. With K randomly selected images for each person from cross-camera image sets, a mini-batch of size $M \times K$ is constructed. We empirically set M and K to 8 and 16 in the training, which results in a mini-batch of size 128. We select *Stochastic Gradient Descent* (SGD) as the optimization method for both stages.

For the testing procedure, we directly utilize the feature embedding network enhanced by MGAT to extract features for probe and gallery sets without any extra post-processing methods, because the enhanced feature embedding network has the capability to generate features of optimal discriminant power. This operation is consistent with the evaluation procedure in baseline, which is both simple and efficient.

Table 1: CHI of baseline and graph optimized features on probe and gallery sets.

Dataset	Methods	probe	gallery
iLIDS-VID	Baseline	140.0140	198.5689
	Optimized	140.0694	198.9442
PRID2011	Baseline	515.4785	275.5601
	Optimized	519.3805	279.3300
MARS	Baseline	276.6798	296.8376
	Optimized	314.6908	322.3207

4.3. Ablation study

In this section, we investigate the effectiveness of our proposed MGAT by conducting a series of experiments on the iLIDS-VID, PRID2011 and MARS datasets.

Clustering characteristic. As the main idea of our method is to improve the clustering characteristic of the learned features, we now investigate how greatly it is achieved with the proposed MGAT.

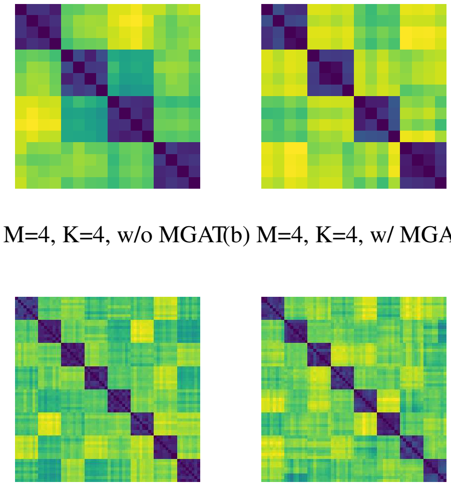
We use the feature-embedding network in baseline and the optimized version to extract the features for the probe and gallery sets of all the datasets. As for metrics to evaluate the clustering characteristic, we adopt the classical Calinski-Harabaz Index (CHI) [6]. When evaluating, a higher Calinski-Harabaz score relates to a model with better defined clusters. The results of CHI are shown in Table 1 respectively. Obviously, our approach has undoubtedly achieved the best performance in both evaluations.

In order to understand the role of MGAT more intuitively, we randomly selected M individuals, each containing K samples, visualized their distance matrix after using MGAT, and compared with that of the baseline. Figure 4 shows the visualization results. We can clearly see that with the use of MGAT, the distances between samples belonging to the same person are smaller, while the distances between samples belonging to different persons are increased visibly.

Comparison with original GAT. The original GAT [21] is proposed to address the problem of node classification on graph. There are some limitations to directly apply it in ReID tasks, for the reason that it only takes advantage of similarity to pull nodes, but neglects the dissimilarity. We conduct experiments to verify that our improved MGAT is better suited for ReID tasks. Results can be found in Figure 5. For all the datasets we use, our MGAT has achieved better performances.

4.4. Comparison with State-of-the-art methods

Results on iLIDS-VID dataset. The results of our proposed MGAT and other state-of-the-art methods on the iLIDS-VID dataset are listed in Table 2. The top-1 accuracy



(a) M=4, K=4, w/o MGAT
(b) M=4, K=4, w/ MGAT

(a) M=8, K=8, w/o MGAT
(b) M=8, K=8, w/ MGAT

Figure 4: Visualization of distance matrices. The use of MGAT can better reduce the intra-class distance and increase the inter-class distance. (the colder the tone, the smaller the value).

