Context-Aware Graph Convolution Network for Target Re-identification

Deyi Ji[1][0000-0001-7561-9789]*, **Haoran Wang**[2]*, **Hanzhe Hu**[3], **Weihao Gan**[1], **Wei Wu**[1], **Junjie Yan**[1]

[1] SenseTime Research   [2] Xidian University   [3] Peking University

{jideyi, ganweihao, wuwei, yanjunjie}@sensetime.com

wanghaoran@stu.xidian.edu.cn,   huhz@pku.edu.cn

Abstract

Most existing re-identification methods focus on learning robust and discriminative features with deep convolution networks. However, many of them consider content similarity separately and fail to utilize the context information of the query and gallery sets, e.g. probe-gallery and gallery-gallery relations, thus hard samples may not be well solved due to the limited or even misleading information. In this paper, we present a novel Context-Aware Graph Convolution Network (CAGCN), where the probe-gallery relations are encoded into the graph nodes and the graph edge connections are well controlled by the gallery-gallery relations. In this way, hard samples can be addressed with the context information flows among other easy samples during the graph reasoning. Specifically, we adopt an effective hard gallery sampler to obtain high recall for positive samples while keeping a reasonable graph size, which can also weaken the imbalanced problem in training process with low computation complexity. Experiments show that the proposed method achieves state-of-the-art performance on both person and vehicle re-identification datasets in a plug and play fashion with limited overhead.

Introduction

Re-identification (Re-ID) problems aim to determine corresponding targets across multiple cameras or different locations in the same camera from the gallery set with a given probe. Existing Re-ID tasks focus on both person and vehicle which play an important role in intelligent surveillance systems. Much process has been made in recent years in solving the specific problems of illumination variations, viewpoint changes and occlusions (Song et al. 2018; Zhou et al. 2019; Hou et al. 2019; Wang et al. 2020). However, the study on the context of the query and gallery sets receives less attention in research community. To further illustrate the point, we conduct an analysis of existing Re-ID methods.

On the one hand, a number of person Re-ID methods attempt to learn robust and discriminative features using well-designed deep CNN (Li et al. 2014; Lin et al. 2017; Su et al. 2017; Song et al. 2018; Schumann and Stiefelhagen 2017; Subramaniam, Chatterjee, and Mittal 2016) or fine-grained

loss functions (Cheng et al. 2016; Yi et al. 2014; Hermans, Beyer, and Leibe 2017; Zheng et al. 2018), and have witnessed great performance improvements in recent years. However, they treat each probe-gallery pair separately, without considering the relations between the query and gallery images. Thus the context information has not been fully explored in these systems, as a result, many advanced works (Zhou et al. 2019; Xia et al. 2019; Chen et al. 2019; Yang et al. 2019; Chen et al. 2020a) achieved the high top-1 accuracy of almost 95% on the Market-1501 dataset, but the mAP is still less than 90%.

On the other hand, several methods have tried to optimize the ranking results by exploiting the context information in person Re-ID. However, most of them operate in an unsupervised or semi-supervised manner as a post-processing operation. Concretely, Karaman et al. (Karaman et al. 2014) used a discriminative Conditional Random Field (CRF) to exploit the local context among the near neighbors in a semi-supervised way. Garcia et al. (Garcia et al. 2015) proposed a post-ranking framework and analyzed context information in an unsupervised way. K-reciprocal reranking (Wang et al. 2019b) adjusted the gallery ranks using Jaccard distance calculated with k-reciprocal nearest neighbors in the whole probe and gallery sets. Generally, these methods have two weaknesses. One is that they are not able to take full advantage of context information among two sets, as they perform as a post-process without training under the supervision. Second, they are of the sizeable computational quantity, since they need to take all query and gallery data for the matrix calculation. The time and memory consumption issues are severe in practice, especially on a large scale dataset. Besides, it is impractical to obtain all query images in a real-life scenario, such as a surveillance camera where the probe persons are dynamically changing in real-time environments.

For the first time, Shen et al. (Shen et al. 2018) proposed a graph model to utilize the context information in a supervised way, named SGGNN. Their purpose was to learn a discriminative feature that can be used for binary classification. With a siamese network, they processed one query image and one gallery image to obtain the probe-gallery feature, then took these features as nodes and used a Graph Neural Network to refine further. The work provided an empirical approach to explore context information, however,

---

there are three main limitations that have not yet been addressed. First, they neglect the differentia in the gallery set, randomly sample the gallery images to form probe-gallery nodes. Thus hard positive and negative ones are not explicitly studied without appropriate sample selection. Second, they build the graph in a fully-connected way, where one node connects to all the other ones. This may incur redundant or even harmful connections between nodes. Actually, the method is essentially a full aggregation of node features, the gallery-gallery relations are used as the aggregate weights rather than to construct a meaningful graph. Third, the SGGNN only used two shallow fully-connected layers for message passing, thus the resulting features may be inadequate for discriminative representation. In conclusion, the probe-gallery and gallery-gallery relations are not well studied in the graph-based network.

