北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：基于多尺度联合学习的行人重识别
作者：谢彭宇，徐新
DOI：10.13700/j.bh.1001-5965.2020.0445
收稿日期：2020-08-24
网络首发日期：2020-10-16
引用格式：谢彭宇，徐新．基于多尺度联合学习的行人重识别．北京航空航天大学学报.
https://doi.org/10.13700/j.bh.1001-5965.2020.0445

基于多尺度联合学习的行人重识别

谢彭宇[1]，徐新[1,2,3]✉

（1. 武汉科技大学 计算机科学与技术学院，武汉 430065；2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室，武汉 430065；3. 上海交通大学 电子信息与电气工程学院，上海 200240）

*通信作者 E-mail：xuxin0336@163.com

摘要 现有的行人重识别方法主要关注于学习行人的局部特征来实现跨摄像机条件下的行人辨识。然而在人体部件存在运动或遮挡、背景干扰等行人数据非完备条件下，会导致行人局部辨识信息丢失概率的增加。针对这个问题，本文提出了一种多尺度联合学习方法来对行人辨识特征进行精细化表达。该方法包含三个分支网络，分别提取行人的粗粒度全局特征、细粒度全局特征和细粒度局部特征。其中粗粒度全局分支通过融合不同层次的语义信息来增强全局特征的丰富性；细粒度全局分支通过联合全部局部特征，在对全局特征进行细粒度描述的同时学习行人局部部件间的相关性；细粒度局部分支则通过遍历局部特征来挖掘行人非显著性的信息以增强局部特征的鲁棒性。为了验证提出方法的有效性，我们在 Market-1501，DukeMTMC-ReID 和 CUHK03 三个公开数据集上开展了对比实验，实验结果表明我们的方法取得了最佳性能。

关键词 行人重识别；多尺度；联合学习；多分支网络；深度学习

中图分类号 V221+.3；TB553 文献标志码 A DOI: 10.13700/j.bh.1001-5965.2020.0445

Multi-scale joint learning for Person Re-Identification

XIE Pengyu[1], XU Xin[1,2,✉]

（1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China；

2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China;

3. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China）

*Tel.: 18971577148 E-mail: xuxin0336@163.com

Abstract Existing person re-identification approaches mainly focus on learning person local features to match a specific pedestrian across different cameras. However, local features may face great challenges in real scenarios, such as human body motion and occlusion, clustered background, etc. This paper presents a multi-scale joint learning method to extract the fine-grained person feature. This method consists of three subnets, i.e. coarse-grained global feature extraction subnet, fine-grained global feature extraction subnet, and fine-grained local feature extraction subnet. The coarse-grained global feature extraction subnet enhances the diversity of the global feature by fusing semantic information at different levels. The fine-grained global feature extraction subnet aims to learn the correlation among local components of a pedestrian. The fine-grained local feature extraction subnet enhance robustness by traversing local features. Extensive experiments have been conducted to evaluate the performance of the proposed method against state-of-the-art methods on Market-1501, DukeMTMC-ReID, and CUHK03 person re-identification datasets.

Key words Person Re-Identification; multi-scale; joint learning; multi-branch network; deep learning

近年来我国安防形势日趋严峻，各类大案需要对嫌疑目标进行跨摄像机的检索、分析、比对。行

收稿日期：2020-08-24

基金项目：国家自然科学基金（U1803262，61602349，61440016）

作者简介：谢彭宇 男，硕士研究生。主要研究方向：计算机视觉、行人重识别。徐新 男，博士，教授，博士生导师。主要研究方向：计算机视觉、机器学习、行人重识别。

Fund： National Natural Science Foundation of China (U1803262，61602349，61440016)

网络首发时间：2020-10-16 16:43:56 网络首发地址：https://kns.cnki.net/kcms/detail/11.2625.V.20201016.1507.002.html

人重识别作为视频侦查的关键技术以处理跨摄像机的行人外观差异为基础，旨在从多摄像机条件下检索特定身份的行人。因此探索具备跨摄像机行人检索能力的行人重识别方法具有重要意义。

随着深度学习[1]的迅速发展，行人重识别方法对行人的表征能力得到了极大提升。然而实际监控场景中的行人姿态、遮挡、背景等方面存在着极大差异，给现有行人重识别方法带来了极大挑战。因此现有行人重识别方法主要关注于提取行人的局部部件特征，进而对行人进行精细化表示。这些方法主要分为两类：第一类方法利用人体姿态估计模型，来分别对各个行人部件进行精细化表达。基于此，Zhao 等人[2]设计了一个树形融合网络，通过选取各个部件最显著的特征进行融合，以增强部件级特征的鉴别性。然而这类方法首先需要对图像进行预处理，增加了方法的复杂性。并且这类方法高度依耐于姿态估计模型的鲁棒性，错误的估计会大大影响识别的结果。第二类方法则假定行人图像已对齐，进而将特征图裁剪为条状或者分块。Sun 等人[3]通过把网络输出的特征图进行水平划分，从而学习不同条状区域的局部特征。此类方法通过提取行人的局部部件特征实现了精细化的行人特征表示。然而这些方法忽略了部件之间的联系，缺乏行人结构性信息。在摄像机拍摄角度受限、人体部件存在运动或遮挡等行人数据非完备条件下，会导致行人局部辨识特征信息丢失概率的增加。

