Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Collaborative Attention Network for Person Re-identification

To cite this article: Wenpeng Li *et al* 2021 *J. Phys.: Conf. Ser.* **1848** 012074

View the article online for updates and enhancements.

Collaborative Attention Network for Person Re-identification

Wenpeng Li[1], Yongli Sun[1*], Jinjun Wang[2], Junliang Cao[1], Han Xu[3] , Xiangru Yang[3] , Guangze Sun[1] , Yangyang Ma[1] and Yilin Long[1]

[1]AI Lab, Nanjing Fiberhome Tiandi CO., LTD, Nanjing, Jiangsu, 210019, China

[2]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China

[3]Xi'an, Shaanxi, 710000, China

[*]Corresponding author's e-mail: sunyongli@fiberhome.com

Abstract. The quality of visual feature representation has always been a key factor in many computer vision tasks. In the person re-identification (Re-ID) problem, combining global and local features to improve model performance is becoming a popular method, because previous works only used global features alone, which is very limited at extracting discriminative local patterns from the obtained representation. Some existing works try to collect local patterns explicitly slice the global feature into several local pieces in a handcrafted way. By adopting the slicing and duplication operation, models can achieve relatively higher accuracy but we argue that it still does not take full advantage of partial patterns because the rule and strategy local slices are defined. In this paper, we show that by firstly over-segmenting the global region by the proposed multi-branch structure, and then by learning to combine local features from neighbourhood regions using the proposed Collaborative Attention Network (CAN), the final feature representation for Re-ID can be further improved. The experiment results on several widely-used public datasets prove that our method outperforms many existing state-of-the-art methods.

1. Introduction

The person re-identification (Re-ID) problem is an important yet challenging computer vision task. The full body images of pedestrian bounding box detected by cameras at different locations are usually the input of a typical person Re-ID system. Then, the person Re-ID system searches the stated person who appears and re-appears from all the bounding box images. Compared with a typical face recognition system that obtains the face image from the constrained environment, the person Re-ID task is usually performed in scenes with occlusion, posture change and illumination variation, as well as inter camera differences.

In recent years, plenty of research shows that deep neural network is an effective method at extracting discriminative features for images classification [2, 5, 7], therefore, it is widely used as base models in many person Re-ID methods. To give an example, ResNet [5] was used as the base model to extract the features of person images in many works [3, 12]. For such network, only the global representation can be obtained, so the performance of the model is limited at distinguishing slightly different body images. For example, some approaches based on [5] can only get obvious global feature such as the overall color of clothing, it is not enough for person Re-ID. As illustrated in figure 1, local features or global features may or may not be sufficient for identifying different person if they are not extracted from proper scale. To improve the performance of global feature, local patterns are reported by many researchers. For example, [14] collects representative local blocks from the input image and build the feature representation models, a good balance is achieved between the discriminative power and generalization ability, so as to better solve this problem. Based on the part detector, [26, 29] firstly located body parts, and then extracted different features of each local parts to obtain explicitly a part-based feature representation for Re-ID. Although the experimental results show that the combination of global and local features can enhance the accuracy, these methods are limited by the manual rule and strategy local regions/slices are defined.

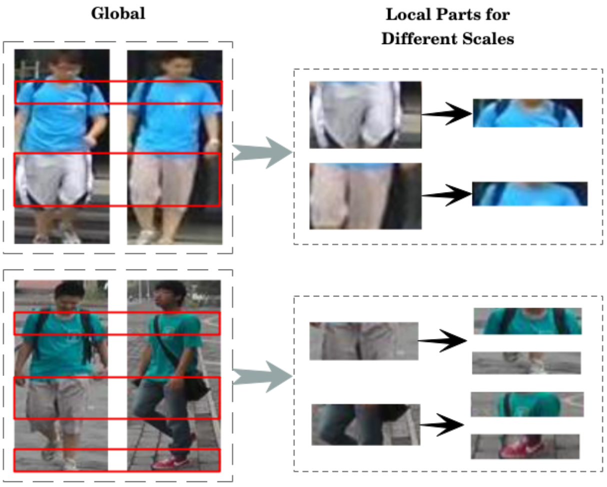

Figure 1. These two pairs of pedestrian images are from the Market-1501 dataset [31]. On the left column are images of various persons with similar clothing and on the right column are local parts partitioned from images by different scales. Models which solely extract global features cannot make a distinction between different persons for each pair. In the upper pair, persons can hardly be distinguished by local features in smaller scale while local parts with larger scale play a more important role in this example. In the lower pair, the colors of shoes which are local details in smaller scale are essential in this case.

To operate in a more data-driven way, researchers also apply the attention mechanism to implicitly infer the importance between local regions. For instance, [1] used the High-Order Attention module which contains high-order statistics to capture the subtle difference information of different person. Part-based methods usually aim at learning the attention between parts, so the end-to-end method based on learning the relevance of each part or multi-parts combination has also been studied. For example, [20] presented a part-based Re-ID models where the final feature representation is obtained by arbitrary combination of local features divided into fixed spatial positions and size from mid-level features. By giving the same attention to each part, this method produces excellent performance which can be viewed as a hard combination mechanism. The challenge is, on the other hand, when cutting the global feature into multiple slices, it becomes difficult to retain the global level information, and hence the problem becomes as how to define the proper parts or part combinations which is usually ad-hoc and may easily overfit to the underline dataset. Our experiments show that as we keep increasing the granularity of feature slices, model accuracy will decrease dramatically once there are too many of them for the dataset.

