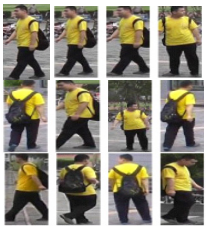
Opto-Electronic Engineering
光 电 工 程

Article

2020 年，第 47 卷，第 11 期

DOI: 10.12086/oee.2020.190628

基于多分区注意力的行人重识别方法



薛丽霞，朱正发，汪荣贵，杨　娟*

合肥工业大学计算机与信息学院，安徽 合肥 230009

摘要：行人重识别是计算机视觉中一项具有挑战性和实际意义的重要任务，具有广泛的应用前景。背景干扰、任意变化的行人姿态和无法控制的摄像机角度等都会给行人重识别研究带来较大的阻碍。为提取更具有辨别力的行人特征，本文提出了基于多分区注意力的网络架构，该网络能同时从全局图像和不同局部图像中学习具有鲁棒性和辨别力的行人特征表示，能高效地提高行人重识别任务的识别能力。此外，在局部分支中设计了一种双重注意力网络，由空间注意力和通道注意力共同组成，优化提取局部特征。实验结果表明，该网络在 Market-1501、DukeMTMC-reID 和 CUHK03 数据集上的平均精度均值分别达到 82.94%、72.17%、71.76%。

关键词：行人重识别；局部特征；双重注意力网络；深度神经网络

中图分类号：TP391.4；TP301.6　　**文献标志码：**A

引用格式：薛丽霞，朱正发，汪荣贵，等. 基于多分区注意力的行人重识别方法[J]. 光电工程，2020，**47**(11): 190628

Person re-identification by multi-division attention

Xue Lixia, Zhu Zhengfa, Wang Ronggui, Yang Juan*

College of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China

Abstract: Person re-identification is significant but a challenging task in the computer visual retrieval, which has a wide range of application prospects. Background clutters, arbitrary human pose, and uncontrollable camera angle will greatly hinder person re-identification research. In order to extract more discerning person features, a network architecture based on multi-division attention is proposed in this paper. The network can learn the robust and discriminative person feature representation from the global image and different local images simultaneously, which can effectively improve the recognition of person re-identification tasks. In addition, a novel dual local attention network is designed in the local branch, which is composed of spatial attention and channel attention and can optimize the extraction of local features. Experimental results show that the mean average precision of the network on the Market-1501, DukeMTMC-reID, and CUHK03 datasets reaches 82.94%, 72.17%, and 71.76%, respectively.

Keywords: person re-identification; local features; dual attention network; deep neural networks

Citation: Xue L X, Zhu Z F, Wang R G, *et al.* Person re-identification by multi-division attention[J]. *Opto-Electronic Engineering*, 2020, **47**(11): 190628

收稿日期：2019-10-17； 收到修改稿日期：2020-03-10

作者简介：薛丽霞(1976-)，女，博士，副教授，硕士生导师，主要从事智能视频处理与分析、视频大数据与云计算、智能视频监控与公共安全、嵌入式多媒体技术等的研究。E-mail：xixzzm@163.com

通信作者：杨娟(1983-)，女，博士，讲师，硕士生导师，主要从事视频信息处理、视频大数据处理技术、深度学习与二进神经网络理论与应用等的研究。E-mail：yangjuan6985@163.com

1 引 言

行人重识别任务是在跨摄像头中进行指定行人检索，即对于给定一个行人图像，在多台不同角度、没有视野重叠覆盖的摄像头不同时间段拍摄的行人图像数据库中找到该行人目标。随着监控摄像头在公共区域的大量普及，行人重识别技术受关注程度越来越高，在视频内容检索、视频监控以及智能安防等领域已成为一项核心技术。

解决行人重识别任务的常见方法是从特征提取和度量学习两个方面考虑，首先是学习特征向量对行人图像进行特征表示[1-3]，然后通过度量学习准确的度量图像间的相似性[4-8]。传统的行人重识别方法[4]依赖于手工提取行人特征，再进行相似性度量。但由于监控摄像头的分辨率低以及光照、角度等影响，同一个行人在不同摄像机中可能有很大差异，而不同的行人在外观上可能很相似，这使得手工提取特征很难应用到复杂的现实环境中。

近年来，由于深度学习强大的拟合和表征能力，在计算机视觉任务中都取得了出色的竞争表现[9-10]。通过深度卷积神经网络提取的行人特征比以前的手工编码特征具有更高的泛化能力，使得应用深度学习模型来解决行人重识别任务的准确率提高到了一个新的水平。与此同时，带有标签的行人重识别数据集(如CUHK03[11]、Market-1501[12]和 DukeMTMC-reID[13])的出现，为深度模型的训练在数据层面上提供了可行性。

在最初的基于深度学习的行人重识别研究方法中，研究者们主要使用最直接的从行人图像的整体上提取识别特征方法，即通过网络模型在图像上提取行人全局特征向量用以相似性检索[14-15]。虽然，这类方法在各大数据集上较传统方法取得了突破性进展，但是由于只考虑到整体图像中捕获最显著的外观特征来表示不同行人之间的区别，忽略了一些不显著或不频繁的细节信息，从而导致获取的行人特征不足以准确表示复杂场景中的行人身份信息。

因此，行人重识别研究并不仅仅只关注在全局特征上，也开始逐渐研究局部特征，并证明了结合局部特征的行人图像表示是最有效的[16-17]。局部特征提取的关键是对整体图像进行分割及局部区域的精确定位。目前，效果较好的行人重识别方法在提取局部特征的功能上有所不同，大致可以概括为两种：一是根据行人固有的身体结构，将图像在水平方向上分割成若干条条带，在其上提取局部特征[18-20]；二是利用人体姿态估计和骨架关键点等先验知识来预测行人身体结构信息以裁剪出更准确的局部区域[21-22]。但是上述方法都有各自的缺陷。第一种水平分块方法没有考虑局部之间不对齐问题；第二种局部划分方法需要一个额外的骨架关键点或者姿态估计的模型，这会带来额外的姿态估计误差。