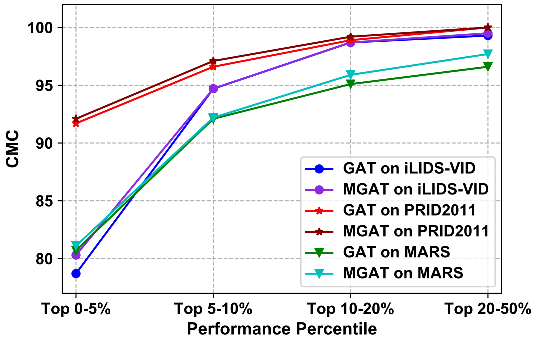


Figure 5: CMC curves for the original GAT and our proposed MGAT.

Table 2: Comparison with related methods on **iLIDS-VID** dataset. (* means additional datasets are used in training)

Methods	top-1	top-5	top-10	top-20
TDL [36]	56.3	87.6	95.6	98.3
CNN+XQDA [42]	53.0	81.4	-	95.1
CNN+RNN [20]	58	84	91	96
JSTRNN [46]	55.2	86.5	-	97.0
ASTPN [33]	62	86	94	98
QAN [17]	68.0	86.8	95.4	97.4
RQEN [24]	77.1	93.2	97.7	99.4
SDM [39]	60.2	84.7	91.7	95.2
DRSA* [15]	80.2	-	-	-
Baseline	76.0	94.0	**98.7**	99.3
MGAT	**80.3**	**94.7**	**98.7**	**99.5**

of our proposed method is 80.3, which outperforms all the compared methods.

To solve the problem of more obscure inter-class difference for video-based re-identification than still-image-based re-identification, TDL [36] integrates a top-push constrain to enforce the optimization on top-rank matching. Our proposed method outperforms TDL by 24% for top-1 accuracy. CNN+XQDA [42] acts as a baseline for combining CNN features with traditional metric to solve the problem of re-identification in the early years. Using better CNN network and feature optimization, our end-to-end approach outperforms CNN+XQDA by 27.3% for top-1 accuracy. CNN+RNN [20] is the first end-to-end method for video-based re-identification, which uses a RNN to fuse frame features obtained by CNN, and then the contrastive loss is used to learn the metric. Our approach outperforms it by 22.3% for top-1 accuracy. JSTRNN [46] handles spatial and temporal information simultaneously by carefully designed spatial recurrent module and temporal attention module. Our approach outperforms it by 25.1% for top-1 accuracy. ASTPN [33] enables the feature extractor to be aware of the current input video sequences, in a way that interdependency from the matching items can directly influence the computation of each other's representation. Our approach outperforms it by 18.3% for top-1 accuracy.

QAN [17] is based on the idea that samples with poor quality will hurt the metric, it implicitly learns the quality of each sample by then fuse features by their quality scores. Our approach outperforms it by 12.3% for top-1 accuracy. RQEN [24] holds the same idea with QAN, it tries to aggregate complementary information from all frames in a sequence for video-based re-identification, using better regions from other frames to compensate the influence of an image region with poor quality. Our approach outperforms it by 3.2% for top-1 accuracy. SDM [39] proposed an interpretable reinforcement learning method to decide whether a pair of images belong to the same or different person. Our approach outperforms it by 20.1% for top-1 accuracy. DRSA [15] introduces multiple spatiotemporal attention model to automatically discovers the diverse set of distinctive body to address the problem of image occlusion. Though it utilizes 6 additional ReID datasets to pretrain the model before training on each dataset we use, our method still outperforms it.

Results on PRID2011 dataset. Table 3 illustrates the performance of our proposed MGAT and other state-of-the-art methods on the PRID2011 dataset. Our proposed MGAT outperforms all the introduced methods above except DRSA [15] for top-1 accuracy. We attribute our fail-

Table 3: Comparison with related methods on **PRID2011** dataset. (* means additional datasets are used in training)

Methods	top-1	top-5	top-10	top-20
TDL [36]	56.7	80.0	87.6	93.6
CNN+XQDA [42]	77.3	93.5	-	99.3
CNN+RNN [20]	70	90	95	97
JSTRNN [46]	79.4	94.4	-	99.3
ASTPN [33]	77	95	99	99
QAN [17]	90.3	98.2	99.32	**100.0**
RQEN [24]	91.8	98.4	99.3	99.8
SDM [39]	85.2	97.1	98.9	99.6
DRSA* [15]	**93.2**	-	-	-
Baseline	87.6	96.6	98.9	**100.0**
MGAT	92.1	**97.1**	**99.2**	**100.0**

ure to the extra ReID datasets DRSA uses. Considering the small gap, we are confident that if the same additional information is used, our approach will be also greatly improved.