The situation has motivated us to introduce the Collaborative Attention Network (CAN) to further improve the performance of combined global and local feature for person Re-ID. Our idea is to firstly over-segment the bounding box image and then apply the proposed CAN to learn to weighted combine the adjacent slices to maintain the intrinsic relation between the adjacent features in the neighbourhood regions, the information is not lost when the global features is divided, and more discriminative features are generated based on the local patterns. There are two contributions in this paper:

- We show that finer local features are beneficial for more discriminative feature representation given proper method to combine it with global level feature.
- We present the Collaborative Attention Network that enhances the popular part-based model by enforcing the dependency information from neighbourhood regions.

2. Related works

With the large-scale development of related research, deep-learned neural network methods is gradually replacing the traditional hand-craft feature extraction in person Re-ID. Based on many public person Re-ID datasets [13, 22, 23, 31], researchers can train Re-ID models on a massive image data. Some deep learning methods with tricky strategies to specifically deal with person Re-ID problems are proposed. Then, partial features are proven to be useful because persons are distinguished between each other by local details along with general global features. So recently, part-based methods and attention mechanism are becoming popular. For multi-task learning, some of the methods used metric loss along with classification loss to form the multiple loss functions. The Triplet Loss is introduced into person Re-ID, which improve the model performance by a large margin [6]. Muti-task learning has become a common strategy when training a person Re-ID model.

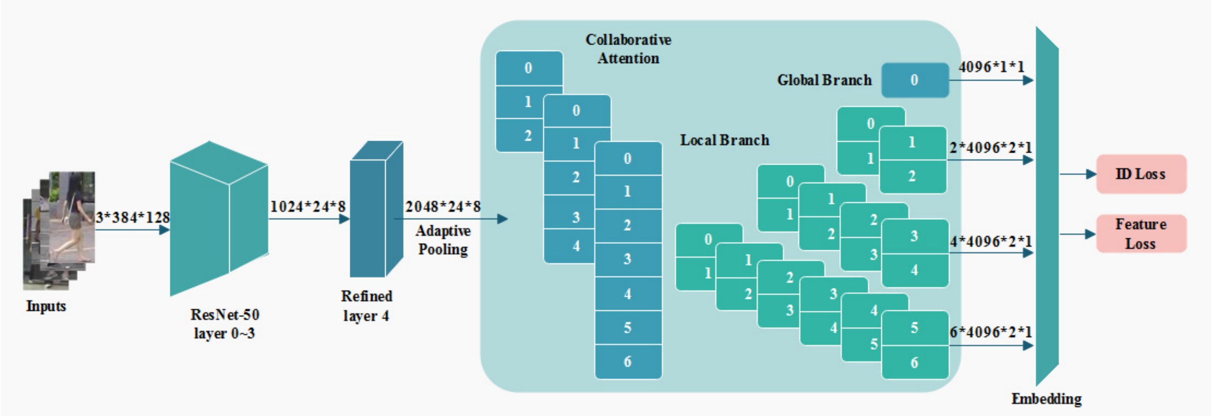

Figure 2. Pipeline of Collaborative Attention Network (CAN) method. We use ResNet50 where the Layer 4 is refined as the base model to get more precise features. Collaborative Attention mechanism is introduced into the partial feature branches and the multi-loss function is used to ensure the feature representation more discriminative.

2.1. Methods with attention mechanism

Developed from deeply-learned methods in the early stage and in order to solve misalignment problem, some models which rely on attention mechanisms were proposed. [10] proposed a content-aware network, which learn discriminative features over local body part and full body. In [29], Zhao et al. reported that body part extractor inspired by attention models can improve the performance combined with feature calculation. Li et al. in [11] addressed the effectiveness of combining attention selection and feature representation in an end-to-end learning process. Also, the hard regional attention and the soft pixel-wise attention were both used to deal with multiple levels of attention subjects. Body landmarks and person pose estimation can also be viewed as an attention mechanism. Landmark detectors and body segmentation network [8, 16, 26] were used as the basis of the main body attention in some methods which also achieved great progress. Nevertheless, the performance of these methods is not ideal since the difference between the datasets of training such detectors and person retrieval datasets.

2.2. Part-based methods

The more effective part-based model is the major contribution of this paper. Inspired by the attention models, some researchers established some simple but strong and useful part-based methods. The Part-Based Convolutional Baseline (PCB) is proposed in [20], where a uniform partition on global features was used. Then these local features were used for classification and this method was verified to be effective yet not to be improved. In [20], an attention mechanism named refined part pooling (RPP) method is also introduced that can re-assign parts by image blocks instead of uniformly slicing the images. Inspired by this part-based approach, the novel Multiple Granularity Network (MGN) is put forward in [21]. As the state-of-the-art method on person Re-ID benchmarks. MGN combines global and local features as the final feature representation with multi-branch structure. The original feature maps are separated into 2 and 3 stripes in branches of local features. Due to the diversity of granularity, different branches have different preference in feature extraction. The effective combination of global features and multi-granularity local features ensures that a more comprehensive feature extractor will be obtained. Our method is inspired by the multi-branch structure but the difference is that our method has more finer feature partition and more effective collaborative attention method.

