Multiple Biological Granularities Network for Person Re-Identification

Shuyuan Tu
Central South University
ChangSha, China
8209190307@csu.edu.cn

Tianzhen Guan
Central South University
ChangSha, China
184712059@csu.edu.cn

Li Kuang[*]
Central South University
ChangSha, China
kuangli@csu.edu.cn

ABSTRACT

The task of person re-identification is to retrieve images of a specific pedestrian among cross-camera person gallery captured in the wild. Previous approaches commonly concentrate on the whole person images and local pre-defined body parts, which are ineffective with diversity of person poses and occlusion. In order to alleviate the problem, researchers began to implement attention mechanisms to their model using local convolutions with limited fields. However, previous attention mechanisms focus on the local feature representations ignoring the exploration of global spatial relation knowledge. The global spatial relation knowledge contains clustering-like topological information which is helpful for overcoming the situation of diversity of person poses and occlusion. In this paper, we propose the Multiple Biological Granularities Network (MBGN) based on Global Spatial Relation Pixel Attention (GSRPA) taking the human body structure and global spatial relation pixels information into account. First, we design an adaptive adjustment algorithm (AABS) based on human body structure, which is complementary to our MBGN. Second, we propose a feature fusion strategy taking multiple biological granularities into account. Our strategy forces the model to learn diversity of person poses by balancing the local semantic human body parts and global spatial relations. Third, we propose the attention mechanism GSRPA. GSRPA enhances the weight of spatial relational pixels, which digs out the person topological information for overcoming occlusion problem. Extensive evaluations on the popular datasets Market-1501 and CUHK03 demonstrate the superiority of MBGN over the state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → *Multimedia and multimodal retrieval.*

KEYWORDS

Computational Imaging, Person Re-Identification, Multiple Biological Granularities, Global Spatial Relation

[*] Li Kuang is the corresponding author.

ICMR '22, June 27–30, 2022, Newark, NJ, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531365

ACM Reference Format:
Shuyuan Tu, Tianzhen Guan, and Li Kuang. 2022. Multiple Biological Granularities Network for Person Re-Identification. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512527.3531365

1 INTRODUCTION

Person re-identification (Re-Id) aims to retrieve images of a specific pedestrian among the gallery captured by different times, places, or cameras device. In Re-Id task, Re-Id model select the top-n person images which are the most similar with the original query image among gallery. Previous studies have proposed plenty of excellent models based on deep convolution network which promote the accuracy to a new level in person re-identification [1, 3, 16]. Researchers recently begin to resort to local informative feature representations located in significant human body parts. The person image is divided into several local regions and each region contains a portion of informative semantics from the whole image, which can alleviate interferes from adjacent non-relational semantic information. According to recent studies, researchers pay more attention to extracting features from multiple granularities. The previous multiple granularities are commonly under the pre-defined principle such as biological concepts about human body structural information [15, 25, 28]. To take more different sizes of granularities into consideration, MGN [28] proposes a new combination strategy among multiple granularities taking comprehensive human body structure into account. However, the granularities combination strategy of MGN are too fine resulting in overfitting. Most of previous multiple granularities-based methods are based on the pre-defined division principle for locating image regions, which potentially result in performance fluctuations with the variable poses condition.

Recently, some studies focused on attention dedication in Re-Id task. In essence, the goal of attention is how to enhance the weight of discriminative features and suppress the weight of irrelevant interferences appropriately. The learnable modeling functions with attention mechanism need to imitate the attention mechanism of human physiological vision [5, 14, 29, 31]. According to the recent implement of attention [30], most of researchers design their attention model based on convolution operation in Re-Id field with limited receptive fields. To alleviate the restriction, researchers try to use convolution layers in larger size of kernel which increases the complexity of the whole network. However, this style of attention ignores the exploration of the spatial relational features in the global scope. Therefore, this style of attention is potentially constrained under the variable poses condition.

We make four contributions: (1) We design an adaptive adjustment algorithm (AABS) based on human body structure. Our AABS forces the network to learn human body structure knowledge. (2) We design the Re-Id model named Multiple Biological Granularities Network (MBGN), which uses ResNet-50 as the backbone and split the pipeline into global branch and local branches for extracting different size of feature maps. The feature fusion strategy of MBGN includes multiple biological granularities taking the balance of feature fusion into account. (3) We design the attention mechanism named Global Spatial Relation Pixel Attention (GSRPA), which uses involution blocks [12] for extracting the global spatial relational vectors. We stack each relational vector and obtain the attention matrix from convolution operations for relation map embedding, forcing the model to learn topological information. (4) Our method boosts the recognition accuracy to a promising level in person Re-Id, which is evaluated on the benchmark datasets Market-1501 [35] and CHUK03 [14]. Our method outperforms the latest popular method CAL [21] model by 1.0% on Rank-1 and 2.8% on mAP on Market-1501 respectively.

2 RELATED WORK

Most of the studies concentrate on background clutter [15, 17] and occlusion problems [27]. Researchers also pay extensive attention to unsupervised learning [13], transformation of online learning and offline learning [37]. Moreover, researchers begin to apply NLP approaches to Re-Id task [8], which mostly refers to Transformer [7]. Besides, Graph neural network begins to occupy the direction of person Re-Id [9, 32], boosting higher matching accuracy in general. However, researchers always design their models under the similar assumption that all of their query and gallery images have the similar postures. The diversity of poses constraints their advanced models. Consequently, occlusion and variable pose Re-Id challenge is still attractive for our studies in this field.