同时，研究者们还提出了针对行人重识别的注意力深度学习模型[23-24]。类似于人类视觉处理的注意力机制，有选择性地倾向于注意图像中的行人部分，而忽略其他不感兴趣的区域，有助于解决行人重识别问题。Li 等[25]为展现不同层次的注意力机制感知和学习行人特征，提出了 HA-CNN 网络模型，用来学习互补的区域级硬注意力特征和像素级软注意力特征，增强柔和和兼容性程度，优化处理未对齐图像的特征提取技术。Liu 等[26]提出了一个多级别注意力模型 HydraPlus-Net，将注意力机制映射到不同的特征层，使其挖掘多级别特征信息。上述此类方法大都将区域注意力网络合并嵌入到深层的行人重识别模型中。大多数现有的行人重识别工作集中于使用全身图像进行注意力学习，忽视了从行人身体的局部部位学习的注意力特征。同时，全局注意力更多地集中在全局信息区域上，这往往会抑制或忽略行人身体部位周围的局部信息区域，从而导致当人的图像出现较大的姿态变化、严重的失调、局部遮挡等情况时，重识别效果不佳。

因而，本文重新考虑了如何利用局部特征和注意力机制学习到更加具有识别力的行人特征，设计了一个基于局部注意力的行人重识别网络，即多分区注意力网络模型(multi-division attention network，MDA)。图 1 展示了 MDA 网络的整体框架图。该网络主要从两个方面解决上述提及的困难：一是同时学习全局特征和不同分块数量的局部特征，兼顾行人的整体信息和局部细节信息，优化深度学习中的行人重识别；二是设计了一种双重局部注意力网络，分为空间注意力网络 SANet 和通道注意力网络 CANet，二者在功能上形成很强的互补性，提高行人重识别模型的性能。

2 方 法

给定 n 张训练图像 $I=\{I_i\}_{i=1}^{n}$，由不重叠的摄像头拍摄的 n_{id} 个有区别的行人，相应的身份标签为 $L=\{L_i\}_{i=1}^{n}$ (其中 $L_i\in[1,\cdots,n_{id}]$)，目标是学习一个行人重识别模型，该模型能够对给定的查询图像重新识别出该图像。如图 1 所示，整个网络包括三个部分：主干

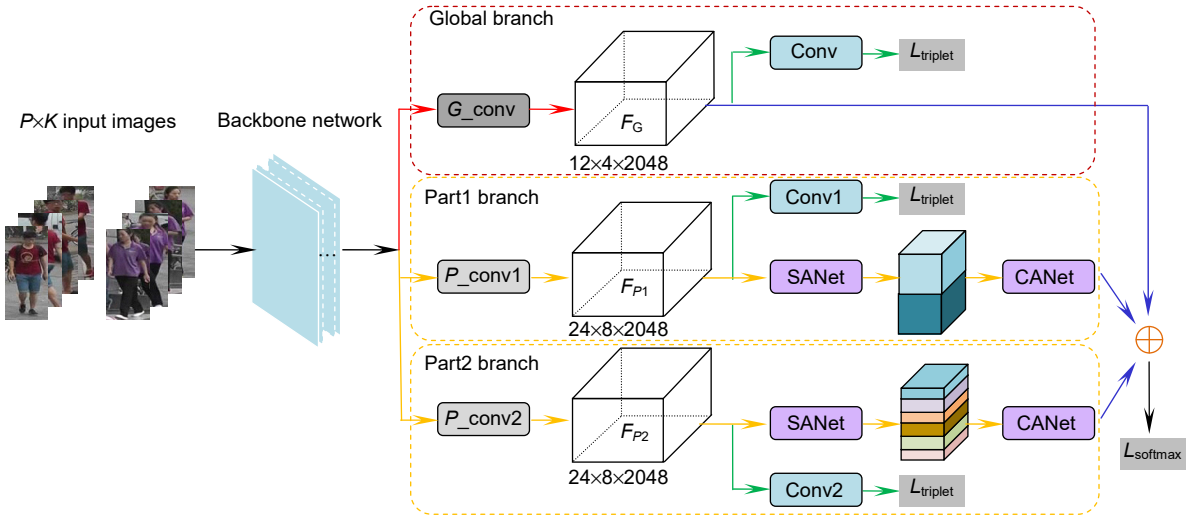


图 1 MDA 模型框架的概述

Fig. 1 Overview of our proposed MDA network for person re-identification

网络(backbone network)，全局和两个局部分支(global、Part1 and Part2 branch)以及双重局部注意力网络(dual local attention network，DLA)。首先，将训练图像传递到主干网络中，以提取深层的视觉特征。在这里，使用 ResNet50 的 conv 4_1 层输出的特征图作为后续网络分支的嵌入输入。然后，将提取的特征图经过 *G*_conv 输入到全局分支以提取全局特征，同时经过 *P*_conv1 和 *P*_conv2 输入到两个局部分支中，分别经过 SANet 和 CANet，输出局部注意力特征。最后，将全局特征和局部注意力特征进行融合，其结果经过一个全连接层后计算 softmax 损失，同时全局分支和两个局部分支都会计算 triplet 损失。

2.1 网络子结构

1) 主干网络

本文中主干网络采取的是 ResNet50 网络，借助其在行人重识别领域的优势。为适应网络模型中全局和局部特征融合需求，在网络层上都对 ResNet50 原始版本进行简化改动，以及只采用 conv 4_1 层之前的网络部分，后面连上 *G*_conv 和 *P*_conv 分别进入三个独立的分支。*G*_conv 和 *P*_conv 结构大体相同，都是由 conv 5_*x* 层组成。不同点在于为了获得更高粒度的特征图，*P*_conv 删除了 conv 5_1 位置的下采样操作，而 *G*_conv 不做任何改变。

