Multi-Granularity Hypergraphs and Adversarial Complementary Learning for Person Re-Identification

Yi Ma
my1996mh@mail.ustc.edu.cn
Tian Bai *
baitian@ustc.edu.cn
Wenyu Zhang
wenyuz@mail.ustc.edu.cn
Jian Hu
hujian1@mail.ustc.edu.cn

School of Software Engineering
University of Science and Technology of China
Hefei, China

Abstract

Many person re-identification methods possess the following two limitations: *(i)* They suffer from missing body parts and occlusions. *(ii)* They fail to get diverse visual cues. To handle these problems, we proposed a Multi-Granularity Hypergraphs and Adversarial Complementary Learning (MGHACL) method. We specifically first uniformly partitioned the input images into several stripes, which were used later to obtain multi-granularity features. Afterward, we used the proposed MGHACL to study the high-order spatial relations between these features (which makes the studied features robust) and the complementary information of these features (which contain different visual cues). Next, we integrated the studied high-order spatial relation information and the complementary information to improve the representation capability of each regional feature. Moreover, a supervision strategy was used to learn to extract more accurate global features. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods on four holistic person ReID datasets and two occluded ReID datasets.

1 Introduction

Person re-identification aims to find all the pictures in the gallery that are most likely to belong to the same identity as a given picture. It plays an important role in many applications, such as video surveillance, human identity validation, and autonomous driving [8, 29, 41]. However, many existing reID methods may achieve a sub-optimal solution due to occlusion and failing to get the diverse visual cues.

In recent years, to enhance the robustness of learned features, researchers have proposed a part-based method, which can be classified roughly into three categories. The first method [19, 27, 45] uses human semantic or posture information to divide pedestrian pictures into

* Corresponding Author.

different parts. The auxiliary information makes input richer, improving the performances. However, it needs to be manually labeled, which brings high costs, or obtained through pose estimation algorithms, which makes the model overly rely on the accuracy of these algorithms. The second method [3, 6, 20, 38] uses the attention network to focus on the informative regions. It can locate discriminative visual cues without additional information, which is very concise and efficient. However, a problem with this method is that it usually tends to extract the most salient features, while the re-identification of a pedestrian may additionally count on numerous cues suppressed by the most salient information. The third method [9, 23, 25, 39] divides the feature map into fix-height horizontal strips. This method generates more discriminative local features, but it considers each local feature separately and ignores the relations between them, which reduces the performance of the model. Speaking of mining different cues as much as possible, some researchers have proposed methods of suppressing salient features obtained by manually setting a threshold [17] or a class activation map (CAM) [28]. An achievable challenge of these methods is that the setting of the threshold is very technical and the CAM is required to retrain the model.

To address the above problems, we propose the Multi-Granularity Hypergraphs and Adversarial Complementary Learning for person re-identification. Specifically, we first divide the feature map horizontally to obtain a fixed number of stripes, which are used to obtain a multi-granularity feature set. Then we use the proposed Multi-Granularity Hypergraphs Learning (MGHL) to explore the high-order spatial relations between features of the feature set, thereby enhancing their robustness. Next, we use the Adversarial Complementary Learning (ACL) to fully mine complementary information of each feature by erasing its corresponding area in the original feature map separately. Finally, we integrate the learned relations and complementary information to improve the representation ability of each regional feature. Moreover, we also use a supervision strategy to extract more accurate global features. The combination of global features and local features makes our model learn better.

In summary, we have made three major contributions: (1) We propose the MGHL to integrate the high-order spatial relations between the multi-granularity regional features, improving their robustness. (2) We propose the ACL to learn complementary information of each regional feature, which can mine diverse visual cues. (3) We also use a supervised strategy to extract more accurate global features and integrate the global features with these regional features, forming the final more powerful representations of pedestrians.

