Multi-branch network with hierarchical bilinear pooling for person reidentification

Wenyu Wei | Wenzhong Yang | Enguang Zuo | Qiuru Ren | Qiuchang Chen

College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, China

Correspondence
Wenzhong Yang, College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, China.
Email: ywz_xy@163.com

Funding information
National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data, Grant/Award Number: XJ201810101; National Natural Science Foundation of China, Grant/Award Number: U1603115; Unassigned; National key R&D plan project, Grant/Award Number: 2017 YF C08207023

Abstract
Because of issues such as viewpoint changes, posture variations, and background cluttering, the task of person reidentification (Re-ID) remains challenging. The model of combining global features and part features has been widely used in person Re-ID technology in recent years, but these efforts ignored feature interaction between the convolutional layers and thus lost detailed information conducive to identifying pedestrians under different cameras. To achieve interaction between hierarchical features, a multibranch network with hierarchical bilinear pooling (MBN-HBP) is proposed. The network consists of a global branch, a part-level branch, and a hierarchical bilinear pooling (HBP) branch. The person features extracted by the network include not only global and part-level features but also detailed HBP features. The final feature representation will be more robust to deal with the complex surveillance environment. By conducting comprehensive experiments, competitive performance on the Market-1501, DukeMTMC-Re-ID, and CUHK03 datasets is obtained.

1 | INTRODUCTION

Person reidentification (Re-ID) has always been a research hotspot in intelligent video surveillance, and this technology can help people find a person with a specified identity across cameras. Therefore, this is of great significance to the police in carrying out tasks such as tracking criminal suspects and finding missing persons. The task of person Re-ID has many inherent challenges, such as viewpoint changes, human pose variations, and lighting changes, making it challenging to match a person with the same identity under other cameras.

The emergence of deep learning provides a new research direction for the research of person Re-ID. In the early development of person Re-ID based on deep learning, some methods [1, 2] used a combination of deep features and handcrafted features to represent the person invariant features. Although the performance of these methods is better than traditional handcrafted features [3] and metric learning [4] approaches, they are still far from reaching the requirements of practical applications.

In recent years, some works [5–8] used person part features to represent robust invariant features. Some part-based models [5, 6] used body keypoints positioning to extract local features of the person, which could introduce errors caused by additional algorithms. Several works [7, 8] exploited the strategy of partitioning the feature map horizontally into several stripes to extract the local features. Even so, posture changes, part occlusions, inaccurate detection, and other issues might cause pedestrian misalignment problems. Some works [9–11] utilized multibranch networks to extract robust person feature descriptors. Therefore, we adopt a multibranch network framework as our Re-ID model.

Before the development of deep learning, some researchers [12] proposed bilinear models. In this way, Lin et al. [13] proposed bilinear pooling and applied it to fine-grained image classification. As shown in Figure 1, the structure combines the output of the two feature extractors by outer product and then performs a pooling operation. This is conducive to capturing the correlation between paired feature channels and modelling part-feature interactions, which is very useful for fine-grained image classification. Afterwards, some studies [14–16] improved the bilinear pooling for fine-grained classification. Considering the interaction between features from different layers, Yu et al. [16] proposed a hierarchical bilinear pooling (HBP) model. In their model, the researchers regarded different convolutional layer feature maps as responses to...
different local properties instead of explicitly defined localizing object parts. Therefore, HBP is used to achieve feature interaction between different convolutional layers. Inspired by the HBP model, we believe that HBP can also realize the interaction of the local features of pedestrians between different convolutional layers, making the information more detailed and less lost, thereby significantly improving the discrimination of features. Therefore, considering the characteristics of the person Re-ID task, we embed the HBP model as one of the branches into a multibranch network framework that can be used for person Re-ID. The features extracted by this model include global features, part-level features, and HBP features with supplementary information. The contributions of this work can be summarized as follows:

1. We propose to apply bilinear pooling to the person Re-ID task, which is one of the few attempts to do so with person Re-ID technology.
2. We embed HBP into a multibranch network framework for Re-ID tasks and design a multibranch network with hierarchical bilinear pooling (MBN-HBP) to learn discriminative person features. To our knowledge, this is the first time that HBP has been used for person Re-ID, and it achieves good results.
3. We conduct a comprehensive evaluation of our model on three Re-ID datasets, and experiments verify that our method is effective.

2 | RELATED WORK

2.1 | Person reidentification

The two main processes of person Re-ID are feature learning and similarity measurement. Early traditional approaches [17, 18] used image colour, texture, and contour to express person representation and designed powerful measurement algorithms [4, 19] to match pedestrian images. However, these works cannot cope with issues such as posture variations, viewpoint changes, and background cluttering in cross-camera images, and thus have not been further developed and promoted. The rise of convolutional neural networks (CNNs) has promoted computer vision technologies such as person Re-ID, and person Re-ID technology based on deep learning was born. The person representation extracted by CNN can cope with the complex camera environment, which is very beneficial for cross-camera pedestrian matching. Simultaneously, deep learning can supervise model training through a variety of loss functions, and the loss function has the role of feature measurement. Some works [20–22] regarded Re-ID as a classification problem, so they used the softmax loss function to optimize the CNN model. There are still some works [23, 24] that regard person Re-ID as a similarity measurement problem, so triplet loss is used to optimize the model. Inspired by multitask learning, there are now some models [25–27] that combine softmax loss and triplet loss, which has been proven to be a very effective method for person Re-ID tasks. Therefore, the proposed approach applies a combination of two loss functions to optimize our multibranch network model.

