Harmonious Multi-branch Network for Person Re-identification with Harder Triplet Loss

ZENGMING TANG and JUN HUANG, Shanghai Advanced Research Institute, Chinese Academy of Sciences, China and University of Chinese Academy of Sciences, China

Recently, advances in person re-identification (Re-ID) has benefitted from use of the popular multi-branch network. However, performing feature learning in a single branch with uniform partitioning is likely to separate meaningful local regions, and correlation among different branches is not well established. In this article, we propose a novel harmonious multi-branch network (HMBN) to relieve these intra-branch and inter-branch problems harmoniously. HMBN is a multi-branch network with various stripes on different branches to learn coarse-to-fine pedestrian information. We first replace the uniform partition with a horizontal overlapped partition to cover meaningful local regions between adjacent stripes in a single branch. We then incorporate a novel attention module to make all branches interact by modeling spatial contextual dependencies across branches. Finally, in order to train the HMBN more effectively, a harder triplet loss is introduced to optimize triplets in a harder manner. Extensive experiments are conducted on three benchmark datasets — DukeMTMC-reID, CUHK03, and Market-1501 — demonstrating the superiority of our proposed HMBN over state-of-the-art methods.

CCS Concepts: • **Computing methodologies** → **Object identification; Image representations; Visual content-based indexing and retrieval;**

Additional Key Words and Phrases: Person re-identification, pooling strategy, attention mechanism, triplet loss

ACM Reference format:

Zengming Tang and Jun Huang. 2022. Harmonious Multi-branch Network for Person Re-identification with Harder Triplet Loss. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 4, Article 98 (February 2022), 21 pages.

<https://doi.org/10.1145/3501405>

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. Re-ID is a popular computer vision task for its giant potential in video surveillance applications. Recently, deep learning methods have pushed the performance of Re-ID to a new level. However, many challenges — such as pose variations, illumination variations, view angle variations, and occlusions — make Re-ID non-trivial.

This work was supported by the National Key R&D Program of China (2020YFC1523202).

Authors' address: Z. Tang and J. Huang (corresponding author), Shanghai Advanced Research Institute, Chinese Academy of Sciences, No.99 Haik Road, Pudong New Area, Shanghai, China, 201210, University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District, Beijing, China, 100049; emails: {tangzengming2019, huangj}@sari.ac.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

1551-6857/2022/02-ART98 \$15.00

<https://doi.org/10.1145/3501405>

Fig. 1. Illustration of intra-branch and inter-branch problems. **For the problem of feature learning in a single branch**, Branch N employs a uniform partition with 6 stripes. The head is divided into two stripes, which diminishes the representational capability in head regions. **For the problem of correlation among different branches**, the multi-branch network shares lower layers to learn strongly correlated features and performs independent feature learning in higher layers for different branches. However, strong relations between branches vanish after the split.

To relieve these issues, many part-based methods [22, 24, 44, 50] with multiple branches have been proposed to learn local features that have achieved promising results. Specifically, these methods combine information in different granularities and learn coarse-to-fine representations in multi-branch networks. Although they achieve state-of-the-art performance, they still suffer from intra-branch and inter-branch problems, that is, the problems of feature learning in a single branch and correlation among different branches.

Feature learning in a single branch. In a single branch, some part-based methods conduct pre-defined horizontal or vertical partitions on feature maps to extract fine-grained information for local feature learning based on the assumption that images are well aligned [5, 9, 41, 45, 57, 58]. Part-based Convolutional Baseline (PCB) [41] achieves competitive results compared with state-of-the-art methods by partitioning feature maps into 6 horizontal stripes. In PCB, as the number of stripes increases, retrieval accuracy improves at first but drops dramatically in the end. The over-increased number of stripes helps to learn fine-grained information but compromises the representational capability in meaningful local regions. We argue that the uniform partition is not optimal, which separates important semantic regions, as shown in Figure 1.

Correlation among different branches. As is shown in Figure 1, a multi-branch network shares lower layers and extracts distinct information at higher layers for different branches. The sharing learning scheme builds branch interaction in lower layers by extracting the same low-level features (e.g., edges, lines) for each branch. In this manner, the strongly correlated information in the low-level layer is exploited. However, the interaction among branches is neglected in higher layers of the network after the split.

Triplet loss [12] is a popular loss function in part-based methods with multiple branches because of its enormous capability to optimize the similarity among samples. Triplet loss aims at reducing and enlarging intra-class and inter-class variations. However, there is still room in optimizing triplets.

Optimizing triplets. A triplet contains one anchor, one positive, and one negative. Given an anchor, mining the hard positive and hard negative is an essential part of learning with triplet loss. Schroff et al. [33] select all anchor-positive pairs, and pick hard negatives by semi-hard negative mining. Hermans et al. [12] propose to choose the hardest positives and hardest negatives within a mini-batch. However, triplets selected by the hardest positive and hardest negative mining are still

not hard enough for models to discriminate without up-weighting anchor-to-positive distance or down-weighting anchor-to-negative distance. In this manner, intra-class and inter-class variations are difficult to further reduce and enlarge.

In this article, we propose a novel model, a *harmonious multi-branch network* (HMBN) with *harder triplet loss* (HTP), to tackle these problems. The HMBN jointly learns pedestrian representations in multi-granularity with three branches called S1B, S2B, and S3B. HMBN adopts S1B to learn global features and applies S2B and S3B to capture fine-grained information. In the single branch, instead of performing a uniform partition, we design a pooling strategy called *horizontal overlapped pooling* (HOP) to conduct a horizontal overlapped partition on feature maps and cover meaningful local regions between adjacent stripes. Furthermore, to learn interactive features among branches, we incorporate the *inter-branch attention module* (IBAM), which involves three *inter-branch attention submodules* (IBASMs). The IBAM enables our HMBN to refine features by aggregating spatial contextual information from different branches in higher layers. In this manner, interaction among branches is preserved in higher layers of the HMBN. In addition, a novel *harder triplet loss* (HTP) is introduced to optimize intra-class and inter-class similarities more effectively by optimizing triplets in a harder manner. HTP up-weights anchor-to-positive distance and down-weights anchor-to-negative distance by a polynomial mapping function and penalizes more in cases in which anchor-to-positive distance is not substantially smaller than the anchor-to-negative distance. In this process, HTP further reduces and enlarges intra-class and inter-class variations.

