PGGANet: Pose Guided Graph Attention Network for Person Re-identification

Zhijun He[a,b], Hongbo Zhao*[a,b], Wenquan Feng[a]

[a] School of Electronics and Information Engineering, Beihang University, Beijing, China 100191;
[b] Hefei Innovation Research Institute of Beihang University, Hefei, China 230012

*bhzhb@buaa.edu.cn

ABSTRACT

Person re-identification (ReID) aims at retrieving a person from images captured by different cameras. For deep-learning-based ReID methods, it has been proved that using local features together with global feature of person image could help to give robust feature representations for person retrieval. Human pose information could provide the locations of human skeleton to effectively guide the network to pay more attention on these key areas and could also help to reduce the noise distractions from background or occlusions. However, methods proposed by previous pose-related works might not be able to fully exploit the benefits of pose information and did not take into consideration the different contributions of different local features. In this paper, we propose a pose guided graph attention network, a multi-branch architecture consisting of one branch for global feature, one branch for mid-granular body features and one branch for fine-granular key point features. We use a pre-trained pose estimator to generate the key-point heatmap for local feature learning and carefully design a graph attention convolution layer to re-evaluate the contribution weights of extracted local features by modeling the similarities relations. Experiments results demonstrate the effectiveness of our approach on discriminative feature learning and we show that our model achieves state-of-the-art performances on several mainstream evaluation datasets. We also conduct a plenty of ablation studies and design different kinds of comparison experiments for our network to prove its effectiveness and robustness, including holistic datasets, partial datasets, occluded datasets and cross-domain tests.

Keywords: Person Re-identification, Convolutional Network, Human Pose Estimation, Deep Learning, Graph Convolution Network.

1. INTRODUCTION

Person re-identification (reID), aiming to retrieve a specific person across different times and cameras, has been getting more and more attention from both industry and academia. With the great advances in deep convolutional neural network over the last decade, person representations could be extracted in a much more discriminative and robust way, which pushed the performance of ReID to a new level. Many existing researches [1,2,3,4,5,6] have shown that CNN-based methods could achieve better identification rates and mean average precision compared with traditional hand-crafted methods. In recent years, more and more studies tend to descript person identities based on local feature mining, which has been confirmed to be effective in many previous works [7,8,9,10,11]. Learning part-level features could help to filtering the useless information and make the networks concentrate more on identity-relevant areas. In this paper, we aim to improve the performance and robustness of person representation by using pose information and Graph Convolutional Network (GCN).

The motivation of this work is threefold: **First**, pose-guided mechanism could make the network pay more attention on identification-related areas from person images. For example, the network should be trained to be sensitive to body parts rather than the bike or trees in background. Generally, when we need to tell whether two images are from the same person, the most natural and correct way is to focus on the person-related parts but not some confusing background factors. With the help of pose cues, the network could effectively get rid of useless background noises and therefore perform better especially for industry application scenarios where we could hardly build a specific new dataset for model training and deployment. **Second**, there have been several methods that make use of it for part-level feature learning [5,14,15,16,17]. Most of them usually obtain the local key-point features by simply multiplying the pose heatmap with the resulting feature maps from backbone network. But we found that such usage may not bring in improvement. On one hand, as shown in Figure 1, global feature from backbone (e.g. ResNet50) are highly abstract that each pixel could correspond to a large area on original image and the feature of key-point pixel would easily be contaminated by unnecessary background clutters. On the other hand, the sparsity of heatmap could cause the loss of useful deep-level abstract information, and the occlusion or different posture view point will also bring in negative influences. Therefore, it is important to found a reasonable method to fully make use of the pose cues. Relevant analysis and description will be shown in Section 3.1. **Third**, almost all of the existing part-based methods obtain the final person representation by directly concatenating the extracted features without considering the different contribution of each local feature. Different from some works [37,38] that use GCN to exploit the relations between different parts, we propose to use GCN to generate an independent attention weight value for each local feature, aiming to enhance the contribution of the ones with more discriminative information.
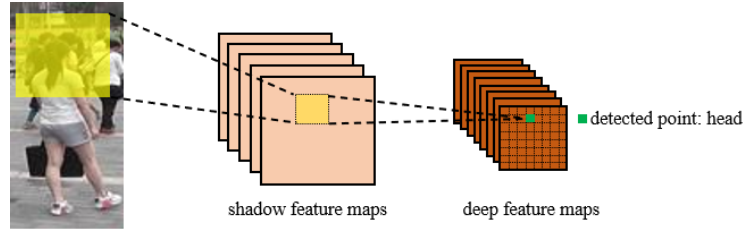
Figure 1. Illustration of reception field on deep feature maps. The detected head point on deep-level feature maps could be approximately corresponding to a relatively large area on input image, and some background noises are easily to be involved (background persons in yellow area).

Base on the analysis above, we present in this work a Pose Guided Graph Attention Network (PGGAN), a multi-branch architecture consisting of one branch for global feature learning, one branch for mid-granular body features learning, one branch for fine-granular key point features learning and a GCN module for feature fusion. We use a pre-trained pose estimator to provide the key-point heatmap for local feature extracting and carefully design a graph convolution layer to learn the attention weight for each extracted local feature. For global branch, we use both average and max pooling (GAP and GMP) to obtain the global vector. For mid-granular branch, we firstly generate the body attention mask matrix based on key-point locations and then add it to low-level shadow feature maps in ResNet50 backbone. Shadow feature maps have less reception field and thus could be more suitable for attention mask to restrain the background noises. For fine-granular branch, we generate the fine-level attention matrices to mask the regions of interest on shadow feature maps in order to make the network focus on semantic key-point areas. For GCN module, we use a learnable adjacency matrix to learn the attention weight values based on the similarities between different local features. For inference procedure, we aggregate all the feature vectors output from GCN as the final representation for person retrieval, respectively, three global and five body-relevant local features. We conduct a plenty of ablation studies of our network to prove its effectiveness and robustness. As will be seen in experiments, our approach shows its competitive advantages for person identity representation compared with other part-based methods. We conduct a plenty of ablation studies and design different kinds of comparison experiments for our network to prove its effectiveness and robustness, including holistic datasets, partial datasets, occluded datasets and cross-domain tests.

The main contributions of this paper can be summarized as follows:

- We design a novel pose-guided attention mechanism based on muti-branch architecture to better exploit the discriminative person representations.

- We introduce a self-adaptive graph attention convolution module to evaluate the contribution weight of different local features.

- Our work outperforms the state-of-the-art on several benchmarks favorably, including holistic datasets test, partial datasets test, occlusion datasets test and cross-domain tests.

2. RELATED WORK

Person reID: With the fast development of deep learning, CNN-based feature learning [33,34,35,20,21,23] is proved to be far more effective than conventional hand-crafted methods [31,32] and thus has been widely used in person reID in recent years. CNN-based methods typically train the reID network by multi-class classification loss and ranking loss [30]. For classification loss [25], each person identitiy is treated as one single class and the prediction logits from fully connect layer are supervised in the training stage. For metric loss, such as contrast loss [26], triplet loss [27,28] and hard-triplet [29] loss, network is trained to maximize the distance between different identities and minimize the distance between same identities at the same time. Like most of current ReID methods, this paper uses both classification and triplet loss to supervise the present network and baseline. Furthermore, based on pose information, this paper proposes an advanced pose-guided hard sample mining method to improve the sampling accuracy in early training phase. Besides, GAN has also been proved to be effective for different kinds of reID tasks [18,19,24], especially for occluded reID and cross-domain reID. Zheng et al [19] start the first attempt to use GCN to improve the model performance by generating extra training images. Wei et al [24] proposed a Person Transfer Generative Adversarial Network (PTGAN) weaken the domain gap between source dataset and target dataset.

