Branch Interaction Network for Person Re-identification

Zengming Tang[0000−0001−5485−1829] and Jun Huang(✉)[0000−0003−4939−3880]

Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China
{tangzengming2019,huangj}@sari.ac.cn

Abstract. Most existing Person Re-identification (Re-ID) models aim to learn global and multi-granularity local features by designing a multi-branch structure and performing a uniform partition with the various number of divisions in different branches. However, the uniform partition is likely to separate meaningful regions in a single branch, and interaction between various branches disappeared after the split. In this paper, we propose the Branch Interaction Network (BIN), a multi-branch network architecture with three branches for learning coarse-to-fine features. Instead of traditional uniform partition, a horizontal overlapped division is employed to make sure essential local areas between adjacent parts are covered. Additionally, a novel attention module called Inter-Branch Attention Module (IBAM) is introduced to model contextual dependencies in the spatial domain across branches and learn better shared and specific representations for each branch. Extensive experiments are conducted on three mainstream datasets, i.e., DukeMTMC-reID, Market-1501 and CUHK03, showing the effectiveness of our approach, which outperforms the state-of-the-art methods. For instance, we achieve a top result of 90.50% mAP and 92.06% rank-1 accuracy on DukeMTMC-reID with re-ranking.

1 Introduction

Person re-identification (Re-ID) aims to retrieve a person of interest across non-overlapping camera views in a large image gallery with a given probe. Recently, deep learning methods dominate this community, which obtain state-of-the-art results. Deeply-learned features provide discriminative representation ability but still are not robust for many challenges like variations in view angle, pose, and illumination.

To relieve these issues, many part-based methods [1–3] are proposed to learn part features and achieve promising results. They can be categorized into two groups by the number of branches. The first group applies single branch methods,

(✉) Corresponding Author.

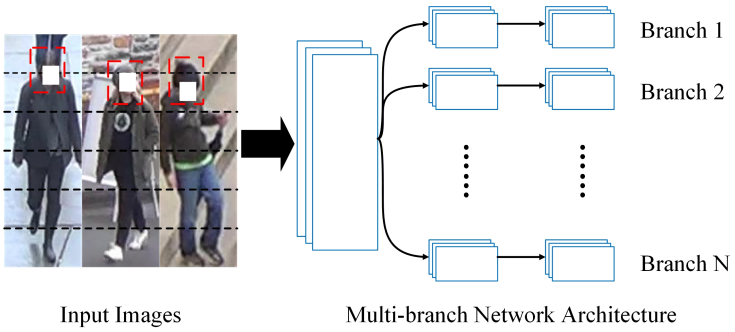

Fig. 1. Overview of problems in the single branch and multiply branch interaction. Left: uniform partition on input images. Heads are divided into two parts, which diminish the representational capability in head regions. Right: multi-branch network architecture. Strong relations between branches vanished after the split.

which split the deep feature maps into several pre-defined patches and promote the network to focus on fine-grained features in each local region. The second group, using multiple branch methods, combines local and global information in different granularities and learns coarse-to-fine representations. Although they push the performance of Re-ID to a new level, they still suffer from the problems of learning in the single branch and ignore the correlation between different branches.

Feature learning in the single branch By the uniform partition, local information is preserved, which is robust for occlusion and partial pedestrian retrieval. For example, Part-based Convolutional Baseline (PCB) [1] is implemented by partitioning feature maps into 6 horizontal stripes. Nevertheless, in PCB, as the number of stripes increases, retrieval accuracy improves at first, but drop dramatically in the end. Over-increased numbers break the balance between learning fine-grained features and extracting meaningful body region information. In other words, the division will separate important semantic parts, as illustrated in Fig. 1.

Correlation between different branches Multi-branch networks have gained state-of-the-art performance by sharing lower layers of network and extracting different granularity features at the higher layers in different branches, as shown in Fig. 1. Since lower layers capture the same low-level features for different branches with the same input image, branch relatedness in lower layers is built. Besides, sharing lower layers keeps the model parameters in a low level. However, during testing, features of all branches are concatenated, but context information between them vanish after the split during training. As a result, interaction among branches is neglected in higher layers of the network.

In this paper, we propose a novel Branch Interaction Network (BIN) to address the above problems. The network learns coarse-to-fine representations for Re-ID in a multi-branch structure. It has three branches. One is for capturing

coarsest information i.e., global information, while others are for learning multi-level fine-grained information with the various number of partitions. In order to preserve the consistency of meaningful regions across equally-sliced parts, Horizontal Overlapped Pooling (HOP) is adopted to extract local features on horizontal overlapped patches of equal size. Furthermore, we propose a new attention module, namely Inter-Branch Attention Module (IBAM), which contains three submodules called Inter-Branch Attention Submodule (IBASM). IBAM aggregates features from three branches and produces three refined corresponding representations complementary to the other two branches. Besides, IBAM, injected in higher layers of the network, promotes all the branches to learn shared features while they are trained in its specific granularity.