To sum up, our main contributions are as follows:

- We propose a novel *harmonious multi-branch network* (HMBN) to learn discriminative pedestrian information by handling intra-branch and inter-branch problems harmoniously.
- We design a new pooling strategy named *horizontal overlapped pooling* (HOP) that helps to keep the balance between learning fine-grained information and extracting features in meaningful local regions.
- We incorporate a compound attention module called the *inter-branch attention module* (IBAM) into the HMBN to learn interactive representations for each branch. To the best of our knowledge, this is the first module that builds strong relations among different branches in higher layers for Re-ID.
- We introduce a generalized triplet loss termed *harder triplet loss* (HTP) to optimize triplets in a harder manner, which is more effective than traditional triplet loss.
- Extensive experiments on three datasets show that the HMBN outperforms state-of-the-art methods. In addition, ablation studies verify that HOP, IBAM, and HTP all contribute to an accuracy gain.

This article is an extended version of our early and preliminary conference work [43]. In this extended journal version, we have four modifications. (1) We introduce a multi-branch architecture for robust intra-branch and inter-branch feature learning explicitly. In our conference version, we mainly introduce two independent components (HOP and IBAM) explicitly but ignore the elaborate system in its entirety. The whole system (HMBN) is also a contribution of our work, as it alleviates intra-branch and inter-branch problems harmoniously. (2) Each component is discussed in more detail, for example, an additional comparison between the original uniform partition and our proposed horizontal overlapped partition is presented. (3) We propose an HTP to optimize triplets more effectively. (4) More comprehensive experiments with parameter analysis and visualizations are conducted. Specifically, we add ablation experiments to validate the HTP and an additional ablation study on CUHK03-Labeled and CUHK03-Detected datasets. In addition, each

component is verified more thoroughly compared with the conference version, for example, HOP is ablated both in single-branch and multi-branch networks.

2 RELATED WORKS

In this section, we discuss recent related works in terms of part-based Re-ID, attention-based Re-ID, and metric learning-based Re-ID.

2.1 Part-Based Re-ID

First, convolutional neural networks (CNNs) are applied in an image classification task [11, 16, 17, 19, 37, 42]. Thus, it is popular to treat the training process of Re-ID as an image classification task and extract global pedestrian representations. However, use of global features is insufficient for capturing fine-grained cues. Many researchers aggregate global and local features. These methods can be divided into two categories according to the number of branches. The first category is single-branch methods [41, 45, 68]. For example, Varior et al. [45] crop images into several horizontal stripes and process the stripes sequentially by long short-term memory (LSTM) [13] cells. In this manner, contextual information among image regions is leveraged to enhance local feature representative capability. Zhou et al. [68] design a novel OSNet, which is composed of stacked convolutional streams with different receptive field sizes for extracting multi-scale features. The second category is multi-branch methods. Multi-branch methods are superior to single-branch methods in aggregating branch-specific information, that is, fine-grained cues [22, 50], human parsing results [18, 38], pose estimation [30, 39, 54], key points estimation [54], and semantic attributes [24, 44]. For example, in the SPReID framework, Kalayeh et al. [18] utilize the parsing model to generate probability maps associated to 5 predefined human parts and extract robust local features. Sarfraz et al. [32] incorporate 14 main body joint keypoints to model pedestrian information.

However, none of them perform feature map partition with overlap. In contrast, HOP is designed to cover meaningful body regions while extracting fine-grained information.

2.2 Attention-Based Re-ID

Attention mechanisms have verified their effectiveness in many tasks, for example, image classification [16, 48, 53, 60], video classification [25, 29], image caption [6, 55], and in generative adversarial networks (GANs) [27, 59]. It is also efficient and effective for Re-ID tasks with dynamically focusing on salient regions. Some methods employ spatially based attention [14, 23, 34, 38] for feature refinement. For example, Li et al. [23] introduce the HA-CNN to learn soft pixel attention and hard regional attention jointly for robust feature extraction. Channel-based attention is also explored in many works [47, 54]. The attention mechanism is also applicable in frame or feature sequences [3, 15, 20, 36]. Si et al. [36] propose a framework termed *DuATM*, which is composed of a dual attention mechanism to learn context-aware information by modeling intra-sequence and inter-sequence dependencies.

These works apply an attention mechanism to focus at certain patterns with greater notice within a single branch. In contrast, our proposed IBAM helps to generate refined representations by aggregating information from all branches.

2.3 Metric Learning-Based Re-ID

Metric learning aims to learn a similarity or a mapping function to minimize the intra-class variation while maximizing the inter-class variation. Triplet loss is a widely used loss function that treats the Re-ID task as a ranking task and optimizes the similarity among anchor, positive, and negative samples. Various methods [12, 35, 40, 52] are proposed to select hard triplets for learning discriminately. Batch hard triplet loss [12] is designed with selecting the hardest

Fig. 2. The overall architecture of the proposed HMBN. The HMBN contains a base module and three independent branches: S1B, S2B, and S3B. IBAM is injected in the higher layer of the network for modeling interactive information in different branches. HMBN learns global features by applying GMP on three branches and extracts local features by employing HOP on S2B and S3B. The whole network is trained with classification loss and HTP. GMP is short for *global max pooling*.

positives and hardest negatives for robust Re-ID learning. In addition, many methods [61, 69, 70] are proposed to improve the gradient backpropagation. Zhou et al. [69] introduce the center point of the positive pair to model all of the pairwise relationships.

In contrast to previous variations of triplet loss, HTP is a generalized triplet loss to optimize triplets in a harder manner dynamically.

3 HARMONIOUS MULTI-BRANCH NETWORK (HMBN)

In this section, we first describe the overall architecture of the HMBN. Then, the coarse-to-fine structure and *horizontal overlapped pooling* (HOP) are discussed, followed by a novel attention module named the *inter-branch attention module* (IBAM). Next, an improved triplet loss called *harder triplet loss* (HTP) is presented. Finally, we discuss the relations between the proposed modules and some existing methods.

3.1 Overall Architecture

As shown in Figure 2, the HMBN is a multi-branch network, including a base module and three independent branches. ResNet-50 [11] is applied for our feature extraction backbone. The base module consists of previous layers before conv4_2, which is capable of generating shared low-level visual features for each branch. Specifically, three branches are directly borrowed from subsequent layers after conv4_1 — stripe 1 branch (S1B), stripe 2 branch (S2B), and stripe 3 branch (S3B) — based on the number of stripes. S1B performs the Re-ID task at the global level, while S2B and S3B both perform feature learning at the global level and part level. In S2B and S3B, we remove the last spatial down-sampling operation to enrich the granularity. As a result, feature tensors T_1 , T_2 , and T_3 , the output of conv5 from S1B, S2B, and S3B, respectively, have different spatial sizes. In order to integrate multi-branch features, we inject the IBAM on the higher layer of the HMBN to exploit complementary information across branches.