3. Method

3.1. Overview

Collaborative Attention Network (CAN) is a globally-and-locally muti-branch network and the pipeline is shown in figure 2. Generally, the pipeline can be separated into: base feature extractor, Collaborative Attention module, feature embedding layer and the loss function. In our proposed structure, global features are firstly sliced into partial features by different scales to form multiple branches and then both local and global features are combined to obtain discriminative feature representations.

3.2. Network structure

In this part, the detail of our network structure will be introduced including the base model, Collaborative Attention module and the feature embedding layer.

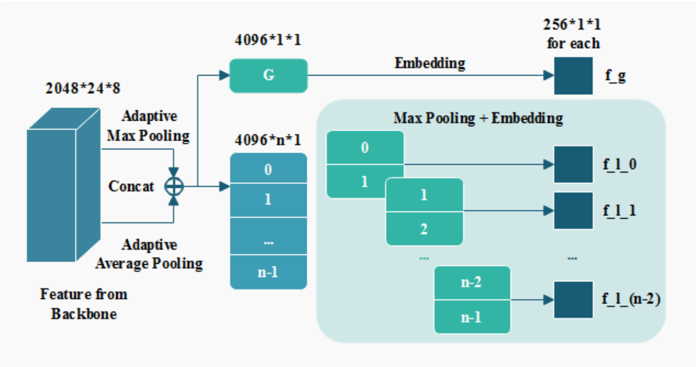

Figure 3. Collaborative Attention module for local branch. The feature map from the base model for each local branch is passed into Adaptive Pooling Layer and outcomes the global block and local feature map with specific size. Then, adjacent local blocks are collaboratively used as local features in our proposed part-based method. Local and global features are used to produce the final feature representation.

3.2.1. Base model for feature extraction. Typically for person Re-ID problem, deep neural networks for image classification are utilized as base models. In our proposed CAN structure, ResNet50 [5] is used as the base model since it gives outstanding performance for image classification. Layer 0 to Layer 3 in ResNet50 are kept the original forms while Layer 4 is modified for the sake of following model extension. Since the global features are evenly separated into several parts after the base feature extraction, features can be enriched if feature maps with larger size can be obtained from base model. Under this consideration, we modify the Bottleneck Layer in Layer 4 of ResNet50 with a stride of 1 for convenience of cutting global features into more pieces. After modified Layer 4, $2048 \times 24 \times 8$ dim features can be obtained.

3.2.2. Collaborative attention module. The pipeline is separated into one global branch and three local branches from modified Layer 4. We also retain the global features in all local branches along with the global branch. Since we argue that global features can reveal distinct attributes in different branches, neural weights of Layer 4 are not shared among all branches. This Collaborative Attention module is shown in figure 3.

After we obtain the feature maps from Layer 4, we need to horizontally slice the feature maps into different parts for each branch. So firstly Adaptive Pooling is introduced to generate feature maps with different scales. In Adaptive Pooling Layer, when the output size is fixed and input size is given by the base model, other hyperparameters of pooling module can be inferred by the following equations (assuming padding in pooling module equals to 0):

$$\begin{cases} padding = 0 \\ stride = \boldsymbol{floor}(IS/OS) \\ kernel\_size = IS - (OS - 1) * stride \end{cases} \tag{1}$$

Where IS is the size of the input feature map and OS represents the size of the output feature map. From the figure 3, we operate both Adaptive Max Pooling and Adaptive Average Pooling and then concatenate the outputs together before passing into next step. Former experiments have reported the effectiveness of such Adaptive Concatenation Pooling module. Because some feature information can be lost either in max pooling or in average pooling, the feature dimension is expanded by concatenating max and average pooling outputs, which can also enrich the feature information. The size of feature maps are 3×1 , 5×1 and 7×1 respectively for each local branch which will be sliced into parts. As what is shown in figure 3, besides the feature map that will be divided into partial features, a global feature block G with size of 1×1 is generated for the reason that the global features learning by the neural network could be varied for different scales of local features. These global features from local and global branches are then delivered into embedding layer which will be introduced in the following subsection. As a result, Table 1 shows the details of feature dimensions for each branch after the Adaptive Pooling Layer in our network.

Table 1. The size and dimension of feature map for each branch after the Adaptive Pooling Layer.