Recently, a considerable amount of models has arisen for reasoning on graph-structured data in Re-ID tasks. Zhou et al. (Zhou et al. 2018) proposed a patch-wise graph matching to handle spatial misalignment caused by large variations in viewpoints and human poses. Wu et al. (Wu et al. 2019) learnt the underlying view-invariant representations by the unsupervised graph association. Though the graph structure fits well to the relation reasoning, these methods only utilized the graph to explore the content similarity of the probe-gallery pairs, the context similarity still remained unsolved. Meanwhile, Graph Convolution Network (GCN) is proposed for the problem of classifying nodes in the graph. The convolution operation defined over graphs is a weighted aggregation of target node features with its neighbor nodes. With the guidance of the adjacency matrix, we can explicitly specify the connections between certain samples. In this way, we intend to use GCN to learn the context information for relation reasoning, considering the Re-ID task as a node classification problem.

Based on the above analysis, we propose a novel Context-Aware Graph Convolution Network (CAGCN) for Re-ID tasks. The framework takes one probe image and a certain number of candidate gallery images as inputs, and outputs the relation between the probe and each gallery image directly. In response to the above concerns, we discuss three main questions:

1. How to select sufficient and meaningful positive samples for the graph construction? In practical Re-ID tasks, the gallery set may be extensive, for one probe image, the positive and negative samples in the gallery may be heavily imbalanced distributed. We propose an intuitive and effective hard gallery sampler instead of randomly selecting candidate gallery images, which guarantees the quality and quantity of positive probe-gallery samples.

2. How to reason the relations between query and gallery sets validly? We build a context-aware graph to explicitly study two relations. On the one hand, we model the probe-gallery relations as node features in a high-dimension space. On the other hand, the gallery-gallery relations are modeled in the edges in a low-dimension space. Especially, we take the easy positive, hard positive and hard negative samples into consideration, making the graph meaningful and reasonable. Moreover, we introduce a deep GCN model for bet-

ter feature learning in the process of relation reasoning.

3. How to design a flexible and adjustable system? Sometimes we need to be more resource-efficient in terms of time, memory, or power. For example, a robot might need to make decisions within a time and power limit. However, existing re-ranking methods have high computational cost, as they need to take all images for the similarity calculation, which is impractical when dealing with large-scale datasets. Fortunately, our method can be easily adjusted in sample number due to the proposed sampler and graph design. For example, we can only take one query image and a fixed number of gallery images according to the resource constraints.

Overall, our contribution is threefold:

- We propose a novel Context-Aware Graph Convolution Network (CAGCN) for target re-identification tasks. To our knowledge, it is the first attempt to introduce the GCN for exploring context of the query and gallery sets in Re-ID tasks. Especially, we build a context-aware graph with relation guided edges, which explicitly model the relations in the two sets, guaranteeing the validity and effectiveness of information flow. Furthermore, the network is trained end-to-end in a supervised way.

- The graph in our framework is constructed from an adjustable set of meaningful gallery candidates, which are selected by the proposed hard gallery sampler strategy. It ensures that we can select only a small number of samples to achieve good results, and the performance also shows a consistent improvement with the increase in resources.

- The proposed framework is practical and can show in a plug-and-play fashion. Experiments show that the framework achieves state-of-the-art results on both person and vehicle benchmark datasets, i.e., Market-1501, DukeMTMC-ReID and VeRi-776. Notably, the mAP accuracy is significantly improved in all cases.

Related Work

Re-identification. Re-identification (Re-ID) tasks focus on both person and vehicle, and have been extensively studied in the passed few years. Mainstream approaches for person re-identification focus on three main aspects: learning a powerful and discriminative feature representation for person images (Li et al. 2014; Lin et al. 2017; Su et al. 2017; Song et al. 2018; Schumann and Stiefelhagen 2017; Subramaniam, Chatterjee, and Mittal 2016; Wang et al. 2016; Guo et al. 2019; Yang et al. 2019; Wang et al. 2019a), designing a fine-grained distance metric for measuring similarity between image pairs (Cheng et al. 2016; Zheng et al. 2018; Hermans, Beyer, and Leibe 2017; Yi et al. 2014), or adopting some post-processing algorithms to refine the retrieval results in an offline way (Leng et al. 2015; Wang et al. 2019b; Li et al. 2012). Feature learning based on deep convolution neural networks was first introduced in (Li et al. 2014; Yi et al. 2014), nowadays the kind of methods focus on both global and part-patch information (Zhang et al. 2017; Sun et al. 2018; Zheng et al. 2018; Hou et al. 2019; Wei et al. 2019). For metric learning (Yu, Wu, and Zheng 2017; Yu et al. 2018), triplet based loss functions are widely used

(Hermans, Beyer, and Leibe 2017). Vehicle Re-ID gains increasing attention recently and shares common frameworks with person Re-ID. Recent vehicle Re-ID methods (Chu et al. 2019; Khorramshahi et al. 2019; Tang et al. 2019; Lou et al. 2019; Feng et al. 2018) mainly focus on embedding spatial and temporal space or viewpoints information. Our method is beneficial to both tasks.

Graph convolution network. Recently GCN draws increasing attention and has proved to be very effective in many computer vision tasks, such as action recognition and node classification. Kipf et al. (Kipf and Welling 2016) proposed a simple and well-behaved layer-wise propagation rule for neural network models and demonstrated its effectiveness in semi-supervised classification tasks. Veličković et al. (Veličković et al. 2017) proposed a graph attention network which incorporates the self-attention mechanism into propagation step. Wang et al. (Zhongdao Wang and Wang 2019) introduced GCN to face clustering tasks. Li et al. (Li et al. 2019) attempted to solve the vanishing gradient problem in deep GCNs. Hu et al. (Hu et al. 2020) proposed a class-wise dynamic GCN for semantic segmentation. However, GCN is few introduced to Re-ID tasks. Shen et al. (Shen et al. 2018) proposed a similarity guided graph neural network, the relations are transformed by two shallow fully-connected layers to update probe-gallery similarity with handcrafted weights, and there is no connection among galleries. Chen et al. (Chen et al. 2020b) aimed to optimize the ranking results with graph models in the embedding set, and introduced the attention mechanism to enhance important channels. By contrast, we are the first to propose to use GCN to explore the context of the query and gallery. Moreover, Re-ID requires high recall for positive samples while reducing the amount of negative samples in graph nodes. Thus the sample strategy for graph nodes needs to be well designed, which has not been studied before.