Multi-granularity feature fusion strategy is a popular solution for occlusion and variable pose Re-Id problems, which has drawn a surge of interests from both industry and academia. Multi-granularity feature fusion strategy aims to find out the balance of coarse-grained level, fine-grained level and medium-grained level feature representations. The strategy forces the model to learn different semantically related features in different perspectives. It is widely acknowledged that PCB [25] is the first approach taking multi-granularity feature fusion into consideration. However, PCB refers to the average feature segmentation which destroys the semantics of the original person images. MGN [28] proposes a new feature fusion strategy among multiple granularities. MGN takes the human body structure into account which boosts the recognition accuracy dramatically. However, the segmentation schema of MGN is based on the local pre-defined body parts which is ineffective with situations of diversity of person poses and occlusion. Therefore, MGN is not suitable to handle with unfriendly person images with poses changes. Then, RMGL [29] performs local partition on the intermediate representations to operate receptive region ranges novelty, rather than current approaches on input images or output features. However, local partition on the intermediate representations may cause misalignment problem. In our work, our multi-granularity feature fusion includes fine-grained level features, medium-grained level features, structure features which has more robustness.

In design of attention in Re-Id, researchers commonly aim to extract informative feature representations using amount of convolution layers to reach a larger receptive field [33]. Moreover, researchers introduce a kind of dedicated attention block which is consist of convolution layers, boosting the performance of extracting spatial significant feature maps [30]. Hu et al. [10] design a kind of encoder-decoder style attention block including numbers of convolution layers, which promote the model to learn more global feature representations. However, above approaches are constrained with the large variance due to the limited receptive fields. Although some of the approaches implement larger kernel size and more numbers of convolution layers to obtain the larger receptive fields, they increase spatial complexity. Zhang et al. [36] proposed a global attention mechanism for mining the global information. However, this global attention mechanism only uses convolution operations. The common characteristics of convolution are spatial agnostic and channel specific, which is not suitable for exploring global spatial relation. In our work, we introduce involution [12] operation which is channel agnostic and spatial specific.

In our work, we intend to dig out relational structural semantics by traversing every feature vector extracted from convolution layers and calculating their affinity. We stack every relational feature vector and calculate the attention matrix from it by our dedicated network. Our network determines the most significant part through the global range comparison.

3 APPROACH

In this section, we propose the Multiple Biological Granularities Network (MBGN) based on Global Spatial Relation Pixel Attention (GSRPA). Besides, we design an adaptive adjustment algorithm (AABS) based on human body structure, which is complementary to our model. Figure 1 illustrates our model architecture, which contains two independent branches with multiple unitary branches in general. Our model takes multiple biological granularities into account novelty, including fine-grained level features, medium-grained level features, structure features.

3.1 Multiple Biological Granularities Network

ResNet-50 is pre-trained on ImageNet dataset as the backbone network. In order to obtain the feature representation map in appropriate size, we remove the full connection layers of ResNet-50. GSRPA modules are embedded after the last four residual blocks in ResNet-50 for digging out global spatial relation pixel information. The network is divided into two branches, including *Branch g* and *Branch l*. *Branch g* extracts the global features of the images. According to human body structure, the original global feature maps in *Branch l* are adjusted to the appropriate size by our dedicated AABS. *Branch l* extracts multi-granularity local features and calculates the similarity matrix between local features. We input similarity matrix into the fully connected network for learning the posture features of human body structure. Posture knowledge can deal with the problem of posture changes of the specific person under the same camera. In testing stage, the features of query images are concatenated, matching against the person features in gallery.

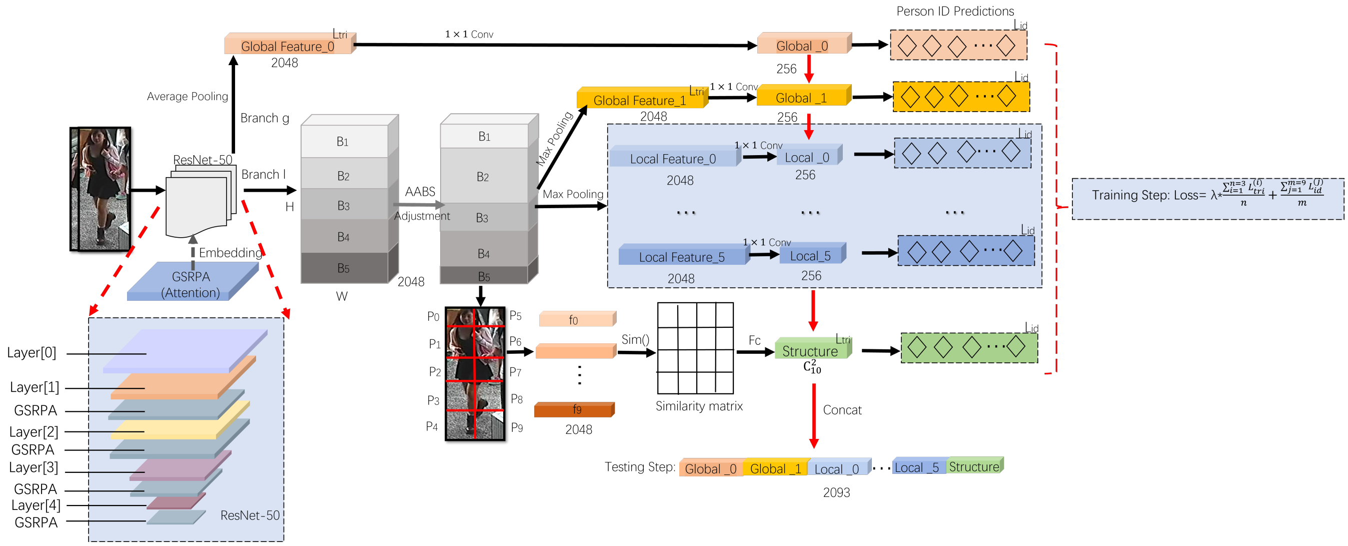

Figure 1: Architecture of MBGN based on GSRPA. L_{id} refers to identification loss and L_{tri} refers to Triplet loss. Pretrained ResNet-50 is used as the backbone, which is equipped with our dedicated attention layer. MBNG contains two independent branch which refers to *Branch g* and *Branch l*. *Branch g* aims to extract global feature maps. *Branch l* aims to extract multiple local feature maps which include fine-grained level features, medium-grained level features, structure features and global features. The global features in *Branch l* are obtained by max pooling, while the global features in *Branch g* are obtained by average pooling.