针对这个问题，近年来研究者开始尝试结合行人的局部和全局特征。例如，Zheng 等人[4]和 Wei 等人[5]在使用姿态估计模型提取局部特征的基础上，结合全局特征来避免模型误检对识别精度的影响。Zheng 等人[6]和 Fu[7]等人通过金字塔模型对特征图进行不同尺度的划分，从而获得多尺度的部件信息。Wang 等人[8]则通过一个全局分支和两个局部分支，对行人进行多粒度描述。这些方法通过结合离散的局部特征与全局特征，能提高行人辨识的性能。然而，实际监控场景中的遮挡、背景等因素会对行人信息带来干扰，进而降低行人的全局特征和局部部件特征的辨识性能。

有鉴于此，本文提出了一种多尺度联合学习方法来对行人辨识特征进行精细化表达。该方法结合了行人的粗粒度全局特征、细粒度全局特征和细粒度局部特征。其中粗粒度全局分支通过融合不同层次的语义信息来增强全局特征的丰富性；细粒度全局分支通过联合全部局部特征，在对全局特征进行细粒度描述的同时学习行人局部部件间的相关性；细粒度局部分支则通过遍历局部特征来挖掘非显著性信息增强行人局部特征的鲁棒性。本文的贡献主要包括三个方面：

（1）提出了一种多尺度联合学习方法，该方法结合了行人的粗粒度全局特征、细粒度全局特征和细粒度局部特征；

（2）结合细粒度全局特征和细粒度局部特征能够有效学习行人局部部件间的相关性并挖掘非显著信息，从而提高遮挡、背景差异等条件下的行人辨识性能；

（3）通过在三个行人重识别数据集上的实验测试,我们综合比较了提出方法与 12 种目前的主流模型的性能。

本文第 1 节回顾现有的行人重识别方法；第 2 节详细介绍多尺度联合学习方法的主要步骤；第 3 节通过实验对比文中方法与现有主流方法的性能；第 4 节是总结。

1 相关工作

由于在真实场景下背景[9]，姿态，照明[10]，视角，相机[11]等条件变化很大，行人重识别在计算机视觉中是一项十分具有挑战性的任务。图 1 展示了在真实场景下受遮挡的行人图像。以前，人们通过手工特征，例如颜色、HOG[12]等，通过 XQDA[13]或 KISSME[14]来学习最佳的相似度量。然而，传统的手工特征描述能力有限。近年来随着深度学习的兴起，深度网络的特征学习已成为行人重识别的常见做法。Zheng 等人[15]提出的(IDE)把 ResNet-50[16]做为骨干网络，将行人重识别当成分类问题，通过 ID-Loss 训练。在此之后，一系列基于深度学习的行人重识别方法被提出。最近一些工作意识到行人局部特征有助于行人精细化表示。