To sum up, our main contributions are three-fold:

- We introduce a new pooling strategy called HOP on multi-branch network architecture. HOP, which employs an overlapped division and Global Max Pooling (GMP) to obtain a vector representation, is shown superior to the combination of original uniform partition and GMP.
- We incorporate a novel attention module into BIN to model spatial contextual, multi-level dependencies across branches. It is found that complementary information efficiently promotes the performance of Re-ID. To the best of our knowledge, this is the first work which builds strong relations between different branches for Re-ID.
- We conduct extensive experiments on three datasets and show that BIN achieves competitive results in comparison with state-of-the-art methods. HOP and IBAM are also verified that each enhances accuracy.

2 Related Works

This section mainly discusses part-based and attention-based Re-ID, which are strongly related to our method.

2.1 Part-based Re-ID

Part-based methods focus on learning local parts information for region-level embeddings of person. It can be divided into two groups, as mentioned in Section 1. In the single branch methods, considering that methods slicing the last feature map horizontally into a small fixed number may not be robust for challenges like low resolution, viewpoint variation, HPM [4] explore a coarse-to-fine pyramid model to discover sufficient part descriptors of a given person. OSNet [5] achieves multi-scale feature learning by designing a omni-scale residual block. Multiple branch methods are proposed to model multiple information such as fine-grained features, pose information in different branches. It is proved that integrating the local and global features can promote the results. In CA^3Net [6], appearance network consisting of three branches is designed to extract global, horizontal human parts and vertical human parts features. In order to overcome

the misaligned problem, pose information is utilized in FEN [7] to match the feature from global and local body parts.

Different from previous part-based methods, we crop the feature maps into overlapped patches to learn local features as well as preserve essential information.

2.2 Attention-based Re-ID

The attention mechanism can enhance features, which helps to locate meaningful regions. Mancs [8] emphasize discriminative features by the proposed fully attention block (FAB). HA-CNN [9] extracts features by jointly learning hard region-level and soft pixel-level attention. Person attributes like gender, handbag can guide attention mechanisms to find meaningful information. In AANet [10], latent attribute regions are located by combining class sensitive activation regions from each attribute in attribute detection task. A^3M [11] proposes an attribute-category reciprocal attention module to leverage attribute information, and it is helpful to select key features for Re-ID.

Previous methods strengthen representational capability by utilizing information from single branch. However, we propose the IBAM to help BIN generate more discriminative features by combining information from different branches.

3 Branch Interaction Network (BIN)

In this section, we first describe the overall architecture of Branch Interaction Network (BIN). Then the proposed Horizontal Overlapped Pooling (HOP) is discussed, followed by a novel attention module named Inter-Branch Attention Module (IBAM). Finally, we discuss the relations between the proposed modules and some existed methods.

3.1 Overview

As is shown in Fig. 2, the BIN is a multi-branch network, including a base network and three independent branches. ResNet-50 [12] is applied for our feature extraction backbone. The base network consists of previous layers before conv4_2, which is capable of generating shared low-level visual features. Specifically, three branches are directly borrowed from subsequent layers after conv4_1, namely Stripe 1 Branch (S1B), Stripe 2 Branch (S2B), Stripe 3 Branch (S3B) based on the number of stripes. S1B performs the global-level person re-identification task, while S2B and S3B perform part-level and global-level feature learning. In S2B and S3B, we remove the last spatial down-sampling operation to enrich the granularity. As a result, feature tensors $\boldsymbol{T}_1$,$\boldsymbol{T}_2$,$\boldsymbol{T}_3$, the output of conv5 from S1B, S2B and S3B have different spatial sizes. In order to integrate multi-branch features, we inject IBAM on the outputs of conv4 to exploit complementary information across branches. Refined feature maps are fed into the following layers.

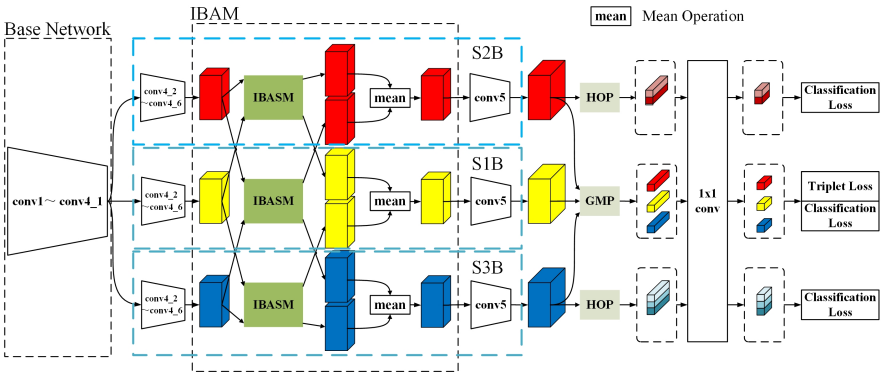

Fig. 2. The overall architecture of the proposed BIN. BIN contains a base network and three independent branches, i.e., S1B, S2B, S3B. IBAM is added after conv4 to capture complementary information among branches. BIN extracts global features by employing GMP on three branches and learns local features by applying HOP on S2B and S3B. The whole network is trained with triplet loss and classification loss.