Concretely, the global feature map in *Branch l* is bisected, and then the upper and lower parts are trisected and bisected respectively to form five local features (*B1, B2, B3, B4, B5*) of different sizes initially. *Bi (i=1, 2...,5)* is shown in Figure 1. The five local features will be input to our dedicated algorithm named AABS. Then we implement our adaptive adjustment algorithm (AABS) based on human body structure to adjust the segmentation for overcoming feature misalignment problem. The details of our adaptive adjustment algorithm (AABS) will be described in the Section 3.2.

According to Figure 1, we also calculate the structure feature to eliminate part of the influence caused by the pedestrian's own posture change. After the adaptive adjustment of AABS, we divide the pedestrian's global feature map into two equal parts in the horizontal direction, thus obtaining 10 local feature maps. We pool these 10 local feature maps into 2048-dimensional feature vectors. The cosine similarity between two feature vectors is used to measure the similarity between features. All local feature similarity forms a similarity matrix with a total of C_{10}^2 similarity values. The similarity matrix is reconstructed into posture feature vectors. We input the reconstructed feature vectors into a learnable network composed of 3 fully connected layers to obtain the structure feature vector.

In order to take the balance of feature fusion into account, we propose a new promising combination strategy among different size of feature representations in *Branch l*. After the segmentation and adjustment of the global feature map, we carry out our multi granularity combination strategy *(B1, B1+B2, B2+B3, B3+B4, B4+B5, B5)*, which obtains 6 local feature vectors and a global feature map (e. g., *B1* refers to fine-grained level feature and *B1+B2* refers to medium-grained level feature). *Bi (i=1, 2...,5)* is shown in Figure 1. Our combination strategy is a process of mutual learning between local features, which can offset the weak semantics of fine-grained local features. Meanwhile, it can avoid over fitting caused by too much attention to local features and alleviate the impact of human semantic misalignment to a certain extent.

In two branches, we both obtain the global feature vectors, which are based on different pooling operation. Average pooling operation can preserve the background semantics from images, while max pooling operation can obtain more discriminative local texture feature. In order to suppress the weight of interfere from background information, our global feature is consisted with two different pooling perspectives.

3.2 Adaptive Body Segmentation Algorithm

In some unfriendly person images, a segment may contain multiple overlapping body parts. To alleviate this problem, we design an adaptive adjustment algorithm (AABS) based on human body structure as the standard of local feature segmentation to replace the traditional average segmentation schema.

The input of AABS is five local features (*B1, B2, B3, B4, B5*) from *Branch l*. *Bi (i=1, 2...,5)* is shown in Figure 1. Each local features forms a segmentation plane. The output of AABS are five adjusted local segmentation planes.

The entire algorithm process is summarized in Algorithm 1. $T < w, h_t, h_b, d >$ denotes width, upper boundary in height, lower boundary in height, depth of the feature map input respectively. $R < r_0, r_1, \ldots, r_n >$ denotes the input human body structure ratio and n is the number of split planes. $S(X, Y)$ denotes the cosine similarity. GAP refers to the global average pooling.

In Algorithm 1, the algorithm initially calculates the positions of all the initial segmentation planes, and then adjusts each segmentation plane successively based on the cosine similarity between the pixel plane near the original segmentation plane and the local feature map. The above process is iterated until the segmentation plane is no longer changed, and the adjusted segmentation plane is the output.

Algorithm 1 Adaptive Human Body Segmentation

```
Input:
    Feature map $T < w, h_t, h_b, d >$, the number of split planes $n$,
    human body structure ratio $R < r_0, r_1, \dots, r_n >$
Output:
    Adjusted local segmentation plane of feature map $L'$
Calculate the initial split plane $L < l_0, l_1, \dots, l_{n-1} >$;
for $i = 0$ to $n - 1$ do
    $l_i \leftarrow (h_b - h_t) * \sum_{j \leftarrow 0}^{i} r_j / \sum_{j \leftarrow 0}^{n} r_j$;
end for
Calculate the initial feature maps $P < p_0, p_1, \dots, p_n >$;
$p_0 \leftarrow \langle w, 0, l_0, d \rangle$;
for $i = 1$ to $n$ do
    $p_i \leftarrow \langle w, l_{i-1}, l_i, d \rangle$;
end for
$L' \leftarrow L$;
while $L' \neq L$ do
    $L \leftarrow L'$;
    for $i = 0$ to $n - 1$ do
        for $j = 0$ to $1$ do
            Pool feature map to feature vector: $f_{i+j} = GAP(p_{i+j})$;
            $p'_{ij} \leftarrow < w, l_i + j - 1, l_i + j, d >$;
            Pool feature plane $p'_{ij}$ to feature vector: $f'_{ij} = GAP(p'_{ij})$;
        end for
        Calculate the Similarity: $S(f'_{i0}, f_i), S(f'_{i0}, f_{i+1})$;
        Calculate the Similarity: $S(f'_{i1}, f_i), S(f'_{i1}, f_{i+1})$;
        if $S(f'_{i0}, f_i) < S(f'_{i0}, f_{i+1})$ and $S(f'_{i1}, f_i) < S(f'_{i1}, f_{i+1})$ then
            $l_i \leftarrow l_i - 1$;
        else if $S(f'_{i0}, f_i) > S(f'_{i0}, f_{i+1})$ and $S(f'_{i1}, f_i) > S(f'_{i1}, f_{i+1})$
        then
            $l_i \leftarrow l_i + 1$;
        else
            $L'_i \leftarrow l_i$;
        end if
    end for
end while
return $L'$;
```

3.3 Global Spatial Relation Pixel Attention

Global Spatial Relation Pixel Attention (GSRPA) aims to calculate the global relation matrix by considering the relational global feature maps with the help of convolution and involution operation [12]. Common convolution kernel has two ordinary characteristics: spatial agnostic and channel specific, while involution is channel agnostic and spatial specific. Therefore, involution is more suitable to extract global relation in spatial perspective.