这样，进入局部分支的特征图尺寸比全局分支的特征图尺寸大一倍，会强制这两个局部分支学习更高粒度的特征和更多的细节信息。*G*_conv 和 *P*_conv 模块独立训练，不共享参数，最小化过度拟合的风险。主干网络的更多细节参数展示在表 1 中。

这样，对于输入训练图像 I，首先使用主干网络提取图像的特征，该特征可以表示为

$$F_I = B\_\mathrm{cnn}(I) \quad , \tag{1}$$

其中：F_I 为主干网络的 conv4_1 层的输出，B_cnn 表示整个主干网络结构。然后，将特征图 $F_I \in R^{24\times8\times1024}$ 输入到全局分支中，经过 *G*_conv 层后可以表示为

$$F_G = G\_\mathrm{conv}(F_I), F_G \in R^{12\times4\times2048} \quad 。 \tag{2}$$

同理，将特征图 F_I 输入到局部分支经过 $P\_\mathrm{conv}i(i\in[1,2])$ 层后可以表示为

$$F_{P_i} = P\_\mathrm{conv}i(F_I), F_{P_i} \in R^{24\times8\times2048} \quad , \tag{3}$$

其中：$P_i(i\in[1,2])$ 分别表示两个不共享权重参数的 Part1 和 Part2 局部分支，F_{P_i} 为各 $P\_\mathrm{conv}i$ 层的输出。

2) 全局分支

全局分支的目的是从整个行人图像中学习最优的全局层次的特征表示。如图 1 所示，F_G 将会经过两条线路：一条是直接与局部分支的输出做特征融合；另一条用于计算 triplet 损失。在计算损失这条线路上，F_G 会经过 conv 层(由核为 12×4 最大全局池化层、1×1 卷积层、BN 层和 ReLU 层组成)，可以表示为

$$y_G = \mathrm{ReLU}(W_G F_G + b_G), y_G \in R^{1\times1\times256} \quad , \tag{4}$$

其中：W_G、b_G 为卷积层的参数权重和偏置。式(4)的目的是将 2048 维的特征降维成 256 维 y_G，用于计算 triplet 损失。

3) 局部分支

MDA 网络模型设计有两个局部分支 $P_i(i\in[1,2])$，目的是综合学习不同分块数量下行人图像区域中最具

表 1 Backbone network 结构

Table 1 Backbone network structure

	Layer name	Share	Patch size	Output size
Backbone	Input data	-	-	384×128,3
	Conv2d	Yes	7×7, 64	192×64,64
	BN	Yes	64	192×64,64
	Max pool	Yes	3×3, 64	96×32, 64
	Conv2_*x*	Yes	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	96×32, 256
	Conv3_*x*	Yes	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	48×16, 512
	Conv4_1	Yes	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}$	24×8, 1024
*G*_conv	Conv5_*x*	No	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	12×4, 2048
*P*_conv *i* (*i*∈[1,2])	Conv5_*x*	No	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	24×8, 2048

有判别能力的视觉特征信息。这两个局部分支具有相同的网络结构，它们的不同点在于其中分块数量 N 不同。在每个局部分支中，为使得每个分块都有合适大小的局部图像区域，将输入特征图映射统一地分成 N 条水平方向的条带分块，独立地学习局部特征表示。本文选取 N=2 和 N=6，即 P_1 分支有条带分块$\{k_{11}, k_{12}\}$和 P_2 分支有条带分块$\{k_{21}, \ldots, k_{26}\}$。

与全局分支的结构相似，局部分支中 F_{P_i} 也会经过两条线路：一条会经过 DLA 网络，学习局部注意力特征 $C_{P_i^k}$。在这里 P_i^k 表示第 P_i 局部分支中第 k 个分块，$C_{P_i^k}$ 表示其学习到的注意力特征，用来和全局分支的 F_G 做特征融合。另一条用于计算 triplet 损失，F_{P_i} 会经过 conv_i(i∈[1,2]) 层(由核为 24×8 最大全局池化层、1×1 卷积层、BN 层和 ReLU 层组成)，可以表示为

$$y_{P_i}=\mathrm{ReLU}(W_{P_i}F_{P_i}+b_{P_i}), y_{P_i}\in R^{1\times1\times256}, i\in[1,2] \quad , \tag{5}$$

其中：W_{P_i}、b_{P_i} 为卷积层的参数权重和偏置。式(5)的目的是将两个局部分支中 2048 维的特征降维成 256 维 y_{P_i}，用于计算 triplet 损失。

4) 特征融合

特征融合部分采用并行策略来实现，同时考虑到融合过程可能会对局部注意力特征向量的某些特定维度产生过大的响应，加入一个非线性激活函数来平衡局部注意力特征响应。融合特征可以定义为

$$f_{P_i^k}=\sigma(C_{P_i^k})\otimes F_G, f_{P_i^k}\in R^{1\times1\times256} \quad , \tag{6}$$

其中：

$$\sigma(x)=\frac{1}{(1+\mathrm{e}^{-x})}$$

是一个 Sigmoid 函数，目的是将 $C_{P_i^k}$ 的值标准化为(0,1)，$\otimes$ 表示每个元素对应相加，$f_{P_i^k}$ 表示第 P_i^k 个分块的注意力特征和全局特征经过特征融合模型后的输出。最后，融合特征 $f_{P_i^k}$ 会经过一个全连接层，将其维度降维成 n_{id} 维，即与训练数据集中行人身份数量相同，输出的结果分别计算 softmax 损失。

2.2 双重局部注意力

在行人重识别任务中引入注意力机制，是希望通过类似于人脑注意力的机制，利用很小的感受野处理图像中特定区域，降低了计算的维度，同时网络学习图像中高响应区域的特征表示，使得该部分区域的特征得到增强。受此思想的影响，本文在局部分支中提出利用注意力机制进一步提取出更具有分辨能力的局部特征，在具体实现过程中，运用双重局部注意力模型，即空间注意力(spatial attention network，SANet)

