计算机科学
COMPUTER SCIENCE

基于渐进式注意力金字塔的行人重识别方法

张帅宇, 彭力, 戴菲菲

引用本文

张帅宇, 彭力, 戴菲菲.基于渐进式注意力金字塔的行人重识别方法[J]. 计算机科学, 2023, 50(6A): 220200084-8.

ZHANG Shuaiyu, PENG Li, DAI Feifei. Person Re-identification Method Based on Progressive Attention Pyramid [J]. Computer Science, 2023, 50(6A): 220200084-8.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于动态时空神经网络的城市交通流量预测方法

City Traffic Flow Prediction Method Based on Dynamic Spatio-Temporal Neural Network

计算机科学, 2023, 50(6A): 220600266-7. https://doi.org/10.11896/jsjkx.220600266

面向交通流量预测的时空Graph-CoordAttention网络

Spatial-Temporal Graph-CoordAttention Network for Traffic Forecasting

计算机科学, 2023, 50(6A): 220200042-7. https://doi.org/10.11896/jsjkx.220200042

基于多模态特征融合的时间序列异常检测

Anomaly Detection of Time-series Based on Multi-modal Feature Fusion

计算机科学, 2023, 50(6A): 220700094-7. https://doi.org/10.11896/jsjkx.220700094

联合人体姿态估计和多目标跟踪的跨数据集学习

Cross-dataset Learning Combining Multi-object Tracking and Human Pose Estimation

计算机科学, 2023, 50(6A): 220400199-7. https://doi.org/10.11896/jsjkx.220400199

基于改进Yolov4-tiny的轻量型目标检测算法

Lightweight Target Detection Algorithm Based on Improved Yolov4-tiny

计算机科学, 2023, 50(6A): 220700006-7. https://doi.org/10.11896/jsjkx.220700006

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220200084

基于渐进式注意力金字塔的行人重识别方法

张帅宇[1]　彭　力[1]　戴菲菲[2]
1 物联网技术应用教育部工程研究中心(江南大学物联网工程学院)　江苏 无锡 214122
2 台州市产品质量安全检测研究院　浙江 台州 318000
(1206856688@qq.com)

摘　要　针对现有行人重识别算法对行人特征提取不充分,导致算法在行人遮挡、姿态变化等场景下准确度较低的问题,提出了基于渐进式注意力金字塔的行人重识别方法。该方法基于注意力机制设计了一种渐进式的特征金字塔结构,将通道和空间两种注意力模块嵌入特征金字塔结构中,并分别应用在特征的通道和空间两个维度上,通道注意力金字塔聚合骨干网络各层级不同通道维度中值得关注的特征,空间注意力金字塔提取不同空间维度中值得关注的特征。金字塔的每一级都按照"切分-关注-合并"的原则,自底向上不断学习行人特征图在不同切分等级下的注意力,让网络充分挖掘到来自不同通道维度和不同空间维度的关键特征。同时,通过级联结构和可变形卷积实现多层级特征对齐,进一步提高模型的重识别精度。分别在 Market-1501 和 DukeMTMC-reID 两个主流数据集上对该方法进行实验,实验结果表明该方法可以让模型关注到更丰富的行人特征,模型的 Rank-1 指标相比基准网络分别提高了 3.2%和 5.8%,mAP 指标分别提高了 6.8%和 6.6%。

关键词:行人重识别;注意力机制;特征金字塔;特征对齐;池化;度量学习
中图法分类号　TP391.4

Person Re-identification Method Based on Progressive Attention Pyramid

ZHANG Shuaiyu[1],PENG Li[1] and DAI Feifei[2]
1 Engineering Research Center of Internet of Things Technology Applications,School of IoT Engineering,Jiangnan University,Wuxi,Jiangsu 214122,China
2 Taizhou Institute of Product Quality and Safety Testing,Taizhou,Zhejiang 318000,China

Abstract　Aiming at the problem that the existing person re-identification algorithms do not fully extract person features,resulting in low accuracy of the algorithm in scenes such as person occlusion and posture change,a person re-identification method based on progressive attention pyramid is proposed. This method designs a progressive feature pyramid structure based on the attention mechanism,embeds the channel and spatial attention modules into the feature pyramid structure,and applies them to the channel and spatial dimensions of the feature. Channel attention pyramid aggregates the noteworthy features in different channel dimensions at each level of the backbone network,and the spatial attention pyramid extracts the noteworthy features in different spatial dimensions. Each level of the pyramid follows the principle of "split-attend-concat",and continuously learns the person feature map under different segmentation levels from the bottom up. Attention allows the network to fully mine key features from different channel dimensions and different spatial dimensions. At the same time,the multi-level feature alignment is realized through the cascade structure and deformable convolution,which further improves the re-identification accuracy of the model. In this paper,the method is tested on two mainstream datasets,Market-1501 and DukeMTMC-reID,respectively. Experimental results show that this method can allow the model to focus on richer person features. Compared with the baseline network,the Rank-1 index of the model increases by 3.2% and 5.8%,and the mAP index increases by 6.8% and 6.6%,respectively.