With the Global Max Pooling (GMP), BIN generates global feature representations $\boldsymbol{g}_i(i = 1, 2, 3)$ for each branch. A parameter shared 1x1 convolution layer, followed with a batch normalization layer and ReLU layer, is applied to reduce the dimension from 2048-dim $\boldsymbol{g}_i(i = 1, 2, 3)$ to 256-dim $\boldsymbol{u}_i(i = 1, 2, 3)$. Finally, each $\boldsymbol{u}_i(i = 1, 2, 3)$ is trained with triplet loss [13] and classification loss. Specifically, triplet loss on global features can be formulated as :

$$L_{tri}^{g} = \sum_{i=1}^{N_g} \left(\frac{1}{N_t} \sum_{j=1}^{N_t} \left[m + \left\| \boldsymbol{u}_i{}^{(j)} - \boldsymbol{u}_i{}^{(j+)} \right\|_2 - \left\| \boldsymbol{u}_i{}^{(j)} - \boldsymbol{u}_i{}^{(j-)} \right\|_2 \right]_+ \right) \tag{1}$$

where N_g and N_t are the numbers of global features and sampled triplets, $\boldsymbol{u}_i{}^{(j)}$, $\boldsymbol{u}_i{}^{(j+)}$, $\boldsymbol{u}_i{}^{(j-)}$ are the feature $\boldsymbol{u}_i$ extracted from anchor, positive and negative samples in j-th triplet respectively, m is the margin parameter, and $[\cdot]_+$ denotes $max(\cdot, 0)$. Classification loss on global features can be formulated as :

$$L_{cls}^{g} = \sum_{i=1}^{N_g} \left(-\frac{1}{N} \sum_{j=1}^{N} log \frac{\exp(((\boldsymbol{W}^i)_{y_j})^T \boldsymbol{u}_i)}{\sum_{k=1}^{C} \exp(((\boldsymbol{W}^i)_k)^T \boldsymbol{u}_i)} \right) \tag{2}$$

where N, C are the number of input images and identities, y_j is the ground truth of j-th input image. $(\boldsymbol{W}^i)_k$ denotes the weight matrix for k-th identity in the fully connected layer whose input is $\boldsymbol{u}_i$.

With our proposed HOP, BIN partitions $\boldsymbol{T}_i(i = 2, 3)$ into 2 and 3 horizontal stripes in S2B and S3B, and pools these stripes to generate column feature vectors i.e., $\boldsymbol{p}_m^n$, where m, n refer to the m-th stripe in Stripe n Branch. The dimension of $\boldsymbol{p}_m^n$ is also reduced to 256 by the 1x1 convolution layer. Finally,

dimension-reduced features $\boldsymbol{v}_m^n$ are only trained with classification loss. Classification loss on local features are formulated as :

$$L_{cls}^l = \sum_{n=2}^{N_b} \sum_{m=1}^{n} \left(-\frac{1}{N} \sum_{j=1}^{N} log \frac{\exp(((\boldsymbol{W}_m^n)_{y_j})^T \boldsymbol{v}_m^n)}{\sum_{k=1}^{C} \exp(((\boldsymbol{W}_m^n)_k)^T \boldsymbol{v}_m^n)} \right) \tag{3}$$

where N_b is the number of branches, $(\boldsymbol{W}_m^n)_k$ is the weight matrix for k-th identity in the fully connected layer whose input is $\boldsymbol{v}_m^n$. And the final loss is defined as following:

$$L = \frac{1}{N_{tri}} L_{tri}^g + \lambda \frac{1}{N_{cls}} \left(L_{cls}^g + L_{cls}^l \right) \tag{4}$$

where N_{tri} and N_{cls} are the numbers of features trained with triplet loss and classification loss, λ is a trade-off parameter. Specifically, we set λ to 2 in the following experiments.

3.2 Horizontal Overlapped Pooling (HOP)

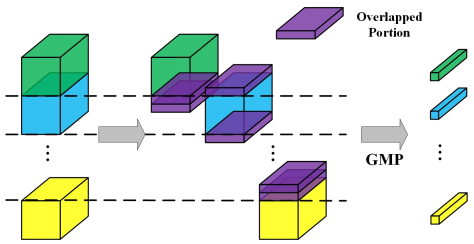

Fig. 3. Illustration of HOP. Firstly, uniform partition is performed on the feature map. Then, each stripe is padded with two overlapped portions. Finally, we pool them by GMP.