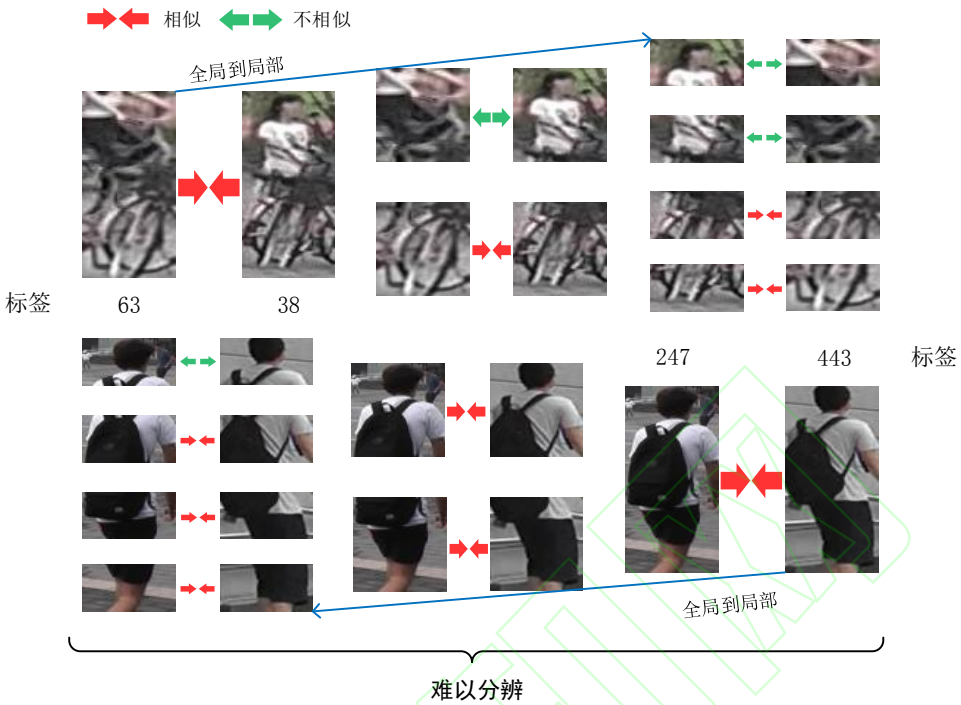

图 1 真实场景下受遮挡的相似行人图像
Fig.1 Occluded pedestrian image in real scene

1.1 局部特征

通过我们总结，基于局部特征主要分为以下两种方式：第一种方式，通过预定义区域对行人进行划分，基于人体姿态估计的方法获得身体部件。Zhao 等人[2]根据姿态估计模型定位的部件区域，对图像进行部件分割，以提取部件级特征。Su 等人[17]则通过学习各个部件级特征的权重，以增强局部特征的鲁棒性。不同于上述方法直接对行人图像进行分割。Suh 等人[18]和 Xu 等人[19]则利用双流网络来实现行人部件匹配，其中上流网络用来获取不同行人部件的特征区域，并引导下流网络来增强指定区域的关注度。Saquib Sarfraz 等人[20]引入了行人的姿态信息进行训练，以获得有鉴别性的特征。然而这种方法由于姿态估计和行人重识别数据集之间的间隙，其精度很大程度上依赖估计器的性能。第二种方式，假设行人已经对齐根据人体结构信息进行划分。一些方法[21]通过把网络输出的特征图进行水平划分获得局部特征，Sun 等人[3]提出了 Part-based Convolutional Baseline(PCB)以获得精细化描述。为了避免行人不对齐，导致各个部件之间产生误匹配。Luo 等人[22]基于动态规划的思想，通过计算各局部特征之间的最短路径，对行人部件进行对齐。Sun 等人[23]则提出了部件感知模型，避免由于部件被遮挡产生的噪声。我们的方法属于此类，与其他方法不同，我们考虑了身体部位之间的相关性，而不是直接的使用局部特征。

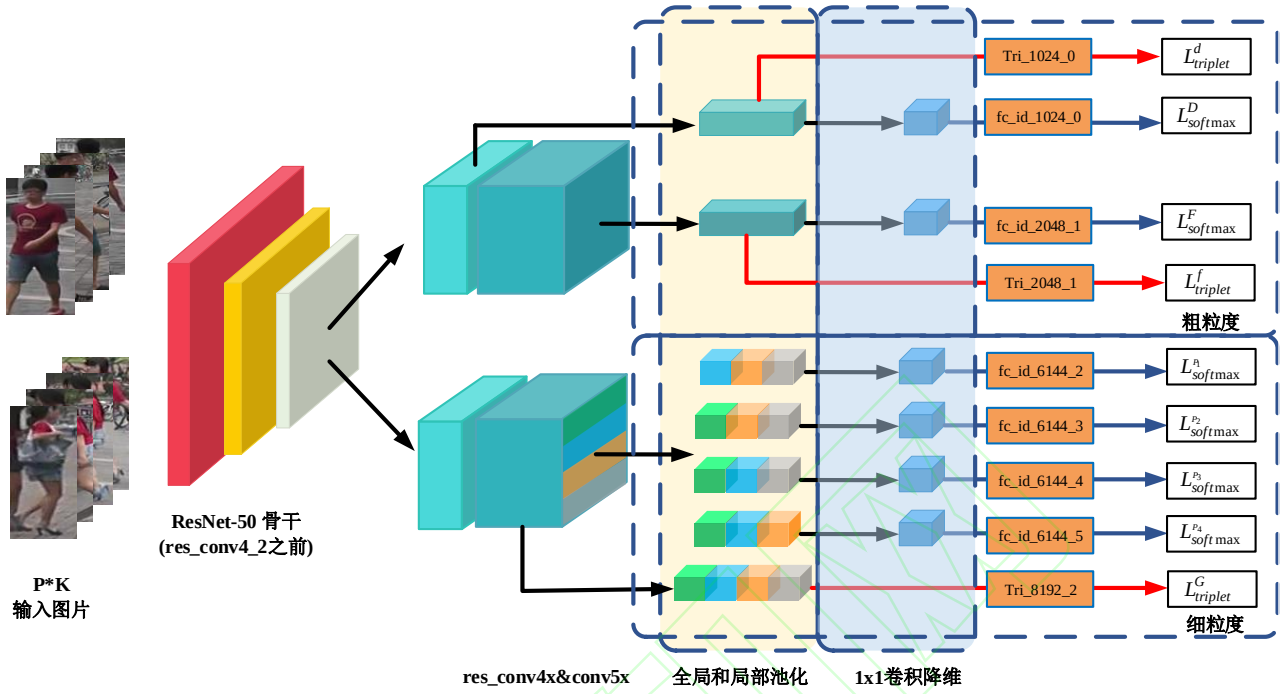

图 2 多尺度联合学习网络框架
Fig.2 Multi-scale joint learning network framework

1.2 联合特征

虽然局部特征能对行人进行精细化表示，但无法对行人显著性特征进行描述。还有一些[24,25]方法联合行人的局部和全局特征，增强行人特征的完整性。Zheng 等人[4]和 Wei 等人[5]通过提取行人较粗部件特征并联合全局特征提高模型的鲁棒性。Li 等人[24]设计了一个注意力模块以增强联合特征的鉴别性。基于 PCB，Wang 等人提出了 MGN[8]通过多分支网络，将全局特征和局部特征相结合，对行人进行多粒度描述。Zheng 等人[6]和 Fu 等人[7]则基于金字塔模型，通过不同尺度的水平划分来学习多尺度的局部和全局特征。

2 多尺度联合学习方法

多尺度联合学习网络由粗粒度全局分支、细粒度全局分支和细粒度局部分支构成。粗粒度全局分支用于增强全局特征的丰富性。而细粒度全局分支在对全局特征精细化描述的同时又学习了相邻部件之间的相关性。细粒度局部分支则通过遍历组合局部特征，加强学习局部特征之间的非显著性信息。图 2 展示了多尺度联合学习网络的框架结构，对其框架和各个模块的详细说明如下。