The deep learning method based on feature representation is categorized into three types: local feature methods, global feature methods, and their combined features [28]. Local part information from humans is often regarded as crucial detailed information that can distinguish various identities. Therefore, an increasing number of researchers have designed rich part-based features as supplements to global features to improve feature robustness. Some methods [5, 6] obtained the local feature representation by locating the semantic parts of the human body, but this most likely introduces additional algorithmic errors. Sun et al. [7] proposed a part-based convolutional baseline (PCB) that extracts part features to divide the features map horizontally from the backbone network into six part-level features. Later, some solutions [8, 29–31] used similar horizontal partitions to obtain local feature representations. Nevertheless, the part-level models ignored the connection between each part and did not achieve alignment well, thus limiting the performance improvement of the model.

To address these problems, Bai et al. [10] used long short-term memory to capture the contextual information between each part to strengthen the connection between the partitions. We regard HBP features as global and local part-level features supplementary to improve the performance of the model.

In addition to the above works, there are some attention mechanism models for person Re-ID. These efforts focus on extracting regions containing highly discriminative features while ignoring other regions with little or no discriminative capability [32]. Yang et al. [33] designed an intraattention and

---

**FIGURE 1** Schematic diagram of bilinear pooling model conducive to capturing the correlation between paired feature channels and modelling part-feature interactions.
interattention network. The intraattention network learns the optimal global and local feature representations and attention mapping. The interattention network automatically learns the optimal weights to achieve the optimal fusion of the intraattention network output. Bryan et al. [34] proposed the idea of combining non-local operations with second-order statistics in a CNN and designed a Second-Order Non-local attention model. Previous attention models mainly focussed on the design of low-level attention mechanisms. Therefore, Chen et al. [35] proposed a high-level attention module to model and utilize complex high-level statistical information to capture the nuances between different pedestrians. To make full use of pedestrian attribute information, Tay et al. [36] proposed the attribute attention network (AANet), which combines identity classification, body part detection, and human attributes and jointly learns features space with strong discriminative abilities. Based on multiscale features, Qian et al. [37] and Li et al. [38] utilized a leader-based attention learning layer and a self-attention model to learn robust and distinguishable pedestrian features, respectively. Unlike these attention mechanism models, the HBP feature in our model can realize the interaction between different convolutional layers to strengthen the distinguishing features to obtain robust pedestrian representations.

### 2.2 Bilinear pooling model

To suppress the impact of background cluttering and extract discriminative features, Lin et al. [13] proposed a bilinear pooling CNN model for fine-grained classification tasks. This model uses two parallel CNN models as feature extractors and then combines the two output features by the outer product and applies pooling to the fusion feature. Since then, some efforts [39–41] have made improvements to bilinear pooling and used it for other computer vision tasks. Considering the high dimensionality of bilinear pooling features, Gao et al. [39] proposed Compact Bilinear Pooling to reduce the feature dimension. This method has universal applicability in computer vision tasks. In the Visual Question Answering task, Yu et al. [40] designed a Multimodal Factorized Bilinear pooling model to fuse multimodal features efficiently. In the task of 3-D object recognition, Yu et al. [41] exploited the relationship between polynomial kernel and bilinear pooling and used the bilinear pooling to aggregate local convolution features to obtain an effective 3-D object representation. In the person Re-ID task, previous works [42, 43] have also introduced bilinear pooling models. Ustinova et al. [42] found that person Re-ID shares considerable similarity with fine-grained categorization and proposed a multiregion bilinear CNN model for person Re-ID. The model divides the image into several regions; each region uses two feature extractors to learning feature maps and then uses bilinear pooling to perform feature fusion in the same region. The part-aligned bilinear representations model proposed by Suh et al. [43] directly inputs the complete image into two feature extractors, and this a siamese network. These two extractors are the appearance map extractor and part map extractor. The features output by these two subnets are finally aggregated through bilinear pooling to generate the final pedestrian image descriptors. Different from their work, our model can learn robust person representations through bilinear pooling operations without two various feature extractors.

### 3 Proposed Framework

For the existing methods [16, 44], HBP is used only to capture the relationship between the convolutional layer features for the fine-grained image classification task. In the task of person Re-ID, this work is the first time that a HBP strategy has been used to make features more refined than global features. For the existing multibranch networks for person Re-ID [9–11], these structures extract only the global and local features of pedestrians and do not consider the more detailed HBP features. The HBP features can be regarded as enhanced global features, representing a more discriminative person feature filtered the background cluttering. In the MBN-HBP, we use a loss function strategy similar to [25, 45, 46] and other frameworks, that is, the combined use of softmax loss and triplet loss to train the deep model. The following is a detailed introduction of the network framework proposed in this article.