Local feature exploration: Local feature learning could enhance the discriminative person representation to a great extent by splitting the images into different parts. One branch of part-based methods is to uniformly divide the images into horizontal grids and extract the corresponding local features respectively [7,13,22,23]. In [10], Sun et al proposed a part-based convolutional network by dividing the input image into several horizontal stripes and using part-level features for inference stage, which has achieved good performance in ReID. In [13], Wang et al proposed a MGN model, which made use of the output of ResNet50 layer3 and trained the network with three branches. Experiment results have shown that MGN could effectively boost the performance of ReID model. This branch of methods usually assumes that the body parts of particular grids on matching images are well-aligned. However, this could be problematic for real-world deployment, where the pre-trained detector tools (e.g. YOLO, SSD) might not always localize the person tightly. Another branch is to use auxiliary pose cues to capture the local features of certain body parts [5,15,16]. These methods used extra skeleton estimation model to provide human pose information. However, the combining manner they used to generate local features is to simply multiply the pose heatmap with feature maps, which has been proved inefficient for both holistic and occluded datasets. Besides, there are also some other researches [11,21,2] focusing on designing attention-based networks which have been proved effective to enhance the key features of different channels or pixels. Motivated by these works mentioned above, we design a novel fusion manner to leverage the pose information by pre-trained estimator and the propose network will be detailed in section 3.
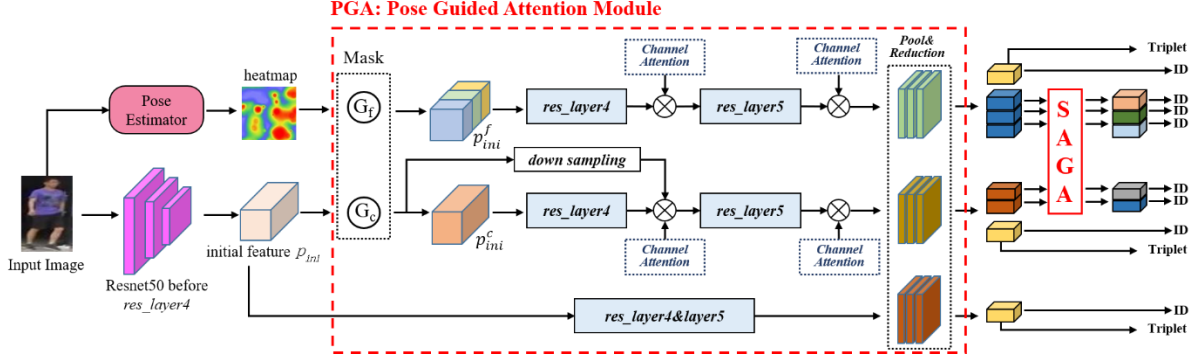
3. METHOD

3.1 Overview

Figure 2. The pipeline of the proposed PGGA network. It mainly consists of two parts: a multi-branch backbone architecture combined with pose-guided attention mask to enable local feature learning of different granular level and a well-designed self-adaptive graph attention module to evaluate the contribution weights for different local features. The final representation contains three global features and five weighted local features respectively from different level. We use triplet loss combined with classification loss for training. All the features are concatenated for inference and testing.

Figure 2 shows the overview of our proposed network architecture. We design our backbone on the basis of *ResNet50* but particularly split the network into two connected parts: a shadow part consisting of the layers before *res_layer4* and a deep part consisting of three branches with each branch sharing the similar architecture with layers after *res_ layer4*. Each input person image is firstly processed by shadow part to obtain the low-level abstract features and the output is regarded as initial feature maps for deep part network. At the same time, we extract the pose information of each input image by using a pose estimator [36], which is pre-trained on MPII dataset with parameters setting fixed. In order to generate the local salient features, we introduce a Pose-Guided Attention (PGA) module based on deep part backbone to jointly extract the local features by three different levels: global feature, coarse-granular and fine-granular features. Subsequently, the resulting features are fed through a carefully designed Self-Adaptive Graph Attention (SAGA) module to generate the final local features of size *1 ×1×C*. Inspired by the excellent ability of Graph Convolutional Network (GCN) on node relation modeling, we use SAGA module to further strengthen the contribution of the ones containing more representative information. We concatenate the global vectors from PGA and all the resulting features from SAGA along the channel dimensions as a final representation for testing procedure. We train the whole network end-to-end by classification loss and triplet loss.

3.2 Pose-Guided Attention Module

Local feature learning by using part-based methods have been proved to be effective for diverse person representations. Meanwhile, pose cues could provide a relative accurate body structure information and thus has been employed by several works [5,14,16] for semantic-level local feature extracting. As mentioned in Section 1, we found from experiments that simply multiplying the pose heatmap with the output features from backbone network may not bring in promotion. The reason of such phenomenon could be summarized three-fold: 1) features from deep layers have large reception field and some background clutters around the key-point on original image could be involved; 2) the

sparsity of heatmap could cause a significant loss of useful deep-level abstract information. 3) occlusion or different posture view point could result in wrong representation of extracted semantic body parts. To alleviate these problems, we design our PGA module by combining the pose information with a multi-branch network architecture, which could effectively improve the network performance on local feature learning.
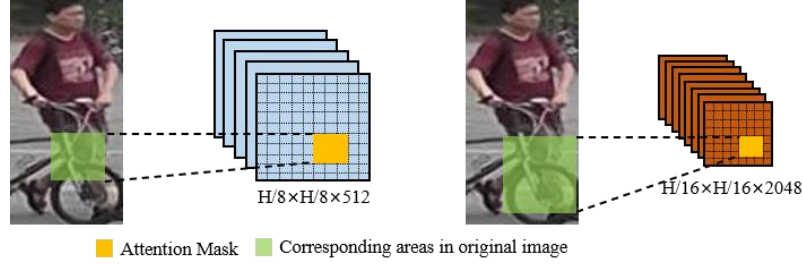

Figure 3. Illustration of the layers to choose for attention mechanism. Compared with deep features, adding attention mask on shadow feature maps of size H/8×W/8×512 could bring in less background clutters (e.g. the bike in the left green area are less involved in attention mask than the right one).

We divide the network into three different branches: one global branch to learn the global feature of person image, one coarse-granular branch to focus on the coarse areas that contains the person body and one fine-granular branch to focus on the fine-grained areas that contains the thirteen defined body parts. We split the Resnet50 model from layer *res_conv4* and use the output feature maps of size *H/8 ×W/8 ×512* as the input of PGA module, where *H* and *W* are the original image height and width. As shown in Figure 3, the reason we choose *H/8 ×W/8* feature maps rather than *H/16 ×W/16* final ones from *res_layer5* for adding attention mechanism is due to the considering that shadow feature maps have a relatively small reception field with more detailed information and thus the features of key-point areas could be separated more easily for better local feature learning. For the given initial feature maps, we then construct our PGA module by leveraging pose information. The pose heatmaps indicate the locations of these detected key-points, which could effectively guide the network pay more attention on human key-point areas that contain more discriminative information than other background areas.