Results on MARS dataset. Being the first large scale video-based re-identification dataset, MARS [42] is a more objective and fair evaluation criteria for video-based methods. The results of our proposed MGAT and other state-of-the-art methods on the MARS dataset are shown in Table 4. Note that all the experiments are conducted in the single-query mode. Our method outperforms all the compared approaches for top-1 accuracy except DRSA [15]. But for top-5 accuracy and mAP, our method outperforms all the methods by a large margin. The reason why DRSA defeats our proposed MGAT is discussed in the analysis for the PRID2011 dataset.

Besides all the approaches introduced above such as CNN+XQDA [42], JSTRNN [46] and QAN [17], our method also outperforms two new approaches for MARS. DCFBLP [14] stacks designed multi-scale context-aware network (MSCAN) to learn powerful features over full body and body parts, and uses spatial transformer networks (STN) to learn and localize deformable person parts. Our method outperforms it by 9.3% and 15.7% for top-1 accuracy and mAP. TriNet [11] uses a variant of the triplet loss to perform deep metric learning in defense of triplet loss. Our method outperforms it by 1.3% and 4.1% for top-1 accuracy and mAP.

Results on Market1501 dataset. HA-CNN [31] designs a lightweight yet deep CNN architecture by devising a holistic attention mechanism for locating the most discriminative pixels and regions in order to identify optimal visual patterns for ReID. Our result outperforms it by 0.8% for mAP and 0.3% for top-1 accuracy. AlignedReID [5] is said to surpass human-level performance in ReID, which outperforms our method by 14.2% and 2.9% for mAP and top-1 accuracy. However, the result is not surprising, since it is a bundle of advanced methods for ReID, including *metric learning*, *feature fusion*, *mutual learning*, *re-ranking*.

Table 4: Comparison with related methods on **MARS** dataset. (* means additional datasets are used in training)

Methods	top-1	top-5	top-20	mAP
CNN+XQDA [42]	65.3	82.0	89.0	47.6
JSTRNN [46]	70.6	90.0	97.6	50.7
DCFBLP [14]	71.8	86.6	93.0	56.1
QAN [17]	73.74	84.9	91.6	51.7
TriNet [11]	79.8	91.4	-	67.7
SDM [39]	71.2	85.7	91.8	-
DRSA* [15]	**82.3**	-	-	65.8
Baseline	79.8	91.9	96.4	71.2
MGAT	81.1	**92.2**	**97.7**	**71.8**

Table 5: Comparison with related methods on **Market1501** dataset.

Methods	mAP	top-1	top-5	top-10
OIM Loss [32]	60.9	82.1	-	-
SpindleNet [40]	-	76.9	91.5	94.6
MSCAN [14]	53.1	76.3	-	-
k-reciprocal [44]	63.6	77.1	-	-
Point 2 Set [45]	44.3	70.7	-	-
SVDNet [25]	62.1	82.3	-	92.3
Part Aligned [41]	63.4	81.0	92.0	94.7
HA-CNN [31]	75.7	91.2	-	-
AlignedReID [5]	**90.7**	**94.4**	-	-
Baseline	76.0	90.3	95.8	97.1
MGAT	76.5	91.5	**97.2**	**98.0**

5. Conclusion

In this paper, we argue that the mainstream methods for person re-identification (ReID) mainly focus on the correspondence between individual sample images and labels, while ignoring rich global mutual information resides in the whole sample set. To address this defect, we propose MGAT module that enables nodes to directionally attend over other nodes' features under the guidance of label information. In this way, the previously neglected global mutual information is exploited to generate an optimized feature space with more discriminant power.