Branch	Feature map size	Feature dimension
Global	1×1	2048×2
Part_3	$3 \times 1 + 1 \times 1(global)$	2048×2
Part_5	$5 \times 1 + 1 \times 1(global)$	2048×2
Part_7	$7 \times 1 + 1 \times 1(global)$	2048×2

After achieving a multi-branch feature structure, Collaborative Attention module is proposed to deal with local features in each branch. Figure 3 shows how the Collaborative Attention works for $part\_n$ branch in which local feature $\boldsymbol{f_L}$ is in size of $n \times 1$. Local features are evenly divided into n blocks $\left\{\boldsymbol{f_{L_0}}, \boldsymbol{f_{L_1}}, \cdots, \boldsymbol{f_{L_{(n-1)}}}\right\}$ with size of 1×1 for every block. As what is mentioned in [20], such partition operation is a hard attention mechanism in which we make the learning process focus on the separated parts in a manual way. In our method, we concatenate two adjacent local blocks $\left[\boldsymbol{f_{L_k}^T}, \boldsymbol{f_{L_{(k+1)}}^T}\right]^{\boldsymbol{T}}$ as the Collaborative Attention mechanism. Then, the Max Pooling is used to downsample local features to obtain partial features $\boldsymbol{f_{LP_k}}$, which make the local features and auxiliary global features have identical feature map sizes. With help of Collaborative Attention module, richer information can be included to enhance the learning process without losing spatial neighbourhood features. We apply this module to each local feature branch, and the experimental results suggest that compared with some part-based methods, it can produce a better performance.

3.2.3. Embedding layer. Global features from Adaptive Pooling and local features from Max Pooling are all have the dimension of 4096, in order to reduce the redundancy of the final feature representation, a 1×1 convolutional layer is used as the embedding layer. The 1×1 Convolutional layer is widely used as feature map encoders while it can also reduce the feature dimensions. In our method, we reduce the 4096-dim features to the dimension of 256 for each feature maps and the weights are shared among all branches. The weights set of this layer is $\boldsymbol{W^T} = \{\boldsymbol{w_1^T}, \boldsymbol{w_2^T}, \cdots, \boldsymbol{w_m^T}\}$ where $m = 256$ is the final feature dimension for each branch. The features from branch $part\_n$ is shown in the following equation (2):

$$\boldsymbol{F_n} = \{\boldsymbol{f_G}, \boldsymbol{f_{LP_0}}, \cdots, \boldsymbol{f_{LP_{n-1}}}\} = \{\boldsymbol{f_{n_0}}, \boldsymbol{f_{n_1}}, \cdots, \boldsymbol{f_{n_{n-1}}}\} \tag{2}$$

in which $n = 3, 5, 7$ represents different branches where the feature maps are divided into 3, 5, and 7 parts. Then the Embedding Layer can be expressed by equation (3):

$$\mathbf{W^T}\boldsymbol{F_n} = \{\boldsymbol{W^T f_G}, \boldsymbol{W^T f_{LP_0}}, \cdots, \boldsymbol{W^T f_{LP_{n-2}}}\} = \{\boldsymbol{W^T f_{n_0}}, \boldsymbol{W^T f_{n_1}}, \cdots, \boldsymbol{W^T f_{n_{n-1}}}\} = \{\boldsymbol{f_{i_0}}, \boldsymbol{f_{i_1}}, \cdots, \boldsymbol{f_{i_{n-1}}}\} \tag{3}$$

For each local branch, the feature vector is $\boldsymbol{f_{i_0}}$ is the embedded global feature and the others are embedded local features. Each feature map $\boldsymbol{f_{i_k}}$ where $k = 0, 1, \cdots, (n-1)$ equals to:

$$\boldsymbol{f_{i_k}} = \boldsymbol{W^T f_{n_k}} = [\boldsymbol{w_1^T f_{n_k}}, \boldsymbol{w_2^T f_{n_k}}, \cdots, \boldsymbol{w_m^T f_{n_k}}] = \left[f_{i_{k_0}}, f_{i_{k_1}}, \cdots, f_{i_{k_{m-1}}}\right] \tag{4}$$

where m = 256 is the dimension of every feature map after embedding. Four 256-dimensional global feature maps and twelve 256-dimensional local features maps are obtained with the processing of Collaborative Attention Network.

3.3. Loss function

As a crucial step in the training process. Combining ID loss and Triplet Loss is the typical strategy of loss function for person Re-ID task. ID loss ensures that the model can classify person well and Triplet Loss is beneficial for obtaining more discriminative features. In addition to the ID loss and Triplet Loss, we also use center loss [12, 25] in our loss function. The final loss function can be expressed as:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{Trip} + \lambda\mathcal{L}_C \tag{5}$$

In the above formula, CrossEntropy is commonly employed in classification problem as the ID loss. After the FC Layer and Softmax Layer, the output vector is probabilities for different classes which can be expressed as $\boldsymbol{q}$ and the ground truth for this feature is $\boldsymbol{p}$ which is a one-hot vector. Then the CrossEntropy can be formulated as:

$$\mathcal{H}(\boldsymbol{p}, \boldsymbol{q}) = -\sum_{i=1}^{k} p_i \log q_i \tag{6}$$

where k is the number of class. The predicted probabilities will be close to the ground truth by minimizing the CrossEntropy.