和通道注意力(channel attention network，CANet)。

SANet 约束每个位置上的所有通道，使得最终输出一个空间维度一致的注意力特征图，如图 2 所示。SANet 模块的输入是 P_conv 的输出 $F_{P_i}(i\in[1,2])$，首先经过 Drop 层，Drop 层的数学描述如下：

$$U=\frac{\sum_{m=1}^{c}1\sim h;1\sim w;m}{w},U\in R^{24\times8\times10}\quad,\qquad(7)$$

式中：U 为 Drop 层的输出，h、w、c 为特征图 F_{P_i} 的高、宽和通道数。可以看出，Drop 层是专门为后续卷积层的输入大小而设计的对参数进行压缩，使得参数量只有原来的 $1/w$。实际上，这种跨通道压缩是合理的，因为在模型设计中，所有通道共享相同的空间注意特征图。

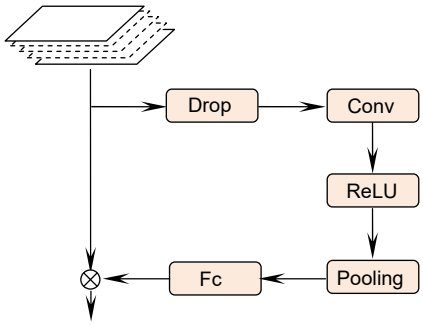


图 2 SANet 网络结构

Fig. 2 Detailed network of the SANet subnet

然后经过一个卷积层和 ReLU 层，其目的是用于提取空间注意力特征，可以表示为

$$S=\mathrm{ReLU}(W_S U+b_S),S\in R^{2\times3\times10}\quad,\qquad(8)$$

其中：W_S、b_S 为卷积层的参数权重和偏置。接着特征图 S 被输入到池化层和全连接层得到 θ_{part}，最后 θ_{part} 和 F_{P_i} 进行双线性插值得到 SANet 的输出特征向量 S_{P_i}。

CANet 是约束每一个通道上的所有特征值，最后输出长度与通道数目相同的一维向量作为特征加权输出。整个 CANet 有两个支路：多通道分支和直连分支，如图 3 所示。

图 3 中①表示提取每个通道注意力特征的网络结构。②和③分别表示 multi_channel 分支和 shortcut 分支，$M(x)$和 $S(x)$是两者的输出。squeeze 层的输出通道为 c/τ，因此在结构①中存在 c/τ 个分支，最后将所有通道连接在一起。

前面空间注意力网络的输出特征向量 $S_{P_i}(i\in[1,2])$ 经过不同数目的切块后，得到 $S_{P_i^k}$ 作为 CANet 的输入，其中 P_i^k 表示第 P_i 局部分支中第 k 个分块，例如本文中 P_1 局部分支会有{k_{11}、k_{12}}两个分块，P_2 局部分支会有{k_{21}，…，k_{26}}六个分块。特征图 $S_{P_i^k}$ 首先经过 squeeze 层(由一个卷积层、BN 层、ReLU 层以及全连接 FC 层组成)，将输入的 2048 维降低成 $2048/\tau$，可被表示为

$$C=\mathrm{ReLU}(W_c S_{P_i^k}+b_c),C\in R^{24\times8\times\frac{2048}{\tau}}\quad,\qquad(9)$$

其中：W_c、b_c 为卷积层的参数权重和偏置，τ 是一个缩放参数，目的是为了减少通道个数从而降低计算量，本文取 $\tau=256$。

在多通道分支的①结构中，对于每个通道都有一个 1×1 卷积层、Sigmoid 激活函数层和全局平均池化层，再将所有 c/τ 的个通道都串接起来。直连分支中只有简单的卷积层。最后，依据下式将多通道分支的输出 $M(x)$和直连分支的输出 $S(x)$计算，得到 CANet 的输出特征向量：

$$C_p=(1+M(x))\times S(x)\quad。\qquad(10)$$

多通道分支中的 Sigmoid 激活函数会导致其结构①输出归一化为 0 到 1 之间，特征图的输出响应变弱，这样多通道叠加结构①会使得最终输出 $M(x)$的特征图每一个点上的值变得很小。因此，式(10)中将 $M(x)$与 1 相加，可以很好地解决降低特征值问题。

2.3 损失函数

为充分发挥多分区注意力网络特征表示学习的能力，使用了两种损失函数(深度学习中最为常见的)，一个是 softmax 损失，另一个是 triplet 损失。给定 n 张训练图像 $I=\{I_i\}_{i=1}^n$，相应的身份标签为 $L=\{L_i\}_{i=1}^n$(其中 $L_i\in[1,\cdots,n_{id}]$)，f_i 为预测层的输入，即各局部注意力特征与全局特征做融合后的输出 $f_{P_i^k}(i\in[1,2],k\in[1,\ldots,6])$，softmax 损失被定义为

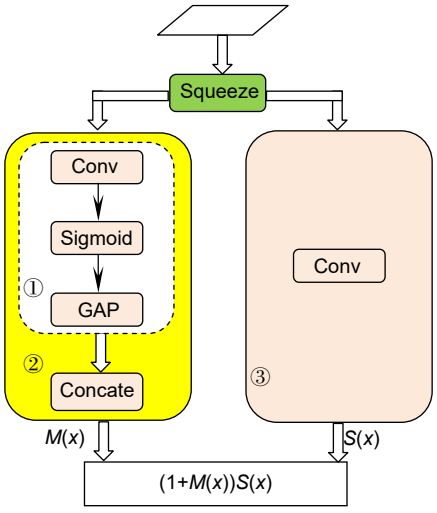


图 3 CANet 网络结构

Fig. 3 Detailed network of the CANet subnet

$$L_{\text{softmax}}=-\sum_{i=1}^{n}\log\frac{\exp(W_{L_i}f_i)}{\sum_{k=1}^{n_{id}}\exp(W_k f_i)}\ , \tag{11}$$

