Pixel-wise Graph Attention Networks for Person Re-identification

Wenyu Zhang,   Qing Ding*,   Jian Hu,   Yi Ma,   Mingzhe Lu

University of Science and Technology of China

Hefei, AnHui, China

wenyuz@mail.ustc.edu.cn,dingqing@ustc.edu.cn,hujian1@mail.ustc.edu.cn

my1996mh@mail.ustc.edu.cn,mingzhe@mail.ustc.edu.cn

ABSTRACT

Graph convolutional networks (GCN) is widely used to handle irregular data since it updates node features by using the structure information of graph. With the help of iterated GCN, high-order information can be obtained to further enhance the representation of nodes. However, how to apply GCN to structured data (such as pictures) has not been deeply studied. In this paper, we explore the application of graph attention networks (GAT) in image feature extraction. First of all, we propose a novel graph generation algorithm to convert images into graphs through matrix transformation. It is one magnitude faster than the algorithm based on K Nearest Neighbors (KNN). Then, GAT is used on the generated graph to update the node features. Thus, a more robust representation is obtained. These two steps are combined into a module called pixel-wise graph attention module (PGA). Since the graph obtained by our graph generation algorithm can still be transformed into a picture after processing, PGA can be well combined with CNN. Based on these two modules, we consulted the ResNet and design a pixel-wise graph attention network (PGANet). The PGANet is applied to the task of person re-identification in the datasets Market1501, DukeMTMC-reID and Occluded-DukeMTMC (outperforms state-of-the-art by 0.8%, 1.1% and 11% respectively, in mAP scores). Experiment results show that it achieves the state-of-the-art performance.

CCS CONCEPTS

• **Computing methodologies → Image representations**; **Visual content-based indexing and retrieval**.

KEYWORDS

graph convolutional networks, graph Generation, self-attention, person re-identification

---

*Corresponding author.

---

1 INTRODUCTION

Person re-identification attracts the attention of scholars. These studies mainly use the image of a specific pedestrian as the query to find all the images of the pedestrian in the datasets (gallery). The main application fields of it include video surveillance, intelligent security and so on. There are many challenges in person re-identification tasks: limited training data, occlusion, lighting, scale, pose, and so on. How to obtain robust features of the image is the key to solving these challenges.

Convolutional neural networks (CNN) has been successfully applied in the field of person re-identification. Existing methods mainly adopt ResNet [1] as the backbone to extract the features of images. However, due to the limitation of CNN's receptive field, the CNN based methods are difficult to identify blurred pictures or small targets. To solve above problems, it is necessary to give higher weight to human body information in the feature extraction stage and extend the receptive field to obtain global information.

Since the attention mechanism [2] can get the correlation between objects, and generate different weights. CNN with attention mechanism [3–5] is a good way to solve the above problems. In the field of person re-recognition, Some typical models [6, 7] apply attention mechanism to the feature maps output by CNN to calculate the importance of each feature in feature maps and assign different weights to highlight important features and suppress unimportant features. However, the existing methods only get the relationship between nodes as one-order relationship [6] or global relationship [7], which can't obtain the implicit high-order relationship.

To expand the receptive field, dilated convolutions can be applied in the feature extraction stage [8], but the continuity of features will be lost due to the discontinuity of convolution kernel. As a result, it is difficult to obtain complete features of small objects.

Human beings get the overall cognition through the continuous expansion of a certain point to the surroundings. Inspired by this ecology, graph attention networks (GAT) [9] can be used to solve the above problems by simulating the characteristics of human vision. The features of nodes are propagated on the edges given different weights, and each node obtains a new representation by aggregating the neighbor features. In this way, besides important features will be assigned a higher weight, the range of receptive field will be scalable.

The premise of using GAT is to convert an image into a graph and obtain its adjacency matrix. As shown in Fig. 1, a single pixel is regarded as a node, and its adjacent pixels are neighbors. Based
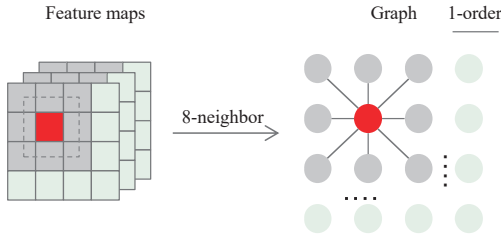
Figure 1: We treat the picture as a regular graph and optimize the features of the picture from the perspective of graph convolution.

on this setting, we propose a novel graph generation algorithm, which is not only faster than the traditional algorithm but also obtains a sparse and strictly local graph. Subsequently, using GAT on the obtained graph, not only can obtain high-order information of neighbor nodes but also get a large receptive field. These two steps are combined into a module called pixel-wise graph attention module (PGA). Combined with ResNet, we build a new model pixel-wise graph attention network (PGANet). And we apply the model to the commonly used data sets Market1501[10], DukeMTMC-reID[11] and Occluded-DukeMTMC[12] for person re-identification.