2 Related work

2.1 Part-based methods

A part-based method is proposed to solve the body part misalignment. Several works [2, 19, 34] in recent years push the performance of person reID to a new level with the aid of pose estimation and human semantic parsing. [19] proposed a model named SPReID, where human semantic parsing is introduced to assist the extraction of local features. Compared with the method based on bounding box part detection, it avoids explicitly detecting human body parts. While this method is difficult to ensure the semantic consistency of different images in terms of the pose variations and occlusion. [46] achieve state-of-the-art performance by proposing a solution called Viewpoint-Aware Loss with Angular Regulariza-tion, which effectively models the distribution on bothidentity-level and viewpoint-level. [34] extract semantic local features using several keypoints detected by a pose estimation model trained on

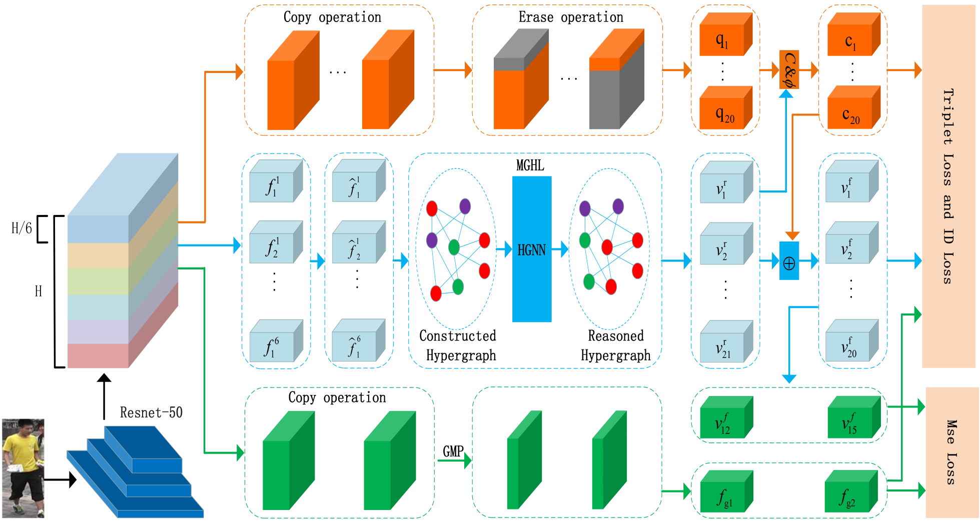

Figure 1: The architecture of our Multi-Granularity Hypergraphs and Adversarial Complementary Learning (MGHACL).

COCO [21]. However, this method makes the performance of the model rely on the accuracy of the estimation algorithm and requires an auxiliary pose dataset.

Benefiting from person structure, other common solutions of the part-based method divide the feature map into several horizontal parts, whose embeddings are trained with non-shared classifiers. These methods can extract robust local features without requiring additional information. To learn part-level features, [31] propose a network named PCB, which conducts uniform partition on the conv-layer rather than the images, and a refined part pooling (RPP) method, which generates refined parts with enhanced within-part consistency. This method is effective except in the case of occlusion. The joint learning of local and global features has been proved to be effective in alleviating occlusion. To this end, the Multiple Granularity Network (MGN) [35] based on PCB combines global features and discriminative regional features, which are learned in different branches. But the performance of MGN is sensitive to the global features extracted by the backbone from the entire picture containing random background information. To extract more robust global and local features, [39] propose a novel pyramidal model that incorporates both the coarse-to-fine features and the gradual cues between them. One advantage of this method is that even in the case of misalignment, it can still work well by matching on different scales. However, this method considers each feature individually and ignores the relations between them, limiting the performance of the model.

2.2 Hypergraph learning

In recent years, many researchers have extended their research on the convolutional neural networks (CNNs) to Graph Convolution Network (GCNs) and achieved promising results in many common computer vision tasks, such as image classification [7], image semantic segmentation [22]. But GCN is few introduced to Re-ID tasks. [30] proposed a similarity-guided graph neural network to update the probe-gallery similarity using the pairwise relations between probe-gallery pairs. [16] proposed a framework named Part-based Hierarchical Graph Convolutional Network, which constructs a hierarchical graph to represent the pairwise relations among different parts and performs both local and global feature learning by the messages passing.