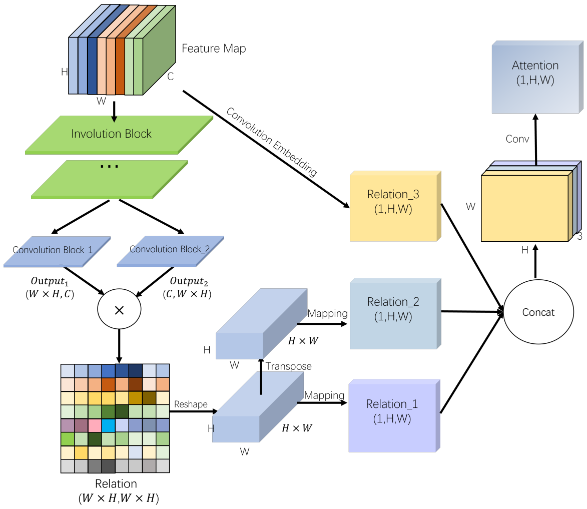

Figure 2: Illustration of GSRPA calculation.

Figure 2 illustrates calculation procedure of Global Spatial Relation Pixel Attention. First, we input the feature maps to the involution block for mining global relation in spatial perspective. Second, the output is input to two convolution blocks for extracting different directional relation respectively. To obtain the total relational map, we multiply the output matrices of two convolution blocks. Third, in order to determine the most significant spatial topological relations from each direction, we concatenate three final maps which refer to Relation_1, Relation_2 and Relation_3 in Figure 2. In the end, we convolute the output of the third step to obtain the attention matrix.

Mathematically, the calculation pipeline is described as Eq.2, Eq.4, Eq.5 and Eq.7 logically. *inv* and f denote the involution operator and input feature maps respectively. $Conv_1$ and $Conv_2$ denote the specific convolution block respectively which is depicted in Figure 2. Eq.1 describes the calculation of $Conv_1$ and $Conv_2$. *BN* denotes batch normalization. The input feature maps can be split into multiple feature vectors. $Output_i$ refers to the Convolution $Block_i$ (i=1, 2) in Figure 2.

Eq.2 obtains the total relational map, which the coordinate of (i, j) in **Relation** matrix refers to the relational information between the i-th feature vector and the j-th feature vector from the original input feature maps. [**Relation**(i,j), **Relation**(j,i)] refers to the integral relation between the i-th feature vector and the j-th feature vector. Therefore, the Relation matrix contains bi-directional relations.

$$\mathrm{Conv}_i(f) = \mathrm{ReLu}(\mathrm{BN}(\mathrm{Conv}(f))) \quad i = 1, 2 \tag{1}$$

$$\mathbf{Relation} = \mathrm{Conv}_1(\mathrm{inv}(f)) \times \mathrm{Conv}_2(\mathrm{inv}(f)) \tag{2}$$

Eq.4 and Eq.5 obtain two direction relation maps by reshaping and transposing. To determine the most significant spatial topological relations from each direction, we input reshaped map and transposed map to the mapping layer respectively. The mapping layer is based on convolution operation. Eq.3 describes the calculation of mapping layer.

$$\mathrm{Map}(X) = \mathrm{ReLu}(\mathrm{BN}(\mathrm{Conv}(X))) \tag{3}$$

$$Relation_1 = \mathrm{Map}(Reshape(Relation)) \tag{4}$$

$$Relation_2 = \mathrm{Map}(Reshape(Relation).T) \tag{5}$$

Eq.7 obtains the attention matrix by concatenating and convolution mapping for taking the global relations among pixels and original feature map itself into consideration. Eq.6 describes the calculation of embedding operation. *Gap* denotes global average pooling.

$$\mathrm{Em}(f) = \mathrm{Gap}\left(\mathrm{Conv}_{1\times1}(f)\right) \tag{6}$$

$$Attention = \mathrm{Conv}\left(Concat\left(\mathrm{Em}(f), Relation_1, Relation_2\right)\right) \tag{7}$$

In Figure 2, we denote C, H, W as the channels, width, and height respectively. The Involution Block contains two involution layers followed by batch normalization. Each involution layer is designed in 7×7 kernel size and 16 group channels, which can get the best capacity to extract spatial relation according to Li et al. [12]. Then, each convolution block is followed by batch normalization (BN) and ReLu layer (ReLu).

In Figure 2, all mapping operation in GSRPA refers to the convolution mapping, which is consisted with Bn, Conv and ReLu. The embedding operation is consisted with convoluting (1×1) and global average pooling in order to overcome the domain gap between the relation information and original feature map.

3.4 Training

The loss function of our Re-Id model contains recognition identification loss and Triplet loss [4]. Eq.8 demonstrates the total loss function. λ denotes the loss proportion coefficient.

$$L = L_{id} + \lambda * L_{tri} \tag{8}$$

(1) The recognition identification loss (L_{id}) is described in Eq.9.

$$L_{id} = -\sum_{i=1}^{N} q_i \log\left(p_i\right), q_i = \begin{cases} 0, y \neq i \\ 1, y = i \end{cases} \tag{9}$$

We denote p_i as the output probability distribution of input image with category i. We denote y as the sample label and denote N as the number of person images in the training dataset.