For a single mini-batch in our training strategy, there are many different IDs and each ID has several bounding box images. Therefore, the Triplet Loss and center loss are designed to guide the training process with feature distance. For Triplet Loss $\mathcal{L}_{Trip}$ in this equation, hard Triplet Loss is adopted in our training:

$$\mathcal{L}_{Trip} = \sum_{i=1}^{N}\left[\alpha + \left\|\boldsymbol{f}_i^a - \boldsymbol{f}_i^p\right\|_2^2 - \|\boldsymbol{f}_i^a - \boldsymbol{f}_i^n\|_2^2\right]_+ \tag{7}$$

where $\boldsymbol{f}_i^a, \boldsymbol{f}_i^p, \boldsymbol{f}_i^n$ are the anchor feature, positive feature and negative feature, respectively. The batch hard positive and negative features are used to improve performance for Triplet Loss, which means:

$$\begin{cases} \boldsymbol{f}_i^p = \arg\max_{f_i^p} \left\|\boldsymbol{f}_i^a - \boldsymbol{f}_i^p\right\|_2^2 \\ \boldsymbol{f}_i^n = \arg\max_{f_i^n} \|\boldsymbol{f}_i^a - \boldsymbol{f}_i^n\|_2^2 \end{cases} \tag{8}$$

Center loss $\mathcal{L}_C$ is designed to make samples with identical ID close to clustering center of this ID, which is formulated as:

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m}\left\|\boldsymbol{x}_i - \boldsymbol{c}_{yi}\right\| \tag{9}$$

where $\boldsymbol{c}_{y_i}$ denotes the feature center for the class of y_i.

Supervised by the multi-loss function with Cross Entropy Loss, hard Triplet Loss and Center Loss, the training process can be more effective and the model can give out more discriminative features. The benefits of utilizing such type of multi-loss function can be seen from the experiments result in the next section.

4. Experiments

In order to verify the effectiveness and robustness of our proposed method Collaborative Attention Network. We conducted a number of comparative experiments on several public datasets which contains the Market-1501 [31], the DukeMTMC-reID [13] and the CUHK03 [22]. The details of the benchmarks and these datasets are elaborated as follows.

4.1. Datasets and benchmarks

4.1.1. Market-1501. This dataset was captured by six cameras (five high-resolution cameras and one low-resolution camera) on campus of Tsinghua University. The dataset was separated into training set and testing set which is further split into a gallery and a query set for testing. There are 1501 person IDs and 12936 bounding box images in total, in which one certain person was captured by at least two cameras. Bounding boxes in the query set are manually labeled while in the gallery set were labeled by DPM detector.

4.1.2. DukeMTMC-reID. This is a subset of DukeMTMC dataset and specificlly for the person Re-ID task. There are 16522 training images from 702 person IDs, 2228 query images from another 702 person IDs and 17661 gallery images from the same person IDs as query set. Images are sampled from video tracks by every 120 frames and all the videos are captured by eight high-resolution cameras on Duke campus. This dataset was manually labeled to ensure the quality bounding boxes.

4.1.3. CUHK03. There are 1467 different person IDs in this dataset collected from 5 pairs of cameras. The set has 13164 images in total with different image size and is separated into three groups: manually-labeled images for training, DPM-detected images for training and testing set. The dataset also suggested two types of testing protocols, and we reported the one that splits the testing set into the query and the gallery part, so as to keep consistency with other datasets.

4.1.4. Benchmarks. The Cumulative Matching Characteristics (CMC) and mean average precision (mAP) are two commonly-used measurements of Re-ID model accuracy. For single-gallery-shot condition, we use the CMC top-k accuracy. The top-k accuracy equals to 1 if top-k ranked gallery samples contain query identity and equals to 0 otherwise. For multi-gallery-shot condition, such top-k accuracy cannot properly describe the model accuracy. Then, the mean average precision (mAP) is used to compute the mean of the precision for every query sample and it is suitable to measure the multi-gallery-shot circumstance.

Table 2. Comparison between different levels of how to partition global features into local features.

Patterns of how to partition	mAP	rank-1	rank-3	rank-5
$Part$-1, 2, 3	86.6	94.4	97.2	97.9
$Part$-1, 2, 3, 4	86.4	94.4	97.2	98.0
$Part$-1, 3, 5	86.1	94.6	97.5	98.1
$Part$-1, 3, 5, 7	85.4	94.3	97.2	98.0
$Part$-1, 3, 5, 7, 9	83.9	93.9	96.9	97.9

Table 3. Comparative experiments of part-based methods without/with Collaborative Attention mechanism.

Method	mAP	rank-1	rank-3	rank-5	rank-10
$Part$-1, 2, 3 with CA	87.4	94.8	97.6	98.4	99.1
$Part$-1, 2, 3, 4 with CA	86.7	94.5	97.5	98.3	99.1
$Part$-1, 3, 5 with CA	87.6	95.0	97.8	98.4	99.2
$Part$-1, 3, 5, 7 with CA	87.9	95.2	97.5	98.3	99.2
$Part$-1, 3, 5, 7, 9 with CA	87.2	94.8	97.4	98.2	98.9
$Part$-1, 3, 5, 7 w/o CA	85.4	94.3	97.2	98.0	98.8
$Part$-1, 3, 5, 7 with CA	**87.9**	**95.2**	**97.5**	**98.3**	**99.2**

4.2. Effectiveness of Collaborative Attention

As we explained above, slicing the global region into multiple parts can help obtaining local features with more details, but over segmenting the global region may on the other hand loose global information and overfit. Hence in this section, we report the model performance under different number of part partition, and the results are shown in Table 2.