2.1 粗粒度全局分支

给定一组图像 $X=\{a_1,a_2,\ldots,a_k\}$ 是监控系统中摄像头所捕获的人，其中 k 是图像的数量。我们使用 ResNet-50[16]作为多尺度联合学习的骨干网络。不同于其它方法只提取网络的高级特征来对行人进行描述。我们考虑到网络的不同阶段的输出所带来的语义差异，通过融合不同层次的语义，来增强行人全局特征的丰富性。我们的做法如下，对于粗粒度全局分支，我们分别将 res_conv4 和 res_conv5 层的特征图进行全局平均池化(GAP)和全局最大池化(GMP)操作，分别得到全局特征 $d_{\max,avg}$ $f_{\max,avg}$ 之后将其数值相加得到特征向量 D 和 F：

$$D=d_{\max}+d_{avg} \tag{1}$$

$$F = f_{\max} + f_{avg} \tag{2}$$

对于 $d_{\max,avg}$　$f_{\max,avg}$ 我们将这 4 个特征向量分别通过三元组损失进行训练。特征向量 D 和 F 我们分别通过 1*1 的卷积减少至 512 维，并通过 softmax 函数进行训练。

2.2 细粒度全局分支

通过对特征图进行水平划分，以获得行人的精细化描述是一种常见的方式。在现有的做法中，只对每个局部特征单独的进行学习，并将其连接起来，以产生行人的描述。虽然这种方式可以使行人获得更精细化的表示，但由于行人图像中存在不对齐的现象，容易使各个部件之间产生误匹配。更重要的是，由于各个局部特征是独立存在的，虽然获得了精细化的行人描述，但是缺少了行人特征的完整性。这会导致在相同部件具有相似属性的不同人难以进行区分，造成误判。为此，我们考虑各个部件之间的相关性，通过对全局特征进行细粒度描述来解决此问题。

具体的，行人图像通过骨干网络可以得到一个大小为 C*H*W 的三维特征张量。其中 C 是特征通道数，H 和 W 分别是特征空间的高度和宽度。我们根据特征图的 H 轴将特征图划分为 n 个部分，每个部分的大小为 C*(H/n) *W，分别对每部分进行平均和最大池化操作。可以得到长度为 C 的特征向量 $g_{\max\ i}(i=1,2,\cdots,n)$ 和 $g_{avg\ i}(i=1,2,\cdots,n)$。我们将 $g_{\max}$ 和 g_{avg} 分别连接起来，得到了长度为 n*C 的向量 $G_{\max}$ 和 G_{avg}。通过将局部特征互相关联，使其在既保证全局特征完整性的同时，又对行人特征进行细粒度描述。对于 $G_{\max}$ 和 G_{avg} 使用三元组损失[26]进行训练。我们通过考虑各部件之间的相关性，以缓解部件之间的误匹配，并增强相似部件之间的判别。

2.3 细粒度局部分支

经过划分的局部特征，通过身体各个部件之间相互联系，可以增强局部特征的鉴别力。虽然细粒度全局分支将两个相邻部件之间相互关联，以学习到相邻部件之间的相关性。但是，对于不相邻的两个部件，是否也存在着潜在的相关性。此外，由于部件相互之间间隔的尺度不同，就会形成不同尺度的局部特征。那么对于局部特征在什么尺度下，能够和全局特征进行有效联合，我们仍需要进一步研究。

我们的做法如下：对于长度为 C 的特征向量 $g_{\max\ i}(i=1,2,\cdots,n)$ 和 $g_{avg\ i}(i=1,2,\cdots,n)$，我们将 n 设置为 4，并分别将每个 $g_{\max\ i}$ 和 $g_{avg\ i}$ 数值相加得到局部特征向量 $g_i(i=1,2,\cdots,n)$。

$$g_i = g_{\max\ i} + g_{avg\ i} \tag{3}$$

为了挖掘不相邻的局部特征之间相关性，我们分别对 g_i 进行丢弃，根据丢弃的 g_i 的不同，可以得到多组包含不同 g_i 的局部特征。现在我们假设每次丢弃 1 个 g_i，并对 g_i 进行遍历的丢弃，那么我们就会获得 4 组由不同的 g_i 组成的局部特征向量 $P_n(n=1,2,\cdots,4)$。

$$P_n = \{g_i \mid i = 1\ldots 4 and i \neq n\} \tag{4}$$

对于每组 P_n 我们将其分别通过多部件相关性进行训练。图 3 展示了细粒度局部分支的丢弃 1 个 g_i 时的示意图。

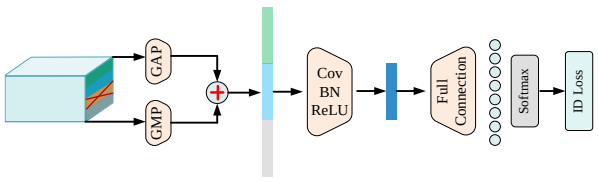

图 3 细粒度局部分支
Fig.3 Fine-grained local branch

对于每组 P_n，由于都缺少了行人的某个关键部件，当我们对 P_n 挖掘相关性时，就会使原本不相邻的两个局部特征之间相互连接。通过利用不相邻部件之间的相关性，从而可以挖掘更多关键但非显著