#### 3.1 Hierarchical bilinear pooling

Traditional bilinear pooling models often require two different feature extractors, which directly increase the complexity of the model. In contrast, HBP [16] requires only a single CNN structure to enhance the ability of representation. The main idea is to treat features output by different convolutional layers as feature maps to different parts, rather than explicitly defined part feature maps. Therefore, without increasing the feature extractor, a certain part of a deep network is layered, the features of each convolutional layer are extracted separately, and the two are merged through bilinear pooling. Finally, all the fused features are connected as the feature vector of the final output of the model. This approach effectively realizes the interlayer feature interaction and improves feature representation ability.

Suppose \( X \in \mathbb{R}^{d \times w \times h}, Y \in \mathbb{R}^{d \times w \times h} \) are feature maps from two different convolutional layers, and the traditional bilinear pooling [47] can be defined as

\[
\alpha_i = x^T W_i y
\]  

where \( x \) and \( y \) are respectively \( c \)-dimensional descriptors at a spatial location on \( X \) and \( Y \), \( W_i \in \mathbb{R}^{c \times c} \) is a projection matrix containing bias, and \( \alpha_i \) is the output of the bilinear model. Considering that \( W_i \) contains a large number of parameters that need to be learnt, Kim et al. [47] proposed factorized bilinear pooling (with fewer parameters for
regularization) to reduce the rank of the projection matrix to avoid learning more parameters.

In a factorized bilinear pooling model, \( W_i \) can be factorized into two low-rank matrices through matrix factorization [48]. Therefore, \( a_i \) can be redefined as

\[
a_i = x^T W_i y = x^T U_i V_i^T y = U_i^T x \odot V_i^T y
\]

(2)

where \( U_i \in \mathbb{R}^{c \times d} \) and \( V_i \in \mathbb{R}^{c \times d} \) are projection matrices of \( x \) and \( y \), respectively, and \( \odot \) is the Hadamard product. Thus, the output features \( \alpha \in \mathbb{R}^k \) of the model can be defined as

\[
\alpha = P^T (U^T x \odot V^T y)
\]

(3)

where \( P \in \mathbb{R}^{d \times k} \) is the classification matrix and \( k \) is the number of categories in the classification model.

In the HBP [16] model, the features from multiple different convolutional layers are input into the bilinear pooling model, and the two are fused through the Hadamard product. Finally, all the fusion features are connected as the final output vector. The formula can be written as

\[
\alpha_{HBP} = P^T \text{cat}(U^T x \odot V^T y, U^T x \odot S^T z, V^T y \odot S^T z, \ldots)
\]

(4)

where \( P \in \mathbb{R}^{d \times k} \) is the classification matrix, \( \text{cat}(\cdot, \cdot) \) represents the concatenation of vectors, \( x, y, z \) are feature maps from different convolution layers, \( U \in \mathbb{R}^{c \times d}, V \in \mathbb{R}^{c \times d}, S \in \mathbb{R}^{c \times d}, \ldots \) are projection matrices sized of \( c \times d \), and \( \alpha_{HBP} \in \mathbb{R}^k \) is the output of the HBP model.

In our proposed approach, person Re-ID is both a classification and a verification task. Therefore, our model is optimized by two loss functions (they will be introduced in Section 3.3). The HBP in our model can be redefined as follows:

\[
\beta_{HBP} = Q^T \text{cat}(U^T x \odot V^T y, U^T x \odot S^T z, V^T y \odot S^T z, \ldots)
\]

(5)

where \( Q \in \mathbb{R}^{d \times n} \) is the discriminant matrix used to discriminate pedestrians with different identities, \( n \) represents the number of identities in the training dataset, and \( \beta_{HBP} \in \mathbb{R}^n \) is the output of the HBP module. The HBP module can be used as a branch of our model to extract some pedestrian discriminative features as supplementary information for global and local features.

3.2 | Multibranch network

This general multibranch network architecture we propose is shown in Figure 2, which employs ResNet-50 as a backbone network and cancels downsampling operations in Layer4 of the backbone network. ResNet-50 is a deep residual CNN that has been used for all kinds of computer vision tasks such as classification, detection, and recognition. Because it can prevent the gradient from disappearing and obtain more reliable and representative features, we use ResNet-50 as the base network and combines the characteristics of the person Re-ID task to design a model that extracted the global features, local features, and bilinear pooling features of pedestrians. In CNNs, the input data is passed through layer by layer, which will inevitably cause some information to be lost. The bilinear pooling operation can make up for this shortcoming in ResNet-50, that is, by using the complementarity between the intermediate convolutional layers to enhance the representation, thereby reducing information loss.