Note that the first row in Table 2 with $part$-1, 2, 3, it is essentially the same model as MGN [21]. $Part$-1, 2, 3 means that the final feature includes the global feature, as well as local features by partitioning the global region into 2 and 3 parts, as explained in Section 3.2. The accuracy with such partition can achieve 86.6% mAP score and 94.4% rank-1 score. However, as we continue to grow the number of local parts, the accuracy decreses, and for $part$-1, 3, 5, 7, 9 the accuracy is 83.9% for mAP and 93.9% for rank-1. It shows that simply increasing the number of partition may not be always effective.

Next we apply the proposed CAN to make use of collaborative local features, and present the results in Table 3. Comparing Table 2, it is clearly seen that the proposed CAN can bring in improvement, where in all combinations of partitions, $part$-1, 3, 5, 7 achieves the best performance in which the CA mechanism increased the model accuracy from 85.4% to 87.9% for mAP and from 94.3% to 95.2% for rank-1.

Table 4. Experimental results of progressively adding different loss functions and comparative results between solely global features and twofold (global and local) features in clustering loss function.

Loss Combination	mAP	rank-1	rank-3	rank-5
CAN+$\mathcal{L}_{CE}$	87.0	95.0	97.8	**98.7**
CAN+$\mathcal{L}_{CE} + \mathcal{L}_{Trip}$	87.9	95.2	97.5	98.3
CAN+$\mathcal{L}_{CE} + \mathcal{L}_{Trip} + \mathcal{L}_{C}$	88.3	95.4	97.6	98.4
Global & local to $\mathcal{L}_{C}$	**89.6**	**95.7**	**97.8**	**98.6**

4.3. Combination of loss functions

In this paper, we use multi-loss functions where ID Loss is to improve the classification abilities in the model while Triplet Loss and Center Loss ensures that output features are more discriminative. In the experiments, we compare the results when we progressively introduce CrossEntropy $\mathcal{L}_{CE}$, triplet Loss $\mathcal{L}_{trip}$ and Center Loss $\lambda\mathcal{L}_{C}$. We also argue that it would be useful if the outputs from both global and local branches are guided by Triplet Loss and Center Loss, unlike the loss function strategy in MGN in which only global features play a part in Triplet Loss. Table 4 shows the experiment results of loss-function-related .

If only ID Loss is employed as loss function, our proposed Collaborative Attention Network (CAN) can achieve the performance mAP/rank-1=87.0%/95.0%. Triplet Loss and Center Loss can enhance features clustering attributes and we name these two loss functions as clustering loss. We firstly take global features into calculation of clustering loss as what the existing methods do. After combining Triplet Loss with ID Loss (CrossEntropy), CAN achieves mAP/rank-1=87.9%/95.2% and when Center Loss is adopted, the accuracy is improved to mAP/rank-1=88.3%/95.4%. Then we furtherly use features from local branches as input of clustering loss to guide the training process, the model achieves the best performance with mAP 89.6% and rank-1 95.7%. So supervising each local feature by loss functions during training process can enhance the feature performance when testing.

4.4. Visualization of the method

To validate the effectiveness of our method, we also visualize the obtained feature maps and make comparisons between different methods in figure 4. The features maps are before the last pooling layer of the network for each method, which can be viewed as the final feature representations. It can be clearly seen that, when we partition the global features into several parts in column two and three without using Collaborative Attention, the highlight in the feature maps breaks into disconnected pieces, and the global information is obviously lost. The results from column four and five from MGN shows continuous highlight due to limited number of local parts defined. By introducing our proposed Collaborative Attention in column six and seven, we can clearly see that the feature maps are focused on the whole human bodies without losing the neighbourhood feature information between adjacent local parts. As a result, our method obtains the most discriminative feature representations which outperforms other existing methods.

4.5. Implement details

In this section, the details of the training strategy for the Collaborative Attention Network is introduced. Because the Triplet Loss and Center Loss are applied to our final loss function, we set the mini-batch to 32 images where 8 different person IDs are randomly picked with 4 bounding boxes images for each ID. Each input image is resized to 384×128 with height of 384 and width of 128. The loss function is a multi-loss function with CrossEntropy, Triplet Loss and Center Loss and the weight of Center Loss in this function is set to 0.0005. Adam optimizer is utilized with default parameters ($\epsilon = 10^{-8}, \beta_1 = 0.9, \beta_2 = 0.999$). The initial learning rate is 3×10^{-4} and the learning rate is decayed to 3×10^{-5} at epoch 250, decayed to 3×10^{-6} at epoch 350 and decayed to 3×10^{-7} at epoch 450. The whole training process stops at epoch 600. We also normalize the features and weights in FC Layer so we use the cosine distance as metrics to measure feature similarity during testing process.