However, GCNs cannot assign different weights to neighbor nodes. [33] introduced an attention-based architecture named Graph Attention Networks (GATs), which can address the shortcomings of prior methods based on graph convolutions using masked self-attentional layers. The studies in [3, 34] show that besides one-order information, high-order one should be imported and may work better for person ReID. While GATs would lead to over-smoothing problems when leveraging higher-order neighbors. To overcome such limitations, the hypergraph neural networks (HGNNs) [1, 10, 18] are proposed recently to learn high-order correlations in hypergraphs by using deep neural networks. [37] propose a HGNN architecture, namely Multi-Granular Hypergraph (MGH), to exploit temporal dependencies in videos. However, this method focuses on exploring multi-granular temporal cues in video-based person re-ID and is not suitable for our image-based person re-ID. Motivated by this work, a hypergraph was constructed in our work. Differently, our method focuses on exploring the high-order spatial cues between multi-granularity features extracted from a person image, resulting in more robust feature representations.

2.3 Diversity feature mining

Deep learning models tend to focus on the most salient information while ignoring some potentially salient but useful cues. However, the re-identification of a person may rely on various clues hidden by the most salient features. To solve this problem, researchers proposed to reversely improve the feature extraction ability of the model by suppressing the high response area of the feature map. [17] propose a Self-attention Guided Adaptive DropBlock Network (SaADB) for person re-ID, which can adaptively erase the most discriminative regions. However, this method returns a drop mask according to a predefined threshold, which is a technical work. [6] propose a Salient Feature Extraction (SFE) unit, which can adaptively extract potential salient features by suppressing the salient features learned in the previous cascaded stage. This strategy can be seen as an explicit drop scheme, which enables the network to discover diverse visual features. While it is limited by the number of cascading layers and causes high coupling of the network. For occluded ReID, diversity features are also important. To obtain more useful information related to the person identity, [45] propose the identity-guided human semantic parsing method for aligned person re-identification, which can locate personal belongings in addition to human parts. [15] presents a novel method named Matching on Sets (MoS), which can positions occluded person re-ID as a set matching task without requiring spatial alignment.

3 The proposed method

Given a set of images $I = \{I_1, I_2, \dots, I_N\}$ containing N images, we use a CNN backbone network $\mathcal{BN}$ to extract individual feature maps. As a result, we get three-dimensional tensors $\{\mathbf{F}_i|_{i=1}^{N}\}$ with shape $C \times H \times W$, where C is the number of channels, and H and W represent the spatial height and width of the tensor, respectively.

Our framework is illustrated in Fig.1. First, we equally divide the feature map F into six horizontal parts, each of which has the size of $C \times (H/6) \times W$, assuming H can be divided by six. In a similar spirit to the pyramidal model proposed in [39], we get a set of 3-dimensional submaps $\mathbf{P} = \{f_k^l | l = 1, \cdots, 6, k = 1, \cdots, 7-l\}$ as shown in Fig.2. We next apply a 1×1 convolution layer with batch normalization and ReLU to each submap, resulting in a set of regional features $\mathcal{F} = \{\hat{f}_k^l \in R^c\}$. We then use the proposed Multi-Granularity Hypergraphs and Adversarial Complementary Learning (MGHACL) to optimize these features.

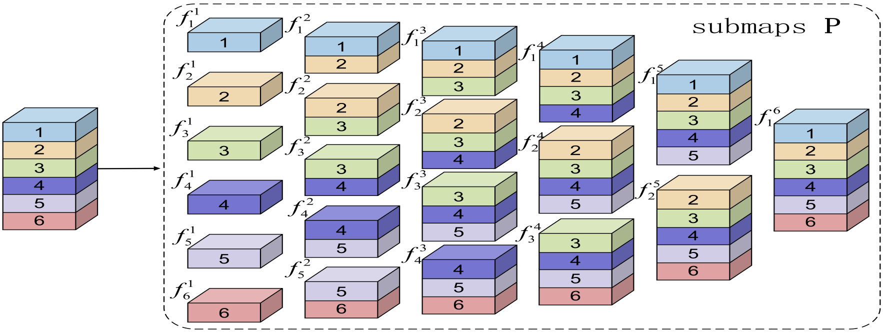

Figure 2: Illustration of the submaps **P**. Each of submap in **P** capture the discriminative information of different spatial scales.