In summary, the contribution of this paper is threefold:

- We propose a novel graph generation algorithm: The algorithm can transform the image into a sparse and strictly local graph, which can make use of the structure information. It is one order of magnitude faster than the traditional KNN based algorithm.
- We propose a PGA module: Multi-layer PGA can simulate human vision. With the continuous iteration of PGA, the receptive field is continuously expanded and high-level information is obtained.
- We propose a new model PGANet: It is composed of CNN and PGA. We apply it to person re-identification to demonstrate the effectiveness of our model.

2 RELATED WORK

In this section, we introduce the existing methods of image feature extraction and the application of graph convolution networks in the field of computer vision.

2.1 Image Feature Extraction Based On CNN

CNN based methods for person re-identification often use ResNet [11, 13, 14] as the backbone to extract image features. However, due to the limited receptive field, CNN can not obtain global information. As a result, it is difficult to distinguish blurred images and small targets. To solve this problem, some methods have been proposed. The part-based models [15? –20], divide the picture into sub-pictures to get the part-level features. So they can provide fine-grained information for better representation. But the processed

images are required to be aligned. Attention-based models [6, 7, 21] assign different importance to different features in the feature extraction process. Therefore, it can obtain the relationship between each pixel in the picture. But, the existing methods only obtain one-order relationships or global information[7] and the implicit high-order information cannot be obtained. Dilated convolutions can be applied in the feature extraction to overcome the limited receptive field problem [8], but the continuity of features will be lost due to the discontinuity of the convolution kernel.

2.2 Image Feature Extraction Based On GCN

GCN aggregates neighbor node data along the direction of the edge to update the new representation of the source node. It is mainly divided into two categories: spectral domain approaches [22–24] and spatial domain approaches [9, 25–27]. Because node neighbors can be flexibly aggregated and it has lower computational complexity, spatial GCN is the mainstream method. Hamilton et al.[26] propose the GraphSAGE framework, which efficiently generates node embeddings for previously invisible data using node feature information. In GAT model [9], attention mechanism is adopted in the propagation stage. Li et al. [27] proposed that GCN can go deeper as CNN by applying residual structure to alleviate the problem of vanishing gradient. Recently, GCN has been widely used in the field of computer vision. In the task of multi-label image recognition, Chen et al. [28] propose a novel reweighting scheme to create an effective label correlation matrix to guide the information dissemination between nodes in GCN. Context-aware graph convolutional network (CAGCN) [29], in which the relationship of the probe-gallery is encoded into the graph, processes the hard samples through graph reasoning by using the context information flow and other simple samples. However, GCN is rarely used in image feature extraction. Defferrard et al. [23] proposed to apply a clustering algorithm to convert the image into the graph, and then use GCN in the generated graph. The computation complexity of the method to build a graph is as high as $O(N^2)$. In the Dual Attention Network (DANet) model[30], attention is applied between each pixel to establish a fully connected graph, which was used to aggregate the features of nodes and applied it in image segmentation. But the graph obtained in this way connects each node to all nodes, ignoring the structural information.

3 METHODOLOGY

PGANet is the backbone of image processing in our method. Its structure is shown in Fig. 2 and 3. Firstly, we design a graph generation module to transform the output of CNN into a graph. Then the graph is processed by GAT. These two modules are built into a multi-layer PGA. Finally, multi-layer PGA is inserted into Conv_x of ResNet. As shown in Fig. 2, a new framework PGANet is constructed by stacking multi-layer PGA and CNN.