式中：n 为训练集的行人图像数量，n_{id} 是训练集行人身份 ID，W_k 为第 k 类训练 ID 的权重参数。

对所有分支的 256 维特征向量使用 triplet 损失进行训练以提高排序性能。由于传统的三元组随机从训练数据中抽样三张图片，会导致抽样出来的大部分都是简单易区分的样本对，不利于网络学习到更好的表征。为提高网络的泛化能力，采用难样本采样三元组损失(triplet loss with batch hard mining，TriHard 损失)[27]。随机挑选 P 个身份标签的行人，每个行人标签随机挑选 K 张不同的图片，作为一个训练批次，即一个批次含有 $P\times K$ 张图片。对于批次中的每一张图片 a，挑选一个最难的正样本和一个最难的负样本以及基图 a 组成一个三元组，这里的正样本和负样本指的是与基图具有相同或不同身份的行人，最难的正样本指距离最远的正样本，最难的负样本指距离基图最近的负样本。最后会计算 a 和批次中的每一张图片在特征空间的欧氏距离，然后选出与 a 距离最远(最不像)的正样本 p 和距离最近(最像)的负样本 q 来计算三元组损失。定义与图片 a 为相同 ID 的图片集为 α，剩下不同 ID 的图片集为 β，则 TriHard 损失表示为

$$L_{\text{triplet}}=\frac{1}{P\times K}\sum_{a\in P\times K}\left[\gamma+\max_{p\in\alpha}\left\|f_a-f_p\right\|_2-\min_{q\in\beta}\left\|f_a-f_q\right\|_2\right]_+ , \tag{12}$$

其中 γ 是边距超参数，用于控制内部距离和内部距离的差异。

3 实　验

本文在 CUHK03[11]、Market-1501[12] 和 DukeMTMC-reID[13]数据集上进行了充分的实验，结果表明，与现有的网络模型相比，本文提出的模型具有更好的鲁棒性和有效性。本文使用首位命中率(Rank-1)和平均精度均值(mean average precision，mAP)作为行人重识别方法的评价指标。同时，为提高结果所反映性能的准确性，使用了 Re-ranking 评估方法[29]。

3.1 实验细节

整个模型的实现是基于 PyTorch 框架来完成的，使用单个 NVIDIA GEFORCE GTX 1080TI GPU 来训练和测试模型。本文在 ImageNet[28]数据集上预训练 ResNet50 网络的权重参数用来初始化主干网络。对于每个最小训练批次，随机从数据集中选取 P 个身份的行人和从每类行人中随机选取 K 张行人图像。在训练阶段，先将训练图像大小调整为 384×128，然后依概率 p=0.5 进行水平翻转，以及使用 Random Erasing 模拟物体遮挡情况进行数据增强。在测试阶段，只是将图像大小调整为 384×128。本文使用随机梯度下降(SGD)进行优化，冲量为 0.9，l_2 正则化的权重衰减因子设为 5E-4，初始学习率设为 2E-3，每训练 80 个迭代次数下降 10%。在每一个预测层之前使用 dropout 层，dropout 比设置为 0.5。

3.2 实验结果

在 Market-1501 数据集上，将本文提出的方法与 9 种有代表性的方法进行比较，实验结果如表 2 所示。可以看出，本文提出的方法取得了较好的识别效果，mAP 和 Rank-1 分别达到了 82.94%和 94.03%，在使用 Re-ranking 技术后更是达到了 90.27%和 94.98%，进一步提高了识别准确率。在这里选取的比较方法有以下几种：水平分割方法(PCB+RPP[18])和借助行人姿态(Spindle[16]、PDC[22])来完成行人局部特征的提取；行人区域对齐方法提出的(Part-Aligned[24])；全局特征和局部特征的联合学习(AlignedReID[31])；结合行人属性解决行人重识别问题(APR[30])；注意力机制的引入

表 2　Market-1501 数据集实验结果

Table 2　Comparison of results on Market-1501

Methods	Rank-1/%	mAP/%	Methods	Rank-1/%	mAP/%
PCB+RPP[18]	93.80	81.60	HA-CNN[25]	91.20	75.70
Spindle[16]	76.90	64.67	Hydraplus-net[26]	91.80	-
PDC[22]	84.14	63.41	DuATM[32]	91.40	76.60
Part-Aligned[24]	81.00	63.40	Ours	**94.03**	**82.94**
AlignedReID[31]	91.80	79.30	Ours(RK)	**94.98**	**90.27**
APR[30]	87.04	66.89			

"RK" refers to implementing re-ranking[29] operation

(HA-CNN[25]、Hydraplus-net[26]、DuATM[32])。

在图 4 中，显示了某些给定行人图像的前 10 个排序结果。可以看出，即使在只有查询图像 4(a)的背影图时，大多数排名结果也是能够保证准确率的。对于具有相似外观的查询图像 4(b)和 4(c)，由于网络可以提取足够的行人特征信息，因此即使待查询图像中存在不对齐情况，也可以获得良好的识别精度。在查询图像 4(d)中的行人存在严重遮挡和姿态问题，本文提出的方法识别性能不是很好。

检索的图像全部来自 Market-1501 数据集中的图像，而不是同一张相机拍摄的图像。其中具有绿色边框的图像与给定查询图像属于同一行人，而具有红色边框的图像则不属于同一行人。

对于更大的和更具有挑战性的 DukeMTMC-reID 数据集，本文方法的重识别性能也很出色，分别与 5 种行人重识别方法进行了比较，表 3 给出了实验结果，其在 Rank-1 和 mAP 上的性能分别达到了 84.68%和 72.17%，在 Rank-1 指标上比 PCB+RPP 和 HA-CNN 分别高出了 1.38%和 4.18%，在 mAP 指标上比 PCB+RPP 和 HA-CNN 分别高出了 2.97%和 8.37%。在这个目前最具挑战性的数据集上，进一步验证了本文方法的优势。