Context. Several works (Li et al. 2012; Su, Zhao, and Lin 2018; Garcia et al. 2015; Karaman et al. 2014; Wang et al. 2019b; Su et al. 2020a,b) have utilized contextual information in their networks. Li et al. (Li et al. 2012) developed a re-ranking model by analyzing the relative information and direct information of near neighbors of each pair of images. Karaman et al. (Karaman et al. 2014) combined discriminative models of person identity with a CRF to exploit the local manifold approximation induced by the nearest neighbor graph in feature space, while a graph topology defined by distances between all person images in feature space leveraged local support for label propagation in the CRF. In (Garcia et al. 2015), an unsupervised re-ranking model was learnt by jointly considering the content and context information in the ranking list. Wang et al. (Wang et al. 2019b) utilized the whole datasets information and adopted a Jaccard distance to refine the gallery ranks. Though these works successfully integrate information from k-nearest neighbors to refine the results, the misleading matches still exist without effective guidance of supervision information. To tackle this problem, we propose to explore the context directly in the training process, during the graph aggregation under the label supervision.

Approach

In this section, we first give the problem definition, then we show the detail of the overall framework, context-aware graph and the process of reasoning on the graph.

Problem Definition Given the probe image p and gallery image set $G = \{g_i | i = 1, 2, 3, ..., n\}$, we use a basic network to extract their features, which are $\boldsymbol{f_p}$ and $\{\boldsymbol{f_{g_i}} | g_i \in G\}$. Then a hard gallery sampler is adopted to construct the gallery candidates set $G_c = \{g_i | i = 1, 2, 3, ..., k\}$. Subsequently, we obtain the probe-gallery features by subtracting gallery features from the probe feature $\boldsymbol{f_p}$, which are used as the node features for graph building. The graph can be formulated as $G(V, E, A)$, where V, E and A denote the nodes, edges and adjacency matrix. E and A are carefully designed to build the context-aware graph. Finally, we apply graph convolutions on the built graph and then output the binary classification likelihood for each node, deciding whether the corresponding gallery is identical to the probe image.

Framework Overview

As illustrated in Figure 1, the proposed framework consists of three parts, including feature extraction, graph construction and graph reasoning. In particular, we adopt the model proposed in (Luo et al. 2019b) as the basic network. In the graph construction, probe-gallery search and gallery-gallery search select meaningful gallery candidates firstly, then the graph nodes are obtained by the subtraction operation of the probe-gallery pairs. The edges are explicitly designed with the guidance of gallery-gallery relations. In the graph reasoning, we propose the GCN block for a deeper network to fully explore the context information based on the built graph. Each node feature is aggregated with the context information from others under the label supervision. Finally, an MLP is adopted for binary classification for each graph node. The framework can be applied to both person and vehicle Re-ID tasks in a plug and play fashion.

Context-Aware Graph Construction

As discussed above, for one specific probe, there may exist heavily imbalanced distribution of positive and negative samples in the gallery set. So during the graph construction, we first propose an effective hard gallery sampler to recall hard gallery samples and weaken the affect of imbalanced distribution. Then we show that in our GCN module, context information flow is well controlled with the relation guided graph edges, where hard probe-gallery features can be optimized with other easy probe-gallery features.

Hard gallery sampler In training stage, hard gallery samples are essential for context aggregation during graph reasoning, as they contain more discriminative information. However, random sampling in the gallery set can not meet the requirements. Thus, we design a Hard Gallery Sampler (HGS) to select hard positive gallery images. For each probe, we sample k gallery candidates and denote them as $G_c = \{g_i | i = 1, 2, 3..., k\}$. A widely used plain sampler calculates the similarity between probe and gallery images and selects the k nearest ones. However, in order to recall more hard positive samples, k needs to be very large, which will
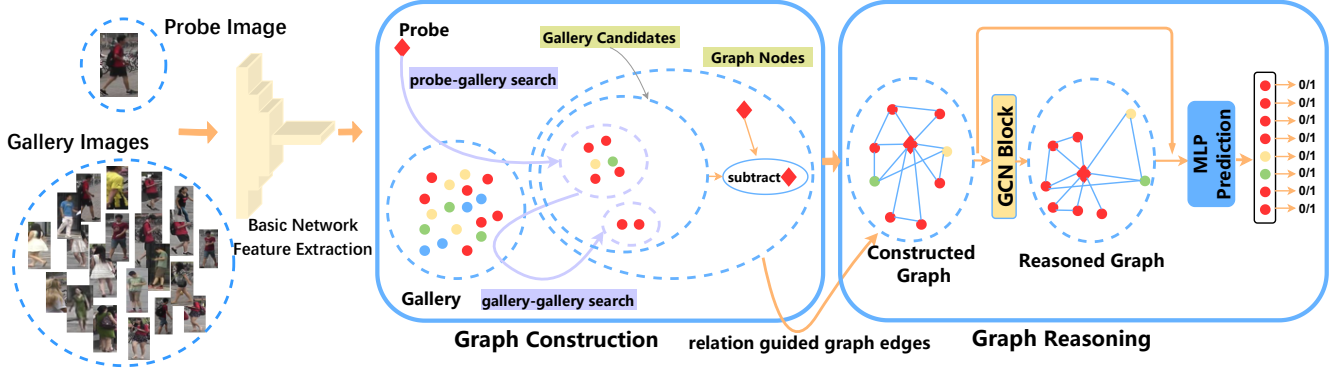
Figure 1: An overview of our proposed framework. Given the probe and gallery images, we first feed them into the basic network for feature extraction. Then we utilize hard gallery sampler (HGS) method to select gallery nodes to construct the graph (different colors refer to different target identities), where probe-gallery search and gallery-gallery search take place to achieve high recall. Then we subtract the features of gallery candidates with probe feature to obtain node features. Then we construct the context-aware graph and GCN block for relation modeling. Finally, we perform an MLP on each node, denoting whether the corresponding gallery is identical to the probe.