3.1 Graph Generation Algorithm

The graph can be formulated as $G = < V, E, A >$, in which V is the set of nodes in the graph, E is the edges between nodes in the graph, and the adjacency matrix A is usually used to represent the connection relationship of each node. The feature maps output by CNN is expressed as $F \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent
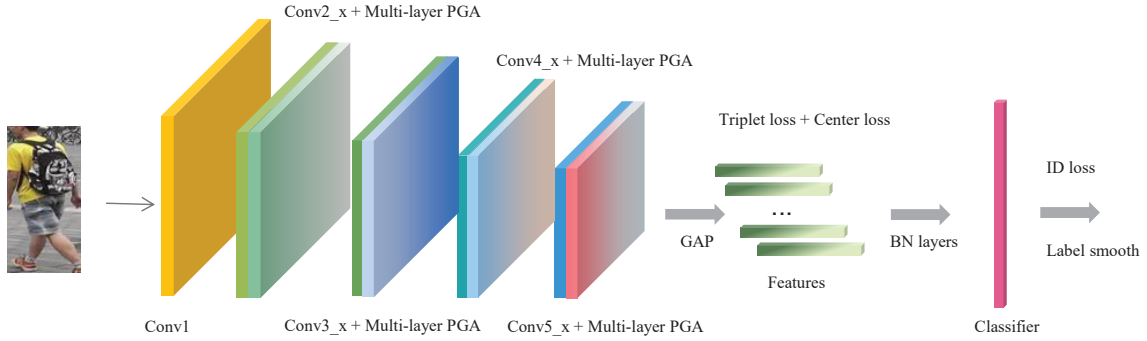
Figure 2: In PGANet, the features extracted by CNN are continuously optimized in the PGA module. Ultimately get a more robust picture representation

the height, width, and channel respectively. We regard the feature maps as regular graph. The pixel in the feature map is regarded as the node in the graph, and its value in every channel is the feature of the node, which is a C-dimensional feature vector.

$$A = \begin{cases} 1, & v_j \text{ is the neighbor of } v_i \\ 0, & v_j \text{ isn't the neighbor of } v_i \end{cases} \quad (1)$$

As shown in Eq. 1, a graph generation algorithm is needed to generate a sparse and strictly local adjacency matrix A. The traditional method uses a clustering algorithm to obtain the adjacency matrix of the graph, the time complexity is as high as $O(N^2)$. We propose a novel graph generation algorithm, which can efficiently find 4-neighbor, 8-neighbor, or 2-neighbor of nodes. In particular, the 2-neighbor's graph is constructed in the channel of the picture, thus each feature map is flattened as the feature of the node. Our algorithm finds neighbors of nodes in parallel and does not need to calculate the distance between nodes. The time complexity of algorithm 1 is $O(kHW)$, since the **for** loop has $O(H)$ complexity, the operation of **addition**, **subtraction**, **slicing**, and **assignment** in each loop has $O(5kW)$ complexity. We know that $N = H * W$ and k is 2, 4, or 8. When N is very large, the time complexity of algorithm 1 is smaller than the traditional algorithm, $O(kHW) << O(N^2)$. As shown in Algorithm 1, the algorithm which generates a 4-neighbor graph is described.

3.2 Pixel-wise Graph Attention Networks

PGANet is composed of multiple PGA modules connected in series. The operation process of PGA is shown in Fig. 3. The self-attention mechanism utilized by PGA closely follows the work of non-local operations [3]. In PGA, the propagation mechanism of node features is modified by GAT [9].

We first describe a single PGA. As shown in Fig. 3, the feature maps $F \in \mathbb{R}^{C \times W \times H}$ output by Conv_x in ResNet are used as the input of PGA. The corresponding adjacency matrix A is generated based on the graph generation algorithm proposed in Section 3.1. Subsequently, the correlation matrix R is obtained by calculating the correlation of each node in the feature maps, as shown in Eq 2.

Algorithm 1 Graph Generation Algorithm

Input: h is the height of feature maps, w is the width of feature maps, n = h × w is the number of pixels in the feature maps, AD-JACENCYGENERATION(**node**, **neighbor**)is a function to generate adjacency matrix based on node pairs.

Output: A is Adjacency matrix of feature maps.

1: **function** GRAPHGENERATION(h, w, n)
2:     $M \leftarrow$ RANGE(n).RESHAPE(h, w)
3:     $r \leftarrow M$.rows ;
4:     **node** $\leftarrow$ [] ; **neighbor** $\leftarrow$ [];
5:     **for** $i = 0 \rightarrow$ h − 1 **do**
6:         **node** $\leftarrow$ $r_i[1:]$
7:         **neighbor** $\leftarrow (r_i - 1)[1:]$
8:         **node** $\leftarrow r_i[:-1]$
9:         **neighbor** $\leftarrow (r_i + 1)[:-1]$
10:        **if** i != h-1 **then**
11:           **node** $\leftarrow r_i$
12:           **neighbor** $\leftarrow (r_i - w)$
13:        **end if**
14:        **if** i != 0 **then**
15:           **node** $\leftarrow r_i$
16:           **neighbor** $\leftarrow (r_i + w)$
17:        **end if**
18:     **end for**
19:     $A \leftarrow$ ADJACENCYGENERATION(**node**, **neighbor**)
20:     **return** A
21: **end function**