With global max pooling (GMP), the HMBN generates global feature representations g_i ($i = 1, 2, 3$) for each branch. A parameter shared 1×1 convolution layer, followed by a batch normalization layer and ReLU layer, is applied to reduce the dimension from 2048-dim g_i ($i = 1, 2, 3$) to 256-dim global feature u_i ($i = 1, 2, 3$).

Fig. 3. Visualization results of activations in three coarse-to-fine networks.

Fig. 4. Horizontal overlapped pooling (HOP) in a general form. GMP is short for global max pooling. The size of the overlapped portion is $C \times h \times W$. l is the total height of overlapped areas in one stripe. k is the number of partitions.

With our proposed HOP, HMBN partitions $T_i (i = 2, 3)$ into 2 and 3 horizontal stripes in S2B and S3B, and pools these stripes to generate column feature vectors, that is, $\mathbf{p}_m^n$, where m, n refer to the m -th stripe in the stripe n branch. The dimension of $\mathbf{p}_m^n$ is also reduced to 256 by the 1×1 convolution layer to acquire a dimension-reduced local feature $\mathbf{v}_m^n$.

3.2 Coarse-to-Fine Structure

The HMBN is a multi-branch network for coarse-to-fine feature learning. S1B is designed to learn the coarse-grained information (i.e., global information), and S2B and S3B are incorporated to learn fine-grained information with different granularities. We compare the activations of the last convolutional feature maps from B + S1B, B + S2B, and B + S3B in Figure 3. B + S1B is short for a model including base module and S1B, and so forth. B + S1B mainly focuses on the most discriminative regions (e.g., shoulder, shoes). With the increase in number of stripes, more detailed regions can be observed. Regions marked by a red eclipse are ignored by B + S1B but are observed by B + S2B and B + S3B. Regions marked by a yellow eclipse are noticed only by B + S3B.

3.3 Horizontal Overlapped Pooling (HOP)

Given a feature map $F \in \mathbb{R}^{C \times H \times W}$, HOP is illuminated in Figure 4. It has two parameters: l and k . l is the total height of overlapped areas in one stripe and k is the number of stripes. When $k = 1$, HOP degrades into GMP, and remains the global information. When $k > 1$, we learn the fine-grained information. Thus, in HMBN, we keep $k = 2$ in S2B and $k = 3$ in S3B.

Fig. 5. Comparison between a uniform partition and horizontal overlapped partition when the number of stripes is 6. (a) Uniform partition. (b) Horizontal overlapped partition.

First, we perform a uniform partition on the feature map F horizontally. With the aim of devoting equal attention to each stripe, parts on the top and bottom are extended in one direction; 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. In this case, $l = 2h$. As a result, we require that l must be an even number. Finally, each horizontal stripe is pooled by GMP to generate a part-level vector.

To highlight the difference between a uniform partition and horizontal overlapped partition, we visualize the region covered in each stripe in Figure 5 when the number of stripes is 6. The uniform partition diminishes the representational capability for partitioning the head into 2 stripes. With the horizontal overlapped partition, information for the head region is preserved well.

3.4 Inter-Branch Attention Module (IBAM)

Features extracted from different branches together help to boost the feature representational capability. In order to make branches interact, an IBAM is applied, as shown in Figure 2. Features from paired branches are fed into an *inter-branch attention submodule* (IBASM), which outputs paired refined features. The HMBN has three branches, which form $C_3^2 = 3$ combinations when we choose paired branches. As a result, each branch is selected twice and has two refined outputs that build interaction between various branches. The mean of two refined outputs is used to update the original feature, which is represented by the mean operation in Figure 2.

Figure 6 depicts the detailed structure of IBASM. Given two feature maps $A \in \mathbb{R}^{C \times H \times W}$, $B \in \mathbb{R}^{C \times H \times W}$ from different branches, a 1×1 convolution layer is employed to generate four new feature maps X , Y , M , and N , where $X, Y, M, N \in \mathbb{R}^{\frac{C}{8} \times H \times W}$. 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 in the spatial domain is calculated by matrix multiplication between transposed X and N . It is then normalized to obtain the spatial attention map $S \in \mathbb{R}^{L \times L}$, as shown here:

$$S_{i,j} = \frac{\exp(m_{i,j})}{\sum_{i=1}^L \exp(m_{i,j})}, m_{ij} = X_i^T N_j, \quad (1)$$

where X_i , N_j denote the i^{th} and j^{th} spatial features of X and N , respectively.

To calculate the output C , the HMBN first predicts A with attention map S and information from input B . The prediction, which is the result of matrix multiplication between transposed S and M , is reshaped to $\mathbb{R}^{C \times H \times W}$. Then, the HMBN performs an element-wise sum between the weighted

Fig. 6. The inter-branch attention submodule (IBASM). “ $\oplus$ ” denotes element-wise sum; “ $\otimes$ ” denotes matrix multiplication.

prediction and the original A . The output C is defined as

$$C_j = \gamma_1 \sum_{i=1}^L S_{i,j}^T M_i + A_j, \quad (2)$$

where γ_1 is a learnable weight that is initialized as 0. The output D is defined as

$$D_j = \gamma_2 \sum_{i=1}^L S_{i,j} Y_i + B_j. \quad (3)$$

In this manner, the refined A , which is denoted as C , contains reciprocal information from B . The refined B , which is denoted as D , contains reciprocal information from A .

The IBAM can be plugged into two positions in the HMBN: the output of conv4 and the output of conv5. Since the inputs of the IBAM need to have the same size, we apply a modification, removing the last spatial down-sampling operation in S1B when injecting the IBAM on the output of conv5. We find that adding the IBAM at the output of conv4 brings more performance improvement, because keeping the down-sample operation in S1B will produce complementary features when the down-sampling operation is removed in S2B and S3B. For this reason, the IBAM is placed on the output layer of conv4.

It is worth noted that the proposed IBAM is pluggable and can be injected into any existing multi-branch network because IBASM does not change the size of the feature map.

Armed with our proposed IBAM, spatial contextual dependencies across branches are well established and the interactive information in multi-granularity is utilized in higher layers.

3.5 Harder Triplet Loss (HTP)

In this subsection, we revisit the traditional triplet loss, then discuss its drawbacks in optimizing triplets. Finally, we propose the *harder triplet loss* (HTP) for making up these deficiencies.