introduce more negative samples and lead to more severe imbalanced distribution problems, as well as higher computation complexity.

To circumvent this issue, we propose an intuitive two-hop HGS method. Though hard positive samples are usually not so similar as the probe in appearance due to the viewpoint changes, they can be retrieved by some easy positive ones as a bridge. Formally, given probe p and gallery set G, we define $N(p, k)$ as k-nearest gallery images of p. In the sampling process, we first select the k_1-nearest gallery images $N(p, k_1) = \{g_1, g_2, g_3, ..., g_{k1}\}$ and initialize $G_c = N(p, k_1)$. Then for each gallery image g_i in $N(p, k_1)$, we select its k_2-nearest gallery images $N(g_i, k_2)$. Noted that there may be overlapped images between $N(p, k_1)$ and $N(g_i, k_2)$, we update $N(g_i, k_2)$ as:

$$N'(g_i, k_2) = N(g_i, k_2) - N(g_i, k_2) \cap N(p, k_1). \quad (1)$$

From $i = 1$ to $i = k_1$, we add $N'(g_i, k_2)$ to G_c until the total number of images in G_c gets to k. In general, both k_1 and k_2 are smaller than k, thus both $N(p, k_1)$ and $N(g_i, k_2)$ have higher precision than $N(p, k)$. In this way, the hard positive samples can be obtained by the easy ones, so G_c has a higher recall rate than the original $N(p, k)$. The whole process is illustrated in Figure 2.

After obtaining k gallery images by the HGS method, we get the node features X of each probe-gallery pair by the subtraction operation, which explore and model the probe-gallery relations through the GCN.

Relation guided graph edges The most straightforward way is to construct a complete graph where any two nodes exist an edge. However, that may incur redundant or even harmful connection, passing messages between totally irrelevant nodes will finally lead to indistinguishable features. Actually, the method of SGGNN (Shen et al. 2018) can be viewed as a fully-connected graph, which may suffer the problem especially without sampling of gallery images.

The edge connection style in graph determines the context information aggregation flow among nodes. We analyze that in the retrieval task, a good information flow needs to ensure that, during graph reasoning, hard positive probe-gallery features (i.e. hard positive nodes) should be "closer" to easy positive nodes, and vice versa. With the help of HGS, we solve this problem by desiging the meaningful edges between nodes. Specifically, given the gallery candidate set $G_c = \{g_i | i = 1, 2, 3, ..., k\}$, we first obtain k' gallery neighbors of g_i using the HGS method, denoted as $H(g_i)$. Then for another node j, if $g_j \in H(g_i)$, there will be an edge between node i and node j. Formally, a sparsely connected graph is obtained by:

$$A_{ij} = \frac{\dagger(g_j \in H(g_i))exp(F(V_i, V_j))}{\sum_{j=1}^{N} \dagger(g_j \in H(g_i))exp(F(V_i, V_j))} \quad (2)$$
$$F(V_i, V_j) = \phi(V_i)^T \phi'(V_j)$$

where A is the adjacency matrix, we perform softmax function on each node so that the sum of all the relation values of one node i will be 1. $\dagger(\cdot)$ is the indicator function, ϕ, ϕ' denote two different linear transformations. $F(V_i, V_j)$ denotes the gallery-gallery relations between two gallery samples with the dot-product similarity.

In this way, the context information aggregation flow is well controlled by the gallery-gallery relations. Without loss of generality, the sampled gallery candidate set G_c is mainly composed of three types of nodes: easy positive, hard positive and hard negative nodes. On one hand, during the edge connection process, hard positive nodes can be connected to easy positive ones, which helps narrow the gap between positive nodes, leading to the higher recall finally. On the other hand, for hard negative nodes, though sampled due to the vicinity to the probe image, the number of connections with positive nodes is limited by the graph formula, resulting in being pushed "farther" from the positive nodes during graph reasoning under the label supervision. Finally, we rep-
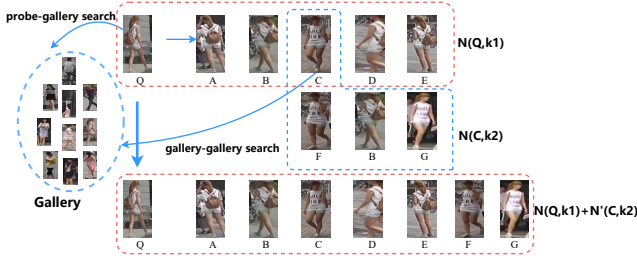
Figure 2: The proposed two-hop hard gallery sampler. For one specific probe image Q, we first do probe-gallery search to select k_1-nearest gallery images $N(Q, k_1)$. Then for each gallery image in $N(Q, k_1)$ (for example C), we perform gallery-gallery search to obtain k_2-nearest gallery images $N(C, k_2)$. Finally, we merge each gallery-gallery results to the probe-gallery results in order, until the total images number increases to k.

resent the constructed graph via a binary adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$, where $|V|$ is the node size.