$$R = \theta(F)\phi(F)^T \quad (2)$$

where $()^T$ stands for transpose. $\phi()$ and $\theta()$ are the same transfer functions with different parameters, which including 1×1 convolutional layers, batch normalization, ReLU activation, and matrix transformation. The transfer function changes the feature maps to new feature space and transforms the new feature maps into a set of nodes through matrix transformation, e.g., $V' = \phi(F)$, it can be
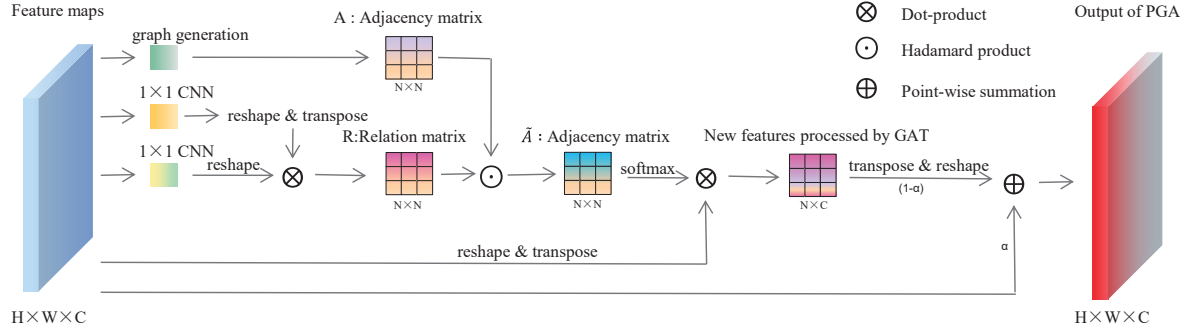
Figure 3: Diagram of our proposed pixel-wise graph attention module (PGA). PGA obtains the attention between each node through non-local operations. We use GAT to update the features of each node.

defined as $V' = \left\{ v_i' \in \mathbb{R}^{C'}, i = 1, \ldots, N \right\}$, where v_i' represents the ith node, C' is the transformed feature dimension, $N = H \times W$ is the number of all nodes in the feature maps. As shown in the Fig. 3, we obtain a matrix of node correlation $R \in \mathbb{R}^{N \times N}$ through the help of matrix change and matrix multiplication.

The previous method [30] regards this correlation matrix R as a graph of feature maps. But R is a fully connected graph, each node is connected to all nodes. As a result, all structural information is dropped [9]. Therefore, we map the correlation matrix R to the adjacency matrix A. We obtain the adjacency matrix with correlation and use the softmax function to normalize it. As shown in Eq. 3:

$$\tilde{A} = \text{softmax}(A \odot R) \quad (3)$$

As shown in Eq.3, different weights are assigned to all the edges on the graph. We get a sparse and strictly local adjacency matrix $\tilde{A} \in \mathbb{R}^{N \times N}$. The next step is to propagate the node features on the graph. As shown in Eq. 4:

$$\tilde{V} = \text{ReL}\,U(\tilde{A}V) \quad (4)$$

The new feature is $\tilde{V} \in \mathbb{R}^{N \times C}$. By using matrix transformation, it can be converted into feature maps again. Essentially, PGA is a variant of GAT. The single-layer PGA can be expressed as:

$$PGA(F) = \text{transf}\left(\text{softmax}\left(A \odot R\right)\text{transf}(F)\right). \quad (5)$$

where the operation of function $transf()$ is related to the form of the input. If the form of the input data is $\mathbb{R}^{C \times H \times W}$, we first transpose it and then execute function $reshape(-1, C)$. If the form is $\mathbb{R}^{N \times C}$, we transpose it and then execute function $reshape(H, W, C)$. As same as GCN, PGA has the problems of vanishing gradient and over-smooth. We add the residual structure to the PGA with regarding to DeepGCNs [27]. As shown in Eq. 6:

$$\tilde{F} = \alpha F + (1 - \alpha)PGA(F) \quad (6)$$

where α is a learnable parameter. Therefore, multi-layer PGA is shown in Eq. 7:

$$\tilde{F}^{l+1} = \alpha F^l + (1 - \alpha)PGA\left(F^l\right) \quad (7)$$

3.3 Loss Function

$$L = L_{ID} + L_{\text{Triplet}} + \beta L_C \quad (8)$$

AS shown in Eq. 8, the loss function of the model mainly includes ID loss, triplet loss[14], and Center loss[?]. Following strong baseline[31], ID Loss use label smoothing strategy.