Concretely, for a given input image of size $H \times W \times 3$, we use a pre-trained pose estimator and the shadow layers before *res_conv4* to simultaneously extract the key-point heatmap and initial feature maps, which are respectively denoted by p_{hmap} and p_{ini}. According to [36], we define thirteen semantic parts (Figure 4) to represent the human structure, including head, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, left hip joint, right hip joint, left knee, right knee, left foot and right foot. The dimensions of the mentioned variables in this paper are summarized in Table 1 in section 4.2.

For global branch, we remain the layers of *Resnet50* from *res_conv4* unchanged for global feature learning of person image. We use the combination of Global Max Pooling (GMP) and Global Average

Pooling (GAP) to squeeze the feature maps to a single vector. Then the <mark>final representation</mark> p_{global} is obtained by the following formulation:

$$p_{global} = R_{global}(GAP(f_{global}(p_{ini})) + GMP(f_{global}(p_{ini})))$$

where $f_{global}(*)$ *is* the branch network; R_{global} is a subsequent sub-module to pooling layer, consisting of a *1 × 1* convolution, batch normalization (BN) and ReLU activation function layer.
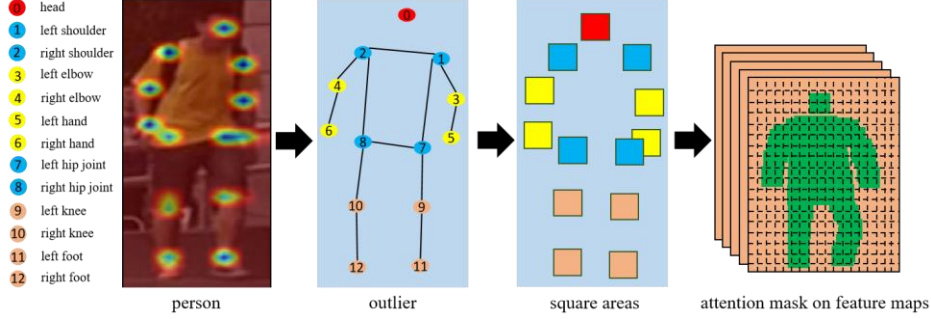

Figure 4. Illustration of the attention mask for coarse branch. The thirteen square areas are firstly generated according to the locations of detected key-points. Then we connect these squares to construct a closed attention area and obtain the mask matrix by filling in the designed coefficient. Such attention mechanism could improve perceptivity of the network on the coarse area of human body as a whole but not just on some limited parts that are vulnerable to posture variance or occlusion, such as shoes, hands and elbows.

For coarse-granular branch, we aim to make the network focus on the coarse areas that contain the human body as a whole. Generally, for a given image, the max value point on each generated heatmap indicates the location of corresponding key-point. That is, by using all the thirteen nodes, one could build an outlier of human body on the image. p_{hmap} is an array containing thirteen key-point probability matrices with value in range of (0,1) and we notice from experiments that for each matrix, only the predicted point has a high value (e.g. 0.8~1.0) while the other pixels have relatively small values (e.g. 0.1~0.3). That is, the heatmap is a <mark>sparse matrix</mark> where most of the elements are close to zero except for the detected key-point ones. Thus, directly multiplying the heatmap with p_{ini} would inevitably results in sparse characteristic of features and severe degradation of detailed information, especially the areas around predicted point. To alleviate this problem, we tend to use the location of each max point rather than the probability value itself to construct a coarse-granular attention mask $Mask_c$ that covers the whole person body for initial feature p_{ini}, formulated as follow:

$$A_i = C_{area}\{(l_i^{row} - \omega, l_i^{col} - \omega), (l_i^{row} + \omega, l_i^{col} + \omega)\}$$

$$Mask_{row,col}^c = \begin{cases} \alpha(1 - \beta), & (row, col) \in A_{outlier} \\ \beta, & else \end{cases}$$

where $C_{area}\{(l_1^{row}, l_1^{col}), (l_2^{row}, l_2^{col})\}$ refers to the square area determined by coordinates of upper-left point l_1 and lower-right point l_2; (row, col) is the element subscript of matrix. We firstly extract a square area A centered on each single peak point by an expansion coefficient ω. Then according to body structure, we connect and aggregate the area A_i of different body parts to generate six closed outliers, including head, upper torso and four limbs. Then we could obtain $Mask_c$ by filling these outliers with the designed weight parameters α and β, which is used to indicate the contrast between attention areas and non-attention ones. Detailed description is shown in Figure 4.

Noticing that for PGA module, each branch consists of a *res_layer4* and a *res_layer5* module, and the product of $Mask_c$ and p_{ini} is used as the input of this coarse branch. In order to fully leverage the pose cues, we also deploy an attention mask on the mid-result output from *res_layer4*. Since the features from *res_layer4* are more concentrated and abstract, we apply bilinear down sampling on $Mask_c$ to slightly lower the contrast degree and make its size accord with the feature size. In addition, considering that pose mask could only provide an attention area range on spatial dimension, we introduce a channel-wise attention sub-module to further enforce the salient features of different channels. Formally, the features of this coarse branch of size *H/16×W/16×2048* is constructed as follows:

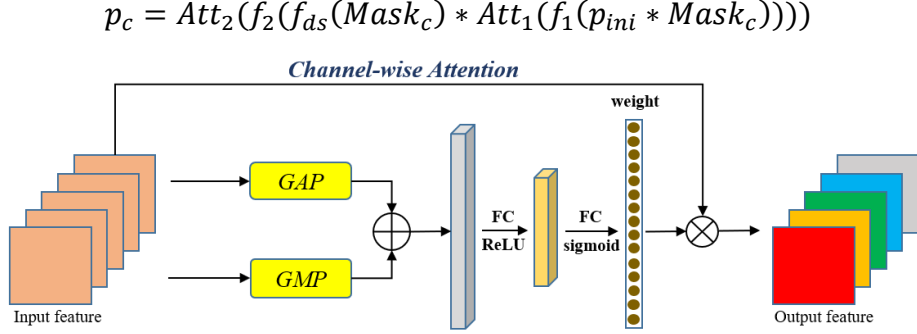$$p_c = Att_2(f_2(f_{ds}(Mask_c) * Att_1(f_1(p_{ini} * Mask_c))))$$

Figure 5. Illustration of channel-wise attention. Since we use attention mask to increase the feature value, using GAP combining with GMP could remain more discriminative features. The output feature is obtained by multiplying the weight value with input feature maps along channel dimension.

where f_{ds} is the down sampling operation; * is element-wise product; f_1 and f_2 are *res_layer3* and *res_layer4* network, respectively. Att is the channel-wise attention sub-module which is deployed twice respectively on the resulting feature of *res_layer3* and *res_layer4*. We show the structure in Figure 5, different from [40], we jointly use GAP and GMP in Att module to better leverage the pose cues. In order for part-level feature learning, we split p_c into two unsymmetrical horizontal grids according to the general ratio of upper body to lower body, and extract one global with two local vector representations of size $1×1×256$ by following formulations:

$$p_g^c = R_g^c(GAP(p_c) + GMP(p_c))$$

$$p_{l_i}^c = R_{l_i}^c\left(GAP\left(p_{grid_i}^c\right)\right), i = 1,2$$

where R_g^c, $R_{l_i}^c$ are reduction module sharing the same architecture with R_{global}; $p_{grid_i}^c$ is the partitioned feature maps from p_c.
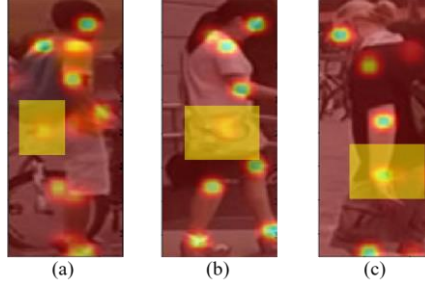
Figure 6. Illustration of occlusion and body posture variance. In picture (a), the left hand is occluded due to side view of person image. In picture (b), the hip joints and left hand are occluded by a backpack. In picture (c), the hip joints, right arm and right hand are all occluded by left side body. Even thought these key points (yellow rectangle areas) are invisible in image, the locations of them could still be predicted by pose estimator.