的潜在信息。基于此，我们进一步探索了，局部尺度特征和全局尺度特征联合判别的有效性。具体的，对于每组 P_n 我们分别通过改变丢弃的尺度进行训练。通过实验我们可以发现，不论丢弃尺度为多少我们的细粒度局部特征都有助于提高精确度。但由于不同数据集行人图像的边界框(bounding box)质量不同，不同尺度的特征嵌入的有效性也不相同，具体实验将在 3.4 部分进行详细阐述。

2.4 损失函数

为了提高该网络学习行人特征表示的辨别能力，我们采用了基于 softmax 的 ID-Loss 和最大三元组损失作为训练阶段的损失函数，这两种方法被广泛的应用于各种行人重识别的方法。

我们首先将行人重识别当作一个分类任务。因此，我们的目标是预测每个人的标签，通过优化分类模型学习行人的具体特征表示。对于第 i 个学习到的特征 f_i，softmax loss 阐述如下：

$$L_{soft\max} = -\sum_{i=1}^{N} \log \frac{\exp(W_{y_i}^T f_i)}{\sum_{k=1}^{C} \exp(W_k^T f_i)} \tag{5}$$

对于 W_k 表示对应于第 k 类的权重。在训练中一个训练批次数量为 N。训练数据集的分类数为 C。不同于传统的 softmax，我们抛弃了线性多分类中的 bias[27]，这有助于获得更好的分类性能。在训练的过程中我们将该损失用于全局特征 D、F，以及局部特征 $P_n(n=1,2,\cdots,4)$。

对于所有的全局特征以及不进行丢弃的局部相关性特征 $\{d_{\max}, d_{avg}, f_{\max}, f_{avg}, G_{\max}, G_{avg}\}$ 通过三元组损失训练来提高性能。我们使用 hard triplet-loss[28]阐述如下：

$$L_{triplet} = -\sum_{i=1}^{P}\sum_{a=1}^{K}[\alpha + \max_{P=1\ldots K} \left\| f_a^i - f_p^i \right\|_2 - \min_{\substack{n=1\ldots K \\ j=1\ldots P \\ j\neq i}} \left\| f_a^i - f_n^i \right\|_2]_+ \tag{6}$$

对于 f_a^i, f_p^i, f_n^i 分别是通过锚(anchor)，积极(positive)和消极(negative)进行采样得到的特征。α 是控制类之间的距离。这里的积极(positive)和消极(negative)是代表和锚(anchor)有相同标签和不同标签的行人。候选的三元组由距离最远的有相同标签的对和最近的有不同标签的对构成。最难的相同标签样本和不同标签样本分别在一个批次中，这个批次有 P 个标签的行人，每个标签有 K 张图片。损失函数鼓励最远的相同标签样本的距离小于最近的不同标签样本的距离。

3 实验

3.1 数据集

Market1501[29]是在清华大学校园中采集的数据集，图像来自于 6 个不同的摄像头，其中有一个摄像头为低分辨率。同时该数据集提供训练集和测试集。训练集包含 12936 张行人图像，测试集包含 19732 张行人图像。图像由检测器自动检测并切割，所以存在一些检测误差。训练数据中一共有 751 人，测试集中有 750 人。所以在训练集中，平均每类(每个人)有 17.2 张训练数据。

CUHK03[30]是在中国香港中文大学采集，行人图像来自 2 个不同的摄像头。该数据集提供机器自动检测和手动检测两个数据集。其中检测数据集包含一些检测误差，更接近实际情况。数据集总共包括 1467 个行人的 14097 张图片，平均每个人有 9.6 张训练数据。最初整个数据集被划分为 20 个随机分组用于交叉验证，但这是针对手工制作的方法而设计的。因此，我们采用提出的新培训/测试协议[31]。

DukeMTMC-reID[32]是在杜克大学内采集，图像来自 8 个不同摄像头，行人图像的边框由人工标注完成。训练集包含 16522 张图像，测试集包含 17661 张图像。训练数据中一共有 702 人，平均每个人有 23.5 张训练数据。

以上三个数据集，我们采用目前行人重识别方法普遍使用的首中准确率(Rank-1)和平均匹配度(mAP)两种评价指标评估方法的性能。所有实验都会使用单一查询方式。此外，为简单起见我们不使用显著改善 mAP 的重新排序算法。

3.2 实施细节

为了从行人图像中获得全局和局部足够的信息，我们将所有图像的大小调整为 288x144，并使用通过 ImageNet[33]分类的 ResNet-50 作为我们的骨干网络。与原始版本不同我们将 res_conv4_1 块之后的后续部分化为两个独立的分支，并与原始的 ResNet-50 共享相似的体系结构。我们设置最后一个卷积层的 stride 从 2 设置为 1，并通过水平移动，标准化和随机擦除来增强训练图像。批量大小设置为 32，P=8，K=4，α=0.3，n=4。我们的模型训练 220 个 epoch。基础学习率设定为 0.03，并在 60 个时期后衰减至 0.003，130 个时期后衰减至 0.0003 直至训练结束。在每个批量中使用具有 0.9 动量的随机梯度下降(SGD)方法以更新参数。我们的方法在 pytorch 框架上实现，并使用单个 GTX1060 GPU 进行训练，所有数据集共享与上述相同的实验设置。