(2) The triplet loss (L_{tri}) is described in Eq.10.

$$L_{tri} = -\sum_{i=1}^{P} \sum_{\substack{anc=1 \\ pos=1..K}}^{K} \left[\left\| f_{anc}^{(i)} - f_{pos}^{(i)} \right\|_2 - \sum_{\substack{n=1..K \\ j=1..P \\ j\neq i}}^{\min} \left\| f_{anc}^{(i)} - f_{neg}^{(j)} \right\|_2 + \alpha \right]_+ \tag{10}$$

We randomly select P persons, and each person randomly takes out K images to form PK-size input batch. $f_{anc}^{(i)}$,$f_{pos}^{(i)}$ and $f_{neg}^{(j)}$ represent the feature vectors extracted from anchor samples, positive samples and negative samples respectively. We denote α as the threshold parameter. In order to learn a low dimensional dense mapping space and enable the network to have better generalization ability, the person features used in metric learning adopt 2048-dimensional human structure features without dimensionality reduction.

4 EXPERIMENTS

4.1 Datasets

We conduct our experiments on the public person Re-Id datasets: Market-1501 [35] and CHUK03 [14]. Market-1501 is divided into training set with 12936 persons images of 751 identities and testing set with 19732 gallery persons images and 3368 query persons images of 750 identities. CUHK03 includes 14097 persons images of 1467 identities from 6 devices, which provides two types of annotations DPM-detected bounding boxes and manually labeled person bounding boxes.

Both two datasets contain the situations of occlusion and the diversity of poses among the same identity. The following visualization experiments validate our proposed model’s superiority for this type of situations.

In the evaluation stage, we conduct Rank-1 (R1) and mean average precision (mAP) to assess the performance of our proposed network.

4.2 Implementation Details

ResNet-50 is removed the full connection layer and classification layer. GSRPA modules are embedded after the last four residual blocks in ResNet-50 which is illustrated in Figure 2. When using hard sample mining, the number of persons in batch is 16. The images selected by each person is 4. The threshold parameter *margin* of triplet is set to 1.2. The proportion of person identity loss and measurement loss (triplet loss) is 1:2.

The probability parameter of dropout is 0.5. Adam with momentum of 0.9 is selected as the optimizer. Adam’s beta1 is 0.9 and beta2 is 0.999. The learning rate is initialized to 2e-4, which the first 40 epochs remain unchanged and the learning rate of more than 40 epochs is decay according to the decay rate of 0.98.

4.3 Comparison with the SOTA Methods

Table 1 demonstrates the performance comparisons of our proposed MBGN based on GSRPA with other state-of-the-art (SOTA) methods on Market-1501 and CUHK03. In comparison, we introduce attention-based methods and other non-attention methods to our experiment for validating our model’ s superiority. Although previous models have obtained high accuracy on Market-1501, our model still outperforms the best competitor MGN by 0.8% on Rank-1 and 5.4% on mAP on Market-1501. The most powerful model on the mAP metric on Market-1501 was CAL. Our proposed model outperforms CAL by 1.0% on Rank-1 and 2.8% on mAP on the Market-1501. On the CUHK03, our model outperforms the best competitor RGA-S by 1.1% on Rank-1 and 1.8% on mAP. The plausible reason is that our model takes multiple granularities from human body structure and global relations among pixels into consideration.

By analyzing of the results, we have the following observations:

(1) Compared with non-attention-based methods, attention-based methods RGA-S [36] and CAL [21] have better performance on Market-1501 and CUHK03 in general. It validates that the suitable attention mechanisms can promote the performance of Re-Id models. The attention mechanisms can enhance the weight of the significant person parts and suppress the irrelevant interferences.

Table 1: Performance (%) comparisons with the state-of-the-arts on Market-1501 and CUHK03.

		Market-1501		CUHK03-Labeled	
	Model	Rank-1	mAP	Rank-1	mAP
Attention-Based	MGCAM [24]	83.8	74.3	50.1	50.2
	AACN [31]	88.7	83.0	-	-
	SPReId [11]	92.5	81.3	-	-
	HA-CNN [16]	91.2	75.7	44.4	41.0
	DuATM [23]	91.4	76.6	-	-
	Mancs [26]	93.1	82.3	69.0	63.9
	MHN-6 [2]	95.1	85.0	77.2	72.4
	RGA-S [36]	95.7	88.0	**79.1**	**75.6**
	CAL [21]	95.5	**89.5**	-	-
Non-Attention-Based	PRB+RPP [25]	93.8	81.6	63.7	57.5
	HPM [6]	94.2	82.7	63.9	57.5
	MGN [28]	**95.7**	86.9	68.0	67.4
	UnityStyle [18]	93.2	89.3	-	-
	ViT-B/16 [8]	94.7	86.8	-	-
	HOReID [8]	94.2	84.9	-	-
	TransReID [9]	95.2	89.5	-	-
	GPS [19]	95.2	87.8	-	-
	FileReID [20]	95.5	89.6	-	-
	DCC [36]	95.4	89.2	-	-
	Ours	**96.5**	**92.3**	**80.2**	**77.4**

Table 2: Results of the comparison with different components of MBGN on Market-1501 datasets. G refers to the global branch from MBGN. CBAM-S is a popular attention block.

Model	Rank-1	mAP
Baseline	92.2	82.0
Baseline+G	93.1	83.3
Baseline+AABS	93.5	83.8
Baseline+GSRPA	95.4	91.6
Baseline+CBAM-S [30]	94.8	85.6
MBGN	**96.5**	**92.3**

(2) In contrast, multi-granularities combination-based network MGN [28], DCC [34], CAL [21] have better performance on Market-1501 and CUHK03. It validates that the multi-granularities combination strategy is helpful for boosting the recognition accuracy. Compared with above previous models, our proposed model achieves significant improvement in the recognition accuracy. It validates that the feature fusion strategy of our dedicated model can effectively address the balance of feature fusion with different sizes. The concrete discussion and analysis on our promising feature fusion strategy will be described in Section 4.4.