Normally, triplet loss is trained on a set of triplet units $\{(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-)\}$, in which $(\mathbf{x}, \mathbf{x}^+)$ represents a positive pair from the same pedestrian and a negative pair $(\mathbf{x}, \mathbf{x}^-)$ represents images from different pedestrians. Given one triplet $(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-)$, triplet loss is formulated as

$$\begin{aligned} L_{tri}(f(\mathbf{x}), f(\mathbf{x}^+), f(\mathbf{x}^-)) &= [m + d_{a,p} - d_{a,n}]_+, \\ d_{a,p} &= d(f(\mathbf{x}), f(\mathbf{x}^+)), \\ d_{a,n} &= d(f(\mathbf{x}), f(\mathbf{x}^-)), \end{aligned} \quad (4)$$

Fig. 7. Empirical distribution of relative distance of a positive pair and negative pair from a converged HMBN model trained with traditional triplet loss and HTP on the DukeMTMC-reID dataset.

Fig. 8. Illustration of polynomial mapping function. (a) Polynomial mapping function on $d_{a,p}$. (b) Polynomial mapping function on $d_{a,n}$.

where m is the margin parameter, $d_{a,p}$ and $d_{a,n}$ are short for anchor-to-positive distance and anchor-to-negative distance, $d(\cdot)$ is the Euclidean distance, $[\cdot]_+$ denotes $\max(\cdot, 0)$, and $f(\mathbf{x})$, $f(\mathbf{x}^+)$, and $f(\mathbf{x}^-)$ are features of sample $\mathbf{x}$, $\mathbf{x}^+$, and $\mathbf{x}^-$, respectively.

The core idea for triplet loss is to optimize the similarity in triplets so that $d_{a,n}$ should be larger than $d_{a,p}$ by the margin m . However, the optimization can still be improved, as shown in Figure 7. To compare traditional triplet loss and HTP, we analyze the empirical distribution of the relative distance of a positive pair and negative pair, which is defined as $[m + d_{a,p} - d_{a,n}]_+$, from a converged HMBN. The number of samples is displayed by log axis because easy samples have extremely large numbers. With our proposed HTP, the majority of samples move left in Figure 7(b) compared with Figure 7(a), indicating that HTP helps intra-class variation and inter-class variation further reduced and enlarged.

For optimizing triplets in a harder manner, HTP penalizes large $d_{a,p}$ and small $d_{a,n}$ with polynomial mapping function, defined as follows:

$$\tilde{d}_{a,p} = (d_{a,p} + 1)^{(1+\alpha)} - 1, \quad (5)$$

$$\tilde{d}_{a,n} = (d_{a,n} + 1)^{(1-\alpha)} - 1, \quad (6)$$

where α is the scale factor. We update $d_{a,p}$ and $d_{a,n}$ with $\tilde{d}_{a,p}$ and $\tilde{d}_{a,n}$. The polynomial mapping function is visualized for several values of α in Figure 8. When $\alpha = 0$, $\tilde{d}_{a,p} = d_{a,p}$, and $\tilde{d}_{a,n} = d_{a,n}$. The larger α is, the more penalty $d_{a,p}$ and $d_{a,n}$ get.

Based on the polynomial mapping function, the HTP is defined as follows:

$$L_{HTP}(f(\mathbf{x}), f(\mathbf{x}^+), f(\mathbf{x}^-)) = [m + \tilde{d}_{a,p} - \tilde{d}_{a,n}]_+. \quad (7)$$

As shown in Figure 2, global features $\mathbf{u}_i (i = 1, 2, 3)$ are trained with HTP and classification loss. Specifically, HTP on global features can be formulated as

$$L_{HTP}^g = \sum_{i=1}^{N_g} \left(\frac{1}{N_t} \sum_{j=1}^{N_t} L_{HTP} \left((\mathbf{u}_i^{(j)}, \mathbf{u}_i^{(j+)}, \mathbf{u}_i^{(j-)}) \right) \right), \quad (8)$$

where N_g and N_t are the numbers of global features and sampled triplets, $\mathbf{u}_i^{(j)}$, $\mathbf{u}_i^{(j+)}$, $\mathbf{u}_i^{(j-)}$ are the feature $\mathbf{u}_i$ extracted from anchor, positive, and negative samples in the j -th triplet, respectively. Classification loss on global features can be formulated as

$$L_{cls}^g = \sum_{i=1}^{N_g} \left(-\frac{1}{N} \sum_{j=1}^N \log \frac{\exp(((\mathbf{W}^i)_{y_j})^T \mathbf{u}_i)}{\sum_{k=1}^C \exp(((\mathbf{W}^i)_k)^T \mathbf{u}_i)} \right), \quad (9)$$

where N , C are the number of input images and identities, and y_j is the ground truth of the j -th input image. $(\mathbf{W}^i)_k$ is the k -th column of the fully connected layer whose input is $\mathbf{u}_i$. Local features $\mathbf{v}_m^n$ are trained only with classification loss. Classification loss on local features is formulated as

$$L_{cls}^l = \sum_{n=2}^{N_b} \sum_{m=1}^n \left(-\frac{1}{N} \sum_{j=1}^N \log \frac{\exp(((\mathbf{W}_m^n)_{y_j})^T \mathbf{v}_m^n)}{\sum_{k=1}^C \exp(((\mathbf{W}_m^n)_k)^T \mathbf{v}_m^n)} \right), \quad (10)$$

where N_b is the number of branches and $(\mathbf{W}_m^n)_k$ is k -th column of the fully connected layer whose input is $\mathbf{v}_m^n$. The final loss is defined as follows:

$$L = \frac{1}{N_{htp}} L_{HTP}^g + \lambda \frac{1}{N_{cls}} (L_{cls}^g + L_{cls}^l), \quad (11)$$

where N_{htp} and N_{cls} are the numbers of features trained with HTP and classification loss and λ is the weight of classification loss. Specifically, we set λ to 2 in the following experiments.

3.6 Discussions

This subsection contains a brief discussion of the proposed modules and some similar existing methods that emphasizes the difference between them. However, our proposed modules and the compared existing methods are designed with different purposes, which means that they can hardly be compared in a fair experimental setting.

Relations between HOP and OBM. The overlapping blocks model (OBM) [4] proposes a multiple overlapping blocks structure to pool features from overlapping regions. The OBM requires multiply scaled horizontal partitions. However, HOP performs on a single scale, which is a light-weight method in the training procedure for its relatively fewer fully connected layers.