Graph Reasoning

After the graph construction, context information can be aggregated among graph nodes with the supervision of each node label. The input of a graph convolution layer are the adjacency matrix A and node feature matrix X, the reasoning process of one GCN layer can be denote as:

$$Z^{(l+1)} = \sigma(A Z^l W^l) \qquad (3)$$

where A is the adjacent matrix. $Z^l \in R^{N \times d}$ is the feature representations of nodes in the lth layer, and $Z^0 = X$. $W^l \in R^{d \times d'}$ is the layer-specific learnable weight matrix, d and d' are the input and output dimension. $\sigma(\cdot)$ denotes an activation function, and we adopt ReLU in this work.

Simply stacking multiple graph convolution layers would lead to the over-smoothing problems. In our framework, we introduce the residual connections (Li et al. 2019) to alleviate the problem. As a result, our network can be deeper with the GCN residual blocks. One GCN residual block consists of one GCN layer followed by one dimension batchnorm layer. Formally, the output of one GCN residual block can be written as:

$$Z_{l+1} = Z_l + \mathcal{G}(Z_l, A). \qquad (4)$$

Where $\mathcal{G}$ denotes Eq. 3 followed by a dimension batchnorm layer. Furthermore, since the graph convolution mainly focuses on the global context information during reasoning process, we concatenate the input feature to maintain both local and global information. Each node feature represents a fully-explored probe-gallery pair, fused with context information in both query and gallery sets. Finally, we perform the binary classification on each node to decide whether the pair is the same person.

In the inference stage, the GCN module outputs the classification probability for each node, which can be viewed as their feature distance. Moreover, the GCN module aims at

learning probe-gallery and gallery-gallery relations, which focuses on the global information between the query and gallery sets. While the original distance (calculated by origin visual feature) still requires attention since it also contains individual probe-gallery information that is conducive to the final feature representation. For each g_i, we denote the original probe-gallery distance as $d_o(p, g_i)$, and the predicted distance by the CAGCN as $d_g(p, g_i)$. Hence, the final distance is obtained by:

$$d(p, g_i) = d_o(p, g_i) + \lambda d_g(p, g_i), \qquad (5)$$

where λ controls the effect of the CAGCN. If $\lambda = 0$, the final distance is just the original distance. By adding the distance from our method, we can achieve higher *mAP* and preserve high *Rank* metrics in all experiments.

Experiments

To verify the effectiveness of the proposed method, we conduct experiments on three large-scale Re-ID datasets, including two person datasets: Market1501 (Zheng et al. 2015) and DukeMTMC-reID (Zheng, Zheng, and Yang 2017), and one vehicle dataset: VeRi-776 (Liu et al. 2016).

Implementation details

We utilize the network proposed in (Luo et al. 2019b) as baseline model and reproduce their results using the same settings, ResNet-50 is used as the backbone. The GCN module in our framework consists of 9 graph convolution layers followed by 3 fully connected layers. We use stochastic gradient descent to optimize our GCN module. The learning rate is 0.01 and the momentum is 0.9. The weight decay rate is 1e-4. The model is trained for 500 epochs with the batch size of 4. Focal loss is adopted as loss function, the α and γ are set to 2 and 0.25. For the hard gallery sampler, we set k_1 to 70, k_2 to 20 and maximum nodes number k to 100. During the graph construction, only top-8 (i.e. $k' = 8$) nodes are connected to construct a sparse graph. All of our experiments are conducted under the single probe setting.

Datasets and evaluation metrics

Market1501 The dataset consists of 32,668 DPM-detected pedestrian image boxes of 1,501 identities. Among these images, 12,936 images from 751 identities are used as the training set, while 3,368 probe images and 19,732 gallery images from the remaining 750 identities are used as the testing set.
DukeMTMC-reID The dataset is a subset of Duke dataset for image-based person reidentification task. This dataset contains 16,522 images of 702 identities for training, 2,228 probe images of the other 702 identities and 17,661 gallery images of 702 identities plus 408 distractor identities for testing.
VeRi-776 The dataset is a large-scale benchmark dateset for vehicle Re-ID. It consists of 49,357 images of 776 distinct vehicles which are captured by 20 non-overlapping cameras from different viewpoints, resolutions and occlusions.
Evaluation metrics We adopt the Cumulative Matching Characteristics (CMC) at Rank-1, Rank-5, Rank-10 and