Acknowledgement This work is partially supported by National Key R&D Program of China (No.2017YFA0700800), Natural Science Foundation of China (NSFC): 61876171 and 61572465.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015. 2

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3

[3] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 2530–2539, 2017. 1, 2

[4] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013. 2

[5] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 8

[6] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *The Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. 6

[7] D. Chung, K. Tahboub, and E. J. Delp. A two stream siamese convolutional neural network for person re-identification. In *The IEEE International Conference on Computer Vision*, pages 1983–1991, 2017. 2

[8] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *The Advances in Neural Information Processing Systems*, pages 2224–2232, 2015. 2

[9] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *The IEEE International Conference on Computer Vision*, pages 1305–1313, 2015. 1, 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[11] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 8

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *The Scandinavian conference on Image analysis*, pages 91–102, 2011. 5

[13] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[14] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017. 8

[15] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018. 2, 7, 8

[16] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2

[17] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799, 2017. 2, 7, 8

[18] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *The IEEE International Conference on Image Processing*, pages 3567–3571, 2013. 1, 2

[19] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995, 2009. 1

[20] N. Mclaughlin, J. M. D. Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *The IEEE International Conference on Computer Vision*, pages 1325–1334, 2016. 2, 7, 8

[21] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *The IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. 2, 6

[22] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang. Deep group-shuffling random walk for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 2265–2274, 2018. 2

[23] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang. Person re-identification with deep similarity-guided graph neural network. In *The European Conference on Computer Vision*, pages 508–526, 2018. 1, 2

[24] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai. Region-based quality estimation network for large-scale person re-identification. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 7, 8

[25] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *The IEEE International Conference on Computer Vision*, pages 3800–3808, 2017. 8

[26] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 2, 3

[27] H. Wang, S. Gong, X. Zhu, and T. Xiang. Human-in-the-loop person re-identification. In *The European Conference on Computer Vision*, pages 405–422, 2016. 2

[28] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *The European Conference on Computer Vision*, pages 688–703, 2014. 5

[29] X. Wang. Intelligent multi-camera video surveillance: A review. *The Pattern Recognition letters*, 34(1):3–19, 2013. 1

[30] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 3

[31] L. Wei, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *The IEEE International Conference on Computer Vision*, pages 2285–2294, 2018. 8

[32] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 8

[33] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks

for video-based person re-identification. In *The IEEE International Conference on Computer Vision*, pages 4743–4752, 2017. 2, 7, 8

[34] M. Ye, C. Liang, Z. Wang, Q. Leng, and J. Chen. Ranking optimization for person re-identification via similarity and dissimilarity. In *The 23rd ACM international conference on Multimedia*, pages 1239–1242, 2015. 2

[35] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *The IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016. 1, 2

[36] J. You, A. Wu, X. Li, and W. S. Zheng. Top-push video-based person re-identification. In *The IEEE International Conference on Computer Vision*, pages 1345–1353, 2016. 2, 7, 8

[37] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai. Hard-aware point-to-set deep metric for person re-identification. *arXiv preprint arXiv:1807.11206*, 2018. 2

[38] S.-I. Yu, Y. Yang, and A. Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3714–3720, 2013. 1

[39] J. Zhang, N. Wang, and L. Zhang. Multi-shot pedestrian re-identification via sequential decision making. *arXiv preprint arXiv:1712.07257*, 2017. 2, 7, 8

[40] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 8

[41] L. Zhao, L. Xi, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *The IEEE International Conference on Computer Vision*, pages 3219–3228, 2017. 8

[42] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *The European Conference on Computer Vision*, pages 868–884, 2016. 2, 5, 7, 8

[43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *The IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 5

[44] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *The IEEE International Conference on Computer Vision*, pages 3652–3661, 2017. 1, 2, 8

[45] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3741–3750, 2017. 8

[46] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 6776–6785, 2017. 2, 7, 8