3.1 Multi-Granularity Hypergraphs Learning

Inspired by HGNN, we propose a novel and powerful hypergraph structure to model the high-order spatial relations between multi-granularity features instead of the pairwise relations. The method can be described in detail as follows.

Hypergraph construction. Without loss of generality, we formulate the hypergraph as $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ represent the set of nodes and hyperedges respectively. We regard the multi-granularity feature set $\mathcal{F}$ as the graph nodes set $\mathcal{V}$, whose bottom-level nodes represent smaller stripes of the image, while higher level-nodes represent combination of the above. The hyperedges are defined in detail as follows.

We define any two nodes v_i and v_j in $\mathcal{V}$ as a neighbor if the area corresponding to one node is a subarea of the area corresponding to another node. For any node $v_i \in \mathcal{V}$, where $i \in \{1, 2, \ldots, 21\}$, we denote all its neighbors as $\mathcal{N}(v_i)$. Then we connect v_i with its neighbor nodes whose layer is lower than v_i by m using a hyperedge defined as follows.

$$e_{i,m} = \left\{v_i, \forall v_j \in \mathcal{N}(v_i)\right\}, \text{ s.t. } l_{v_i} - l_{v_j} = m, \tag{1}$$

where l_{v_i} and l_{v_j} represent the number of layers of v_i and v_j respectively. For node v_i, we set $m \in \{1, 2, \ldots, l_{v_i} - 1\}$. In this way, we can get $l_{v_i} - 1$ hyperedges containing the node v_i, which we denote as $\mathcal{H}(v_i) = \{e_{i,m}|_{m=1}^{l_{v_i}-1}\}$.

Hypergraph reasoning. After getting the hypergraph, we can update the nodes by propagating the information between them. To this end, we designed a hypergraph neural network, which takes the nodes set and hyperedges set as input. Specifically, for node v_i, we perform a hyperedge pooling operation on each of its hyperedges separately to obtain corresponding hyperedge features, which can be formulated as:

$$\mathbf{h}_{i,m}^l = \sum_{j \neq i}^{v_j \in e_{i,m}} \mathbf{v}_j^{l-1}, \forall e_{i,m} \in \mathcal{H}(v_i), \tag{2}$$

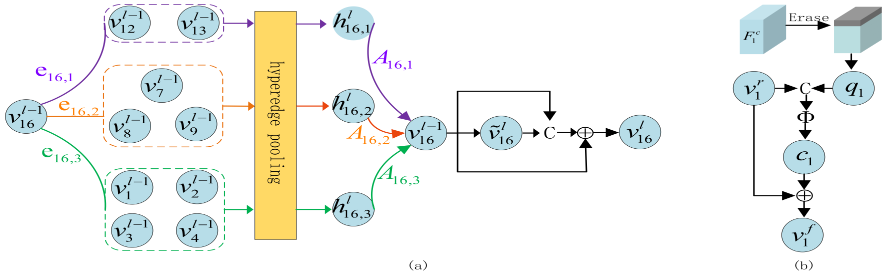

Figure 3: (a) Illustration of Multi-Granularity Hypergraph Learning. Here, we take v_{16} as an example to show node feature propagation. (b) Illustration of Adversarial Complementary Learning. Here, we show a process of extracting the complementary feature of v_1^r. Other features are similarly computed. C is concatenation operation, ϕ is a 1×1 convolution layer with batch normalization and ReLU, and $\oplus$ is element-wise addition.