For fine-granular branch, we aim to make the network focus on the fine local areas of thirteen defined semantic body parts. An intuitive way is directly extracting the vector at the detected points or multiplying the heatmap with feature p_{ini} and obtain the vectors of corresponding body parts by max-pooling, which is also the common usage in most of pose-based ReID methods. However, preliminary experiments show that such usage might sometimes misguide the network especially when it comes to occlusion or body posture variance. For example, as shown in Figure 6 (a), the left hand is occluded by body torso, but the pose estimator could still predict its location. In this circumstance, the features of predicted hand area on p_{ini} actually represent the torso but not the hand. Considering that the occlusion of certain body parts is inevitable, we believe that it is unreasonable to separately extract feature vector for each single semantic part. On the contrary, we tend to guide the network learn by itself which part to focus and obtain the local features by horizontal splitting in this branch.
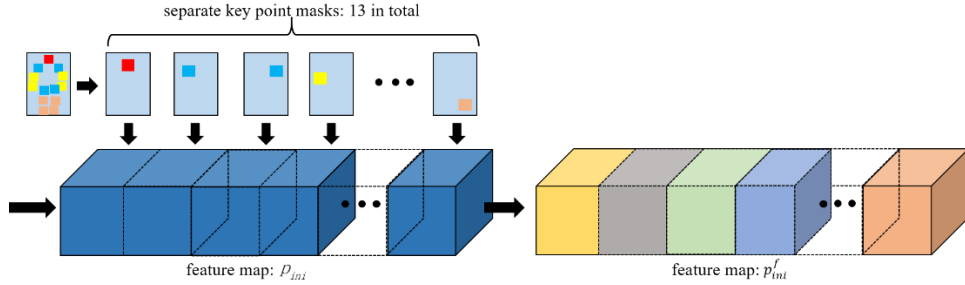

Figure 7. Illustration of the mask attention for fine granular branch. We extract thirteen different attention masks according to the locations of detected key-points, and then multiply them with the corresponding thirteen groups in channel dimension to generate the input feature of this branch.

Different from coarse-granular branch, we do not aggregate A_i but directly generate a separate attention mask matrix for each predicted key-point in a similar way as equation(xx):

$$Mask_{row,col}^{fn} = \begin{cases} conf_n * \alpha(1-\beta), & (row,col) \in A_i \\ \beta, & else \end{cases}, n = 1,2,3 \dots 13$$

where $conf_n$ is the confidence score of key-point, which could indicate the feature quality to some extent. Then we divide the initial feature p_{ini} into thirteen groups in channel dimension with each group multiplying the corresponding defined mask matrix. Figure 7 shows the detail and the formulation are as follows:

$$p_{ini}^f = p_{ini} * T\left(Mask_{group_1}, Mask_{group_2}, \dots, Mask_{group_n}\right), n = 13$$

$$Mask_{group_i} = \mathrm{T}(Repeat(Mask_{key_i}, n_i))$$

where p_{ini}^f is used as the input of *res_layer3* in this branch. We denote by T the concatenate operation in channel dimension. $Repeat(*, n_i)$ is the duplication of * by n_i times and n_i refers to the assigned channel numbers for each key-point mask, which is 512 exactly in total. In this way, feature maps of different channel groups are enhanced by different areas of key-points. Then we use channel-wise attention to automatically learn the weights of different channels. Similarly, we split the resulting feature maps p_f into three unsymmetrical horizontal grids, and extract one global with three local vector representations of size $1 \times 1 \times 256$ by following formulations:

$$p_f = Att_2(f_2(Att_1(f_1(p_{ini}^f))))$$

$$p_g^f = R_g^f(\mathrm{GAP}(\mathrm{p_f}) + \mathrm{GMP}(\mathrm{p_f}))$$

$$p_{l_i}^f = R_{l_i}^f\left(GAP\left(p_{grid_i}^f\right)\right), i = 1,2,3$$

In summary, by introducing pose attention, our proposed PGA module is able to fully develop the person-related features and effectively suppress those non-salient ones from background noise, which consequently leads to a more discriminative and robust person representation.

3.3 Self-Adaptive Graph Attention Module

In PGA module, by taking advantage of pose information, we obtain eight feature vectors: three global ones, two coarse-granular local ones and three fine-granular local ones. A common practice for inference is to directly concatenate all these vectors in channel dimension and matching the candidate image pairs by Euclidean or cosine distance, which is also used in this work for fair comparison with other works. In other words, each of these extracted vectors is considered to have the same contribution on similarity distance. However, since different vector contains different information accounting for person image, it is reasonable to re-evaluate the significance of each vector.

Besides, re-weighted features are robust against misalignment problem to some extent. For some datasets such as Market-1501 and DukeMTMC-reid, there are only a few samples are horizontally misaligned with others since the bounding box of the person image is annotated manually. But for real-world applications or dataset like MSMT17, bounding boxes are usually generated by object detector automatically, which may not match the person tightly, and therefore brings in unnecessary extra background noises. In such circumstance, misalignment would harm the local feature representation of matching images and thus lower the retrieval accuracy, especially for part-based methods. By re-weighting the part-level local features, the weight of misalignment part with less person relevant information would be lowered and the contribution of those parts with discriminative features would be enhanced. Thus, even though the extracted local features are not tightly aligned, the matching results would still be accurate. Detailed discussion will be shown in section 4.3.

In recent years, Graph Convolutional Network (GCN) have been applied by several works [14,37,38] for ReID tasks. These works use GCN to model the spatial of temporal relations between different

node features by building the adjacency matrix. However, for multi-branch networks like our PGA module, the features from different branches are not in the same feature space because of the unshared convolution kernel weight of each branch. That is, directly using GCN to generate the fused features may cause sub-optimal results, which is also proved by our preliminary experiments. Therefore, different from existing methods, we propose to apply GCN to generate the significance weight coefficient for each node features. Figure 8 shows the structure of our SAGA module.
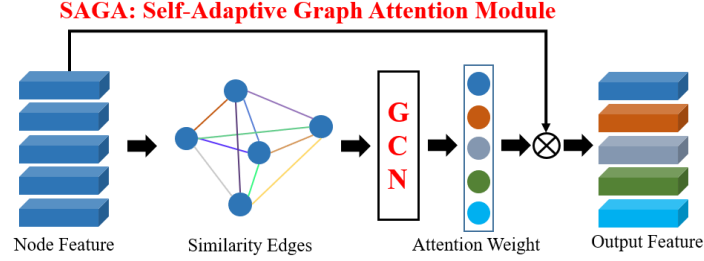

Figure 8. Structure of proposed SAGA module. Different from other GCN-based methods, we use two learnable linear matrices to obtain the similarity edges. Due to the strong ability of GCN on relation modeling, the network could automatically learn the contribution weights of different input node features.