To learn a comprehensive and robust representation of person features, the structure is divided into three branches, namely Branch-1, Branch-2, and Branch-3. These three branches extract global features, local features, and HBP features, respectively. Specifically, Branch-2 draws on PCB [7] that is part-level architecture to extract robust and effective local features. In this network, the feature maps output Layer3 of the backbone network are split into two courses for processing. The feature maps directly enter Branch-1 and Branch-2 through Layer4 in one course. Meanwhile, HBP is performed in Layer4 after the feature maps enter Branch-3 in another course. The feature vectors in Branch-1 and Branch-3 can be reduced from 2048-dimension to 512-dimension through 1 × 1 Conv, and feature vectors after pooling in Branch-2 can be decreased from 2048-dimension through 1 × 1 Conv to 256-dimension. Finally, we apply softmax loss and triplet loss to the features of each branch output at the same time to learn more comprehensive information. All the feature vectors after 1 × 1 Conv are concatenated together as the final feature representation of the original pedestrian image in the testing phase.

3.3 | Loss functions

The person Re-ID task is not regarded as a single classification task instead of multitask learning. Multitask learning can improve model learning the ability to learn discriminative features, so this model utilizes softmax loss for the multiclass classification task and triplet loss for metric learning.

Softmax cross-entropy loss is abbreviated as softmax loss, which is a combination of softmax and cross-entropy loss. The softmax function is widely used to calculate the probability distribution, and cross-entropy loss is used to calculate the similarity between two probability distributions. If the probability distribution in cross-entropy loss is calculated by softmax, then the cross-entropy loss at this time is softmax cross-entropy loss. The advantage of this loss combination is that it will make the numerical calculation more stable. For output \( f_i \) by the full connected layer, the softmax cross-entropy loss function can be described as
**Equation 6**

\[ L_{CE} = -\sum_{i=1}^{N_{m}} \log \frac{e^{W_{yi}f_i}}{\sum_{k=1}^{N_{id}} e^{W_{yk}f_i}} \]

where \( N_{m} \) is the number of images in the mini-batch, \( N_{id} \) denotes the number of identities in the training dataset, \( W_{yk} \) represents the weight of the classifier for \( k \)-th identity, \( y_i \) is the ground truth identity of \( i \)-th image, and \( W_{yi} \) is the weight of the classifier for \( y_i \) identity.

The role of the triplet loss function is to increase the interclass distance and reduce the intraclass distance, which is very effective for distinguishing different pedestrians under the same camera and identifying the same pedestrian under different cameras. To further improve the ability of the triplet loss to distinguish difficult samples, Hermans et al. [24] proposed batch hard triplet loss, which aims to encourage the hardest positive pairs \( (f_a, f_p) \) to be closer in the learnt feature space than the hardest negative pairs \( (f_n, f_n) \) by a margin \( m \). For \( PK \) images (\( P \) is the number of person identities, and \( K \) is the number of images for each identity) in a mini-batch, batch hard triplet loss can be formulated as follows:

\[ L_{BH} = \sum_{i=1}^{P} \sum_{a=1}^{K} \left[ \max_{1 \leq n \leq K} D(f_{a}^{(i)}, f_{n}^{(i)}) - \min_{j=1 \ldots P} D(f_{a}^{(i)}, f_{n}^{(j)}) + m \right] + \]

where \([ \cdot ]_+ = \max(\cdot, 0)\) is a hinge function, \( f_{a}^{(i)} \), \( f_{p}^{(i)} \), and \( f_{n}^{(j)} \) are feature vectors extracted from anchor, positive, and negative samples, respectively, and the distance metric \( D(a, b) = ||a - b||_2 \) is the Euclidean distance. We use the aforementioned batch hard triplet loss as the triplet loss in our model.

In our approach, we apply softmax and triplet loss simultaneously to the features of the three-branch outputs to capture more comprehensive information. Triplet loss is used to reduce the distance between similar samples and increase the distance between different samples. This metric strategy is more suitable for global features than local features. Assuming that different people wear similar black pants, after the triplet loss training, the local feature of ‘black pants’ of different
people will be closer, which will significantly reduce the performance of the model. Therefore, we only use triplet loss for global features, including global features from Branch-1, connection features of local features from Branch-2, and HBP features from Branch-3 in our model. We believe that HBP features are a more refined global feature. Softmax loss can be used for both global and part features in classification tasks. In the classification of local features, softmax loss is conducive to capturing the specific information in the local part to learn obvious interclass representations. During training, softmax loss is employed on all output features directly, including global features from Branch-1, local features from Branch-2, and HBP features from Branch-3. Two loss functions work together to obtain model performance optimal in model training. To balance the contribution of the two loss functions in model training, we integrate two tasks via two parameters $\lambda_1$, $\lambda_2$. The total loss can be defined as

$$L_{total} = \lambda_1 L_{CE} + \lambda_2 L_{BH}$$

In the following Section 4.4, we will test and select the optimal parameter settings for our model.