对于 CUHK03 数据集提供的两种类型的标签，CUHK03-Labeled 表示为手动标记行人边界框，CUHK03-Detected 表示为 DPM[33]检测边界框。本文提出的方法在 CUHK03-Labeled 上的 Rank-1 和 mAP 达到了 75.36%和 71.76%。同时，在 CUHK03-Detected 上的 Rank-1 和 mAP 达到了 73.53%和 65.91%。另外从表 4 中可以观察到 CUHK03-Labeled 和 CUHK03-Detected 之间有明显的差距。这足以证明行人图像标签的标注对行人重识别性能的重要影响，强调了高性能行人检测器的重要性。

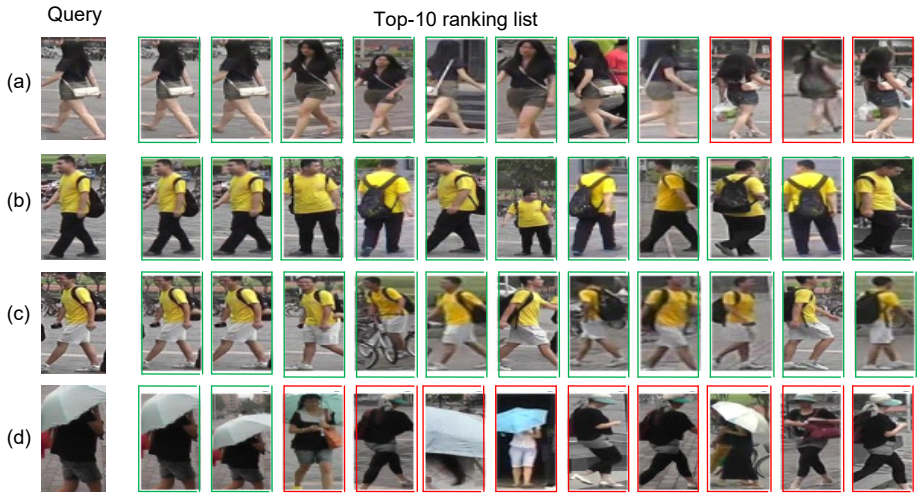


图 4 行人图像前 10 个排序结果

Fig. 4 Top-10 ranking list for some query images

表 3 DukeMTMC-ReID 数据集实验结果

Table 3 Comparison of results on DukeMTMC-ReID

Methods	Rank-1/%	mAP/%
PAN[19]	71.59	51.51
PCB+RPP[18]	83.30	69.20
APR[30]	73.92	55.56
HA-CNN[25]	80.50	63.80
DuATM[32]	81.16	67.73
Ours	**84.68**	**72.17**

表 4 CUHK03 数据集实验结果

Table 4 Comparison of results on CUHK03

Methods	CUHK03-Labeled		CUHK03-Detected	
	Rank-1/%	mAP/%	Rank-1/%	mAP/%
PAN[19]	36.86	35.03	36.29	34.00
PCB+RPP[18]	-	-	63.70	57.50
PDC[22]	**88.70**	-	**78.29**	-
PAN[19]	36.30	34.0	36.90	35.0
Part-Aligned[24]	68.90	-	65.64	-
HA-CNN[25]	44.40	41.0	41.70	38.60
Ours	75.36	**71.76**	73.53	**65.91**

3.3 分析与讨论

为验证本文设计的三分支网络结构的有效性，我们在 Market-1501 数据集上进行了一系列不同分支设置策略的对比实验，图 5 展示了各分支不同组合的比较结果。将各分支的不同组合方法分为两类，一是单个分支(全局分支、Part1 局部分支和 Part2 局部分支)，二是将各分支两两进行自由组合(全局分支和 Part1 局部分支、全局分支和 Part2 局部分支、Part1 局部分支和 Part2 局部分支)。从图中可以直观地看出，一方面，与所有的单个分支实验结果相比，本文提出的分支组合方法效果更好。在仅保留全局分支时，行人重识别结果最差，Rank-1 和 mAP 只达到 84.89%和 69.12%。在单个局部分支的对比实验中，Part2 分支比 Part1 分支在 Rank-1 指标上高 2.65%，在 mAP 指标上高 4.81%，这说明在一定程度上随着局部划分数量的增加，行人重识别效果越来越好。另一方面，基于本文提出的多分区注意力网络，可以增加或减少局部分支的数量，即将三个分支自由组合，则会发现性能显着下降。原因是提出的多分区注意力网络的三个分支之间存在重叠，并且可以引入不同分区之间的相关性，从而可以学习更多差异信息。

我们进一步评估提出的双重局部注意力 DLA 的效果，同样的是在 Market-1501 数据集进行的对比实验，实验结果如图 6 所示。由图可知，在没有加任何注意力机制的全局和局部结合的纯网络 GP 中，Rank-1 和 mAP 分别为 85.33%和 76.40%。在此基础上开始引入注意力机制，结合纯网络 GP 和空间注意力网络 SANet 可以使得 Rank-1 和 mAP 分别达到 89.56%和 80.52%，结合纯网络 GP 和通道注意力网络 CANet 可以使得 Rank-1 和 mAP 分别达到 91.07%和 81.16%。这说明在网络中嵌入注意力机制能提高行人重识别效果，但也反映出加入单一的注意力网络对重识别结果影响不显著。将本文提出的方法与前面三种网络进行比较，证实了提出的双重局部注意力网络能有效地帮助改进行人重识别的性能，以及说明了空间注意力和通道注意力结合的优势和有效性。