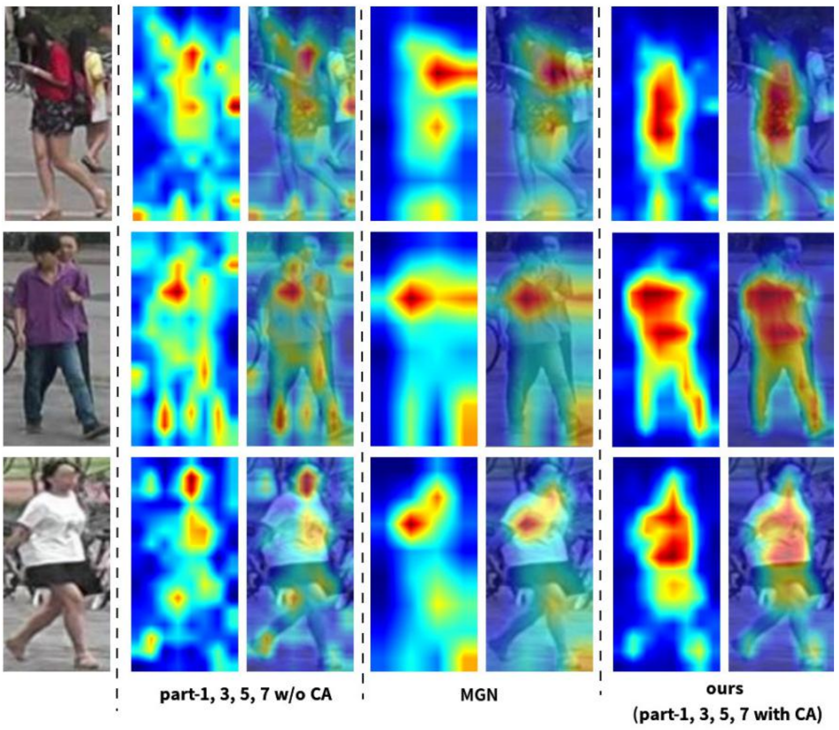

Figure 4. Visualization of feature maps before the last pooling layer in the networks of different methods. The left column of feature maps are from the model of part-1,3,5,7 without introducing Collaborative Attention. The middle column is from the output of MGN while the right column are the feature maps from our proposed method.

4.6. Comparison with the state-of-the-art methods

In this part, based on the dataset of Market-1501, DukeMTMC-reID and CUHK03, the results on our proposed CAN method is compared with some existing state-of-the-art person Re-ID methods respectively. Our experimental results do not use the re-ranking. The comparison with existing image-based person Re-ID on Market-1501 is shown in Table 5 and the comparison results on DukeMTMC-reID and CUHK03 datasets is shown in Table 6. As we can see from Table 5, the pyramidal approach proposed in [30] whose accuracy is mAP/rank-1=88.2%/95.7%, but our proposed method achieves mAP/rank-1=90.6%/96.4% which outperforms the previous best approach by 2.4% in mAP and 0.7% in rank-1.

It can be seen from Table 6 that compared with all the existing approaches the pyramidal method still achieved the best performance with mAP 82.3% and rank-1 91.6% on DukeMTMC-reID, while our proposed method achieves mAP 82.8% and rank-1 86.1%, which outperforms all published methods by a noticeable margin. The superior performance from our proposed Collaborative Attention Network validates the effectiveness and robustness of our method.

Table 5. Comparing with current state-of-the-art methods on Market-1501.

Existing Methods	mAP	rank-1	rank-5	rank-10
Spindle [28]	-	76.9	91.5	94.6
SVDNet [19]	62.1	82.3	92.3	95.2
PDC [18]	63.4	84.1	92.7	94.9
PSE [15]	69.0	87.7	94.5	96.8
Cam-style [33]	71.6	89.5	-	-
GLAD [24]	73.9	89.9	-	-
HA-CNN [11]	75.7	91.2	-	-
CNN-Base [3]	79.8	92.5	-	-
PCB+RPP [20]	81.6	93.8	97.5	98.5
HPM [4]	82.7	94.2	97.5	98.5
SGGNN [17]	82.8	92.3	96.1	97.4
SPRe-ID [9]	83.4	93.7	97.6	98.4
MHN [1]	85.0	95.1	98.1	98.9
DG-Net [32]	86.0	94.8	-	-
MGN [21]	86.9	95.7	-	-
DSA-Re-ID [27]	87.6	95.7	-	-
Pyramid [30]	88.2	95.7	98.4	99.0
CAN(ours)	**90.6**	**96.4**	**98.8**	**99.3**

Table 6. Comparing with existing image-based Re-ID approaches on DukeMTMC-reID and CUHK03.