Specifically, we set the weight of all the three global vectors to 1 and use the other five local ones as input node features, denoted by $v_i \in R^{1 \times 1 \times 256}, i = 1,2,3,4,5$. Motivated by [43], we use similarity distance to generate the edges by:

$$e(v_i, v_j) = \varphi_{v_i}^T \cdot \varphi_{v_j}$$

where φ represent the linear transformation of node pairs with its parameter being learnable via back propagation. Then we construct the adjacency matrix by employing L2-normalization on each row of the affinity matrix and obtain the establish the GCN module by:

$$A_{i,j} = \frac{e(v_i, v_j)}{\|e\|_2}$$

$$\theta = AVW$$

where $W \in R^{d \times 1}$, $V \in R^{d \times 5}V$ and d is the dimension of each node feature; θ is the resulting weight coefficient representing the contribution of each local feature. Then by multiplying it with input vectors, we finally obtain the weighted local features as:

$$Feature_{local_i} = \theta_i * v_i$$

3.4 Loss Function

For both PGGA Net and PA baseline, we use Softmax Cross-Entropy loss (ID loss) for classification and hard-triplet loss for metric learning as the supervising function in training stage. Different from triplet loss, hard-triplet loss aims at maximizing the distance between hardest negative pair while minimizing the distance between hardest positive pair by a margin at the same time. The ID loss and hard-triplet loss in a mini-batch can be described as:

$$L_{id} = -\sum_{i=1}^{P \times K} log \frac{e^{w_{y_i}^T x_i}}{\sum_{c=1}^{N_{id}} e^{w_c^T x_i}}$$

$$L_{tri} = \sum_{i=1}^{P} \sum_{j=1}^{K} \left[max_{p=1,2...K} \|A_i - P_i\|_2 - min_{\substack{n=1,2...K \\ j=1,2...K \\ j \neq i}} \|A_i - N_i\|_2 + \gamma \right]_+$$

In equation (1), N_{id} denotes the number of identities in the whole training set; W denotes the weight parameters of fully connected (FC) layer; y_i is the ground truth identity of input image and x_i is the output of FC layer for ith identity. In equation (2), A_i, P_i and N_i are the feature vectors of anchor, positive and negative samples respectively; γ is the margin to the distance between negative and positive pair. For better performance, we set the bias terms in linear multi-class classifiers to zero. We employ the ID loss on global features and all the part-level local feature vectors. Considering that part-level local features are not suitable for selecting out the hardest positive and negative pairs, triplet loss is only deployed on global features. The total loss for training stage is constructed by a weight parameter τ as follow:

$$L = L_{tri} + \tau * L_{id}$$

4. EXPERIMENT

4.1 Dataset

We conducted experiments on four holistic datasets and three occluded datasets: 1) The **Market-1501** [48] dataset contains 1,501 person identities and 32,217 images from 6 different cameras. The whole dataset is split into training set and testing set. Training set consists of 12,936 images of 751 identities. Testing set consists of 3,368 query images and 19,732 gallery images of 750 identities. 2) The **DukeMTMC-reID** [49] dataset contains 1,812 person identities and 36,411 images from 8 high resolution cameras without any cross-view. Training set consists of 16,522 images of 702 person identities which are randomly selected from dataset. Testing set consists of 2,228 query images and 17,661 gallery images of the other 702 persons in total. 3) The **CUHK03** [50] dataset contains 1,467 person identities and 14,097 images from 6 different cameras. Two kinds of annotations are provided in the dataset: manually labeled bounding boxes and DPM-detected bounding boxes. This dataset provides two split protocols: classic split protocol and new protocol. Considering that the classic protocol is time-consuming and not suitable for deep learning ReID methods, we apply the new one to evaluate the performance of our proposed baseline on both labeled and detected datasets. 4) The **MSMT17** [51] dataset contains 4,101 person identities and 126,441 images from 15 different cameras. Training set consists of 32,621 persons of 1,041 identities. Testing set consists of 93,820 persons of 3,060 identities. The bounding boxes are labeled by faster RCNN automatically. For testing, 11,659 boxes are randomly selected as query while the other 82,161 ones regarded as gallery. 5) The **Occluded-Duke** [18] dataset is selected from DukeMTMC-reID, containing 15,618 training images, 17,661 gallery images and 2,210 occluded query images. 6) The **Occluded-ReID** [55] dataset contains 2,000 images of 200 occluded persons, captured by mobile camera. Each person has five full-body images and five occluded images. 7) The **Partial-REID** [9] dataset consist of 600 images from 60 persons. Each person has five full-body images and five partial images.

4.2 Implementation Details

Table 1. Dimensions of network variables.

Description	Variable	Dimension (Height×Width×Channel)
Input image	P_{input}	H×W×3
Detected heatmap	P_{hmap}	H/8×W/8×13
Feature maps resulting from *res_layer2*	P_{ini}	H/8×W/8×512
Input feature maps for coarse branch	p_{ini}^c	H/8×W/8×512
Input feature maps for fine branch	p_{ini}^f	H/8×W/8×512
Resulting feature vector from global branch	P_{global}	1×1×256
Resulting feature vectors from coarse branch	$p_g^c, p_{l_i}^c$	1×1×256
Resulting feature vectors from fine branch	$p_g^f, p_{l_i}^f$	1×1×256

We summarize the dimensions of different variables in Table 1. The implementation settings are listed below. Resnet50 backbone network is initialized by the weights pre-trained on ImageNet. The FC and pooling layers of standard Resnet50 network are removed and the last stride of *res_layer5* is set to 1. We use HRNet [41] as pose estimator with fixed pre-trained weights on MPII dataset. Each input image is resized to 384×128. We deploy random horizontal flipping and random erasing as data augmentation methods in training stage. The initial learning rate is set to 0.01 and 0.001 for the backbone and other parts. The margin parameter for hard triplet loss function is set to 1.2. SGD optimizer is used with momentum set to 0.9 and weight decay factor set to 0.0005. The expansion coefficient ω is set to 2. The mask weight parameter α and β are set to 2 and 0.5 respectively. The loss weight parameter τ is set to 2. In order to fully demonstrate the effectiveness of our network, we do not apply some training tricks that could lead to slight advance on testing results, such as warmup strategy, label smooth, circle loss etc. In inference stage, all the vectors are concatenated to represent the person feature for retrieval. Our work is implemented on four NVIDIA RTX 1080ti GPUs based on PyTorch framework (version 1.6.0). All the experiments on different datasets follow the same settings mentioned above.

4.3 Comparison with State-of-the-Art Methods

Results on holistic datasets.

We show in Table 2 the quantitative comparisons (without reranking) of our network with the state-of-the-art works. We mainly choose part-based, attention-based and GCN-based methods that are proposed in recent three years to make the comparisons. We use mean average precision (mAP)(%) and rank-1 accuracy (%) to evaluate the network. As shown, our work gives the best results on three datasets in terms of mAP and rank-1. We achieve mAP of 90.5% and rank-1 accuracy of 96.5% on Market-1501, mAP of 80.5% and rank-1 accuracy of 90.4% on DukeMTMC-reID, mAP of 61.9% and rank-1 accuracy of 84.8% on MSMT17. Particularly, compared with other pose-based works, such as [14, 16], our method is the only one that achieves the SOTA results on holistic datasets.