### 4 | EXPERIMENTS

#### 4.1 | Datasets

We implemented experiments to evaluate our proposed framework on three datasets: Market-1501 [49], DukeMTMC-reID [50] and CUHK03 [20, 51]. Basic information about these three datasets is shown in Table 1, wherein $N_{id}$, $N_{im}$, and $N_{cam}$ represent the number of identities, images, and cameras, respectively. The Market-1501 dataset contains images of 1501 pedestrians collected from six different cameras and a total of 32,217 pedestrian images annotated by DPM [52] detector. There are single-query and multiple-query modes in evaluation, and our experiment is implemented in single-query mode. DukeMTMC-reID is a dataset consisting of images of 1812 pedestrians captured by eight cameras and a total of 36,411 person images with handcrafted labelled pedestrian bounding boxes. From the data in Table 1 and previous workers, it can be seen that DukeMTMC-reID is more challenging than Market-1501. The CUHK03 dataset contains two datasets: Detected with manually labelled person bounding boxes and Labelled with DPM detected person bounding boxes. The two datasets contain images of 1467 identities captured by 10 (five pairs) different cameras. CUHK03 involves two protocols. One is a single-shot setting [20], while the other is a new training/testing protocol [51]. All experiments in this work follow the new training/testing protocol. For evaluation protocols, we follow the common practices and use the cumulative matching characteristics at Rank-n and mean average precision (mAP) to evaluate the accuracy of the performance. By the way, we do not use reranking to improve the performance of our method.

#### 4.2 | Implementation details

For the baseline network, we use the ResNet-50 model parameters pretrained on ImageNet to initialize our network. In Layer4, we cancel the original downsampling and split Layer4 into three portions: Layer4_0, Layer4_1, and Layer4_2 in the third branch. We uniformly resize the input images of all training stages to $384 \times 128$ and perform three simple data enhancement methods: random horizontal flipping, normalization, and random erasing. In the training phase, Adam optimizer is used to optimize the entire model, and a total of 700 epochs are trained. The initial learning rate is $2 \times 10^{-4}$, which decreases to $2 \times 10^{-5}$ at epoch 520 and further decays to $2 \times 10^{-6}$ at epoch 580. For the triple loss function, we set the margin to 1.2. The whole training procedure is implemented on the PyTorch framework with a single NVIDIA 2080Ti GPU.

#### 4.3 | Comparison with state-of-the-art methods

On the Market-1501 and DukeMTMC-reID datasets, we select 13 methods to compare with our approach, and the results are shown in Table 2. On the CUHK03 dataset, we compare 12 methods with our solution, and the results are shown in Table 3. These various state-of-the-art methods include local feature models [7, 57, 60, 63], attention mechanism models [35, 36, 55], multibranch networks [8, 25, 30], multiscale networks [56, 62, 64], and other deep models [45, 49, 53, 54, 58, 61]. It

---

**Table 1** Statistics of Market-1501, DukeMTMC-reID, and CUHK03 datasets, wherein $N_{id}$, $N_{im}$, and $N_{cam}$ represent the number of the identity, image, and camera, respectively. Because the identities in the query set and the gallery overlap, the total $N_{id}$ is only equal to training set $N_{id}$ plus gallery $N_{id}$. In addition, the total $N_{im}$ is equal to the sum of query $N_{im}$, training set $N_{im}$, and gallery $N_{im}$.

| Dataset                | Year | Query | Training set | Gallery | Total | Detector   |
|------------------------|------|-------|--------------|---------|-------|------------|
| CUHK03(Detected)       | 2014 | 700   | 1400         | 767     | 7365  | 1467       | 14,097     | 10 | DPM       |
| CUHK03(Labelled)       | 2014 | 700   | 1400         | 767     | 7368  | 1467       | 14,097     | 10 | Handcrafted |
| Market-1501            | 2015 | 750   | 3368         | 751     | 12,936| 1501       | 32,217     | 6  | DPM       |
| DukeMTMC-reID          | 2017 | 702   | 2228         | 702     | 16,522| 1,110      | 17,661     | 8  | Handcrafted |

Abbreviations: DPM, deformable part model; Re-ID, reidentification.
| Methods          | Year | Market-1501 mAP | Rank-1 | Rank-5 | Rank-10  | DukeMTMC-reID mAP | Rank-1 | Rank-5 | Rank-10 |
|------------------|------|-----------------|--------|--------|----------|-------------------|--------|--------|--------|
| SVDNet [53]      | 2017 | 62.1            | 82.3   | 92.3   | 95.2     | 56.8             | 76.7   | 86.4   | 89.9   |
| MLFN [54]        | 2018 | 74.4            | 90.0   |        | -        | 62.8             | 81.2   | 83.3   | -      |
| HA-CNN [55]      | 2018 | 75.7            | 91.2   |        | -        | 63.8             | 80.5   | -      | -      |
| PCB + RPP [7]    | 2018 | 81.6            | 93.8   | 97.5   | 98.5     | 69.2             | 83.3   | -      | -      |
| AANet-50 [36]    | 2019 | 82.5            | 93.9   | -      | 98.6     | 72.6             | 86.4   | -      | -      |
| HPM [9]          | 2019 | 82.7            | 94.2   | 97.5   | 98.5     | 74.3             | 86.6   | 93.0   | 95.1   |
| OSNet [56]       | 2019 | 84.9            | 94.8   | -      | -        | 73.5             | 88.6   | -      | -      |
| Auto-ReID [57]   | 2019 | 85.1            | 94.5   | -      | -        | 75.1             | 88.5   | -      | -      |
| PPS [30]         | 2019 | 85.3            | 94.3   | 97.7   | 98.7     | 76.0             | 88.2   | 95.3   | 96.2   |
| BoT [45]         | 2019 | 85.9            | 94.5   | -      | -        | 76.7             | 86.4   | -      | -      |
| CBN + BoT [58]   | 2020 | 83.6            | 94.3   | 97.9   | 98.7     | 70.1             | 84.8   | 92.5   | 95.2   |
| SNR [59]         | 2020 | 84.7            | 94.4   | -      | -        | 72.9             | 84.4   | -      | -      |
| HOReID [60]      | 2020 | 84.9            | 94.2   | -      | -        | 75.6             | 86.9   | -      | -      |
| MBN-HBP          | 2020 | 86.3            | 94.2   | 97.9   | 98.8     | 79.2             | 89.1   | 95.1   | 96.7   |