3.3 现有方法的对比实验

在本节中我们将所提出的方法与现有的最先进的方法进行比较，以表示我们对于其它方法的优势，这些方法大多都是最近发布的方法，具体情况如表格 1 所示，针对每个数据集详细说明如下。

3.3.1 Market1501 数据集

我们的方法在该数据集上实现了 95.9%Rank-1 和 89.1%mAP。对比仅仅使用了单一局部特征的 PCB[3]，我们的方法分别在 Rank-1 和 mAP 提高了 3.6%和 11.7%。MGN 虽然考虑了多个分支结构，但是却忽略了局部信息之间的联系，作为该数据集上最好的方法我们分别提高了 0.2%Rank-1，2.2%mAP。

图 4 展示了一些查询前 10 名的结果。我们分别选择了行人被遮挡，背景复杂，图片模糊等复杂条件下情况。第一、二排行人的查询图像在被小包遮挡和背景杂乱的情况下，我们网络仍然可以健壮的表示其判别信息。第三个查询图像是在低分辨率下进行捕获的，丢失了大量精细的信息，但我们的网络却可以通过挖掘局部特征之间的潜在联系，找出正确的行人图像。最后一位行人，背景十分杂乱，身体大部分被自行车遮挡且照片也十分的模糊。但我们的方法仍然有较好的效果。我们可以看到，在 Rank-9 中即使行人出现了较大的不对齐现象，我们仍然可以将其正确的找出，这得益于细粒度局部分支对网络的影响。

表 1 多尺度联合学习方法和其它方法性能对比

Table 1 Performance comparison of the multi-scale joint learning method and other methods

方法		CUHK03				Market1501		DukeMTMC-reID	
		Labeled		Detected					
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
基于部件	IDE[15]	22.0	21.0	21.3	19.7	72.5	46.0	67.7	47.1
	MGN[8]	68.0	67.4	66.8	66.0	95.7	86.9	88.7	78.4
	PCB[3]	61.9	56.8	60.6	54.4	92.3	77.4	81.7	66.1
	Pyramid[6]	78.9	76.9	78.9	74.8	95.7	88.2	89.0	79.0
	GFLF-S[34]	76.6	73.5	74.4	69.6	94.8	88.0	89.3	77.1
基于注意力机制	CASN[35]	73.7	68.0	71.5	64.4	94.4	82.8	87.7	73.7
	M1tB[36]	70.1	66.5	66.6	64.2	94.7	84.5	85.8	72.9
	Mancs[37]	69.0	63.9	65.5	60.5	93.1	82.3	84.9	71.8
	HACNN[24]	44.4	41.0	41.7	38.6	91.2	75.7	80.5	63.9
其它	DPFL[38]	43.0	40.5	40.7	37.0	88.9	73.1	79.2	60.0
	BDB[39]	73.6	71.7	72.8	69.3	94.2	84.3	86.8	72.1
	SVDNet[40]	40.9	37.8	41.5	37.3	82.3	62.1	76.7	56.8
我们的	多尺度联合	80.7	77.0	78.0	73.4	95.9	89.1	90.0	80.4

图 4 Market1501 数据集部分图像查询结果

Fig.4 Market1501 Dataset Partial Image Query Results

表2 多尺度联合学习方法消融实验

Table 2　Ablation Experiment of multi-scale joint learning method

方法	CUHK03				Market1501		DukeMTMC-reID	
	Labeled		Detected					
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
基线	59.1	54.2	55.1	50.2	93.5	82.4	85.3	72.0
基线+CG	69.8	66.1	66.9	62.6	94.8	86.9	87.9	76.7
基线+FG	70.9	67.1	68.2	63.3	95.1	87.3	88.2	77.9
基线+CG+FG	76.4	72.1	73.0	68.4	95.3	88.7	88.7	79.1
基线+CG+ FP1	78.4	75.1	76.0	72.2	95.6	88.7	89.1	78.5
基线+CG+ FP2	78.7	75.2	75.5	71.5	95.6	89.0	89.2	79.6
基线+ FG + FP1	77.3	73.1	76.4	71.8	95.6	88.5	89.5	78.9
基线+ FG + FP2	77.6	74.2	75.0	71.4	95.7	88.8	89.5	79.8
基线+CG+FG+FP1	80.7	77.0	78.0	73.4	95.9	88.8	89.6	79.2
基线+CG+FG+FP2	80.8	76.7	76.0	71.8	95.9	89.1	90.0	80.4

3.3.2 DukeMTMC-reID 数据集

我们可以看到我们的方法仍然在 Rank-1 和 mAP 达到了最好的效果分别为，90.0%和 80.4%。采用金字塔模型的 Pyramid 是最接近我们的方法，但仍然低于 0.7%Rank-1 和 1.4%mAP。和只考虑局部特征的 PCB 相比，我们的模型分别超过 PCB 的 8.3% Rank-1 和 14.3%mAP。

3.3.3 CUHK03 数据集

在新协议下的该数据集是现在最具有挑战的数据集之一，该数据集的边界框用两种不同的方法进行标注分别为 CUHK03 Labeled 和 CUHK03 Detected。该方法在 Labeled 上达到了 80.7%Rank-1 和 77.0%mAP，在 Detected 上达到了 78.0%Rank-1 和 73.4%mAP。我们的方法在 Labeled 上相比于其它方法，提高了 1.8%的 Rank-1。对比同样采用多分支结构 MGN 我们将其原有性能提高了大约 20%，这得益于我们的多尺度联合学习方法，增强了不同粒度联合判别的有效性。