Given an input feature map $\boldsymbol{F} \in \mathbb{R}^{C \times H \times W}$, locating meaningful parts by original uniform partition may cause within-part inconsistency, and introduce many outliers near division lines. HOP is proposed to solve this problem by making meaningful regions covered in adjacent parts. It has two parameters, which are l and k. l is the total height of overlapped areas in one stripe. k is the number of partitions. When k=1, we remain the global information. In BIN, we keep the k=2 in S2B and k=3 in S3B.

The HOP is illuminated in Fig. 3. Firstly, we perform a uniform partition on the feature map horizontally. With the aim of devoting equal attention to each stripe, parts on the top or bottom are extended in one direction while others are extended in two directions to keep the same spatial size. An overlapped portion is a smaller 3D tensor whose size is $C \times h \times W$, where h refers to its height. As a result, $l = 2h$. However, it is obvious that l can be an odd number when k=2. To

make HOP universal, we require that l must be an even number. Finally, each horizontal stripe is pooled by GMP to generate a part-level vector.

3.3 Inter-Branch Attention Module (IBAM)

Features extracted from different branches help together to boost the feature representative capability. In order to model spatial contextual dependencies between branches, an IBAM is applied, as shown in Fig. 2. Features from paired branches are fed into an Inter-Branch Attention Submodule (IBASM), and output paired refined features. BIN has three branches, and form $C_3^2 = 3$ combinations when we choose paired branches. As a result, each branch is selected twice and has two refined outputs which build contextual dependencies between features from various branches. A mean operation on these two outputs is performed to update the original features.

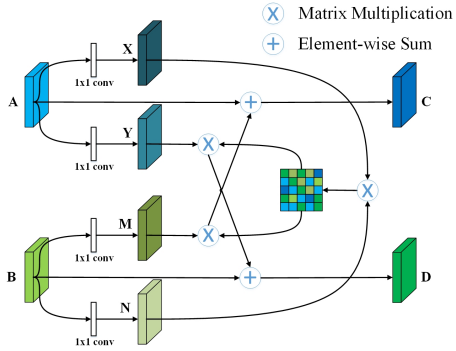

Fig. 4. The detail architecture of Inter-Branch Attention Submodule (IBASM).

Specifically, IBASM captures the similarity between input paired feature maps and aggregate together to produce more discriminative features. Fig. 4 depicts the detail structure of IBASM. Given two feature maps $\boldsymbol{A} \in \mathbb{R}^{C\times H\times W}$, $\boldsymbol{B} \in \mathbb{R}^{C\times H\times W}$ from different branches, a 1x1 convolution layer is employed to generate four new feature maps $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{M}$, and $\boldsymbol{N}$, where $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{N} \in \mathbb{R}^{\frac{C}{8}\times H\times W}$. These four feature maps are reshaped to $\mathbb{R}^{\frac{C}{8}\times L}$, where $L = H \times W$ is the number of feature locations. Pixel-wise similarity is calculated by matrix multiplication between transposed $\boldsymbol{X}$ and $\boldsymbol{N}$, and then normalized to obtain the spatial attention map $\boldsymbol{S} \in \mathbb{R}^{L\times L}$, as shown below:

$$S_{i,j} = \frac{\exp\left(m_{i,j}\right)}{\sum_{i=1}^{L}\exp\left(m_{i,j}\right)}, m_{ij} = \boldsymbol{X}_i^T \boldsymbol{N}_j \tag{5}$$

where $\boldsymbol{X}_i$, $\boldsymbol{N}_j$ denote the i^{th} and j^{th} spatial features of $\boldsymbol{X}$ and $\boldsymbol{N}$ respectively.

To calculate the output $\boldsymbol{C}$, BIN first predicts $\boldsymbol{A}$ by exploiting the information from input $\boldsymbol{B}$ based on spatial attention map $\boldsymbol{S}$. The prediction is reshaped to $\mathbb{R}^{C\times H\times W}$, then BIN performs an element-wise sum between weighted prediction and original $\boldsymbol{A}$. The output $\boldsymbol{C}$ denotes refined $\boldsymbol{A}$ guided by $\boldsymbol{B}$, which is defined as :

$$\boldsymbol{C}_j = \gamma_1 \sum_{i=1}^{L} S^T{}_{i,j} \boldsymbol{M}_i + \boldsymbol{A}_j \tag{6}$$

where γ_1 is a learnable weight which is initialized as 0. The output $\boldsymbol{D}$ denotes refined $\boldsymbol{B}$ guided by $\boldsymbol{A}$, which is defined as :

$$\boldsymbol{D}_j = \gamma_2 \sum_{i=1}^{L} S_{i,j} \boldsymbol{Y}_i + \boldsymbol{B}_j \tag{7}$$

As a result, IBASM keeps the size unchanged. With this property, IBAM and IBASM can be incorporated into any existing multi-branch architecture.