Note: The bold represents the maximum value in a column of data.
Abbreviations: AANet-50, attribute attention network 50; HA-CNN, harmonious attention-convolutional neural network; HPM, horizontal pyramid matching; mAP, mean average precision; MBN-HBP, multibranch network with hierarchical bilinear pooling; MLFN, multilayer feedforward neural; OSNet, omni-scale network; PCB, part-based convolutional baseline; PPS, peri-person space; Re-ID, reidentification; RPP, residual pyramid pooling; SNR, signal-to-noise ratio.

**Table 3** Comparison results (%) with state-of-the-art methods on CUHK03 dataset

| Methods          | Year | Detected mAP | Rank-1 | Rank-5 | Rank-10  | Labelled mAP | Rank-1 | Rank-5 | Rank-10 |
|------------------|------|--------------|--------|--------|----------|--------------|--------|--------|--------|
| SVDNet [53]      | 2017 | 37.3         | 41.5   | -      | -        | 37.8         | 40.9   | -      | -      |
| MLFN [54]        | 2018 | 47.8         | 52.8   | -      | -        | 49.2         | 54.7   | -      | -      |
| HA-CNN [55]      | 2018 | 38.6         | 71.4   | -      | -        | 41.0         | 44.4   | -      | -      |
| PCB + RPP [7]    | 2018 | 57.5         | 63.7   | -      | -        | -            | -      | -      | -      |
| MGN [25]         | 2018 | 66.0         | 68.0   | -      | -        | 67.4         | 68.0   | -      | -      |
| HPM [8]          | 2019 | 57.5         | 63.9   | 79.7   | 86.1     | -            | -      | -      | -      |
| MHN-6 (PCB) [35] | 2019 | 65.4         | 71.7   | -      | -        | 72.4         | 77.2   | -      | -      |
| OSNet [56]       | 2019 | 67.8         | 72.3   | -      | -        | -            | -      | -      | -      |
| Auto-ReID [57]   | 2019 | 69.3         | 73.3   | -      | -        | 73.0         | 77.9   | -      | -      |
| BDB + Cut [61]   | 2019 | 73.5         | 76.4   | -      | -        | 76.7         | 79.4   | -      | -      |
| Multiscale [62]  | 2020 | 67.2         | 70.1   | -      | -        | 68.5         | 70.4   | -      | -      |
| GraftedNet [63]  | 2020 | 71.6         | 76.2   | -      | -        | -            | -      | -      | -      |
| Ms-Mb [64]       | 2020 | 72.9         | 75.4   | -      | -        | -            | -      | -      | -      |
| MBN-HBP          | -    | 74.2         | 77.1   | 89.9   | 93.9     | 78.0         | 80.3   | 91.9   | 95.3   |

Note: The bold represents the maximum value in a column of data.
Abbreviations: AANet-50, attribute attention network 50; HA-CNN, harmonious attention-convolutional neural network; HPM, horizontal pyramid matching; mAP, mean average precision; MBN-HBP, multibranch network with hierarchical bilinear pooling; MLFN, multilayer feedforward neural; OSNet, omni-scale network; PCB, part-based convolutional baseline; PPS, peri-person space; Re-ID, reidentification; RPP, residual pyramid pooling; SNR, signal-to-noise ratio.
can be seen from Table 2 that our approach can achieve the best mAP and Rank-1 performance on both datasets. From Table 3, MBN-HBP can achieve the best mAP and Rank-n performance amongst various state-of-the-art methods. These results prove the superiority of multibranch network structure and pedestrian representation that incorporates HBP features from our work.