where $\mathbf{v}_j^{l-1}$ represents the node feature of v_j in layer $l-1$ of our proposed network. Then we can know the importance of each hyperedge associated with the node by comparing the similarity between the hyperedge feature and the node feature:

$$A_{i,m}^l = \frac{\exp\left(S\left(\mathbf{v}_i^{l-1}, \mathbf{h}_{i,m}^l\right)\right)}{\sum_{m=1}^{l_{v_i}-1} \exp\left(S\left(\mathbf{v}_i^{l-1}, \mathbf{h}_{i,m}^l\right)\right)}, \tag{3}$$

where S is a similarity estimation function, which is achieved by cosine similarity in our work. Then we can integrate the hyperedge features as follows:

$$\tilde{v}_i^l = \sum_{m=1}^{l_{v_i}-1} A_{i,m}^l \mathbf{h}_{i,m}^l, \tag{4}$$

Finally, we concatenate the aggregated hyperedge feature and the previous node feature along the channel dimension and fed the combination to a fully connected layer to update the node feature. The whole process can be described by the following formula:

$$v_i^l = W^l\left(C\left(v_i^{l-1}, \tilde{v}_i^l\right)\right), \tag{5}$$

where W^l is the learnable weight matrix and C is a concatenation operation. We take node v_{16} as an example to illustrate the above process, as shown in Fig. 3(a). The above updating process will repeat for the preset r times, resulting in an output node feature set $\mathcal{V}_{hl} = \{v_i^r|_{i=1}^{21}\}$.

3.2 Adversarial Complementary Learning

Neural networks tend to focus on the most salient information while may ignore some potentially salient but useful cues. To alleviate this problem, we propose Adversarial Complementary Learning (ACL), which can mine the diverse discriminative visual cues by learning complementary information of each feature in the $\mathcal{V}_{hl}$. Note that it is unnecessary to learn the complementary information of v_{21}^r as its corresponding area in F is the entire feature map. Specifically, we copy F 20 times, getting 20 copied feature maps $\{\tilde{F}_i|_{i=1}^{20}\}$. To learn the complementary cues c_i of $v_i^r \in \mathcal{V}_{hl}$, we erase the area corresponding to v_i^r in the $\tilde{F}_i$, and obtain q_i by feeding the erased feature map to a fully connected layer. Then we concatenate v_i^r and q_i, and apply a 1×1 spatial convolutional layer to it, getting the complementary cues c_i. We then use skip connections to integrate the v_i^r and c_i, resulting in the final feature v_i^f. Finally, we get an enhanced local feature set $\{v_i^f|_{i=1}^{20}\}$. As shown in Fig. 3(b), we take the generation of v_1^f as an example to illustrate the working process of ACL.

3.3 Supervised global features

A lot of research has proved that combining global and local features is an effective solution to improve performances in a person re-identification task. To extract more precise global features, we use a supervised strategy. Specifically, we copy F two times and feed the copied feature maps into two 1×1 convolution layers to generate global features $f_G = \{f_{g1}, f_{g2}\}$ with size of c. Then we use two specific local features $f_L = \{v_{12}^f, v_{15}^f\}$, which uniformly divide the feature map into two parts in the training stage, to supervise the global features f_G. In this way, the first c-channel global features f_{g1} should be closer to the upper part feature v_{12}^f and in the same way, the second c-channel global feature f_{g2} should be closer to the bottom part feature v_{15}^f, which makes final global features more robust.

3.4 Training and inference

During the training stage, the overall loss function is a combination of cross-entropy loss with label smoothing [32], batch-hard triplet loss [13], and mean squared error(mse) loss, which is formulated in following Eq. 6.

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{triplet} + \mathcal{L}_{mse}. \tag{6}$$

When inferring, enhanced local features $\{v_i^f|_{i=1}^{20}, v_{21}^r\}$ generated by MGHACL and more accurate global features $\{f_{gi}|_{i=1}^{2}\}$ are concatenated together as the final image representation.

3.5 Extension to different numbers of parts

So far, we describe our method assuming that the feature map F is uniformly divided into six parts. Without loss of generality, we can use different numbers of horizontal parts for the person representation, such as $\{6,4,2\}$ indicating that the initial feature map is uniformly divided into six, four, and two horizontal parts. Then our method is applied to do similar processing for different divided cases, obtaining different local features and global features, which are concatenated together later to form the final features for re-ID. Note that these features share the same backbone network with the same parameters.

4 Experiments

4.1 Datasets and evaluation metrics

Datasets. We evaluate our proposed method on four holistic person ReID datasets, including Market-1501 [40], CUHK03 [43], DukeMTMC-reID [42] and MSMT17 [36], and two occluded ReID datasets, including Occluded-Duke [24] and Occluded-ReID [47]. The details of these datasets are summarized in Table 1.