Relations between IBASM and non-local block. In some ways, the IBASM can be regarded as a variation of the non-local block [51]. The IBASM differs from the non-local block as follows: (1) The IBASM takes two input features while the non-local block takes one input feature. The IBASM performs non-local operations on two features. This modification helps the model refine one input feature with the consideration of the other input feature. (2) The IBASM produces two output features corresponding to two refined input features containing reciprocal information from each other. ‘‘Encoder-decoder attention’’ layers [46] and pair-wise non-local operation [10] both take two input features to compute non-local operations and produce one output feature corresponding to one refined input feature containing the reciprocal information from the other input feature. The IBASM is the first module to build relations between two branches for Re-ID.

Relations between IBAM and PS-MCNN. The IBAM has some similarities with the partially shared multi-task convolutional neural network (PS-MCNN) [1] because both are designed to make

Table 1. Dataset Statistics

	DukeMTMC-reID	CUHK03	Market-1501
Train (IDs/Images)	702/16,522	767/7,365	751/12,936
Gallery (IDs/Images)	1,110/17,661	700/5,332	751/15,913
Query (IDs/Images)	702/2,228	700/1,400	750/3,368
Cameras	8	6	6

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

4 EXPERIMENTS

In this section, we first describe three datasets and evaluation protocols in our experiments. Then, implementation details are introduced. Next, we compare the retrieval accuracy of the HMBN with state-of-the-art methods on these three datasets. Finally, we carry out ablation studies on DukeMTMC-reID, CUHK03-Labeled, and CUHK03-Detected to verify the effectiveness of each component. Parameter analysis and visualization are also included.

4.1 Datasets and Evaluation Protocols

We conduct experiments on three popular Re-ID datasets: DukeMTMC-reID [31, 65], CUHK03 [21], and Market-1501 [63]. Dataset statistics are shown in Table 1. **DukeMTMC-reID** is a subset of the DukeMTMC dataset [31] for Re-ID. It is also one of the largest datasets in the Re-ID task. It contains 36,411 images of 1,812 pedestrians captured from 8 high-resolution cameras. There are 702 pedestrians with 16,522 images randomly divided for a training set. The other 702 pedestrians are included in the testing set, in which 2,228 images and 17,661 images are included in the query set and gallery set, respectively. The remaining 408 pedestrians are distractors. **CUHK03** is a relatively small dataset compared with DukeMTMC-reID. It has 1,467 pedestrians with 14,097 images captured by 6 cameras on the CUHK campus. Both manually annotated and DPM-detected bounding boxes are provided, which are denoted as CUHK03-Labeled and CUHK03-Detected. In this article, we use the setting of both of them. **Market-1501** is another large dataset, which is collected by the Deformable Part Model (DPM) detector [8] from 6 cameras. The whole dataset is separated into a training set with 12,936 of 751 pedestrians and a testing set including 3,368 query images of 750 pedestrians and 15,913 gallery images of 751 pedestrians.

We report Cumulative Matching Characteristics (CMCs) at rank-1, and the mean Average Precision (mAP) with the single-shot setting on all candidate datasets.

4.2 Implementation Details

The proposed HMBN is implemented using the Pytorch [28] framework on a single NVIDIA GTX 1080 Ti GPU. The weights of ResNet-50 [11] pretrained on ImageNet [7] are adopted to initialize parameters of the HMBN.

In the training phase, we resize the input images to 384×128 . Then, they are augmented by random horizontal flip, normalization, and random erasing [67]. The total training phase takes

Table 2. Comparison of HMBN with State-of-the-Art Methods for DukeMTMC-reID Dataset

Methods		DukeMTMC-reID	
		Rank-1	mAP
S	MLFN [2] (CVPR2018)	81.00	62.80
	PCB+RPP [41] (ECCV2018)	83.30	69.20
	HPM [9] (AAAI2019)	86.60	74.30
	OSNet [68] (ICCV2019)	88.60	73.50
	BoT [26] (CVPRW2019)	86.40	76.40
M	PSE [32] (CVPR2018)	79.80	62.00
	HA-CNN [23] (CVPR2018)	80.50	63.80
	CA ³ Net [24] (ACM MM2018)	84.60	70.20
	CAMA [56] (CVPR2019)	85.80	72.90
	HOReID [49] (CVPR2020)	86.90	75.60
	MGN [50] (ACM MM2018)	88.70	78.40
	PISNet [62] (ECCV2020)	88.80	78.70
A	DuATM [36] (CVPR2018)	81.82	64.58
	Manacs [47] (ECCV2018)	84.90	71.80
	AANet-50 [44] (CVPR2019)	86.42	72.56
	CASN [64] (CVPR2019)	87.70	73.70
HMBN		89.86	79.68
HMBN (RK)		92.19	90.44

The top results are in bold. “RK” means re-ranking.

500 epochs. The initial learning rate is set to $2e-4$, and then decays to $2e-5$, $2e-6$ after 320, 380 epochs. An Adam optimizer is used to update the weight parameters with weight decay $5e-4$. The batch size is set to 32, in which each identity contains 4 images. The margin in the HTP is set to 1.2 in the following experiments. It should be emphasized that the current experimental setting is slightly different from the conference version in two ways. (1) Different batch sizes —The batch size is set to 32 in this journal version while the batch size in the conference version is 16. (2) Different machines —Experiments from the conference and journal versions are conducted on two independent machines with different hardware and software, that is, CPUs, GPUs, and operation systems. Based on different experimental settings, the results reported in this article are slightly different from the conference version.

In the testing phase, the input images are resized to 384×128 and augmented only by normalization. All dimension-reduced (256-dim) global and local features are concatenated as the final embedding vector of a pedestrian.

4.3 Comparison with State-of-the-Art Methods

More than 10 existing state-of-the-art methods are compared with our proposed HMBN on DukeMTMC-reID, CUHK03, and Market-1501 datasets in Table 2, Table 3, and Table 4, respectively. We separate these compared methods into three groups: single-branch methods (S), multi-branch methods (M), and attention-based methods regardless of the number of branches (A).

DukeMTMC-reID. Our proposed HMBN achieves the best result of a Rank-1 accuracy of 89.86% and a mAP of 79.68% on the DukeMTMC-reID dataset. We should emphasize the following. (1) The gaps between HMBN and single-branch methods (MLFN [2], PCB+RPP [41], HPM [9],

Fig. 9. Example results of our HMBN on the DukeMTMC-reID dataset. Given a query image, the top-10 ranking list is presented. **Correct** and **incorrect** matches are highlighted green and red, respectively.