Armed with our proposed IBASM, spatial contextual, multi-level dependencies across branches are well established, and the shared information in multi-granularity features are utilized in higher layers.

3.4 Discussion

To highlight the difference between our proposed modules and other related methods, we provide a brief discussion on the relations between them.

Relations between HOP and OBM OBM [14] propose a multiple overlapping blocks structure to pool features from overlapping regions. OBM requires pyramid-like horizontal partitions. However, HOP performs on a single scale, which is a lightweight method in the training procedure for its relatively fewer fully connected layers.

Relations between HOP and RPP RPP [1] is proposed to relocate outliers incurred by uniform partition to the parts they are closest to. In other words, RPP aims to address problems brought by "hard" partition. However, HOP focuses on keeping the balance between learning fine-grained features and extracting meaningful region information. HOP is a new kind of "hard" partition, and uniform partition can be seen a special case of HOP when l=0.

Relations between IBASM and non-local block In some ways, IBASM can be regarded as a variation of the non-local block [15]. IBASM differs from non-local block in: (1) IBASM takes two input features while non-local block takes one input feature. IBASM performs non-local operation on two features. This modification helps model refine one input feature with the consideration of the other input feature. (2) IBASM produces two output features correspond to two refined input features by the guidance from each other. "Encoder-decoder attention" layers from [16] and pairwise non-local operation from [17] both take two input features to compute non-local operation and produce one output feature corresponds to one refined input feature guided by the other input feature.

Relations between IBAM and dual attention block We make the comparison with IBAM and dual attention block [18] on image input. Given an image input, feature sequence is formed by rearranging feature vectors by locations. Dual attention block contains inter-sequence attention and intra-sequence attention. They have similarities because they both seek to find relations between branches or sequences. IBAM and dual attention block differ in: For the same image, intra-sequence attention refined itself by focusing on context-aware information in the single scale, and inter-sequence attention generates aligned counterpart by focusing on consistent regions from the opposite image. However, IBAM is designed to model spatial contextual dependencies from features with multiple granularity from the same image. IBAM has the capability to aggregate complementary information in multiple granularity.

Relations between IBAM and PS-MCNN IBAM has some similarities with PS-MCNN [19], because both are designed to interact with different branches. However, our IBAM is different from PS-MCNN in three aspects. (1) IBAM aims to build relations between different branches with various granularities while PS-MCNN focuses on building relations between different branches with various attribute groups. (2) IBAM builds interactions between all branches by modeling the relations of paired branches while PS-MCNN introduces a new Shared Network (SNet) to learn shared information for all branches. Besides, IBAM considers the spatial information in the process of interaction, which is ignored by PS-MCNN. (3) IBAM is a module that can be easily embedded into any multi-branch network architecture, while PS-MCNN is a network designed for building interactions among different branches with various attribute groups specifically. Our IBAM is more general than PS-MCNN.

4 Experiments

We conduct experiments on three Re-ID datasets: DukeMTMC-reID [20], Market-1501 [21] and CUHK03 [22]. First, we compare the retrieval accuracy of BIN with state-of-the-art methods on these three datasets. Then, we carry out ablation studies on DukeMTMC-reID dataset to verify the effectiveness of each component.

4.1 Datasets and Evaluation Protocol

DukeMTMC-reID This dataset is a subset of the DukeMTMC for Re-ID. It contains 36,411 images of 1,812 persons from 8 cameras. There are 1,404 identities appear in more than two cameras, and the other 408 identities appear in only one camera, which are regarded as distractors. There are 16,522 images of 702 persons in the training set, and the rest 702 persons are included in the testing set, which consists of 2,228 query images and 17,661 gallery images.

Market-1501 This dataset includes 32,668 images of 1,501 identities detected by the Deformable Part Model (DPM) detector from 6 cameras. Specifically, the training set contains 12,936 images of 751 persons, and the testing set includes 3,368 query images and 19,732 gallery images of 750 persons.

CUHK03 This dataset consists of 14,097 images of 1,467 identifies captured by 6 cameras. It provides two types of annotations, which are manually labeled pedestrian bounding boxes and DPM-detected bounding boxes. We perform experiments on both of them.

Evaluation Protocol In our experiments, we adopt standard Re-ID metrics: Cumulative Matching Characteristic (CMC) at rank-1, and the mean Average Precision (mAP) on all candidate datasets. All the experiments are conducted under the single query model.

4.2 Implementation Details

The implementation of our proposed BIN is based on the Pytorch framework. We initialize parameters of the BIN with the weights of ResNet-50 [12] pretrained on ImageNet.