Existing Methods	DukeMTMC-reID		CUHK03	
	mAP	rank-1	mAP	rank-1
CNN-Base [3]	68.5	83.5	59.0	63.5
PSE+ECN [15]	62.0	79.8	30.2	27.3
Cam-style [33]	57.6	78.3	-	-
GLAD [24]	-	-	-	85.0
HA-CNN [11]	63.8	80.5	41.0	44.4
PCB+RPP [20]	69.2	83.3	57.5	63.7
HPM [4]	74.3	86.6	63.9	57.2
MHN [1]	77.2	89.1	65.4	77.2
DG-Net [32]	74.8	86.6	-	-
MGN [21]	78.4	88.7	67.4	68.0
DSA-Re-ID [27]	74.3	86.2	75.2	78.9

Pyramid [30]	79.0	89.0	76.9	78.9
CAN(ours)	**82.3**	**91.6**	**82.8**	**86.1**

5. Conclusion

In this paper, we propose the multi-branch Collaborative Attention Network (CAN) for feature extraction in the person Re-ID task. The method confirms that a combination of both global and local feature can lead to outstanding Re-ID performance, and at the same time, the proposed network utilizes the collaborative attention mechanism to overcome the overfitting problem when excessive local parts are defined. The method also includes a modified loss functions where both global and local branches are imported into the traditional Triplet Loss and Center Loss. Comparison with other approaches shows that our model can achieve superior performance than existing state-of-the-art methods on public datasets.

Acknowledgments

We gratefully acknowledge the patient and professional guidance of Professor Wang in thesis writing.

References

[1] Chen B, Deng W, Hu J. (2019) Mixed high-Order attention network for person re-Identification. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul. pp. 371-381.

[2] Deng, J., Dong, W., Socher, R., Li, L.J., Li, F.F. (2009) Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida. pp. 248-255.

[3] Fu, X., Yang, X., Cao, Z., Gong, K., Zhou, J.T. (2018) Towards good practices on building effective cnn baseline model for person re-identification. arXiv preprint arXiv:1807.11042.

[4] Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T. (2019) Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii. pp. 8295-9302.

[5] He, K., Zhang, X., Ren, S., Sun, J. (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV. pp. 770–778.

[6] Hermans, A., Beyer, L., Leibe, B. (2017) In defense of the triplet loss for person re-identifification. arXiv preprint arXiv:1703.07737.

[7] Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E. (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 7132-7141.

[8] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B. (2016) Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. Switzerland. pp. 34-50.

[9] Kalayeh, M.M., Basaran, E., Gokmen, M., Kamasak, M.E., Shah, M. (2018) Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 1062-1072.

[10] Li, D., Chen, X., Zhang, Z., Huang, K. (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. pp. 384-393.

[11] Li, W., Zhu, X., Gong, S. (2018) Harmonious attention network for person re-Identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 2285-2294.

[12] Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W. (2019) Bags of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles. pp. 0-0.

[13] Cucchiara, R., Tomasi, C. (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking. Switzerland. pp. 17-35.

[14] Zhao, R., Ouyang, W., Wang, X. (2014) Learning mid-level filters for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio. pp. 144-151.

[15] Sarfraz, M.S., Schumann, A., Eberle, A., Stiefelhagen, R. (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 420-429.

[16] Shelhamer, E., Long, J., Darrell, T. (2017) Fully Convolutional Networks for Semantic Segmentation. IEEE transactions on pattern analysis and machine intelligence. 39(4): 640-651.

[17] Shen, Y., Li, H., Yi, S., Chen, D., Wang, X. (2018) Person re-identification with deep similarity-guided graph neural network. In: Proceedings of the European Conference on Computer Vision. Munich. pp. 486-504.

[18] Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q. (2017) Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. Venice. pp. 3960-3969.

[19] Sun, Y., Liang, Z., Deng, W., Wang, S. (2017) Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. Venice. pp. 3800-3808.

[20] Sun, Y., Liang, Z., Yi, Y., Qi, T., Wang, S. (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision. Munich. pp. 480-496.

[21] Wang, G., Yuan, Y., Xiong, C., Li, J., Xi, Z. (2018) Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia. Seoul. pp. 274-282.

[22] Li, W., Zhao, R., Xiao, T., Wang, X.G. (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio. pp. 152-159.

[23] Wei, L., Zhang, S., Wen, G., Qi, T. (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 79-88.

[24] Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q. (2017) Glad: Global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia. California. pp. 420-428.

[25] Wen, Y., Zhang, K., Li, Z., Yu, Q. (2016) A Discriminative Feature Learning Approach for Deep Face Recognition. In: European Conference on Computer Vision. Switzerland. pp. 499-515.

[26] Yao, H., Zhang, S., Zhang, Y., Li, J., Qi, T. (2019) Deep representation learning with part loss for person re-identification. IEEE Transactions on Image Processing, 28(6): 2860-2871.

[27] Zhang, Z., Lan, C., Zeng, W., Chen, Z. (2019) Densely semantically aligned person re-Identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles. pp. 667-676.

[28] Zhao, H., Tian, M., Sun, S., Jing, S., Tang, X. (2017) Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. pp. 1077-1085.

[29] Zhao, L., Li, X., Zhuang, Y., Wang, J. (2017) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. Venice. pp. 3219-3228.

[30] Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R. (2019) Pyramidal person re-identification via multi-loss dynamic training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles. pp. 8514-8522.

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

[32] Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J. (2019) Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles. pp. 2138-2147.

[33] Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y. (2018) Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 5157-5166.