OSNet [68], and BoT [26]) demonstrate the effectiveness of the multi-branch structure, for example, the HMBN surpasses the BoT by Rank-1/mAP = 3.46%/3.28%. (2) Multi-branch methods (PSE [32], HA-CNN [23], CA³Net [24], CAMA [56], HOREID [49], MGN [50], and PISNet [62]) integrate complementary information (e.g., pose estimation, human parsing results, attribute information) into final pedestrian representations, for example, HOREID [49] aligns local features with key-points estimation. Without injecting prior knowledge, such as attributes or poses, the HMBN exceeds the MGN and achieves the best results in this group, by 1.16% in Rank-1 accuracy and 1.28% in mAP. We argue that these methods ignore the interaction among branches in the multi-branch network. On the contrary, the HMBN builds interaction among branches in higher layers of the network by injecting the IBAM. (3) The HMBN outperforms the CASN [64], which achieves the top result in attention-based methods (DuATM [36], Mancs [47], AANet-50 [44], CASN [64]), by 2.16% in Rank-1 and 5.98% in mAP. Instead of modeling intra-branch contextual dependency in attention-based methods, our designed IBAM builds inter-branch dependency. With the help of re-ranking [66], we achieve a higher result of 92.19% in Rank-1 accuracy and 90.44% in mAP.

Figure 9 shows the top-10 ranking results with four query images on the DukeMTMC-reID dataset. Given a query image, the HMBN can retrieve the correct pedestrian under severe visual recognition problems such as view angle variations, illumination variations, and occlusion.

CUHK03. We report a clear winner case on CUHK03-Labeled and CUHK03-Detected in Table 3. The HMBN achieves the top result of Rank-1 accuracy 78.07%, mAP 75.63% on CUHK03-Labeled and Rank-1 accuracy 75.43%, mAP 73.05% on CUHK03-Detected. The HMBN outperforms the CASN [64] and achieves the best result of all previous existing methods by 4.37% in Rank-1 accuracy, 7.63% in mAP, and 3.93% in Rank-1, 8.65% in mAP, respectively.

Market-1501. As illustrated in Table 4, our proposed HMBN achieves competitive results of 94.86% Rank-1 accuracy and 87.45% mAP on Market-1501. Although the MGN and PISNet outperform the HMBN by 0.84% and 0.74% in Rank-1, respectively, the HMBN clearly exceeds all existing methods in terms of mAP (87.45%).

Table 3. Comparison of HMBN with State-of-the-Art Methods for CUHK03-Labeled and CUHK03-Detected Datasets

Methods		CUHK03-Labeled		CUHK03-Detected	
		Rank-1	mAP	Rank-1	mAP
S	MLFN [2] (CVPR2018)	54.70	49.20	52.80	47.80
	PCB+RPP [41] (ECCV2018)	-	-	63.70	57.50
	HPM [9] (AAAI2019)	-	-	63.90	57.50
	OSNet [68] (ICCV2019)	-	-	72.30	67.80
M	HA-CNN [23] (CVPR2018)	44.40	41.00	41.70	38.60
	CAMA [56] (CVPR2019)	70.10	66.50	66.60	64.20
	MGN [50] (ACM MM2018)	68.00	67.40	66.80	66.00
A	MGCAM [38] (CVPR2018)	50.14	50.21	46.71	46.87
	Manacs [47] (ECCV2018)	69.00	63.90	65.50	60.50
	CASN [64] (CVPR2019)	73.70	68.00	71.50	64.40
HMBN		78.07	75.63	75.43	73.05
HMBN (RK)		84.86	86.48	84.16	82.64

The top results are in bold. “-” means not available, “RK” means re-ranking.

Table 4. Comparison of HMBN with State-of-the-art Methods for Market-1501 Dataset

Methods		Market-1501	
		Rank-1	mAP
S	MLFN [2] (CVPR2018)	90.00	74.30
	PCB+RPP [41] (ECCV2018)	93.80	81.60
	HPM [9] (AAAI2019)	94.20	82.70
	OSNet [68] (ICCV2019)	94.80	94.90
	BoT [26] (CVPRW2019)	94.50	85.90
M	PSE [32] (CVPR2018)	87.70	69.00
	HA-CNN [23] (CVPR2018)	91.20	75.70
	CA ³ Net [24] (ACM MM2018)	93.20	80.00
	CAMA [56] (CVPR2019)	94.70	84.50
	HOReID [49] (CVPR2020)	94.20	84.90
	MGN [50] (ACM MM2018)	95.70	86.90
	PISNet [62] (ECCV2020)	95.60	87.10
A	MGCAM [38] (CVPR2018)	83.79	74.33
	DuATM [36] (CVPR2018)	91.42	76.62
	Manacs [47] (ECCV2018)	93.10	82.30
	AANet-50 [44] (CVPR2019)	93.89	82.45
	CASN [64] (CVPR2019)	94.40	82.80
HMBN		94.86	87.45
HMBN (RK)		95.58	94.21

The top results are in bold. “RK” means re-ranking.

4.4 Ablation Studies

To further verify the effectiveness of each component in the MBN, we present ablation analysis on the DukeMTMC-reID, CUHK03-Labeled and CUHK03-Detected datasets. Parameter analysis and visualization are performed on the DukeMTMC-reID and CUHK03-Labeled datasets.

Table 5. Ablation Studies of HMBN on DukeMTMC-reID, CUHK03-Labeled, CUHK03-Detected Datasets

	Model	HTP	DukeMTMC-reID		CUHK03-Labeled		CUHK03-Detected	
			Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1	B + S1B	$\alpha = 0$	82.23	65.19	53.07	49.31	47.43	44.57
2	B + S2B ($l = 0$)	$\alpha = 0$	85.50	73.12	62.50	60.02	61.14	56.83
3	B + S3B ($l = 0$)	$\alpha = 0$	86.40	75.10	65.93	63.86	61.79	58.70
4	B + S1B + S2B ($l = 0$)	$\alpha = 0$	87.66	76.31	69.36	66.68	64.71	62.29
5	B + S1B + S3B ($l = 0$)	$\alpha = 0$	87.25	76.97	71.07	68.07	65.57	63.92
6	B + S1B + S2B ($l = 0$) + S3B ($l = 0$)	$\alpha = 0$	88.42	78.93	74.29	72.11	70.21	68.76
7	B + S1B + S2B ($l = \text{X}$) + S3B ($l = \text{Y}$)	$\alpha = 0$	88.96	79.22	74.50	72.29	71.36	69.06
8	B + S1B + S2B ($l = \text{X}$) + S3B ($l = \text{Y}$) + IBAM	$\alpha = 0$	89.45	79.64	76.29	73.55	72.29	70.79
9	B + S1B + S2B ($l = \text{X}$) + S3B ($l = \text{Y}$) + IBAM	$\alpha = 0.01$	89.86	79.68	78.07	75.63	75.43	73.05

l means the parameter l of HOP in corresponding branch, α means the parameter α of HTP. **X**, **Y** refers to the recommended hyperparameter for different datasets. (**X**, **Y**) equals (2, 0), (2, 0), (2, 2) for DukeMTMC-reID, CUHK03-Labeled, CUHK03-Detected.