4 EXPERIMENT

We introduced the implementation details of our model in Section 4.1. The datasets and evaluation metrics adopted in the experiment are introduced in section 4.2. In section 4.3, we compare our model with several state-of-the-art methods in the task of person re-identification to verify the superiority of it. In section 4.4, the effectiveness of the graph generation algorithm and multi-layer PGA module is proved through ablation study.

4.1 Implementation Details

We use ResNet [1] to build the model. Refer to [7], we insert multi-layer PGA module after each Conv_x module (including Conv_2, Conv_3, Conv_4 and Conv_5). In our model, the number of layers of PGA is set to 3. The input image size is 256X128, and the padding value is set to 10. We use all the tricks summarized in Strong Baseline [31]. To expand the size of the feature maps output by the model, the last spatial down-sampling operation is set to 1. The batch Standardization (BN) layer is added before the classifier. For data augmentation, we apply random cropping, horizontal flipping, and random erasing. The random probability for image horizontal flip and random erasing is set to 0.5 respectively. In our experiments includes three losses as following: Identification loss[32], Triplet loss[33] and Center loss[?]. We adopt Adam optimizer with the weight decay of 5×10^{-4} and the learning rate as $3e^{-4}$. The warm-up strategy is used for the learning rate, and iterations of warm-up are set to 500. Finally, at the re-rank stage, we use the k-reciprocal method[34].

4.2 Datasets and Evaluation Metrics

We verify PGANet on the common datasets including Market1501[10], DukeMTMC-reID[11] and Occluded-DukeMTMC[12].

Market1501 It contains 1501 pedestrians captured by 6 cameras (5 high-definition cameras and 1 low-definition camera) and 32668 detected pedestrian rectangular boxes. Each pedestrian is captured by at least 2 cameras and may have multiple images in one camera. The training set contains 751 people, with 12936 images, and each person has an average of 17.2 training data. In the test set, there are 750 people with 19732 images, and each person has an average of 26.3 test data.

DukeMTMC-reID It is a new type large high definition (HD) video data set recorded by 8 synchronous cameras. The video is sampled every 120 frames to form an image, and total 36, 411 images are obtained. A total of 1404 people appeared under more than two cameras, and 408 people appeared under only one camera. 702 people are randomly sampled into the training set and 1110 people in the test set. In the test set, one photo under each camera of each ID is sampled as a query image. The remaining images are added to the search gallery of the test, and the previous 408 people are also added to the gallery as interference items.

Occluded-DukeMTMC It derived from DukeMTMC-reID. In this dataset, all query images are occluded by large variety of objects (e.g., trees, cars, other persons). The training, query, and gallery set contain 14%/15%/10% occluded images, respectively. The training set of Occluded DukeMTMC contains 15,618 images covering 702 identities in total. The testing set contains 1,110 identities, with 17, 661 gallery images and 2, 210 query images.

Evaluation Metrics Since cumulative matching characteristics (CMC) and mean average precision (mAP) are commonly used metrics for person re-identification tasks, they are also considered as evaluation indicators in this paper.

Table 1: Performance (%) comparisons with the state-of-the-arts methods on Market1501 and DukeMTMC-reID.

| Type | Method | Market1501 | | DukeMTMC | |
		mAP	Rank-1	mAP	Rank-1
Pose-guided	PGFA [12]	76.8	91.2	65.5	82.6
	HOReI [35]	84.9	94.2	75.6	86.9
Part-feature	BDB [36]	84.3	94.2	72.1	86.8
	Pyramid[37]	88.2	95.7	79	89
	MGN[20]	86.9	95.7	78.4	88.7
	AANet-152 [38]	83.41	93.93	74.29	87.65
	ISP [39]	88.6	95.3	80	89.6
Attention	ABD-Net[40]	88.2	95.6	78.5	89
	RGA-SC [7]	88.1	95.8	74.9	86.1
	Mancs [41]	82.3	93.1	71.8	84.9
Other	Baseline [31]	85.9	94.5	76.4	86.4
	CAGCN [29]	**91.7**	**95.9**	**85.9**	**91.3**
Ours	**PGANet-50**	88.5	95.4	78.8	89.4
	PGANet-152	89.3	95.4	79.6	89.1

4.3 Comparison With State-of-the-Art Methods

PGANet versus Other Model. To prove the effectiveness of PGANet. We compare it with several state-of-the-art models in person re-identification task. As shown in Table 1, our model achieves competitive performance in rank stage. We tested the effects of PGANet using ResNet50 and ResNet152 as backbones respectively. Compared with the attention-based model, PGANet mixes graph structure and attention to optimize features, thus can obtain hidden high-order information and highlight important features well. Compared with part-based models, in PGANet, global features can be obtained to achieve competitive performance. With the help of the part-based method, we can achieve better performance. Without any additional models, PGA module can more easily obtain important features than Pose-guided models. CAGCN can obtain better performance by using probe-gallery relations with the help of picture-wise GCN. So we can draw the conclusion that GCN has great potential in the field of computer vision.