Methods	Reference	Market1501			DukeMTMC-reID		
		mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
SFT (Luo et al. 2019a)	ICCV 2019	82.7	93.4	97.4	73.2	86.9	93.9
P^2Net (Guo et al. 2019)	CVPR 2019	83.4	94	98	70.8	84.9	92.1
BDB (Dai et al. 2019)	ICCV 2019	84.3	94.2	-	72.1	86.8	-
CAMA (Yang et al. 2019)	CVPR 2019	84.5	94.7	98.1	72.9	85.8	-
DCDS (Alemu, Shah, and Pelillo 2019)	ICCV 2019	85.8	94.8	98.1	75.5	87.5	-
SCSN (Chen et al. 2020a)	CVPR 2020	88.5	95.7	-	79.0	90.1	-
ISP (Zhu et al. 2020)	ECCV 2020	88.6	95.3	**98.6**	80.0	89.6	**95.5**
CAGCN		**91.7**	**95.9**	98.2	**85.9**	**91.3**	94.3
CAGCN+RR		94	96.2	98	88.9	92.3	94.5

Table 1: Comparison results (%) on Market1501 and DukeMTMC-reID dataset at 3 evaluation metrics: *mAP*, *Rank-1* and *Rank-5*. The RR means the results with the re-ranking (Wang et al. 2019b) post-processing. The bold font denotes the best result without re-ranking.

mean Average Precision (mAP) on all the datasets to evaluate the performance of the model.

Comparison with State-of-the-Art Methods

We compare the proposed CAGCN with other methods on the above three datasets to demonstrate its effectiveness and generalization. Experiments show that our framework achieves the state-of-the-art performance on both person and vehicle tasks. In Table 1, we compare CAGCN with state-of-the-arts on Market1501 and DukeMTMC-reID. CAGCN achieves the best performance on both Rank-1 and *mAP* evaluation metrics. Especially, our model surpasses the previous methods by a large margin on *mAP*. The experimental results show that the proposed framework can exploit the context information among probe and gallery images effectively to learn more discriminative node features. When using the re-ranking scheme as post-processing, our method achieves the highest accuracy of 96.2% and 94% on Rank-1 and *mAP*. Table 2 shows the results of CAGCN on VeRi-776 dataset compared with other state-of-the-art methods for vehicle re-identification. Our model also outperforms other methods on all the evaluation metrics, especially on *mAP*.

Ablation study

We utilize the network proposed in (Luo et al. 2019b) as baseline model and reproduce the results in our framework. **The impact of different samplers.** Table 3 shows the performance of different sampling strategies on three datasets. Here we set the node and layer number in GCN module to 100 and 9 respectively. The plain sampler simply calculates the similarity between probe and gallery images and selects the k nearest ones, while HGS refers to the proposed hard gallery sampler. Experiments show that a plain sampler can obtain most easy galleries and gain considerable improvements on all the metrics, especially *mAP*. By applying hard gallery sampler, the *mAP* can be further improved (1.2% on Market1501, 2.3% on DukeMTMC-reID and 1.1% on VeRi-776), that is to say, more hard positive galleries are sampled and the recall rate of positive samples is further increased. It shows that the HGS effectively mitigates the affect of imbalanced problem discussed above. More intuitively, we set the samples number k to 80, 100, 120, 150, 180 respectively, and show the precision-recall curves of the two samplers during

Methods	Reference	mAP	Rank-1	Rank-5
FDA-Net(Lou et al. 2019)	CVPR 2019	55.5	84.3	92.4
VANet (Chu et al. 2019)	CVPR 2019	66.3	89.8	96.0
AAVER (Khorramshahi et al. 2019)	ICCV 2019	61.2	89.0	94.7
PAMTRI (Tang et al. 2019)	ICCV 2019	71.9	92.9	96.9
PNVR (He et al. 2019)	CVPR 2019	74.3	94.3	**98.7**
CAGCN		**79.6**	**95.8**	**98.7**

Table 2: Comparison results(%) on VeRi-776 dataset at 3 evaluation metrics: *mAP*, *Rank-1* and *Rank-5*. Re-ranking is not used here. The bold font denotes the best result.

the gallery sampling process before graph construction. As shown in Figure 3, compared to the plain sampler, HGS improves the recall more than 1.1% for each k.

Sampler	Market1501		DukeMTMC-reID		VeRi-776	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
baseline	85.4	94.1	75.6	86.3	74.1	93.9
Plain	90.4	95.0	83.1	89.1	78.1	94.7
HGS	91.6	95.0	85.4	89.6	79.2	95.2

Table 3: Ablation study of the effect of different samplers in the proposed method on the three datasets.
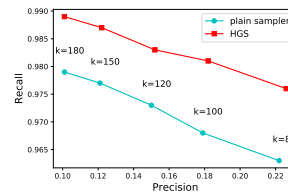

Figure 3: The precision-recall curves for the plain sampler and HGS with different k on Market1501.
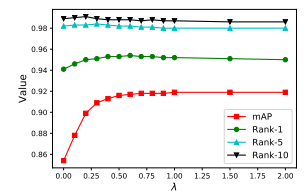

Figure 4: The impact of λ in final distance on Market1501.

The impact of λ in final distance. λ is used to balance the effect of two distances in Eq. 5. The greater the value, the greater the impact of CAGCN. As shown in Figure 4, with λ gradually increasing, *mAP* is gradually increasing and stay still around 0.92 when λ gets to 1.0, which indicates that more and more context information is mined in this process. Meanwhile, *Rank-1*, *Rank-5* and *Rank-10* are relatively sta-

ble, since the individual probe-gallery information is also preserved with the original distance, which has more influence in these indicators.