(3) Compared to the best competitor on Market-1501 (CAL [21]) that learns the attention with counterfactual causality, our model learns the attention with spatial relation information in global scope. The spatial relation information contains clustering-like topological information which is more helpful for overcoming the problem of diversity of poses. The problem of diversity of poses which is existed on Market-1501 may drop the accuracy of CAL.

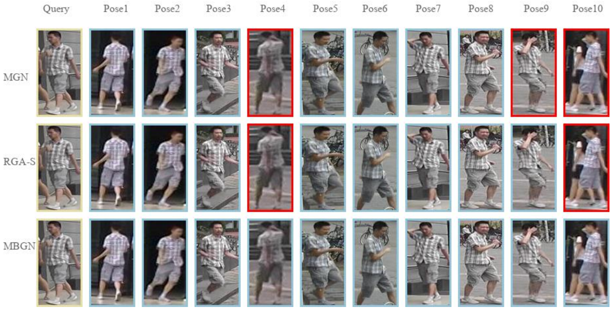

Figure 3: Retrieval visualization of Rank-10 covering results. The images on the first column indicate the query samples and the followed 10 columns are the groundtruth images. The images with red border are the mismatching results, and the images with blue border are the results with correct matching.

(4) Compared to the best competitor on CUHK03 (RGA-S [36]) that also learns the attention with global relation information, our model remains a better method for exploring the global relation information. RGA-S only uses convolution operations which are spatial agnostic and channel specific. Our model uses involution [12] to enhance significant global spatial relation novelty. Involution is channel agnostic and spatial specific which is more suitable to extract global spatial relation information. Our model also contains serval mapping layers for determining the significant topological relations, while RGA-S only concatenates its original feature maps.

(5) In order to validate the superiority of our model in variant pose situation, we visualize retrieve for a set of persons with the

same ID (1186) from Market-1501. Figure 3 shows the Rank-10 covering result among groundtruth images. To show the recognition performance clearly, we only remain 10 specific person images with different types of poses in this ID group from the original test gallery. By observing whether the retrieval result of rank-10 covers the total 10 groundtruth images in this ID group, the performance of both MGN and RGA-S is influenced by multiple pose changes. MGN is based on local pre-defined human body structure. Variant pose may cause person images not to conform to pre-defined principle. RGA-S has better performance compared with MGN due to the exploitation of global relation. However, RGA-S is only based on convolution which is sensitive to channel changes. Variant pose may lead to dramatical changes in the channel, especially when the camera angle is shifted from the front of the person to the back. In contrast, our model has the best performance. The plausible reason is that our model uses involution operator for global relation extraction. Involution is channel agnostic and spatial specific which can alleviate the channel fluctuation influence from variant pose. Our model also takes the human body structure into account by our AABS instead of pre-defined principle. Dynamically adjusting the segmentation of the body is beneficial to handle with the variant pose situation.

4.4 Ablation Study

In the baseline model, we use ResNet-50 as backbone for extracting the person image feature map. Then, we implement PCB [25] to the ResNet-50 for person ID prediction. We set the average segmentation schema as the local feature segmentation schema of our baseline.

According to Table 2, Baseline+G outperforms Baseline by 0.9% on Rank-1 and by 1.3% on mPA. Baseline+AABS outperforms Baseline by 1.3% on Rank-1 and by 1.8% on mAP. Baseline+GSRPA outperforms Baseline by 3.2% on Rank-1 and 9.6% on mPA. Therefore, our GSRPA and MBGN have the promising potential in boosting recognition performance due to mining the global structural information from spatial topological relations.

In Table 2, we implement attention module CBAM-S to our baseline for fairness of comparison with our proposed GSRPA. CBAM-S obtains the attention by a large filter size of 7×7. CBAM-S ignores the relational information in global scope and simply extracts feature maps without processing to mine global relationships. In contrast, Baseline+GSRPA outperforms Baseline+CBAM-S by 0.6% on Rank-1 and by 6% on mAP. It validates that our GSRPA has more potential to boost the performance of the baseline. The plausible reason is that our GSRPA explores the global spatial relational information with the help of learnable modeling functions. The global spatial relational information forces the model to learn the topological and relational person information. The result also validates that our GSRPA has obtained valuable knowledge which refers to global spatial relational knowledge.

By analyzing of the results, we have the following observations:(1) The implementation of global branch can slightly promote the performance of baseline. The plausible reason is that global branch forces the model to extract different size of feature maps, taking the balance of feature fusion into account. Excessive local branches may cause the model to trap into local features which is not suitable for coping with scene transformation or pose transformation situations. (2) Replacing the traditional average segmentation schema with our dedicated AABS also slightly promotes the performance of baseline. It validates that semantic based dynamic segmentation takes more comprehensive human body structure into account. (3) Our dedicated GSRPA is the main factor for performance improvement of baseline. The plausible reason is that GSRPA digs out the person topological and relational information in global range, which is suitable to alleviate the problem of occlusion and the diversity of poses on Market-1501 and CHUK03.

Table 3: Results of comparison with different fusion strategies on Market-1501.

Fusion Strategy	Rank-1(%)	mAP(%)
PCB [25]	90.4	74.6
AABS+PCB	90.8	76.1
AABS+M	90.0	75.4
AABS+M+F	**92.8**	**81.0**
AABS+M+C	87.2	69.3
MGN [28]	91.3	78.7

4.5 Analysis and Discussion on Fusion Strategy

Table 3 demonstrates the results of comparison with different feature fusion strategy. C, F, M refer to the coarse-grained level, fine-grained level and medium-grained level respectively. We denote *Bi* as the i-th local feature unit which is depicted in Figure 1. The M+C refers to the combination of three adjacent local features(*B1+B2,B2+B3,B3+B4,B4+B5, B1+B2+B3,B2+B3+B4,B3+B4+B5*). The M+F refers to our fusion strategy(*B1,B1+B2,B2+B3,B3+B4,B4+B5 ,B5*). The M refers to the combination of two adjacent local features(*B1+B2, B2+B3,B3+B4,B4+B5*). It is necessary to note that each local feature unit *Bi* has obtained after the adjustment of AABS. PCB [25] refers to the average fusion strategy. MGN [28] divides the feature map twice into two equal parts and three equal parts respectively. The architecture of the baseline is the same with ablation experiment part.