### 4.4 Network analysis

We conduct many experiments to evaluate the effectiveness of each component in MBN-HBP, and separate analyses as follows:

#### 4.4.1 Branches

This section of the experiment mainly verifies the effectiveness of MBN-HBP. The results are shown in Table 4. In this experiment, we set the pooling strategy to max pooling and the loss function weight parameter $\lambda_1 = 1$, $\lambda_2 = 2$, and all the results follow the same experimental settings. It is not difficult to see that with the increase in the number of network branches, the mAP and Rank-1 on the two datasets have effectively improved. According to the strategy of extracting local features in PCB [5], we divide the feature maps into six horizontal stripes of the same size in Branch-2. Obviously, from ‘B1’ to ‘B1 + B2’, the performance of the model has been greatly improved, which shows that adding local features to the human representation is of great help in identifying different pedestrians. When adding Branch-3, the HBP part, the mAP on the two datasets increased by 1.01% and 1.32%, respectively, indicating the hierarchical feature interaction in the HBP not only useful for fine-grained classification but also effective for person Re-ID tasks. This kind of interlayer feature interaction information can be used as supplementary information in the ‘global feature + local feature’ mode to improve the discrimination of features effectively. As shown in Figure 3, we visualize the features $fe_0$, $fe_1$, and $fe_2$ output from Layer4_0, Layer4_2, and Layer4_3 and the features $pro_0$, $pro_1$, and $pro_2$ obtained by outer product and project them in pairs as corresponding feature maps. We can observe that, compared with $fe_0$, $fe_1$, and $fe_2$, their product features $pro_0$, $pro_1$, and $pro_2$ can capture more detailed and precise information. Simultaneously, the outer product operation filters out the background noise and more prominently represents the semantic part of the pedestrian image, such as hair, T-shirt, shorts, legs, face etc. These semantic details are vital for identifying different pedestrians. In the original HBP [16] for fine-grained classification tasks, the final output features use only softmax loss. Unlike their work, we consider the characteristics of the person Re-ID task and use triplet loss for the HBP feature. This manner increases mAP on the two datasets by 0.31% and 0.17%, respectively. It is evident that triplet loss is equally effective for HBP features.

#### 4.4.2 Pooling schemes

In CNNs, pooling is generally used to compress features and reduce the number of parameters, which helps to remove redundant information and extract the most important features. Pooling can also reduce the amount of calculation and improve the generalization ability of the model. The most common pooling method is max pooling and average pooling. To prove which pooling strategy is most suitable for our model, we conduct some experiments about three pooling strategies, and the results are shown in Table 5. In this section of the experiment, we set the network structure to a three-branch framework. The loss function weight parameter $\lambda_1 = 1$, $\lambda_2 = 2$, and all the results follow the same experimental settings. In some pyramid networks [8, 30] for Re-ID tasks, the combination of average and max pooling through addition helps improve model performance, so this paper also tried this pooling method. The results show that max pooling is better than the other two pooling methods, which shows that the pooling combination strategy is not suitable for our model. For our multibranch network, max pooling is proving to be the best pooling method.

#### 4.4.3 Parameters of loss functions

We set the weight parameters $\lambda_1$ and $\lambda_2$ for the two loss functions separately in $L_{total}$ to balance contributions of softmax loss and triplet loss in multitask learning. In this section of the experiment, we select the optimal weight parameters for our model. We set the network structure to a three-branch structure with max pooling, and all the results follow the same experimental settings. The results are shown in Figure 4. When $\lambda_2/\lambda_1 = 2/1$, our model achieves the best performance. This part of the work shows that the contribution of each loss function when supervising the network training at the same time is different in the multitask learning model framework. Therefore, it is necessary to integrate these loss functions in different multiloss function frameworks by setting weights to make them play the best role in model training.

| Architecture | Market-1501 (mAP) | Market-1501 (Rank-1) | DukeMTMC-reID (mAP) | DukeMTMC-reID (Rank-1) |
|--------------|------------------|----------------------|---------------------|------------------------|
| B1           | 78.82            | 90.74                | 69.84               | 82.85                  |
| B1 + B2      | 84.94            | 93.88                | 77.68               | 88.54                  |
| B1 + B2 + B3 (no-triplet) | 85.95            | 94.09                | 79.00               | 88.82                  |
| B1 + B2 + B3 | 86.26            | 94.21                | 79.17               | 89.06                  |

Note: The bold represents the maximum value in a column of data. Abbreviations: reID, reidentification; mAP, mean average precision.

| Architecture | Market-1501 (mAP) | Market-1501 (Rank-1) | DukeMTMC-reID (mAP) | DukeMTMC-reID (Rank-1) |
|--------------|------------------|----------------------|---------------------|------------------------|
| B1           | 78.82            | 90.74                | 69.84               | 82.85                  |
| B1 + B2      | 84.94            | 93.88                | 77.68               | 88.54                  |
| B1 + B2 + B3 (no-triplet) | 85.95            | 94.09                | 79.00               | 88.82                  |
| B1 + B2 + B3 | 86.26            | 94.21                | 79.17               | 89.06                  |

Note: The bold represents the maximum value in a column of data. Abbreviations: reID, reidentification; mAP, mean average precision.
5 | CONCLUSION

We find the similarities between fine-grained classification and person Re-ID and apply bilinear pooling for fine-grained classification to the person Re-ID task. By performing numerous experiments, we prove that bilinear pooling is equally effective for person Re-ID tasks. In addition, we adopt a three-branch framework as our model. The three parts extract the global, part-level, and hierarchical bilinear features of pedestrians. This scheme equips the person features with stable discrimination when facing a complex and changeable monitoring environment. We use two loss functions to train our model to learn rich and comprehensive personal information. Whether it is a multibranch structure network, a training method of multitask learning, or the introduction of bilinear pooling features, each scheme plays an indispensable role for our entire model. These elements will give worthy direction to Re-ID tasks in future research.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. U1603115), National key R&D plan project (2017 YF C08207023) and National Engineering
Laboratory for Public Safety Risk Perception and Control by Big Data (XJ201810101). The authors would like to thank the anonymous reviewers for their valuable remarks and comments.