The impact of graph node number. Table 4 shows the influence of nodes number in the GCN module on Market1501. Here the hard gallery sampler is adopted. Experiments show that the *mAP* and *Rank-1* grow with the increase of nodes number, since more positive galleries can be sampled with larger number of nodes. Noted that the *mAP* does not gain much improvement when we increase node number from 100 to 500, that indicates the recall rate of positive samples is saturated, and only sampling 100 nodes can achieve the almost same performance as 500 ones. Actually, it demonstrates the outstanding effectiveness of the proposed sampler from another aspect. It also shows that we can dynamically adjust the size of the model, to balance the resource constraints and performance in practice.

The impact of graph layer number. Table 5 shows the influence of layer number in the GCN module on Market1501. The nodes number is 100. Most previous GCNs in various tasks use no deeper than 4 layers. Benefiting from the residual connection, we can extend the network deeper to obtain more powerful node representations. We compare the results from shallow to deep network (layers=3, 6, 9). Experiments show the *mAP* improves with the growth of layer numbers.

The impact of k' for edge connection. We show the impact of k' during graph edge connection in Table 6, *mAP* is basically stable when k' is very small. However, if k' is set too large, we find that *mAP* decreases rapidly, since large k' will lead to dense negative edge connections in the graph where the information aggregation is adversely affected.

Node Number	mAP	Rank-1	Rank-5	Rank-10
baseline	85.4	94.1	98.2	98.9
50	87.1	94.4	98.2	98.9
80	88.3	94.4	98.3	98.9
100	91.6	95.0	98.0	98.7
500	91.7	95.9	98.2	98.8

Table 4: The performances of the GCN module with different nodes number on Market1501.

Layer Number	mAP	Rank-1	Rank-5	Rank-10
baseline	85.4	94.1	98.2	98.9
3	88.4	94.1	98.0	98.8
6	89.0	94.5	98.2	98.9
9	91.6	95.0	98.0	98.7

Table 5: The performances of our GCN module with different number of layers on Market1501.

Qualitative results. Finally, we show the probe image and its related top-10 gallery images obtained by baseline model and CAGCN framework respectively. As shown in Figure 5, the proposed framework is able to address more hard positive samples and discard hard negative samples even if the probe images are in low quality. More comprehensively, we can notice that, despite that the person in the probe image

k'	mAP	Rank-1	Rank-5	Rank-10
baseline	85.4	94.1	98.2	98.9
5	91.2	95.1	98.1	98.8
8	91.6	95.0	98.0	98.7
10	91.4	94.9	98.0	98.8
15	91.0	94.9	98.1	98.9
50	89.9	94.7	98.0	98.8

Table 6: The impact of k' for graph edge connection on Market1501.

Figure 5: Qualitative results of three probes on Market1501. For each probe, the first row and second row correspond to the ranking results produced by baseline model and CAGCN framework. The blue, green and red box denotes the probe, true positive and false positive respectively.

is facing front, the framework can also retrieve the images where the person is facing back (line 2), and vice versa (line 4). Besides, the framework is more robust to color variance compared with the baseline (line 6). The same phenomenon can be found in other examples, and all cases illustrate that the proposed CAGCN framework can exploit the context information effectively and thus improve the performance.

Conclusion

In this work, we propose a novel CAGCN framework based on deep graph convolution network to utilize the context information among probe and gallery images for target re-identification. CAGCN tackles the challenges existing in graph construction and effective information aggregation, and utilizes the novel hard gallery sampler to obtain high recall for positive samples and keep a reasonable graph size. Additionally, edges among graph nodes are controlled by the original gallery-gallery relations to avoid redundant connections. The proposed framework achieves state-of-the-art performances on both person and vehicle benchmarks.

References

Alemu, L. T.; Shah, M.; and Pelillo, M. 2019. Deep Constrained Dominant Sets for Person Re-identification. In *ICCV*.

Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; and Wang, Z. 2019. ABD-Net: Attentive but Diverse Person Re-Identification. In *The IEEE International Conference on Computer Vision (ICCV)*.

Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; and Yang, Y. 2020a. Salience-Guided Cascaded Suppression Network for Person Re-Identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, X.; Zheng, L.; Zhao, C.; Wang, Q.; and Li, M. 2020b. RRGCCAN: Re-ranking via graph convolution channel attention network for person re-identification. *IEEE Access* 8: 131352–131360.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. 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*, 1335–1344.

Chu, R.; Sun, Y.; Li, Y.; Liu, Z.; Zhang, C.; and Wei, Y. 2019. Vehicle Re-identification with Viewpoint-aware Metric Learning. In *CVPR*.

Dai, Z.; Chen, M.; Zhu, S.; and Tan, P. 2019. Batch DropBlock Network for Person Re-identification and Beyond. In *ICCV*.

Feng, W.; Ji, D.; Wang, Y.; Chang, S.; Ren, H.; and Gan, W. 2018. Challenges on Large Scale Surveillance Video Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Garcia, J.; Martinel, N.; Micheloni, C.; and Gardel, A. 2015. Person re-identification ranking optimisation by discriminant context information analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, 1305–1313.

Guo, J.; Yuan, Y.; Huang, L.; Zhang, C.; Yao, J.; and Han, K. 2019. Beyond Human Parts: Dual Part-Aligned Representations for Person Re-Identification. In *CVPR*.

He, B.; Li, J.; Zhao, Y.; and Tian, Y. 2019. Part-Regularized Near-Duplicate Vehicle Re-Identification. In *CVPR*.