Table 2. Performance caparisons with SOTA on holistic datasets.

Method	Market-1501		CUHK03-Labeled		CUHK03-Detected		DukeMTMC		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
MGN[13]	86.9	95.7	67.4	68.0	66.0	66.8	78.4	88.7	-	-
PCB+RPP[7]	81.6	93.8	57.5	63.7	-	-	69.2	83.3	40.4	68.2
RelationNet[39]	88.9	95.2	75.6	77.9	69.6	74.4	78.6	89.7	-	-
RGA-SC[46]	88.4	96.1	77.4	81.1	74.5	79.6	-	-	57.5	80.3
SCAL[47]	89.3	95.8	72.3	74.8	68.6	71.1	79.1	88.9	-	-
SONA[48]	88.83	95.58	79.23	81.40	77.27	79.9	78.18	89.55	-	-
HOReID[14]	84.9	94.2	-	-	-	-	75.6	86.9	-	-
MGCAM[49]	74.33	83.79	50.21	50.14	46.87	46.71	-	-	-	-
Pyramid[50]	88.2	95.7	76.9	78.9	74.8	78.9	79.0	89.0	-	-
DenseS[51]	87.6	95.7	75.2	78.9	73.1	78.2	74.3	86.2	-	-
ABD[52]	88.28	95.6	-	-	-	-	78.6	89.0	60.8	82.3
SGCS[53]	88.5	95.7	**83.3**	**86.3**	**80.2**	**84.1**	79.0	90.1	-	-
ours-PGGA	**90.5**	**96.5**	74.0	75.9	70.5	72.3	**80.5**	**90.4**	**61.9**	**84.8**

Results on occluded datasets.

We show in Table 3 the experiment results on occluded datasets comparing with the state-of-the-art works. For Occluded-Duke, the model is pre-trained on DukeMTMC-reID. For Occluded-ReID and Partial-ReID, the model is pre-trained on Market-1501. We take all the occluded person images as query set and the full body images as gallery set on all the three datasets. It should be noticed that some of the works listed are specially designed for occluded ReID task [16,54,5,14] while our network is designed for general purpose. Comparison results show that our method significantly outperforms the existing SOTA works on Occluded-Duke with a performance of 58.8%(mAP)/70.9%(Rank-1). Note that images in Occluded-Duke are much more difficult to retrieve due to the large pose variations, occlusions and confusing noises, which further demonstrates the advantage of our work. For the other two datasets, we achieve a close performance to SOTA on Occluded-ReID and the best mAP on Partial-ReId.

Table 3. Performance caparisons with SOTA on occluded datasets.

Method	Occluded-Duke		Occluded-ReID		Partial-ReID	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PCB [7]	33.7	42.6	38.9	41.3	64.8	66.3
PGFA [16]	37.3	55.1	-	-	56.2	68.0
Teacher-S [54]	22.4	18.8	59.8	55.0	73.1	69.2
PVPM [5]	29.2	51.5	61.2	70.4	72.3	78.3
HOReID [14]	43.8	55.1	**70.2**	**80.3**	-	**85.3**
DSR [45]	30.4	40.8	62.8	72.8	-	50.7
Ours-PGGA	**58.8**	**70.9**	69.7	74.1	**74.3**	76.3

4.4 Discussion

Ablation Study.

We conduct several ablation experiments on the different components of our model in Table 4. We respectively denote by G, C, F, POSE, ATT, SAGA the global branch, coarse-granular branch, fine-granular branch, pose mechanism, channel wise attention sub-module and proposed GCN module. Our baseline is a Resnet50 model followed by a reduction layer to squeeze the feature dimension to 256. The 2nd and 3rd rows are designed to separately evaluate the performance of each single branch. The 4th to 6th row are designed to evaluate the effectiveness of the pose attention and channel attention. Specifically, for the 4th row, the absence of POSE means that the network of C-branch and F-branch are constructed without pose mask but only consists of the original *res_layer4* and *res_layer5*. The final row is designed to evaluate the proposed SAGA module. All the experiments follow the same settings mentioned in section 4.2. From the results, we can observe that: 1) either C-branch or F-branch significantly outperforms the baseline. For example, F-branch brings in an improvement of 5.5%(mAP)/2.2%(Rank-1) on Market-1501 and 7.1%(mAP)/3.5%(Rank-1) on DukeMTMC-reID dataset. 2) The results in 5th row demonstrate the effect of our proposed pose attention mechanism. With the help of key-point locations, the network could pay more attention on person-relevant areas. On DukeMTMC-reID, this gives a performance gain of 1.7%(mAP)/0.9%(Rank-1). On Market-1501, even thought the performance is already very high, it still brings in an improvement of 1.9%(mAP)/1.2%(Rank-1). 3) Channel attention is proved to be an effective complement module for our pose mechanism. For example, given a feature of size $48 \times 16 \times 512$, we apply the pose masks on all the 512 maps. But we notice that some of these maps contain salient feature patterns in mask areas while some others not. Thus, our designed ATT module could help to enhance the maps with more useful information and suppress the ones with background clutters. 4) The results in final row indicate the effectiveness of our proposed SAGA module, which respectively brings in a gain of 0.5%(mAP)/0.4%(Rank-1) on Market-1501 and 0.4%(mAP)/0.5%(Rank-1) on DukeMTMC-reID. Detailed discussion with visual proof is shown in latter paragraph.

Table 4. Quantitative comparison of different network components.

Component	POSE	ATT	SAGA	F-dim	Market-1501		DukeMTMC	
					mAP	Rank-1	mAP	Rank-1
Baseline: single-G	×	×	×	256	83.4	92.7	72.1	85.8
G + C	√	√	×	1024	88.5	94.8	78.6	89.4
G + F	√	√	×	1280	88.9	94.9	79.2	89.3
G + C + F	×	×	×	2048	87.7	94.5	78.1	88.8
G + C + F	√	×	×	2048	89.6	95.7	79.8	89.7
G + C + F	√	√	×	2048	90.0	96.1	80.1	89.9
PGGA: G + C + F	√	√	√	2048	90.5	96.5	80.5	90.4

Pose Attention Mechanism.

Figure 8. Class activation map extracted from *res_layer5_conv3* of each branch by Grad-CAM [55]. Images are sampled from Market-1501 dataset. 1st column: activation heatmaps from global branch; 2nd to 3rd column: activation heatmaps of the two horizontal grids from coarse-granular branch; 4th to 6th column: activation heatmaps of the three horizontal grids from fine-granular branch. Dark color denotes higher value.