Table 2: Performance (%) comparisons with the state-of-the-arts methods on Occluded-Duke.

| Method | Occluded-Duke | |
	mAP	Rank-1
Ad-Occluded [42] (CVPR2018)	32.2	44.5
PGFA [12] (ICCV2019)	37.3	51.4
HOReID [35] (CVPR2020)	43.8	55.1
PGANet-50	50.5	59.7
PGANet-152	**54.8**	**65.2**

Experiment on Occluded-Duke As shown in Table2, PGANet achieves mAP 54.8% and Rank-1 65.2% on Occluded-Duke. Compared with HOReID, it has an increase of 11 points. The results show that PGANet can give higher attention to important features, thus obtain more robust representation even in complex scenes

Table 3: Performance (%) comparisons with the state-of-the-arts in rerank

| Method | Market1501 | | DukeMTMC | |
	mAP	Rank-1	mAP	Rank-1
Strong Baseline+RR [31] (CVPR2019)	94.2	95.2	89.1	90.3
AANet-152+RR [38] (CVPR2020)	92.38	95.1	86.87	90.36
CAGCN+RR [29] (AAAI2021)	94.0	96.2	88.9	**92.3**
PGANet-50+RR	94.6	96.0	89.9	91.5
PGANet-152+RR	**94.8**	96.2	**90.0**	92.0

Re-rank experiment As shown in Table 3, our model also achieves excellent performance (mAP 94.6% and Rank-1 96.0%). In the re-rank stage, the k-reciprocal method [34] is adopted as post-processing in our model, which optimizes the results by making effective use of the relationship between the ranking results and the query images. Why does our model achieve state-of-the-art performance after using the re-rank strategy? We compare our model with the best model CAGCN in Table 1. The core idea of CAGCN is to take advantage of the probe-gallery relations, which is consistent with the idea of k-reciprocal method. It essentially uses

information between pictures. Our model focuses on optimizing the features of images and ignores the relationship between images. Therefore, Once the relationship between pictures is also used, the model effect has been greatly improved. The experimental results also prove that our model can obtain more robust features.

4.4 Ablation Study

Advantages of Graph Generation algorithm To test the efficiency of our algorithm, we compare the algorithm with the traditional KNN based algorithm. The size of the feature map used for testing is [16,8,512]. We evaluate the time required to create three graphs in which nodes have different numbers of neighbors, ie. 2-neighbor, 4-neighbor, and 8-neighbor. In 2-neighbor, the graph is built on the channel of the picture, and in 4-neighbor and 8-neighbor, is built on the pixels of the picture. The values are obtained as the average of running time of the algorithm, and drawn as a line graph. As shown in Fig. 4, the algorithm proposed in this paper increases the speed of graph generation by an order of magnitude, at least 12 times faster than a traditional KNN based algorithm.
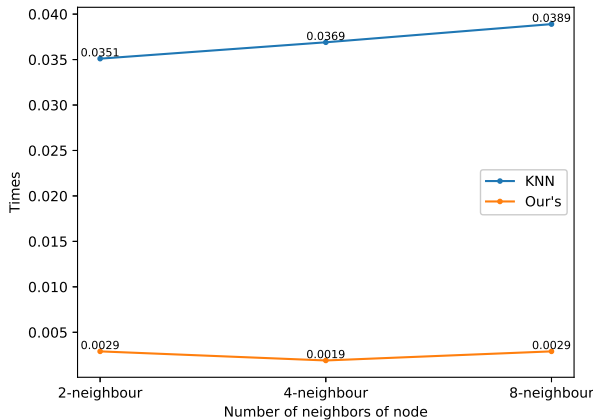

Figure 5: The effect of different layers of multi-layer PGA on the results.

Figure 4: The speed of algorithm.

Figure 6: The impact of using different graphs on the model.