During training, the input images are resized to 384×128 to keep more detailed information. We deploy random horizontal flip, normalization and random erasing [23] for data augmentation. A mini-batch is randomly sampled with 4 identities, and each identity contains 4 images. The margin in the triplet loss is 1.2 in all our experiments. The model is trained for 500 epochs. Adam optimizer is utilized to update the weight parameters with weight decay 5^{-4}. The initial learning rate is 2^{-4}, then decayed to 2^{-5}, 2^{-6} after 320, 380 epochs.

During testing, images are resized to 384×128 and normalized before fed into the network. Global features from all branches and local features from horizontally sliced parts are concatenated as the final pedestrian representation.

4.3 Comparison with State-of-the-art Methods

BIN is compared with 14 existing state-of-the-art methods on three datasets: DukeMTMC-reID, Market-1501 and CUHK03 in Tab. 1. The compared methods are categorized into single branch methods (S), multi-branch methods (M) and attention-based methods regardless of the number of branches (A). Results in detail are discussed as follows.

DukeMTMC-reID The proposed BIN achieves 89.36% Rank-1 accuracy and 79.60% mAP, which outperforms all published methods by a large margin. Note that : (1) The gaps between our method and single branch methods indicate that multi-branch structure is necessary: about 0.76% and 6.10% improvement in Rank-1 accuracy and mAP respectively. These methods focus on global information, local details or both of them in the single branch, which is insufficient for Re-ID. In contrast with single branch methods, our method can capture robust features in multiple granularities from various branches. (2) Although multi-branch methods integrate complementary information into final pedestrian representations, e.g., AANet [10] integrates key attribute information in a unified framework. BIN surpasses them, exceeding the MGN [2], which achieves the best results in this category, by 0.66% in Rank-1 accuracy and 1.20% in mAP. We argue that these methods neglect the interaction among branches.

Table 1. Comparison with state-of-the-art methods on three mainstream datasets. Red and Blue indicate our results and the best results of previous methods respectively. Best results of all methods are marked in bold. "-" denotes not available, "RK" denotes re-ranking operation.

	Methods	DukeMTMC-reID		Market-1501		CUHK03			
						Labeled		Detected	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
S	MLFN [24] (CVPR2018)	81.00	62.80	90.00	74.30	54.70	49.20	52.80	47.80
	PCB+RPP [1] (ECCV2018)	83.30	69.20	93.80	81.60	-	-	63.70	57.50
	HPM [4] (AAAI2019)	86.60	74.30	94.20	82.70	-	-	63.90	57.50
	OSNet [5] (ICCV2019)	88.60	73.50	94.80	84.90	-	-	72.30	67.80
M	PSE [25] (CVPR2018)	79.80	62.00	87.70	69.00	-	-	-	-
	HA-CNN [9] (CVPR2018)	80.50	63.80	91.20	75.70	44.40	41.00	41.70	38.60
	CA^{3}Net [6] (ACM MM2018)	84.60	70.20	93.20	80.00	-	-	-	-
	CAMA [26] (CVPR2019)	85.80	72.90	94.70	84.50	70.10	66.50	66.60	64.20
	MGN [2] (ACM MM2018)	88.70	78.40	**95.70**	86.90	68.00	67.40	66.80	66.00
A	MGCAM [27] (CVPR2018)	-	-	83.79	74.33	50.14	50.21	46.71	46.87
	DuATM [18] (CVPR2018)	81.82	64.58	91.42	76.62	-	-	-	-
	Mancs [8] (ECCV2018)	84.90	71.80	93.10	82.30	69.00	63.90	65.50	60.50
	AANet-50 [10] (CVPR2019)	86.42	72.56	93.89	82.45	-	-	-	-
	CASN [28] (CVPR2019)	87.70	73.70	94.40	82.80	73.70	68.00	71.50	64.40
BIN		**89.36**	**79.60**	94.80	**87.27**	**74.29**	**72.43**	**72.57**	**69.83**
BIN (RK)		**92.06**	**90.50**	95.69	**94.07**	**83.66**	**81.71**	**79.43**	**81.66**

On the contrary, our method remains the strong relations between different branches, which is proved efficient. (3) Compared to attention-based methods, our methods boost CASN [28] by 1.66% in Rank-1 accuracy and 5.90% in mAP. Most attention-based methods build intra-branch contextual dependencies in the spatial or channel dimension. However, we model inter-branch non-local dependencies, which is more competent. With the help of re-ranking [29], we achieve a top result of 92.06% rank-1 accuracy and 90.50% mAP, which is a giant breakthrough.

Some visual examples of BIN on DukeMTMC-reID dataset are illustrated in Fig. 5. Given a query pedestrian image, BIN can retrieve the same person images in low-resolution, view angle variation and occlusion, which shows its robustness for most exiting challenges.