In order to evaluate the effectiveness of our proposed pose attention mechanism, we visualize the class activation maps from *res_layer5_conv3* of each branch by using Grad-CAM [55]. Figure 8 shows some typical examples selected from Market-1501 dataset. For global branch, we can notice that some salient features are easily ignored and sometimes the irrelevant background clutter could mislead the network. In the global map of the 2nd row, only the shoes are highly preferred and the information from shirt, pants or legs is commonly ignored. In the global case of the 3rd row, the area next to the head is actually background noise that contain no useful cues about identity for pedestrian but it is still be highly concentrated on. For local branches, we can see from the figure that by taking advantages of pose information, our proposed PGA module could fully exploit the local semantic features and filter most of the confusing background noises at the same time. For example: 1) in the cases of the 1st row, C-branch mainly focus on upper and lower torso areas, such as the T-shirt and pant, while the responses of F-branch concentrates more on body parts such as the elbow, knees and shoes. 2) on global response map of the 3rd row, by applying pose attention masks, the misleading area at the right of person's head is effectively filtered in C-branch and F-branch. Therefore, the visualization proofs indicate that our proposed PGA module allows the network to fully mine the diverse and discriminative local features on different branches.
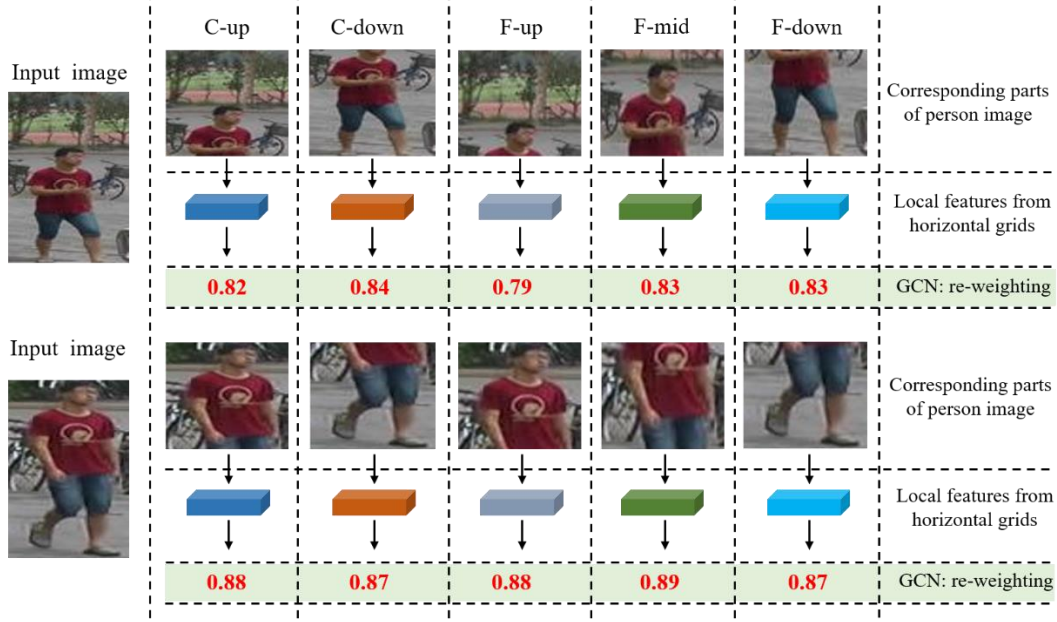
Self-Adaptive Graph Attention Module.

Figure 9. Samples of weight coefficient in proposed SAGA layer. The two images are from a same person identity of Market1501 dataset. Weight coefficients shown in red color are extracted and printed out from testing procedure. Each smaller part picture is cut from the input image according to each split horizontal grid with reception field taken into consideration.

We show in Figure 9 some weight coefficient results extracted from testing procedure. We select two typical person images of a same identity as inputs to illustrate the rationality of our proposed SAGA layer. That is, the first image contains more background interference and the person body in two images are not aligned tightly. Based on a large amount of observation experiments, we found that: 1) features containing less discriminative information tend to be assigned a lower weight. For example, in the first image, obviously there are lots of background noises (e.g. bikes and trees) in the pictures of C-up and F-up parts due to the imprecise annotation bounding box and resulting weights of them are subsequently lower (0.82/0.79), while the other parts that contain useful features (e.g. holistic shirt and pants) are assigned higher weights (0.84/0.83/0.83). 2) features from high-quality image samples are usually learned to have higher contribution weights. In particular, high-quality sample refers to the person image in which the bounding box is perfectly annotated, less background clutters exists and no occlusion occurs. For example, the average feature weights of the second image is a little larger than the first one. 3) the values of weights are usually trained to be in an approximate range of (0.75,0.95). We infer that such phenomenon is to some extent related to the number of horizontal grids. We split the feature maps into two/three stripes in our method and each grid feature still contains useful information more or less, even though the image sample are pictured with a bad misalignment problem.

Cross Domain Test.

Table 5. Quantitative comparison of different network components.

Method	Market-1501→DukeMTMC-reID				DukeMTMC-reID→Market-1501			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
BagTricks [29]	25.8	45.9	59.8	66.2	26.7	56.1	74.4	79.3
RelationNet [39]	33.5	52.7	68.5	72.9	40.8	68.5	81.6	86.2
PCB [7]	26.5	46.1	60.9	66.6	27.1	56.9	75.3	81.9
MGN [13]	30.5	48.8	66.0	70.1	38.4	65.3	78.9	84.2
EANet [60]	36.0	56.1	-	-	35.8	66.1	-	-
SPGAN+LMP [56]	26.2	46.4	62.3	68	26.7	57.7	75.8	82.4
HHL [57]	27.2	46.9	61.0	66.7	31.4	62.2	78.8	84.0
UDAP [58]	49.0	68.4	80.1	83.5	53.7	75.8	89.5	93.2
AD-Cluster [59]	54.1	72.6	82.5	85.5	68.3	86.7	94.4	96.5
PGGA w/o pose	32.2	50.2	66.9	71.2	39.9	66.6	80.3	85.4
PGGA with pose	38.2	58.3	71.1	76.2	43.3	72.3	84.5	89.1

Cross-domain person retrieval has always been a challenging task in which the model trained on a certain dataset would suffer a significant drop on another dataset. Such phenomenon is due to the data-bias existing between source and target datasets. For example, images from Market1501 and DukeMTMC datasets are captured under different camera environment factors including light, camera resolution, proportion of background noise, etc. Thus we conduct experiments to evaluate the network performance on cross-domain ReID task, the results are shown in Table 5. "Market1501 → DukeMTMC" in Table 5 means that the network is trained on Market1501 and directly test on DukeMTMC dataset without other training. According to paper [29], we remove Random Erase Augmentation (REA) since it may lead to a severe degradation on cross-domain performance. The methods are categorized into three groups: 1) three typical part-based methods are re-implemented, including BagTricks, PCB and MGN. 2) EANet, SPGAN+LMP and HHL are unsupervised domain adaptive (UDA-based) methods that employed unsupervised learning. 3) UDAP and AD-Cluster are UDA methods that further employed clustering. All of the methods listed are proposed in recent years.

From the experiments, we can observe that: 1) the effective usage of pose cues makes our PGGA model largely outperform the other part-based works. 2) compared with the second group, even though our work is not designed for cross-domain task in particular, our PGGA model could still achieve a higher performance over most of UDA-based works. What's more, it should be noticed that the methods in second or third group apply unsupervised training on target dataset, while we directly transfer the model for test without any extra training procedures. 3) the performance of UDAP and AD-Cluster is much higher than ours due to the employment of clustering methods, which use the clustered pseudo-labels to train the model on target dataset. 4) we separately evaluate the effect of our pose attention mechanism in the fourth group. We found from experiments that the commonality and high accuracy of pose detector could help to alleviate the gap-bias between source and target domain. That is, by such priori knowledge, the key parts of person in a target domain image could be enhanced and thus lead to a more discriminative feature representation.