Influence of neighbors In PGA, graph convolution to optimize feature maps for better representation. In fully connected graphs, the non local method is used for nodes to obtain the one-to-all relationship. As shown in Fig. 6, better performance is gotten since structural information can be obtained by the model, with the help of the better sparse and strict local graph produced by our algorithm.

5 CONCLUSION

In this paper, we explored the application of GCN in the image feature extraction stage. A new image feature extraction framework PGANet is proposed, which can obtain a larger receptive field and make good use of structural information to obtain hidden high-order information. PGANet uses an alternate structure of CNN and multi-layer PGA.CNN is used to extract image features, multi-layer PGA is used for the obtained feature maps to optimize features and

Advantage of multi-layer PGA In this section, we test the influence of the number of layers of PGA on the model. Fig. 5 shows continuous enhancement of the effect with the increase of the number of PGA layers. Compared with strong baseline, the mAP of the 3-layer model has increased by 2.6%, and the accuracy of Rnk-1 has increased by 0.9% (based on ResNet50). The experimental results show that the larger receptive field effectively improves the model's representation ability and can obtain hidden high-level information as the number of PGA layers increases. As shown in Fig. 5 and 6, even if the model has been trained for many rounds, we can not find the phenomenon of vanishing gradient and over smooth. It is noticeable that PGA with residual structure can effectively alleviate the vanishing gradients problem.
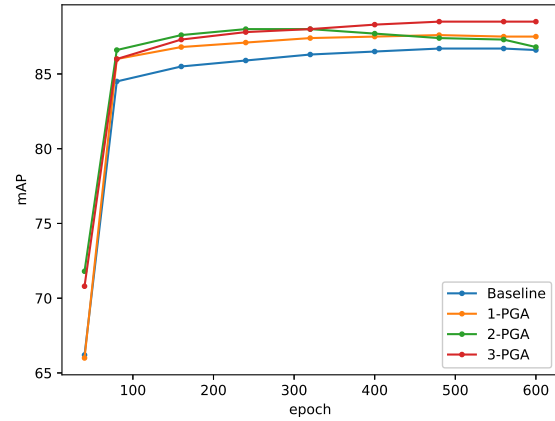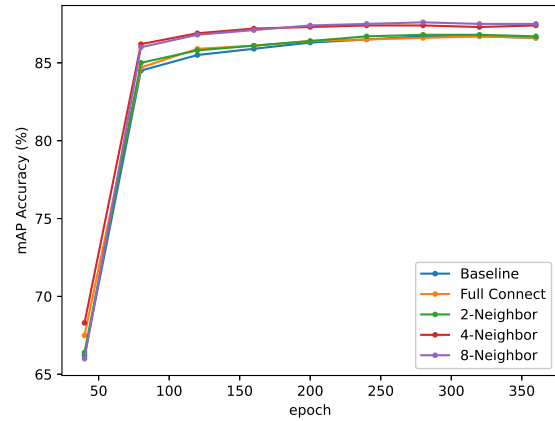
expand the receptive field. Therefore, PGANet can get a more robust representation of the picture. To apply GCN in the feature extraction stage of the picture, a brand-new graph generation algorithm is proposed which can efficiently generate a sparse and strictly local graph corresponding to the picture. Combine our algorithm with the modified GAT, we get the PGA module. With iteration of PGA, high-ord information can be obtained. Extensive experiments on person re-identification demonstrate the effectiveness of our proposed framework.

ACKNOWLEDGMENTS

REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December, 770 – 778. http://dx.doi.org/10.1109/CVPR.2016.90

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 2017-December, 5999 – 6009.

[3] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), 7794 – 7803.

[4] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 8 (2020), 2011 – 2023.

[5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[6] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. 2019. Self-critical attention learning for person re-identification. *Proceedings of the IEEE International Conference on Computer Vision* 2019-October, 9636 – 9645. http://dx.doi.org/10.1109/ICCV.2019.00973

[7] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-Aware Global Attention for Person Re-Identification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3183 – 3192. http://dx.doi.org/10.1109/CVPR42600.2020.00325

[8] Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

[9] Petar Velikovi, Cucurul Guillem, Casanova Arantxa, Romero Adriana, and Bengio Pietro, Lio'and Yoshua. 2017. Graph Attention Networks. *ICLR* (2017).