3.4 消融实验

为了验证我们的方法每个组件的有效性以及探索细粒度局部分支丢弃尺度问题。我们使用单一查询模式在 Market-1501，DukeMTMC-reID 和 CUHK03 这三个数据集上设计了几个不同设置的消融实验。各个指标的结果：Rank-1、mAP 显示在表 2 中。其中 CG 表示粗粒度全局分支，FG 表示细粒度全局分支，FP 表示细粒度局部分支，其中 1、2 分别表示丢弃尺度为 1、2。所有结果只更改一个设置，其余设置均与默认设置相同。

首先，我们在表 2 中显示了不同组件对我们模型的消融实验。在表格的前三行我们可以看到，为了验证单个分支的有效性，我们单独使用粗粒度全局分支或细粒度全局分支，都显著提高了行人重识别的性能。特别的在 CUHK03 数据集上，使用细粒度全局分支分别使 mAP 和 Rank-1 的精确度提高了 12.9%，11.8%。第四行当我们将它们联合使用，进一步提高了精确度。为了验证不同粗细粒度分支的组合对模型性能的影响，我们从第五行开始增加了细粒度局部分支，分别对不同分支相互组合进行实验。可以看出不论以哪一种组合方式相比于不使用该分支，都有助于模型性能的提升。由于细粒度局部分支利用了不相邻的部件间的相关性，进一步挖掘了各局部关键但不显著的信息。相比于不同粒度的全局分支，细粒度局部分支更有助于模型鲁棒性的提高，对模型的性能有更加显著的提升。

其次，我们分析了细粒度局部分支丢弃不同尺度的部件情况。我们可以发现，不论我们以何种尺度丢弃部件来挖掘更多关键但非显著的潜在信息。相比于只使用细粒度全局分支，都有助于增强和全局尺度特征联合判别的有效性。特别的，相比于不使用细粒度局部特征，我们在 DukeMTMC-reID 数

据集上进一步提高了 1.3% Rank-1 和 1.3% mAP。根据实验我们可以发现当丢弃尺度为 2 时，在 Market1501，DukeMTMC-reID，CUHK03 Labeled 数据集上更有助于增强局部尺度特征和全局尺度特征联合判别的有效性。对于 CUHK03 Detected 数据集上丢弃尺度为 1 更有效。我们认为在行人图像对齐较好的 Market1501，DukeMTMC-reID，CUHK03 Labeled 数据集。相比于丢弃尺度为 1，丢弃尺度为 2 时由于相关联的行人部件更少的，更有助于模型挖掘更多关键但非显著的信息，使得行人特征更有鲁棒性。而 CUHK03 Detected 数据集由于行人图像存在大量不对齐的现象，减小丢弃尺度有助于避免具有相似部件的不同人容易混淆的问题。

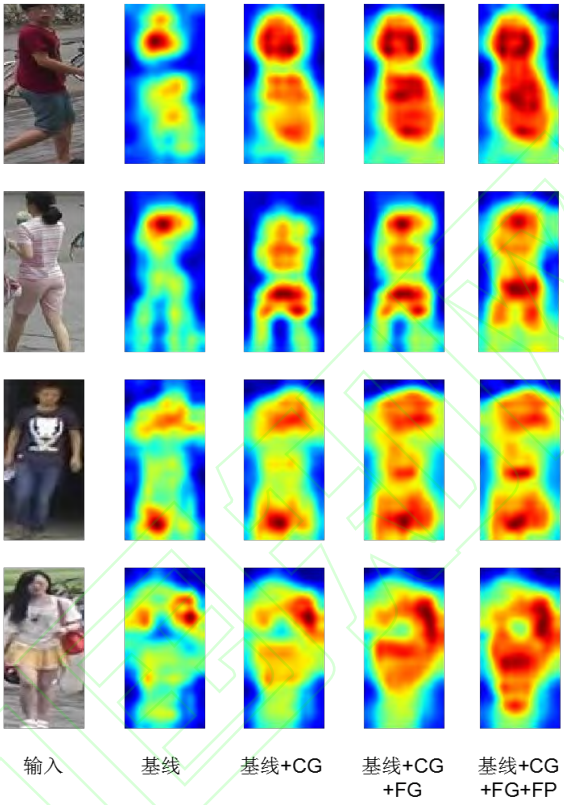

图 5 Market1501 数据集部分图像热力图
Fig.5 Market1501 Dataset Partial Image Heatmap

接着，我们基于热力图可视化了一些图像在不同分支下模型关注的区域情况。如图 5 所示，第 1 列为输入到模型的原始图像，第 2 至 5 列我们以基线为标准并依次递增不同的分支，展示了不同情况下的热力图。可以观察到当添加了粗粒度全局分支后，由于融合了不同层次的语义信息，增强了模型关注区域的丰富性。当进一步增加了细粒度全局分支后，模型将相邻部件之间相互关联，在保证区域丰富性的同时增强了相邻部件之间的关注度以增强相似部件间的辨别。最后当同时利用三个分支时，我们通过利用不相邻部件之间的相关性，挖掘了更多关键但非显著的信息，增强了关注区域的鲁棒性。