Evaluation metrics. We adopt the Cumulative Matching Characteristics (CMC) at Rank-1 and mean Average Precision (mAP) as the evaluation metrics on all datasets.

Dataset	image	Camera	Identity	train	query	gallery
Market-1501	32,668	6	1,501	12,936	3,368	15,913
CUHK03-labeled	14,096	2	1,467	7,368	1,400	5,328
CUHK03-detected	14,097	2	1,467	7,365	1,400	5,332
DukeMTMC-reID	36,411	8	1,812	16,522	2,228	17,611
MSMT17	126,441	15	4,101	32,621	11,659	82,161
Occluded-Duke	35,489	8	1,812	15,618	2,210	17,661
Occluded-ReID	2,000	-	2,00	-	1,000	1,000

Table 1: Statistics of datasets used in our work

4.2 Implementation details

We use the PyTorch [26] to implement our proposed MGHACL model. The backbone network is the ResNet-50 [11]. All images are resized into a resolution of 384 × 128. The feature dimensions C and c are set to 2048 and 256, respectively. In the training stage, the mini-batch size is set to 64, in which we randomly select 16 identities and 4 images for each identity and the hypergraph layer r = 1. We use the Stochastic gradient descent(SGD) as our optimizer with the momentum 0.9 and the weight decay factor 0.0005. We train our model for 90 epochs. A learning rate for the backbone network and other parts are initially set to 1e-3 and 1e-2 respectively, then divided by 10 after 40, 60 epochs, respectively.

4.3 Comparison with state-of-the-art methods

We compare our proposed method with current state-of-the-art methods on person reID datasets in Table 2. Results in detail are shown as follow:

Results on Holistic Datasets. On Market1501 and DukeMTMC-reID, our method is the second best results and the best result is VA-reid [46] on metrics Rank-1 and mAP respectively. Compared with our method, VA-reid [46] using a more strong bockbone SeResnext with a large number of model parameters. As reported in Table 2, MGHACL exceeds RGA [38]/Pyramid [39] by 5.1%/3.0% on metrics Rank-1/mAP on the CUHK03-detected dataset and exceeds RGA [38] by 6.8%/4.7% on metrics Rank-1/mAP for the CUHK03-labeled dataset, showing a significant improvement over the current best state-of-the-art method. Notice that Pyramid [39] uses a more strong backbone resnet101 and also generates a fine-to-coarse feature set. While our model surpasses Pyramid by 9.0%/5.2% on metric Rank-1/mAP on the CUHK03-labeled dataset. On the large scale MSMT17, in comparison with all other approaches, our MGHACL achieves the best performance which outperforms the second best approaches [4] by 0.7%/0.5% in mAP/Rank-1 accuracy, respectively.

Results on Occluded Datasets. As we can see in Table 3, our method sets a new state-of-the-art performance and outperforms other four kinds of methods on the two occluded datasets. Specially, our method significantly outperforms the second best approaches by 4.8%/2.5% on the occluded-Duke dataset, and 9.3%/1.1% scores on the occluded-Reid dataset in terms of mAP/Rank-1.

Method	Market1501		CUHK03				DukeMTMC-reID		MSMT17	
			Labeled		Detected					
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
RGA [38]	88.4	96.1	77.4	81.1	74.5	79.6	-	-	57.5	80.3
PCB [31]	77.4	92.3	-	-	54.2	61.3	66.1	81.8	-	-
PCB+RPP [5]	81.6	93.8	-	-	57.5	63.7	69.2	83.3	40.4	68.2
MHN(PCB)[3]	85.0	95.1	72.4	77.2	65.4	71.7	77.2	89.1	-	-
MGN[3]	86.9	95.7	67.4	68.0	66.0	68.0	78.4	88.7	-	-
RN [25]	88.9	95.2	75.6	77.9	69.6	74.4	78.6	89.7	-	-
ABDNet [4]	88.2	95.6	-	-	-	-	78.6	89.0	60.8	82.3
IANet[14]	83.1	94.4	-	-	-	-	73.4	83.1	46.8	75.5
Pyramid[39]	88.2	95.7	76.9	78.9	74.8	78.9	79.0	89.0	-	-
SPReID[19]	81.3	92.5	-	-	-	-	70.9	84.4	-	-
OSNet [44]	84.9	94.8	-	-	67.8	72.3	73.5	88.6	52.9	78.7
ISPReID [45]	88.6	95.3	74.1	76.5	71.4	75.2	80.0	89.6	-	-
VA-reid [46]	91.7	96.3	-	-	-	-	84.5	91.6	-	-
MGHACL(ours)	89.4	96.2	82.1	87.9	77.8	84.7	80.9	90.1	61.5	82.8