4 结 论

行人重识别是一个具有挑战性和实际意义的计算机视觉问题，本文将卷积神经网络和注意力思想引入行人重识别任务，提出了一种基于多分区注意力的网络模型，取得了显著性的进展。与大多数容易产生局部匹配错位问题或利用全局的注意力机制的现有行人重识别方法相比，本文提出的网络能够以端到端的形式，运用双重注意力机制提取具有互补效果的行人局部注意力特征，并与全局特征进行融合，从而获得具有更好的行人重识别效果性能。同时，也注意到尽管均匀分块方法简单有效，但有待改进，在接下来的工作中，可以结合行人姿态估计和骨架关键点分析，提取更加有效的局部特征，继续研究准确率更高、鲁棒性更好的行人重识别模型。

参考文献

[1] Sun R, Fang W, Gao J, *et al*. Person Re-identification in foggy weather based on dark channel prior and metric learning[J]. *Opto-Electronic Engineering*, 2016, **43**(12): 142–146.
孙锐, 方蔚, 高隽. 暗通道和测度学习的雾天行人再识别[J]. 光电工程, 2016, **43**(12): 142–146.

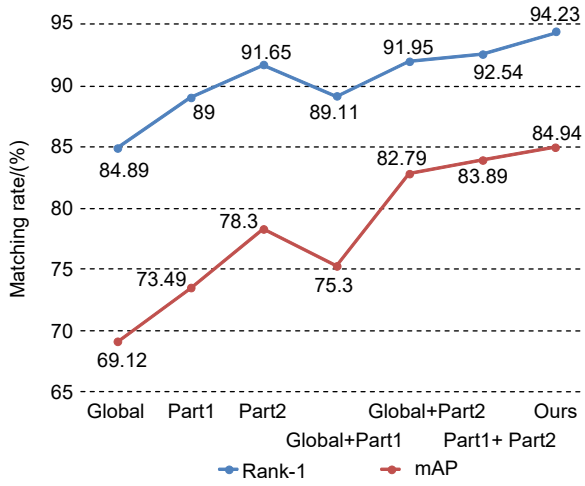


图 5 不同分支组合比较结果图

Fig. 5 Comparison of different branch combination

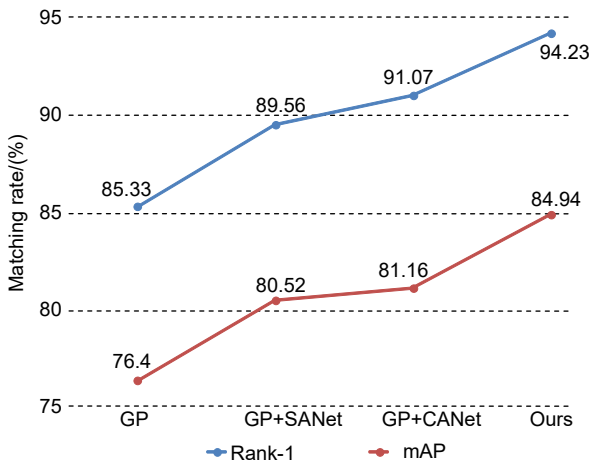


图 6 DLA 效果图

Fig. 6 Evaluations on how DLA enhances person re-identification

[2] Su C, Zhang S L, Yang F, *et al*. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping[J]. *Pattern Recognition*, 2017, **66**: 4–15.

[3] Matsukawa T, Okabe T, Suzuki E, *et al*. Hierarchical gaussian descriptor for person Re-identification[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016: 1363–1372.

[4] Zhao R, Ouyang W L, Wang X G. Person Re-identification by salience matching[C]//*2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 2013: 2528–2535.

[5] Chen D P, Yuan Z J, Hua G, *et al*. Similarity learning on an explicit polynomial kernel feature map for person re-identification[C]//*2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015: 1565–1573.

[6] Sun Y F, Zheng L, Deng W J, *et al*. SVDNet for pedestrian retrieval[C]//*2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 3820–3828.

[7] Yang X, Wang M, Tao D C. Person Re-identification with metric learning using privileged information[J]. *IEEE Transactions on Image Processing*, 2018, **27**(2): 791–805.

[8] Zhang L, Xiang T, Gong S G. Learning a discriminative null space for person Re-identification[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016: 1239–1248.

[9] Liu H, Peng L, Wen J W. Multi-occluded pedestrian real-time detection algorithm based on preprocessing R-FCN[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 180606.
刘辉, 彭力, 闻继伟. 基于改进 R-FCN 的多遮挡行人实时检测算法[J]. 光电工程, 2019, **46**(9): 180606.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//*Proceedings of the 25th International Conference on Neural Information Processing System*, Red Hook, NY, United States, 2012, **25**: 1097–1105.

[11] Li W, Zhao R, Xiao T, *et al*. DeepReID: deep filter pairing neural network for person Re-identification[C]//*2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014: 152–159.

[12] Zheng L, Shen L Y, Tian L, *et al*. Scalable person Re-identification: a benchmark[C]//*2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015: 1116–1124.

[13] Zheng Z D, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person Re-identification baseline in Vitro[C]//*2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 3774–3782.

[14] Sudowe P, Spitzer H, Leibe B. Person attribute recognition with a jointly-trained holistic CNN model[C]//*2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, Chile, 2015: 329–337.

[15] Cheng D Q, Tang S X, Feng C C, *et al*. Extended HOG-CLBC for pedstrain detection[J]. *Opto-Electronic Engineering*, 2018, **45**(8): 180111.
程德强, 唐世轩, 冯晨晨, 等. 改进的 HOG-CLBC 的行人检测方法[J]. 光电工程, 2018, **45**(8): 180111.

[16] Zhao H Y, Tian M Q, Sun S Y, *et al*. Spindle net: person Re-identification with human body region guided feature decomposition and fusion[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017: 907–915.

[17] Wei L H, Zhang S L, Yao H T, *et al*. GLAD: global-local-alignment descriptor for pedestrian retrieval[C]//*Proceedings of the 25th ACM International Conference on Multimedia*, California, Mountain View, USA, 2017: 420–428.