5. CONCLUSIONS

In this paper, we have presented a pose guided graph attention network, which is a multi-branch architecture consisting of three separate branches designed on purpose. We carefully design a PGA module and a SAGA module to fully exploit the pose information of human body. Experiments results demonstrate the effectiveness of our approach on discriminative feature learning and we show that our model achieves state-of-the-art performances on several mainstream evaluation datasets. We also conduct a plenty of ablation studies and design different kinds of comparison experiments for our network to prove its commonality and robustness, includind partial datasets, occluded datasets and cross-domain testing. Furthermore, our unique way of utilizing pose information could be easily extended to some existing state-of-the-art methods and can promote ReID study for both industry and academia.

REFERENCES

[1] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. 2018. Multi-Level Factorisation Net for Person Re-Identification. In CVPR. 2109–2118.

[2] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious Attention Network for Person Re-Identification. In CVPR. 2285–2294.

[3] Dong W , Zhang Z , Song C , et al. Instance Guided Proposal Network for Person Search[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[4] Hou R , Ma B , Chang H , et al. Interaction-And-Aggregation Network for Person Re-Identification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[5] Pose-guided Visible Part Matching for Occluded Person ReID[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[6] Zheng Z , Yang X , Yu Z , et al. Joint Discriminative and Generative Learning for Person Re-Identification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[7] Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018b. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In ECCV.

[8] Sun, C.; Shrivastava, A.; Vondrick, C.; Murphy, K.; Sukthankar, R.; and Schmid, C. 2018a. Actor-centric relation network. In ECCV.

[9] Zheng W S , Xiang L , Tao X , et al. Partial Person Re-Identification[C]// IEEE International Conference on Computer Vision. IEEE, 2016.

[10] Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; and Tang, X. 2017a. Spindle Net: Person re-identification with human body region guided feature decomposition and fusion. In CVPR.

[11] Zhao, L.; Li, X.; Zhuang, Y.; and Wang, J. 2017b. Deeply-learned part-aligned representations for person reidentification. In ICCV.

[12] Ye M，Shen J，Lin G，et al. Deep Learning for Person Re-identification: A Survey and Outlook[J]. 2020.

[13] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou，"Learning discriminative features with multiple granularities for person re-identification，" in 2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018, pp. 274–282.

[14] Wang G , Yang S , Liu H , et al. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification[J]. 2020.

[15] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person reidentification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4099–4108, 2018.

[16] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In ICCV, 2019.

[17] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing,Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In 2017 IEEE International Conference on computer Vision (ICCV), pages 3980–3989, 2017.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[19] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: The IEEE International Conference on Computer Vision, 2017, pp. 3774–3782.

[20] Qin Zhou, Heng Fan, Shibao Zheng, Hang Su, Xinzhe Li, Shuang Wu, and Haibin Ling. Graph correspondence transfer for person re-identification. In AAAI, 2018

[21] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In CVPR, 2018.

[22] Fu Y , Wei Y , Zhou Y , et al. Horizontal Pyramid Matching for Person Re-identification[J]. 2018. In AAAI.

[23] Xuan Z , Hao L , Xing F , et al. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification[J]. 2017.

[24] Wei L , Zhang S , Wen G , et al. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[25] Hadsell R , Chopra S , Lecun Y . Dimensionality Reduction by Learning an Invariant Mapping[C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006.

[26] F Schroff, Kalenichenko D , Philbin J . FaceNet: A Unified Embedding for Face Recognition and Clustering[J]. IEEE, 2015.

[27] Liu H , Feng J , Qi M , et al. End-to-End Comparative Attention Networks for Person Re-Identification[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2017, 26(99):3492-3506.

[28] Luo H , Jiang W , Zhang X , et al. AlignedReID++: Dynamically Matching Local Information for Person Re-Identification[J]. Pattern Recognition, 2019, 94.

[29] Luo H , Jiang W , Gu Y , et al. A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification[J]. IEEE Transactions on Multimedia, 2019.

[30] Zheng, Z.; Zheng, L.; and Yang, Y. 2018. A discriminatively learned CNN embedding for person re-identification. In ACM TOMM

[31] Ma B , Yu S , Jurie F . Covariance Descriptor based on Bio-inspired Features for Person Re-identification and Face Verification[J]. Image & Vision Computing, 2014, 32(6-7):379-390.

[32] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2197–2206, 2015.

[33] Lei Q , Jing H , Lei W , et al. MaskReID: A Mask Based Deep Ranking Neural Network for Person Re-identification[J]. 2018.

[34] Wang G , Zhang T , Cheng J , et al. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment[J]. 2019.

[35] Hermans A , Beyer L , Leibe B . In Defense of the Triplet Loss for Person Re-Identification[J]. 2017.

[36] Sun K , Xiao B , Liu D , et al. Deep High-Resolution Representation Learning for Human Pose Estimation[J]. arXiv e-prints, 2019.

[37] Shen Y , Li H , Yi S , et al. Person Re-identification with Deep Similarity-Guided Graph Neural Network[J]. arXiv, 2018.

[38] Ye M , Ma A J , Liang Z , et al. Dynamic Label Graph Matching for Unsupervised Video Re-Identification[J]. IEEE Computer Society, 2017.

[39] H Park, Ham B . Relation Network for Person Re-identification[J]. 2019.

[40] Jie H , Li S , Gang S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[41] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person reidentification. In CVPR, 2014. 2, 6

[42] Jia Deng,Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In CVPR. 248–255.

[43] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In Proceedings of the European Conference on Computer Vision (ECCV), pages 399–417, 2018. 2, 3

[44] Zhuo J , Chen Z , Lai J , et al. Occluded Person Re-identification[J]. IEEE, 2018.

[45] He L , Liang J , Li H , et al. Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-Free Approach[J]. IEEE, 2018.

[46] Zhang Z , Lan C , Zeng W , et al. Relation-Aware Global Attention for Person Re-Identification[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[47] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in ICCV, 2019, pp. 9637–9646.

[48] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second order non-local attention networks for person re-identification," in ICCV, 2019, pp. 3760–3769.

[49] Song C , Yan H , Ouyang W , et al. Mask-Guided Contrastive Attention Model for Person Re-identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[50] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in CVPR, 2019, pp. 8514–8522.

[51] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in CVPR, 2019, pp. 667–676.

[52] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in ICCV, 2019, pp. 8351–8361.

[53] Chen X , Fu C , Zhao Y , et al. Salience-Guided Cascaded Suppression Network for Person Re-Identification[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[54] Zhuo J , Lai J , Chen P . A Novel Teacher-Student Learning Framework For Occluded Person Re-Identification[J]. 2019.

[55] Selvaraju R R , Cogswell M , Das A , et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization[J]. International Journal of Computer Vision, 2020, 128(2):336-359.

[56] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In IEEE CVPR, 2018.

[57] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In ECCV, pages 176–192, 2018.

[58] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. CoRR, abs/1807.11334, 2018.

[59] Zhai Y , Lu S , Ye Q , et al. AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification[J]. IEEE, 2020.

[60] Huang H , Yang W , Chen X , et al. EANet: Enhancing Alignment for Cross-Domain Person Re-identification[J]. 2018.