Table 2: Performance (%) comparisons of our models with the state-of-the-art results on Market1501, CUHK03, DukeMTMC-reID and MSMT17 datasets.

Method	Occluded-Duke		Occluded-ReID	
	mAP	Rank-1	mAP	Rank-1
DSR [12]	30.4	40.8	62.8	72.8
PCB [31]	33.7	42.6	38.9	41.3
PGFA [24]	37.3	51.4	-	-
HOReID [34]	43.8	55.1	70.2	80.3
MGHACL(ours)	48.6	57.6	79.5	81.4

Table 3: Performance (%) comparisons of our models with the state-of-the-art results on Occluded-Duke and Occluded-ReID datasets.

4.4 Ablation study

In Table 4, Baseline is trained only on the backbone Resnet-50. After the backbone, a feature set containing 21 256-dimension local features is extracted in a similar spirit to the pyramidal model proposed in [39]. Then classification loss and triplet loss are combined to train the model. To demonstrate the effects of every component in our method, several variants are then conducted based on the baseline. Table 4 shows the ablation study results, from which several observations could be drawn:

(1) The first row shows the result of the Baseline. From the second row to the fourth row, we can see that the MGHL, ACL, and SG can all improve the performance of the Baseline. For example, The results in the second row demonstrate the effect of our MGHL, which gives the performance gains of 0.3%/0.6% and 1.4%/1.2% for Rank-1/mAP accuracy on the Market1501 and CUHK03-labeled datasets respectively.

(2) The results in the next three rows demonstrate the effect of combining two components. For example, The fifth row shows that the Rank-1/mAP can be further improved (0.6%/0.9% on Market1501, 2.2%/2.4% on CUHK03-labeled) by combining the MGHL with ACL. This proves that it is effective to explore the high-order spatial relations between multi-granularity features and their complementary information. Besides, the results in the sixth and seventh rows show combining SG with ACL or MGHL is better than using them alone, which reveals that better results can be achieved by combining two components.

Model	MGHL	ACL	SG	Ext	Market-1501				CUHK03-labeled			
					mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Baseline	✗	✗	✗	✗	87.9	95.2	98.1	98.5	77.9	84.4	96.2	97.9
Baseline+	✓				88.5	95.5	98.3	98.7	79.1	85.8	96.7	98.2
		✓			88.4	95.4	98.3	98.6	79.6	85.9	96.7	98.2
			✓		88.4	95.4	98.4	98.7	79.5	85.3	96.5	98.1
	✓	✓			88.8	95.8	98.5	98.6	80.3	86.6	96.9	98.1
		✓	✓		88.7	95.7	98.6	98.9	80.9	86.7	97.0	98.4
	✓		✓		88.6	95.7	98.6	98.8	80.1	86.6	96.9	98.3
	✓	✓	✓		89.0	96.1	98.6	98.9	81.5	87.1	97.1	98.4
	✓	✓		✓	89.2	96.1	98.6	98.9	81.9	87.7	97.2	98.4
MGHACL	✓	✓	✓	✓	89.4	96.2	98.7	99.0	82.1	87.9	97.4	98.4

Table 4: Ablation studies of MGHACL on Market-1501 and CUHK03-labeled in terms of Rank-1, Rank-5, Rank-10 accuracy(%) and mAP(%).