[18] Sun Y F, Zheng L, Yang Y, *et al*. Beyond part models: person retrieval with refined part pooling[Z]. arXiv:1711.09349[cs:CV], 2017.

[19] Zheng Z D, Zheng L, Yang Y. Pedestrian alignment network for large-scale person Re-identification[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, **29**(10): 3037–3045.

[20] Cheng D, Gong Y H, Zhou S P, *et al*. Person Re-identification by multi-channel parts-based CNN with improved triplet loss function[C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016: 1335–1344.

[21] Zheng L, Huang Y J, Lu H C, *et al*. Pose-invariant embedding for deep person Re-identification[J]. *IEEE Transactions on Image Processing*, 2019, **28**(9): 4500–4509.

[22] Su C, Li J N, Zhang S L, *et al*. Pose-driven deep convolutional model for person Re-identification[C]//*2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 3980–3989.

[23] Li D W, Chen X T, Zhang Z, *et al*. Learning deep context-aware features over body and latent parts for person Re-identification[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017: 7398–7407.

[24] Zhao L M, Li X, Zhuang Y T, *et al*. Deeply-learned part-aligned representations for person Re-identification[C]//*2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 3239–3248.

[25] Li W, Zhu X T, Gong S G. Harmonious attention network for person Re-identification[C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 2285–2294.

[26] Liu X H, Zhao H Y, Tian M Q, *et al*. HydraPlus-Net: attentive deep features for pedestrian analysis[C]//*2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 350–359.

[27] Hermans A, Beyer L, Leibe B. In Defense of the triplet loss for person Re-Identification[Z]. arXiv: 1703.07737[cs:CV], 2017.

[28] Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database[C]//*2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009: 248–255.

[29] Zhong Z, Zheng L, Cao D L, *et al*. Re-ranking person Re-identification with k-reciprocal encoding[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017: 3652–3661.

[30] Lin Y T, Zheng L, Zheng Z D, *et al*. Improving person Re-identification by attribute and identity learning[Z]. arXiv: 1703.07220[cs:CV], 2017.

[31] Zhang X, Luo H, Fan X, *et al*. AlignedReID: surpassing human-level performance in person Re-identification[Z]. arXiv: 1711.08184[cs:CV], 2017.

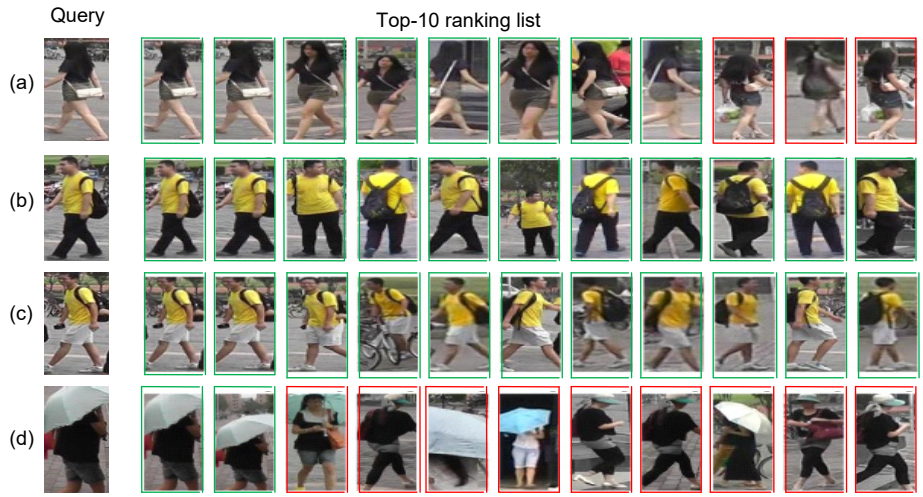
[32] Si J L, Zhang H G, Li C G, *et al*. Dual attention matching network for context-aware feature sequence based person Re-identification[C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 5363–5372.

[33] Felzenszwalb P F, McAllester D A, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//*2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008: 1–8.

Person re-identification by multi-division attention

Xue Lixia, Zhu Zhengfa, Wang Ronggui, Yang Juan*

College of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China



Top-10 ranking list for some query images

Overview: With the popularity of surveillance cameras in public areas, person re-identification has become more and more important, and has become a core technology in video content retrieval, video surveillance, and intelligent security. However, in actual application scenarios, due to factors such as camera shooting angle, complex lighting changes, and changing pedestrian poses, occlusions, clothes, and background clutter in person images. It makes even the same person target have significant differences in different cameras, which poses a great challenge for person re-identification research. Therefore, in this paper we propose a research method based on deep convolutional networks, which combines global and local person feature and attention mechanisms to solve the problem of person re-identification. First, unlike traditional methods, we use ResNet50 network to initially extract person image features with more discriminating ability. Then, according to the person inherent body structure, the image is divided into several bands in the horizontal direction, and it is input into the local branch of the built-in attention mechanism to extract the person local attention features. At the same time, the global image is input to the global branch to extract the person global features. Finally, the person global features and local attention features are fused to calculate the loss function. In the network, in order to better extract the person local features, we design two local branches to segment the person images into different numbers of local area images. With the increase of the number of blocks, the network will learn more detailed and discriminative local features in each different local area, and at the same time, it can filter irrelevant information in local images to a large extent by combining the attention mechanism. Our proposed attention mechanism can make the network focus on the areas that need to be identified. The output person attention features usually have a stronger response than the non-target areas. Therefore, the attention networks we design include spatial attention networks and channel attention networks, which complement each other to learn the optimal attention feature, thereby extracting more discriminative local features. Experimental results show that the method proposed in this paper can effectively improve the performance of person re-identification.

Citation: Xue L X, Zhu Z F, Wang R G, *et al*. Person re-identification by multi-division attention[J]. *Opto-Electronic Engineering*, 2020, **47**(11): 190628

* E-mail: yangjuan6985@163.com