Market-1501 We report the 94.80% Rank-1 accuracy and 87.27% mAP, which significantly surpass most of the recent start-of-the-art methods. Although the Rank-1 accuracy of BIN is slightly lower than MGN(95.7%), BIN achieves the top mAP, outperforms the MGN by a large margin of 0.37%.

CUHK03 As is illustrated in Table 1, we compare the BIN against other methods on the CUHK03 dataset in two types of annotation settings, i.e., labeled and detected. BIN still achieves the best result of Rank-1 accuracy 74.29%, mAP 72.43% on the labeled setting, and Rank-1 accuracy 72.57%, mAP 69.83% on the detected setting, which surpasses the 1st best-compared method by Rank-1/mAP=0.59%/4.43% and 1.07%/5.43% respectively.

Fig. 5. Top-6 ranking list for given query images on DukeMTMC-reID dataset from BIN. Correct and false matches are highlighted by green and red borders respectively.

4.4 Ablation Study

To investigate the effectiveness of each component in our proposed BIN, we conduct a series of ablation experiments on DukeMTMC-reID dataset.

Multi-branch Structure Tab. 2 compares single branch and multi-branch models. For single branch models, with the increase of horizontal stripes, we extract more detail-rich representations, and the accuracy is increased as well. S1B+S2B(l=0) means multi-branch network architecture with two branches of S1B and S2B, and so as S1B+S3B(l=0), S1B+S2B(l=0)+S3B(l=0). The multi-branch structure is superior to each composed single branch and gains further improvements, e.g., S1B+S2B outperform S1B and S2B in Rank-1/mAP by 8.39%/12.90% and 1.89%/2.66%. However, with the increase of k in single branch, the improvement seems to be marginal but enlarge the model parameters, e.g., S2B(l=0) outperforms S1B in mAP by 10.24% but S3B(l=0) outperforms S2B(l=0) in mAP by 0.9%. With the increase of the number of branches, the performance shows the same trend, e.g., S1B+S3B(l=0) outperforms S1B in mAP by 13.20% but S1B+S2B(l=0)+S3B(l=0) outperforms S1B+S2B(l=0) in mAP by 2.61%. As a result, we adopt S1B+S2B(l=0)+S3B(l=0) as the multi-branch model for the following experiments.

Effectiveness of Triplet Loss Our proposed BIN is trained with triplet loss and classification loss. Triplet loss plays a vital role for surpervising the whole network. Tab. 3 shows the effectiveness of triplet loss.

Effectiveness of HOP We define l of HOP in S2B as l_2 and l of HOP in S3B as l_3. Fig. 6 compares the HOP operations with different l. As is shown in Fig. 6a, when l_3 is set to 0, there is an improvement in accuracy when l_2 is increased, indicating the effectiveness of HOP. However, the retrieval accuracy drop dramatically when l_2 is further increased. We also increase l_3 when l_2 is set to 0 in Fig. 6b. They keep the same trend in performance, i.e., rise first, then decrease. With the increase of l in single branch, HOP helps to cover meaningful regions between adjacent parts, but will damage representational capability in

Table 2. Comparison of single branch and multi-branch models. l means the parameter l of HOP in corresponding branch.

Model	Rank-1	mAP
S1B	78.64	61.62
S2B(l=0)	85.14	71.86
S3B(l=0)	86.09	72.76
S1B+S2B(l=0)	87.03	74.52
S1B+S3B(l=0)	87.15	74.82
S1B+S2B(l=0)+S3B(l=0)	**88.15**	**77.13**

Table 3. Evaluation of the effectiveness of triplet loss. "Triplet" refers to the triplet loss.

Model	Rank-1	mAP
S1B+S2B(l=0)+S3B(l=0)	**88.15**	**77.13**
S1B+S2B(l=0)+S3B(l=0) w/o Triplet	85.01	71.45

Table 4. Comparison of different l in different branches.

Model	Rank-1	mAP
S1B+S2B(l=0)+S3B(l=0)	88.15	77.13
S1B+S2B(l=2)+S3B(l=2)	**88.87**	**77.77**
S1B+S2B(l=2)+S3B(l=4)	88.69	77.28
S1B+S2B(l=4)+S3B(l=2)	88.73	77.57
S1B+S2B(l=4)+S3B(l=4)	88.46	77.19

Table 5. Comparison on adding IBAM in different positions.

Model	Rank-1	mAP
S1B+S2B(l=2)+S3B(l=2)	88.87	77.77
S1B+S2B(l=2)+S3B(l=2))+IBAM(conv4) (Our proposed BIN)	**89.36**	**79.60**
S1B+S2B(l=2)+S3B(l=2)+IBAM(conv5)	89.09	79.20

local information slightly. As a result, a proper l is needed. We find that l_2 and l_3 both equal 2 can achieve the best results, as shown in Tab. 4.