Multi-branch Structure. The effectiveness of uniform partition and multi-branch structure are shown in Table 5 with the comparison of model 1, 2, 3, 4, 5, 6. B + S1B is short for a model including a base module and S1B, B + S1B + S2B ($l = 0$) is a multi-branch model composed of a base module, S1B, S2B, and so forth. For uniform partition in the single branch, the number of horizontal stripes controls the granularity of the local feature. With the increase of number of stripes, the performance boosts as well. However, as the number of stripes further increases, the improvement seems to be marginal but enlarges the model parameters: for example, B + S2B ($l = 0$) outperforms B + S1B in mAP by 7.93% but B + S3B ($l = 0$) outperforms B + S2B ($l = 0$) in mAP by 1.98% on DukeMTMC-reID. For the multi-branch structure, a model with a multi-branch structure is better than each composed branch: for example, B + S1B + S2B ($l = 0$) + S3B ($l = 0$) beats B + S1B, B + S2B ($l = 0$), B + S3B ($l = 0$), B + S1B + S2B ($l = 0$) in Rank-1/mAP by 6.19%/13.74%, 2.92%/5.81%, 2.02%/3.83%, and 0.76%/2.62% on DukeMTMC-reID. As the number of branches further increases, the performance increases as well but marginal: B + S1B + S3B ($l = 0$) outperforms B + S1B in mAP by 11.78% but B + S1B + S2B ($l = 0$) + S3B ($l = 0$) outperforms B + S1B + S2B ($l = 0$) in mAP by 2.62% on DukeMTMC-reID. To keep the balance between high retrieval accuracy and low model parameter, we recommend using B + S1B + S2B ($l = 0$) + S3B ($l = 0$) as the architecture to verify the effectiveness of HOP, IBAM, and HTP.

Effectiveness of HOP. Figure 10 shows the Rank-1 accuracy and mAP change with parameter l in HOP, in which l is the total height of overlapped areas in one stripe from one branch. To simplify the notation, l of HOP in S2B is denoted as l_2 and l of HOP in S3B is denoted as l_3 . In the model with one branch S2B (B + S2B), as illustrated in Figure 10(a), the performance rises as well when l_2 is increased, indicating that HOP is essential. However, the accuracy is not always growing with l_2 . In the model with one branch S3B (B + S3B), as illustrated in Figure 10(b), with the increase of l_3 , the accuracy show the same trend. It is noted that over-increased l helps to cover meaningful local regions between adjacent parts but diminishes the feature learning in fine-grained cue. A proper l helps to achieve a good balance between learning fine-grained information and extracting features in meaningful local regions. As is shown in Table 5, (2, 0), (2, 0), (2, 2) are recommended for (l_2 , l_3) on DukeMTMC-reID, CUHK03-Labeled, and CUHK03-Detected datasets.

Effectiveness of IBAM. The comparison of models 7 and 8 in Table 5 shows the effectiveness of the IBAM. Empirically, the IBAM improves by 0.45%/0.42%, 1.79%/1.26%, and 0.93%/1.73% in Rank-1/mAP on the DukeMTMC-reID, CUHK03-Labeled, and CUHK03-Detected datasets, respectively (model 7 vs. model 8). To verify whether the IBAM builds interaction among branches, we visualize the activations of the last convolutional feature maps from three branches on DukeMTMC-reID

Fig. 10. Parameter analysis for l in S2B and S3B. (a) Rank-1 and mAP changes with l_2 on DukeMTMC-reID. (b) Rank-1 and mAP changes with l_3 on DukeMTMC-reID. (c) Rank-1 and mAP changes with l_3 on CUHK03-Labeled. (d) Rank-1 and mAP changes with l_3 on CUHK03-Labeled.

Fig. 11. Visualization results of activations from three branches in the HMBN. The top and bottom two input images are come from DukeMTMC-reID and CUHK03-Labeled, respectively. For each input image, the activations from the first row and second row are returned from the HMBN without and with the IBAM.

and CUHK03-Labeled in Figure 11. It is noted that (1) S1B performs the Re-ID task at the global level, which is likely to ignore detailed information. The IBAM helps S2B and S3B interact with S1B, and S1B learns the global information with the consideration of local information. Red eclipses mark the detailed information ignored in S1B from the HMBN without the IBAM but considered by S1B from HMBN with the IBAM. (2) S2B and S3B perform the Re-ID task at the part level, which

Fig. 12. Parameter analysis for α in the HMBN on the DukeMTMC-reID and CUHK03-Labeled datasets.

fails in learning consecutive local regions. S2B and S3B successfully keep an eye on larger local areas with the injection of IBAM. Yellow eclipses mark the consecutive local region ignored in S2B or S3B from the HMBN without the IBAM but considered by S2B or S3B from the HMBN with the IBAM.

Effectiveness of Harder Triplet Loss. The HMBN is trained with classification loss and HTP. Traditional triplet loss can be seen in the special case of HTP if $\alpha = 0$. As shown in Table 5, the HTP outperforms traditional triplet loss in Rank-1/mAP by 0.41%/0.04%, 1.78%/2.08%, and 3.14%/2.26% on the DukeMTMC-reID, CUHK03-Labeled, and CUHK03-Detected datasets, respectively (model 8 vs. model 9). Parameter analysis for α in HTP is illustrated in Figure 12. Parameter α controls the scale factor of anchor-to-positive distance and anchor-to-negative distance. We can see that the performance of THE HMBN is sensitive to α , and parameter α helps the HMBN achieve a higher retrieval accuracy when $\alpha = 0.01$.