(3) From the eighth and ninth row, we can see that combining three components performs better than combining two components in terms of Rank-1 and mAP accuracy. Particularly, the ninth row show that the improvement by our method is primarily owing to the MGHACL. More robust local features and more accurate global features can further improve the performance of person re-identification. The last row shows that our model that using all components performs best, which suggests that exploiting part-level features of multiple scales is also important.

5 Conclusion

In this paper, we propose a novel framework Multi-Granularity Hypergraphs and Adversarial Complementary Learning to address person reID. The proposed framework can not only effectively explore the high-order spatial relations between multi-granularity features by specifically designing a hypergraph neural network but also fully mine potentially salient but useful cues by explicitly exploring complementary information of multi-granularity features. Moreover, we also use a supervised strategy to extract more accurate global features, which combined with local features can improve the performance of the model. A large number of experiments were conducted on four holistic person ReID datasets and two occluded ReID datasets, where the proposed framework outperformed recent state-of-the-art methods.

References

[1] Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017.

[3] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF ICCV*, pages 371–381, 2019.

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou

Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF ICCV*, pages 8351–8361, 2019.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017.

[6] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *CVPR*, pages 3300–3310, 2020.

[7] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[8] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3691–3701, 2019.

[9] Xing Fan, Hao Luo, Xuan Zhang, Lingxiao He, Chi Zhang, and Wei Jiang. Scp-net: Spatial-channel parallelism network for joint holistic and partial person re-identification. In *ACCV*, pages 19–34. Springer, 2018.

[10] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3558–3565, 2019.

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[12] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on CC*, pages 7073–7082, 2018.

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017.

[14] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, pages 9317–9326, 2019.

[15] Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. Matching on sets: Conquer occluded person re-identification without alignment. In *AAAI*, volume 35, pages 1673–1681, 2021.

[16] Bo Jiang, Xixi Wang, and Bin Luo. Ph-gcn: Person re-identification with part-based hierarchical graph convolutional network. *arXiv preprint arXiv:1907.08822*, 2019.

[17] Bo Jiang, Sheng Wang, Xiao Wang, and Aihua Zheng. Saadb: A self-attention guided adb network for person re-identification. *arXiv preprint arXiv:2007.03584*, 2020.

[18] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. Dynamic hypergraph neural networks. In *IJCAI*, pages 2635–2641, 2019.

[19] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on CVPR*, pages 1062–1071, 2018.

[20] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on CVPR*, pages 2285–2294, 2018.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] Yi Lu, Yaran Chen, Dongbin Zhao, and Jianxin Chen. Graph-fcn for image semantic segmentation. In *ISNN*, pages 97–105. Springer, 2019.

[23] Xiaofei Mao, Jiahao Cao, Dongfang Li, Xia Jia, and Qingfang Zheng. Integrating coarse granularity part-level features with supervised global-level features for person re-identification. *arXiv preprint arXiv:2010.07675*, 2020.

[24] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF ICCV*, pages 542–551, 2019.

[25] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11839–11847, 2020.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[27] Rodolfo Quispe and Helio Pedrini. Improved person re-identification based on saliency and semantic parsing with deep neural network models. *Image and Vision Computing*, 92:103809, 2019.

[28] Rodolfo Quispe and Helio Pedrini. Top-db-net: Top dropblock for activation enhancement in person re-identification. In *2020 25th ICPR)*, pages 2980–2987. IEEE, 2021.

[29] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on CVPR*, pages 6036–6046, 2018.

[30] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 486–504, 2018.

[31] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[34] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020.

[35] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018.

[36] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[37] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 2899–2908, 2020.

[38] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020.

[39] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, pages 8514–8522, 2019.

[40] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[41] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[42] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *CVPR*, pages 3754–3762, 2017.

[43] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017.

[44] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF ICCV*, pages 3702–3712, 2019.

[45] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. *ECCV*, 2020.

[46] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Xing Sun, and Weishi Zheng. Viewpoint-aware loss with angular regularization for person re-identification. In *Proceedings of the AAAI*, pages 13114–13121, 2020.

[47] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *2018 IEEE ICME*, pages 1–6. IEEE, 2018.