According to the description of Table 3, AABS+M+F has the best performance. It validates the superiority of our proposed feature fusion strategy. By analyzing of the results, we have the following observations: (1) Compared with PCB and PCB+AABS, our AABS can promote the model performance due to the consideration of human body structure. (2) The segmentation granularity should not be too large or too small. During the process of feature fusion, coarse-grained level may cause the model not to learn local features thoroughly, while fine-grained level may cause the model to overlearn local features. The combination of medium-grained level and fine-grained level has better performance.

4.6 Evaluation and Retrieval Visualization

Evaluation on GSRPA. According to Figure 1, we implement 4 GSRPA modules to the ResNet-50 for enhancing the significant feature representations. Our design is based on a hypothesis that the number of GSRPA modules is positively correlated with the performance of the network. To address the rationality of our network

Table 4: Results of comparison with the specific network (ResNet-50) equipped with different numbers of GSRPA on CUHK03.

Number of GSRPA	Rank-1(%)	mAP (%)
1	76.2	73.3
2	77.1	74.0
3	78.3	74.2
4	**79.2**	**75.4**

structure design, we will verify the hypothesis in below experiments.

Except for the number of GSRPA modules, the architecture of the model is the same as ResNet-50. Table 4 demonstrates the comparison with the specific network equipped with different numbers of GSRPA modules. We can clearly see that the network equipped with more numbers of GSRPA modules obtains higher recognition accuracy than the network equipped with less numbers of GSRPA modules. Therefore, the number of GSRPA modules is positively correlated with the performance of the network, which confirms the rationality of our network design.

Retrieval Visualization. we select four query samples with Query IDS 0003, 0025, 0210 and 0218 for testing on the test set. The ground-truths corresponding to query samples contains diversity of poses and occlusions such as bikes and railings. Figure 4 illustrates the visualization of Rank-10 results on four specific query samples. The images with blue border on the first column indicate four original query samples and the followed 10 columns are the candidate images sorted from high to low according to the similarity with the query images. The images with red border are the result with wrong matching, and the persons images without border are the result with correct matching.

By analyzing of the results, the plausible reason why our model mismatches specific person is that the person images contain dramatical background change or situation where the pedestrian is facing away from the camera.

4.7 Visualization of Attention

To visualize our proposed attention module, we set Grad-CAM [22] as our baseline, which conducts the qualitative analysis for GSRPA with the comparison. Figure 5 illustrates the comparison with baseline and our GSRPA, which demonstrates the regions covered by our GSRPA masks have more informative semantics than the baseline masks. Our GSRPA pay more attention to the human body structure in global scope and has powerful anti-interference capacity. Consequently, our GSRPA leads the model to concentrate on discriminative body parts taking spatial relational pixel information into account. It is necessary to note that the head of person is always omitted by our GSRPA masks. The plausible reason is that the resolution of head is always low which is not helpful for extracting discriminative feature representations in the global scope.

In order to validate our GSRPA's ability to dig out topological information for alleviating occlusion problem, we select one query sample with occlusion which refers to (a) in Figure 5. According to Figure 5, the person's main body is covered by GSRPA red masks and the occlusion in the lower right corner is not covered by the red masks. It validates that our GSRPA alleviates the constraint of occlusion due to extracting the discriminative features from global spatial relation pixels.

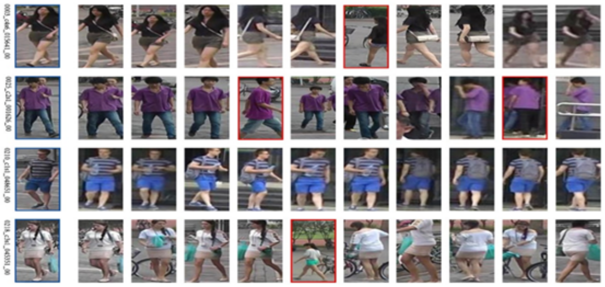

Figure 4: Retrieval visualization of Rank-10 results on four specific query samples.

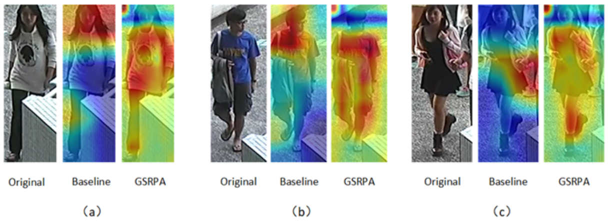

Figure 5: Visualization of the comparison with baseline and our GSRPA. In (a), our GSRPA doesn't pay attention to the occlusion. In (b), our GSRPA pay more attention to the whole-body relational structure according to the red region. In (c), the baseline is disturbed by the secondary person, while our GSRPA still pay attention to the original main pedestrian.

5 CONCLUSION

In this paper, we propose the Multiple Biological Granularities Network (MBGN) based on Global Spatial Relation Pixel Attention (GSRPA) for overcoming the limitations from the occlusion and the diversity of poses in Re-Id task. We design an adaptive adjustment algorithm (AABS) based on human body structure, which is complementary to our MBGN. In some unfriendly person images, a segment may contain multiple overlapping body parts. In order to alleviate the problem, our dedicated AABS dynamically adjusts feature segmentation planes by taking the human body structure into account. With the help of MBGN, we take medium and fine granularities into consideration which overcomes the mismatching problem due to the position misalignment. With the help of our proposed attention layer, we extract the discriminative features from global spatial relation pixels, which boosts the model to resist the interference from occlusion and the diversity of poses. Extensive ablation studies validate the promising power of our model.