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

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019. Interaction-And-Aggregation Network for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, H.; Ji, D.; Gan, W.; Bai, S.; Wu, W.; and Yan, J. 2020. Class-wise Dynamic Graph Convolution for Semantic Segmentation. *arXiv preprint arXiv:2007.09690* .

Karaman, S.; Lisanti, G.; Bagdanov, A. D.; and Del Bimbo, A. 2014. Leveraging local neighborhood topology for large scale person re-identification. *Pattern Recognition* 47(12): 3767–3778.

Khorramshahi, P.; Kumar, A.; Peri, N.; Rambhatla, S. S.; Chen, J.; and Chellappa, R. 2019. A Dual-Path Model With Adaptive Attention For Vehicle Re-Identification. In *CVPR*.

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

Leng, Q.; Hu, R.; Liang, C.; Wang, Y.; and Chen, J. 2015. Person re-identification with content and context re-ranking. *Multimedia Tools and Applications* 74(17): 6989–7014.

Li, G.; Müller, M.; Thabet, A. K.; and Ghanem, B. 2019. Can GCNs Go as Deep as CNNs? In *ICCV*.

Li, W.; Wu, Y.; Mukunoki, M.; and Minoh, M. 2012. Common-near-neighbor analysis for person re-identification. In *2012 19th IEEE International Conference on Image Processing*, 1621–1624. IEEE.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. Deep-reid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 152–159.

Lin, J.; Ren, L.; Lu, J.; Feng, J.; and Zhou, J. 2017. Consistent-aware deep learning for person re-identification in a camera network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5771–5780.

Liu, X.; Liu, W.; Mei, T.; and Ma, H. 2016. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In *ECCV*.

Lou, Y.; Bai, Y.; Liu, J.; Wang, S.; and Duan, L. 2019. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3235–3243.

Luo, C.; Chen, Y.; Wang, N.; and Zhang, Z. 2019a. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 4976–4985.

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

Schumann, A.; and Stiefelhagen, R. 2017. Person re-identification by deep learning attribute-complementary information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 20–28.

Shen, Y.; Li, H.; Yi, S.; Chen, D.; and Wang, X. 2018. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 486–504.

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

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

Su, H.; Gan, W.; Wu, W.; Yan, J.; and Qiao, Y. 2020a. BSN++: Complementary Boundary Regressor with Scale-Balanced Relation Modeling for Temporal Action Proposal Generation. *arXiv preprint arXiv:2009.07641* .

Su, H.; Zhao, X.; and Lin, T. 2018. Cascaded pyramid mining network for weakly supervised temporal action localization. In *Asian Conference on Computer Vision*, 558–574. Springer.

Su, H.; Zhao, X.; Lin, T.; Liu, S.; and Hu, Z. 2020b. Transferable Knowledge-Based Multi-Granularity Fusion Network for Weakly Supervised Temporal Action Detection. *IEEE Transactions on Multimedia* .

Subramaniam, A.; Chatterjee, M.; and Mittal, A. 2016. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, 2667–2675.

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

Tang, Z.; Naphade, M.; Birchfield, S.; Tremblay, J.; Hodge, W.; Kumar, R.; Wang, S.; and Yang, X. 2019. PAMTRI: Pose-Aware Multi-Task Learning for Vehicle Re-Identification Using Highly Randomized Synthetic Data. In *The IEEE International Conference on Computer Vision (ICCV)*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* .

Wang, F.; Zuo, W.; Lin, L.; Zhang, D.; and Zhang, L. 2016. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1288–1296.

Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; and Sun, J. 2020. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019a. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, 3623–3632.

Wang, X.; Hua, Y.; Kodirov, E.; Hu, G.; Garnier, R.; and Robertson, N. M. 2019b. Ranked List Loss for Deep Metric Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial Fine-Grained Composition Learning for Unseen Attribute-Object Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3741–3749.

Wu, J.; Yang, Y.; Liu, H.; Liao, S.; Lei, Z.; and Li, S. Z. 2019. Unsupervised graph association for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 8321–8330.

Xia, B. N.; Gong, Y.; Zhang, Y.; and Poellabauer, C. 2019. Second-Order Non-Local Attention Networks for Person Re-Identification. In *The IEEE International Conference on Computer Vision (ICCV)*.

Yang, W.; Huang, H.; Zhang, Z.; Chen, X.; Huang, K.; and Zhang, S. 2019. Towards Rich Feature Discovery With Class Activation Maps Augmentation for Person Re-Identification. In *CVPR*.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, 34–39. IEEE.

Yu, H.-X.; Wu, A.; and Zheng, W.-S. 2017. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 994–1002.

Yu, R.; Dou, Z.; Bai, S.; Zhang, Z.; Xu, Y.; and Bai, X. 2018. Hard-aware point-to-set deep metric for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 188–204.

Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; and Sun, J. 2017. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184* .

Zheng, F.; Deng, C.; Sun, X.; Jiang, X.; Guo, X.; Yu, Z.; Huang, F.; and Ji, R. 2018. Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training. *arXiv preprint arXiv:1810.12193* .

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *ICCV*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *ICCV*.

Zhongdao Wang, Liang Zheng, Y. L.; and Wang, S. 2019. Linkage-based Face Clustering via Graph Convolution Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-Scale Feature Learning for Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhou, Q.; Fan, H.; Zheng, S.; Su, H.; Li, X.; Wu, S.; and Ling, H. 2018. Graph correspondence transfer for person re-identification. *arXiv preprint arXiv:1804.00242* .

Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; and Wang, J. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.