ORCID

Wenyu Wei https://orcid.org/0000-0003-0995-7063

REFERENCES

1. Li, Y., et al.: A combined feature representation of deep feature and hand-crafted features for person re-identification. In: Proceedings of the 2016 International Conference on Progress in Informatics and Computing (PIC), pp. 224-227. Shanghai, China (2016)
2. Wu, S., et al.: An enhanced deep feature representation for person re-identification. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8. Lake Placid, NY, USA (2016)
3. Liao, S., et al.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2197–2206. Boston, MA, USA (2015)
4. Kostinger, M., et al.: Large scale metric learning from equivalence constraints. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2288–2295. Providence, RI, USA (2012)
5. Su, C.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3980–3989. Venice, Italy (2017)
6. Wei, L., et al.: GLAD: global-local-alignment descriptor for scalable person re-identification. IEEE Trans. Multimed. 21(4), 986–999 (2019)
7. Sun, Y., et al.: Beyond Part Models: person retrieval with refined Part Pooling (and A strong convolutional baseline). European Conference on Computer Vision (ECCV), pp. 501–518. Munich, Germany (2018)
8. Fu, Y., et al.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33. Honolulu, Hawaii, USA (2018)
9. Zhang, S., et al.: Person re-identification with joint verification and identification of identity-attribute labels. IEEE Access. 7, 126116–126126 (2019)
10. Bai, X., et al.: Deep-person: learning discriminative deep features for person re-identification. Pattern Recogn. 98, 107036 (2020)
11. Zhang, J., et al.: Hierarchical and Efficient Learning for Person Re-identification. CoRR abs/2005.08812 (2020)
12. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Comput. 12(6), 1247–1283 (2000)
13. Lin, T., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1449–1457. Santiago, Chile (2015)
14. Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7025–7034. Honolulu, HI, USA (2017)
15. Wei, X., et al.: Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. European Conference on Computer Vision (ECCV), pp. 365–380. Munich, Germany (2018)
52. Ramanan, D., Felzenszwalb, P., McAllester, D.: A discriminatively trained, multiscale, deformable part model. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern recognition (CVPR), pp. 1–8. IEEE Computer Society, Los Alamitos (2008)

53. Sun, Y., et al.: Svdnet for pedestrian retrieval. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3820–3828. IEEE Computer Society, Los Alamitos (2017)

54. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern recognition, CVPR 2018, pp. 2109–2118. IEEE Computer Society, Salt Lake City, 18–22 June 2018

55. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern recognition, CVPR 2018, pp. 2285–2294. IEEE Computer Society, Salt Lake City, 18–22 June 2018

56. Zhou, K., et al.: Omni-scale feature learning for person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on computer vision, ICCV 2019, pp. 3701–3711. IEEE, Seoul (2019, 27 Oct–2 Nov 2019)

57. Quan, R., et al.: Auto-reid: searching for a part-aware convnet for person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 3749–3758 Seoul, Korea (South) (2019)

58. Zhuang, Z., et al.: Rethinking the distribution gap of person re-identification with camera-based batch normalization. European Conference on Computer Vision (ECCV), pp.140-157. Glasgow, UK (2020)

59. Jin, X., et al.: Style normalisation and restitution for generalisable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern recognition (CVPR 2020), pp. 3140-3149. Seattle, WA, USA (2020)

60. Wang, G., et al.: High-order information matters: learning relation and topology for occluded person re-identification. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6448–6457. Seattle, WA, USA (2020)

61. Dai, Z., et al.: Batch dropblock network for person re-identification and beyond. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3690–3700. Seoul, Korea (South) (2019)

62. Wang, C., et al.: Multi-scale multi-patch person re-identification with exclusivity regularised softmax. Neurocomputing. 382, 64–70. https://www.sciencedirect.com/science/article/pii/S0925231219316649 (2020)

63. Wang, J., et al.: Grafted network for person re-identification. Signal Process. Image Commun. 80, 115674. https://www.sciencedirect.com/science/article/pii/S092359651930431X (2020)

64. Jiao, S., et al.: Multi-scale and multi-branch feature representation for person re-identification. Neurocomputing. 414, 120–130. https://www.sciencedirect.com/science/article/pii/S0925231220310626 (2020)

How to cite this article: Wei W, Yang W, Zuo E, Ren Q, Chen Q. Multi-branch network with hierarchical bilinear pooling for person reidentification. IET Biometrics. 2021;1–12. https://doi.org/10.1049/bme2.12040