[10] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. *Proceedings of the IEEE International Conference on Computer Vision* 2015 International Conference on Computer Vision, ICCV 2015, 1116 – 1124. http://dx.doi.org/10.1109/ICCV.2015.133

[11] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9914 LNCS, 17 – 35. http://dx.doi.org/10.1007/978-3-319-48881-3_2

[12] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang DIng, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. *Proceedings of the IEEE International Conference on Computer Vision* 2019-October, 542 – 551. http://dx.doi.org/10.1109/ICCV.2019.00063

[13] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. 2017. SVDNet for Pedestrian Retrieval. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October (2017), 3820 – 3828. http://dx.doi.org/10.1109/ICCV.2017.410

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. arXiv:1703.07737 [cs.CV]

[15] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11208 LNCS (2018), 501 – 518. http://dx.doi.org/10.1007/978-3-030-01225-0_30

[16] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. 2017. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October (2017), 350 – 359.

[17] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-Learned Part-Aligned Representations for Person Re-identification. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October, 3239 – 3248. http://dx.doi.org/10.1109/ICCV.2017.349

[18] Yumin Tian, Qiang Li, Di Wang, and Bo Wan. 2018. Robust joint learning network: improved deep representation learning for person re-identification. *Multimedia Tools and Applications* 78, 17 (2018), 24187–24203.

[19] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. 2019. Horizontal pyramid matching for person re-identification. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019* (2019), 8295 – 8302.

[20] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, 274 – 282. http://dx.doi.org/10.1145/3240508.3240552

[21] Chaoqun Wan, Yue Wu, Xinmei Tian, Jianqiang Huang, and Xian-Sheng Hua. 2020. Concentrated Local Part Discovery with Fine-Grained Part Representation for Person Re-Identification. *IEEE Transactions on Multimedia* 22, 6 (2020), 1605 – 1618.

[22] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral networks and deep locally connected networks on graphs. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (2014).

[23] Michael Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems* 0, 3844 – 3852.

[24] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (2017).

[25] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M. Bronstein. 2017. Geometric deep learning on graphs and manifolds using mixture model CNNs. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January, 5425 – 5434. http://dx.doi.org/10.1109/CVPR.2017.576

[26] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems* 2017-December, 1025 – 1035.

[27] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. 2019. DeepGCNs: Can GCNs go as deep as CNNs? *Proceedings of the IEEE International Conference on Computer Vision* 2019-October, 9266 – 9275. http://dx.doi.org/10.1109/ICCV.2019.00936

[28] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June, 5172 – 5181. http://dx.doi.org/10.1109/CVPR.2019.00532

[29] Deyi Ji, Haoran Wang, Hanzhe Hu, Weihao Gan, Wei Wu, and Junjie Yan. 2021. Context-Aware Graph Convolution Network for Target Re-identification. *AAAI 2021* (2021). arXiv:2012.04298 [cs.CV]

[30] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June, 3141 – 3149. http://dx.doi.org/10.1109/CVPR.2019.00326

[31] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2019-June, 1487 – 1495. http://dx.doi.org/10.1109/CVPRW.2019.00190

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December (2016), 2818 – 2826. http://dx.doi.org/10.1109/CVPR.2016.308

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June-2015, 815 – 823. http://dx.doi.org/10.1109/CVPR.2015.7298682

[34] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January, 3652 – 3661. http://dx.doi.org/10.1109/CVPR.2017.389

[35] Wang Guan'an, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6448 – 6457. http://dx.doi.org/10.1109/CVPR42600.2020.00648

[36] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. 2019. Batch dropblock network for person re-identification and beyond. *Proceedings of*

the IEEE International Conference on Computer Vision 2019-October, 3690 − 3700. http://dx.doi.org/10.1109/ICCV.2019.00379

[37] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. 2019. Pyramidal person re-identification via multi-loss dynamic training. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June, 8506 − 8514. http://dx. doi.org/10.1109/CVPR.2019.00871

[38] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. 2019. Aanet: Attribute attention network for person re-identifications. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June, 7127 − 7136.

[39] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-Guided Human Semantic Parsing for Person Re-identification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12348 LNCS, 346 − 363. http://dx.doi.org/10.

1007/978-3-030-58580-8_21

[40] Tianlong Chen, Shaojin DIng, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. 2019. ABD-net: Attentive but diverse person re-identification. *Proceedings of the IEEE International Conference on Computer Vision* 2019-October, 8350 − 8360. http://dx.doi.org/10.1109/ICCV.2019.00844

[41] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2018. *Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification*. Vol. 11208 LNCS. Munich, Germany. 384 − 400 pages. http://dx.doi.org/10.1007/978-3-030-01225-0_23

[42] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. 2018. Adversarially Occluded Samples for Person Re-identification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 5098 − 5107. http://dx.doi.org/10.1109/CVPR.2018.00535