4 结论

在本文中，我们提出了多尺度联合学习方法，通过三个分支网络，分别提取行人的粗粒度全局特征、细粒度全局特征和细粒度局部特征，对行人不同粒度下的信息联合学习，使其特征更具有区分性。此外我们通过挖掘各个部件之间不同尺度下的潜在关系，联合全局特征形成了更有鉴别性的行人特征。大量实验证明，我们的方法不仅可以在三个主流的行人重识别数据集上实现最好的结果，而且和现有方法相比可以将性能大幅度提高。

参考文献（References）

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.
[2] ZHAO H, TIAN M, SUN S, et al. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 907–915.
[3] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European Conference on Computer Vision. 2018: 480–496.
[4] ZHENG L, HUANG Y, LU H, et al. Pose-Invariant Embedding for Deep Person Re-Identification[J]. IEEE Transactions on Image Processing, 2019, 28(9): 4500–4509.
[5] WEI L, ZHANG S, YAO H, et al. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval[C]//Proceedings of the 25th ACM international conference on Multimedia. 2017: 420–428.
[6] ZHENG F, DENG C, SUN X, et al. Pyramidal person re-identification via multi-loss dynamic training[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8514–8522.
[7] FU Y, WEI Y, ZHOU Y, et al. Horizontal pyramid matching for person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 8295–8302.
[8] WANG G, YUAN Y, CHEN X, et al. Learning Discriminative Features with Multiple Granularities for Person Re-Identification[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 274–282.
[9] WANG Z, JIANG J, WU Y, et al. Learning sparse and identity-preserved hidden attributes for person re-identification[J]. IEEE Transactions on Image Processing, IEEE, 2019, 29(1): 2013–2025.
[10] ZENG Z, WANG Z, WANG Z, et al. Illumination-Adaptive Person Re-identification[J]. IEEE Transactions on Multimedia, 2020: 1–1.
[11] WANG Z, WANG Z, ZHENG Y, et al. Beyond intra-modality: A survey of heterogeneous person re-identification[J]. arXiv preprint arXiv:1905.10048, 2019.
[12] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2005: 886–893.
[13] LIAO S, HU Y, ZHU X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2197–2206.
[14] KOESTINGER M, HIRZER M, WOHLHART P, et al. Large scale metric learning from equivalence constraints[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2288–2295.
[15] ZHENG L, YANG Y, HAUPTMANN A G. Person Re-identification: Past, Present and Future[J]. arXiv preprint arXiv:1610.02984, 2016.
[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770–778.
[17] SU C, LI J, ZHANG S, et al. Pose-Driven Deep Convolutional Model for Person Re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision.. 2017: 3980–3989.
[18] SUH Y, WANG J, TANG S, et al. Part-Aligned Bilinear Representations for Person Re-Identification[C]//Proceedings of the European Conference on Computer Vision. 2018: 418–437.
[19] XU J, ZHAO R, ZHU F, et al. Attention-Aware Compositional Network for Person Re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2119–2128.
[20] SAQUIB SARFRAZ M, SCHUMANN A, EBERLE A, et al. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 420–429.
[21] ZHENG W-S, LI X, XIANG T, et al. Partial person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4678–4686.
[22] LUO H, JIANG W, ZHANG X, et al. AlignedReID++: Dynamically matching local information for person re-identification[J]. Pattern Recognition, 2019, 94: 53–61.
[23] SUN Y, XU Q, LI Y, et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 393–402.
[24] LI W, ZHU X, GONG S. Harmonious attention network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2285–2294.
[25] LIU X, ZHAO H, TIAN M, et al. Hydraplus-net: Attentive deep features for pedestrian analysis[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 350–359.
[26] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 815–823.
[27] WANG F, XIANG X, CHENG J, et al. Normface: L2 hypersphere embedding for face verification[C]//Proceedings of the 25th ACM international conference on Multimedia. 2017: 1041–1049.
[28] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.
[29] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116–1124.
[30] LI W, ZHAO R, XIAO T, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 152–159.
[31] ZHONG Z, ZHENG L, CAO D, et al. Re-ranking person re-identification with k-reciprocal encoding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1318–1327.
[32] ZHENG Z, ZHENG L, YANG Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3754–3762.
[33] 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. IEEE, 2009: 248–255.
[34] PARK H, HAM B. Relation Network for Person Re-identification[J]. arXiv preprint arXiv:1911.09318, 2019.
[35] ZHENG M, KARANAM S, WU Z, et al. Re-identification with consistent attentive siamese networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 5735–5744.
[36] YANG W, HUANG H, ZHANG Z, et al. Towards rich feature discovery with class activation maps augmentation for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1389–1398.
[37] WANG C, ZHANG Q, HUANG C, et al. Mancs: A multi-task attentional network with curriculum sampling for person re-

identification[C]//Proceedings of the European Conference on Computer Vision. 2018: 365–381.
[38] CHEN Y, ZHU X, GONG S. Person re-identification by deep learning multi-scale representations[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 2590–2600.
[39] DAI Z, CHEN M, GU X, et al. Batch DropBlock Network for Person Re-Identification and Beyond[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 3690–3700.
[40] SUN Y, ZHENG L, DENG W, et al. SVDNet for Pedestrian Retrieval[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3820–3828.