Effectiveness of IBAM Tab. 5 reports the effectiveness of IBAM. We add the IBAM on the output of conv4 and conv5 to compare the performance. Specifically, since the inputs of IBAM need to have the same size, we remove the last spatial down-sample operation in the conv5 layer of S1B when adding the IBAM following ResNet conv5. Multi-branch models with IBAM in various positions lead a significant performance improvement. The growth of the IBAM after conv4 is greater than after conv5. Although PCB finds that removing the last spatial down-sample operation in ResNet increases person retrieval accuracy, we argue that the remaining down-sample operation in S1B will produce complementary features. To better understand the IBAM used in our BIN, we visualize the activation maps extracted from the output of each branch in Fig. 7. First, BIN w/o IBAM is not sufficient to capture robust information about input image, e.g., shins in the first input image are ignored. Comparing the activation maps, we see that BIN can extract more discriminative features in each branch with the help of IBAM. Second, the activation maps from S1B in BIN w/o IBAM mainly cover the main body of pedestrians but ignore some detailed regions, e.g., arms in the second image. With the help of IBAM, S1B from BIN focus on the main body and local parts because S1B interacts with S2B and S3B. Third, the distribution of the activation maps from S2B and S3B in BIN w/o IBAM is too scattered, which means BIN w/o IBAM fails in modeling consecutive local areas in S2B and S3B. With the help of IBAM, S2B and S3B can concentrate

Fig. 6. Parameter analysis for l in S2B and S3B. (a) Rank-1 and mAP changes with l_2 while l_3 is set to 0. (b) Rank-1 and mAP changes with l_3 while l_2 is set to 0.

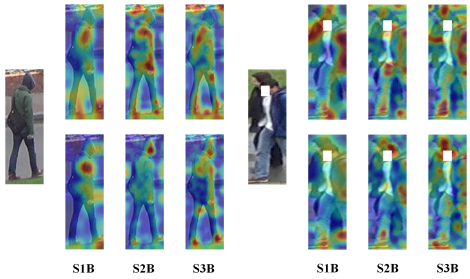

Fig. 7. Visualizaion of activation maps extracted from each branch. For each spatial position, the maximum of all channels is assigned for this part in activation maps. For each input image, the activation maps in first row are generated from our proposed BIN while the activation maps in second row are produced from BIN w/o IBAM.

more on meaningful local regions because of complementary information from S1B.

5 Conclusions

This paper proposes the BIN, a multi-branch network for Re-ID. The multi-branch structure is necessary to capture coarse-to-fine information. HOP is an improvement on the traditional uniform partition and GMP, while IBAM is an extension of attention mechanism. Each component is verified in boosting the robustness of BIN. Extensive experiments on three datasets demonstrate that BIN achieves the state-of-the-art performance. In the future, we will explore the correlation between inter-branch attention mechanism and intra-branch attention mechanism.

Acknowledgments.

This paper was supported by National Key R&D Program of China (2019YFC1521204).

References

1. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 480–496
2. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference, ACM (2018) 274–282
3. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press (2017) 2194–2200
4. Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8295–8302
5. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. arXiv preprint arXiv:1905.00953 (2019)
6. Liu, J., Zha, Z.J., Xie, H., Xiong, Z., Zhang, Y.: Ca 3 net: Contextual-attentional attribute-appearance network for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference, ACM (2018) 737–745
7. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3960–3969
8. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 365–381
9. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2285–2294
10. Tay, C.P., Roy, S., Yap, K.H.: Aanet: Attribute attention network for person re-identifications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7134–7143
11. Han, K., Guo, J., Zhang, C., Zhu, M.: Attribute-aware attention model for fine-grained representation learning. In: 2018 ACM Multimedia Conference on Multimedia Conference, ACM (2018) 2040–2048
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
14. Chen, Y., Zhao, C., Sun, T.: Single image based metric learning via overlapping blocks model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019) 0–0
15. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7794–7803
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. (2017) 5998–6008
17. Fu, Z., Chen, Y., Yong, H., Jiang, R., Zhang, L., Hua, X.S.: Foreground gating and background refining network for surveillance object detection. IEEE Transactions on Image Processing **28** (2019) 6077–6090

18. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: 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. (2018) 5363–5372
19. Cao, J., Li, Y., Zhang, Z.: Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 4290–4299
20. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, Springer (2016) 17–35
21. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. (2015) 1116–1124
22. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 152–159
23. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI. (2020) 13001–13008
24. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2109–2118
25. Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: 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. (2018) 420–429
26. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1389–1398
27. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1179–1188
28. Zheng, M., Karanam, S., Wu, Z., Radke, R.J.: Re-identification with consistent attentive siamese networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5735–5744
29. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1318–1327