6 ACKNOWLEDGMENTS

This work has been supported by the Foundation item: National Key R&D Program of China (2018YFB1402800); National Natural Science Foundation of China (61772560).

REFERENCES

[1] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. 2018. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 2109–2118. https://doi.org/10.1109/CVPR.2018.00225

[2] Binghui Chen, Weihong Deng, and Jiani Hu. 2019. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Seoul, Korea (South), 371–381. https://doi.org/10.1109/ICCV.2019.00046

[3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. 2018. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, 8649–8658. https://doi.org/10.1109/CVPR.2018.00902

[4] Xingping Dong and Jianbing Shen. 2018. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*. Springer International Publishing, Cham, 459–474. https://doi.org/10.1007/978-3-030-01261-8_28

[5] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. 2019. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. AAAI Press, Hilton Hawaiian Village, Waikiki Beach, Honolulu, Hawaii, United States, 8287–8294. https://doi.org/10.1609/aaai.v33i01.33018287

[6] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. 2019. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. AAAI Press, 8295–8302. https://doi.org/10.1609/aaai.v33i01.33018295

[7] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. 2021. LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. *arXiv:2104.01136* (May 2021). http://arxiv.org/abs/2104.01136 arXiv: 2104.01136.

[8] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. Transreid: Transformer-based object re-identification. *arXiv:2102.04378* (March 2021). http://arxiv.org/abs/2102.04378 arXiv: 2102.04378.

[9] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. Transreid: Transformer-based object re-identification. *arXiv:2102.04378* (March 2021). http://arxiv.org/abs/2102.04378 arXiv: 2102.04378.

[10] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 8 (Aug. 2020), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

[11] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, 1062–1071. https://doi.org/10.1109/CVPR.2018.00117

[12] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. 2021. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Nashville, TN, USA, 12321–12330. https://doi.org/10.1109/CVPR46437.2021.01214

[13] Minxian Li, Xiatian Zhu, and Shaogang Gong. 2018. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*, Vol. 11208. Springer International Publishing, Cham, 737–753. https://doi.org/10.1007/978-3-030-01225-0_45 Series Title: Lecture Notes in Computer Science.

[14] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Columbus, OH, USA, 152–159. https://doi.org/10.1109/CVPR.2014.27

[15] Wei Li, Xiatian Zhu, and Shaogang Gong. 2017. Person re-identification by deep joint learning of multi-loss classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 2194–2200. https://doi.org/10.24963/ijcai.2017/305

[16] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, USA, 2285–2294. https://doi.org/10.1109/CVPR.2018.00243

[17] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. 2019. Recover and identify: A generative dual model for cross-resolution person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Seoul, Korea (South), 8090–8099. https://doi.org/10.1109/ICCV.2019.00818

[18] Chong Liu, Xiaojun Chang, and Yi-Dong Shen. 2020. Unity style transfer for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Seattle, WA, USA, 6887–6896. https://doi.org/10.1109/CVPR42600.2020.00692

[19] Binh X Nguyen, Binh D Nguyen, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. 2021. Graph-based Person Signature for Person Re-Identifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Nashville, TN, USA, 3492–3501. https://doi.org/10.1109/CVPRW53098.2021.00388

[20] Xingyang Ni and Esa Rahtu. 2021. FlipReID: Closing the Gap between Training and Inference in Person Re-Identification. *arXiv:2105.05639* (May 2021). http://arxiv.org/abs/2105.05639 arXiv: 2105.05639.

[21] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Nashville, TN, USA, 1025–1034.

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. IEEE, Venice, 618–626. https://doi.org/10.1109/ICCV.2017.74

[23] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, 5363–5372. https://doi.org/10.1109/CVPR.2018.00562

[24] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, 1179–1188. https://doi.org/10.1109/CVPR.2018.00129

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11208. Springer International Publishing, Cham, 480–496. https://doi.org/10.1007/978-3-030-01225-0_30 Series Title: Lecture Notes in Computer Science.

[26] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2018. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. 11208. Springer International Publishing, Cham, 365–381. https://doi.org/10.1007/978-3-030-01225-0_23

[27] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, USA, 6449–6458. https://doi.org/10.1109/CVPR42600.2020.00648

[28] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*. ACM, Seoul Republic of Korea, 274–282. https://doi.org/10.1145/3240508.3240552

[29] Guanshuo Wang, Yufeng Yuan, Jiwei Li, Shiming Ge, and Xi Zhou. 2020. Receptive multi-granularity representation for person re-identification. *IEEE Transactions on Image Processing* 29 (2020), 6096–6109. https://doi.org/10.1109/TIP.2020.2986878

[30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, Vol. 11211. Springer International Publishing, Cham, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1 Series Title: Lecture Notes in Computer Science.

[31] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, 2119–2128. https://doi.org/10.1109/CVPR.2018.00226

[32] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. 2021. Anchor-Free Person Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, TN, USA, 7690–7699. https://doi.org/10.1109/CVPR46437.2021.00760

[33] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. 2019. Attention driven person re-identification. *Pattern Recognition* 86 (Feb. 2019), 143–155. https://doi.org/10.1016/j.patcog.2018.08.015

[34] Hantao Yao and Changsheng Xu. 2021. Dual Cluster Contrastive learning for Person Re-Identification. *arXiv:2112.04662* (Dec. 2021). http://arxiv.org/abs/2112.04662 arXiv: 2112.04662.

[35] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. 2019. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing* 28, 6 (June 2019), 2860–2871. https://doi.org/10.1109/TIP.2019.2891888

[36] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. IEEE, Seattle, WA, USA, 3186–3195. https://doi.org/10.1109/CVPR42600.2020.00325

[37] Jiahuan Zhou, Bing Su, and Ying Wu. 2020. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, USA, 2909–2918. https://doi.org/10.1109/CVPR42600.2020.00298