5 CONCLUSIONS

In this article, we pay attention to widely used multi-branch methods with different stripes and propose a harmonious multi-branch network for Re-ID with HTP. Unlike previous methods that design more branches or more stripes for extracting coarse-to-fine pyramid representations, we analyze how to improve feature learning in a single branch and build interaction among different branches. For feature learning in a single branch, we design the HOP to enhance representational capability in meaningful local regions while extracting fine-grained information. For the interaction among different branches, we incorporate the IBAM to refine representation within a single branch by integrating information from other branches. In addition, we analyze the deficiencies of the commonly applied triplet loss and propose the generalized triplet loss, namely, HTP. Our HTP optimizes triplets in a harder manner, further reducing and enlarging intra-class and inter-class variations. Each component is verified thoroughly in extensive ablation experiments. In addition, the HMBN achieves superior performance compared with state-of-the-art Re-ID methods. In the future, we will explore our idea of harmonious multi-branch learning in more computer vision tasks, that is, image retrieval.

REFERENCES

- [1] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. 2018. 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*. IEEE, Salt Lake City, USA, 4290–4299.
- [2] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. 2018. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 2109–2118.
- [3] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. 2019. Spatial-temporal attention-aware learning for video-based person re-identification. *IEEE Transactions on Image Processing* 28, 9 (2019), 4192–4205.

- [4] Yipeng Chen, Cairong Zhao, and Tianli Sun. 2019. 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*. IEEE, Long Beach, USA, 0–0.
- [5] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Las Vegas, USA, 1335–1344.
- [6] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, USA, 8307–8316.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, USA, 248–255.
- [8] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Anchorage, USA, 1–8.
- [9] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. 2019. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. AAAI Press, Honolulu, USA, 8295–8302.
- [10] Zhihang Fu, Yaowu Chen, Hongwei Yong, Rongxin Jiang, Lei Zhang, and Xian-Sheng Hua. 2019. Foreground gating and background refining network for surveillance object detection. *IEEE Transactions on Image Processing* 28, 12 (2019), 6077–6090.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Las Vegas, USA, 770–778.
- [12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*
- [13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- [14] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, USA, 9317–9326.
- [15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. VRSTC: Occlusion-free video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, USA, 7183–7192.
- [16] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 7132–7141.
- [17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Honolulu, USA, 4700–4708.
- [18] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 1062–1071.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 25. MIT Press, Stateline, USA, 1097–1105.
- [20] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 369–378.
- [21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, USA, 152–159.
- [22] Wei Li, Xiatian Zhu, and Shaogang Gong. 2017. Person re-identification by deep joint learning of multi-loss classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, San Francisco, USA, 2194–2200.
- [23] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 2285–2294.
- [24] Jiawei Liu, Zheng-Jun Zha, Hongtao Xie, Zhiwei Xiong, and Yongdong Zhang. 2018. CA 3 Net: Contextual-attentional attribute-appearance network for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, Seoul, South Korea, 737–745.
- [25] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. 2018. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Munich, Germany, 7834–7843.

- [26] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Virtual Seattle, USA, 0–0.
- [27] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. 2018. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 5657–5666.
- [28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. *arXiv:1912.01703*
- [29] Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. 2018. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 3 (2018), 773–786.
- [30] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, Munich, Germany, 650–667.
- [31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*. Springer, Springer, Amsterdam, the Netherlands, 17–35.
- [32] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 420–429.
- [33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Boston, USA, 815–823.
- [34] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. 2018. End-to-end deep Kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 6886–6895.
- [35] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z. Li. 2016. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*. Springer, Springer, Amsterdam, the Netherlands, 732–748.
- [36] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 5363–5372.
- [37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*
- [38] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 1179–1188.
- [39] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2017. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Venice, Italy, 3960–3969.
- [40] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV’18)*. Springer, Munich, Germany, 402–419.
- [41] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV’18)*. Springer, Munich, Germany, 480–496.
- [42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Boston, USA, 1–9.
- [43] Zengming Tang and Jun Huang. 2020. Branch interaction network for person re-identification. In *Proceedings of the Asian Conference on Computer Vision (ACCV’20)*. Springer, Virtual Kyoto, Japan, 322–337.
- [44] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. 2019. AANet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, USA, 7134–7143.
- [45] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. 2016. A Siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*. Springer, Amsterdam, the Netherlands, 135–153.

- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. MIT Press, Long Beach, USA, 5998–6008.
- [47] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2018. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. Springer, Munich, Germany, 365–381.
- [48] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Honolulu, USA, 3156–3164.
- [49] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Virtual Nashville, USA, 6449–6458.
- [50] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, Seoul, South Korea, 274–282.
- [51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 7794–7803.
- [52] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. 2018. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 8042–8051.
- [53] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. Springer, Munich, Germany, 3–19.
- [54] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 2119–2128.
- [55] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. PMLR, Lille, France, 2048–2057.
- [56] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. 2019. 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*. IEEE, Long Beach, USA, 1389–1398.
- [57] Hantao Yao, Shiliang Zhang, Dongming Zhang, Yongdong Zhang, Jintao Li, Yu Wang, and Qi Tian. 2017. Large-scale person re-identification as retrieval. In *IEEE International Conference on Multimedia and Expo (ICME'17)*. IEEE, Hong Kong, China, 1440–1445.
- [58] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. Deep metric learning for person re-identification. In *22nd International Conference on Pattern Recognition*. IEEE, Stockholm, Sweden, 34–39.
- [59] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *International Conference on Machine Learning*. PMLR, Long Beach, USA, 7354–7363.
- [60] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. 2020. ResNeSt: Split-attention networks. *arXiv:2004.08955*
- [61] Shun Zhang, Jia-Bin Huang, Jongwoo Lim, Yihong Gong, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang. 2020. Tracking persons-of-interest via unsupervised representation adaptation. *International Journal of Computer Vision* 128, 1 (2020), 96–120.
- [62] Shizhen Zhao, Changxin Gao, Jun Zhang, Hao Cheng, Chuchu Han, Xinyang Jiang, Xiaowei Guo, Wei-Shi Zheng, Nong Sang, and Xing Sun. 2020. Do not disturb me: Person re-identification under the interference of other pedestrians. In *European Conference on Computer Vision*. Springer, Virtual Glasgow, UK, 647–663.
- [63] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Santiago, Chile, 1116–1124.
- [64] Meng Zheng, Srikrishna Karanam, Ziyang Wu, and Richard J. Radke. 2019. Re-identification with consistent attentive Siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, USA, 5735–5744.
- [65] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, Venice, Italy, 3754–3762.

- [66] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Honolulu, USA, 1318–1327.
- [67] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, New York, USA, 13001–13008.
- [68] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-scale feature learning for person re-identification. *arXiv:1905.00953*
- [69] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. 2019. Discriminative feature learning with consistent attention regularization for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Long Beach, USA, 8040–8049.
- [70] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. 2017. Point to set similarity based deep feature learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Honolulu, USA, 3741–3750.

Received June 2021; accepted November 2021