Keywords　Person re-identification,Attention mechanism,Feature pyramid,Feature alignment,Pooling,Metric learning

1　引言

行人重识别[1]是利用计算机视觉算法进行跨设备检索特定行人的技术,它通过将给定的目标行人图片与摄像头收集的行人图片相关联,从而实现从多个摄像头中检索出目标行人。这门技术被广泛应用于智能安全和行人搜索等领域,是计算机视觉领域研究的热点之一。

行人重识别主要依靠从给定行人图片中提取到的特征来匹配目标行人,因此提取到的特征的好坏在某种程度上就决定着匹配精度的高低。在早期的行人重识别研究中,诸如 HOG[2],SIFT[3],LOMO 等[4]算法通过手工设计的特征存在局限性,难以适应数据量较大的任务。近年来,随着深度学习的不断发展,利用卷积神经网络提取到的行人特征具有较高的辨识力,从而大幅提高了行人重识别精度。

基金项目:国家自然科学基金(61873112,61802107)
This work was supported by the National Natural Science Foundation of China(61873112,61802107).
通信作者:彭力(jnpengli@outlook.com)

Luo 等[5]提出的 BagTricks 算法，以 ResNet-50 为骨干网络，联合多损失对网络进行训练，其提出的 BNNeck 可以有效加快损失的收敛，但该方法仅从行人图片中提取全局特征，模型在复杂环境和行人被遮挡的场景下鲁棒性较差。提取局部特征可以抑制遮挡带来的不利影响，Sun 等[6]提出了 PCB 方法，该方法将特征图水平分成 6 块，每一块单独预测行人身份；Fan 等[7]提出了一种空间通道并行的网络——SCPNet，该算法将特征图均匀分块后，通过空间维度和通道维度的对应关系，监督网络学习具有辨识力的特征；Wang 等[8]也借鉴均匀分块的思想，提出了 MGN 算法，联合行人的全局特征和不同粒度等级的局部特征对网络进行训练。对特征图均匀分块以提取行人局部特征的方式，虽然在一定程度上缓解了遮挡场景下的信息缺失问题，但这种直接切分的方式仍存在以下两方面问题：一方面，由于切分带来的信息损失，导致模型难以聚焦于局部关键特征；另一方面，行人不对齐将会影响局部特征的有效性。

通过引入注意力机制来模拟人类视觉感知，可以让模型关注图片上的重点区域，Zhang 等[9]设计了关系感知全局注意力(Relation-aware Global Attention)模型来捕获全局结构信息，提取更有区分度的行人特征；Li[10]为网络引入和谐注意力(Harmonious Attention)，联合学习软像素注意力和硬区域注意力；Ding 等[11]提出了 AMFC 算法，该算法使用注意力模块级联不同层级特征，有效利用了低层特征中的细节信息。这些行人重识别算法仅为全局特征引入注意力，虽然提高了重要特征的权重，但同时也让模型忽视行人的细粒度特征。

针对上述方法存在的问题，本文提出了基于渐进式注意力金字塔的行人重识别方法(PAPNet)，该方法受人类从局部到整体渐进式观察事物的视觉模式的启发，通过渐进式注意力金字塔结构捕获特征图在不同切分等级下的关键信息，从而增强网络关注行人特征的丰富性和多样性，利用特征对齐结构指导多层级特征融合，再对融合后的特征图进行空间维度上的切分，使得提取到的行人局部特征可以有效对齐。渐进式注意力金字塔结构使用均匀分块方式将特征在通道和空间两个维度进行不同等级切分，按照“切分-关注-合并”的原则处理输入特征：首先在相应的维度上按等级均匀切分特征图，然后分别为不同切分等级的特征图引入注意力模块，单独学习每个部分的注意力，最后合并注意力特征图作为特征金字塔下一级的输入。通过从局部到整体、从细粒度到粗粒度的逐级关注，聚合特征在各个切分等级下的显著信息，从而让模型在通道和空间两个维度上对行人获得综合、全面、完整的感知。

此外，在池化层方面，本文引入广义平均池化(Generalized Mean Pooling，GeM Pooling)来代替传统的池化层，使得网络在训练过程中自动学习最优的池化超参数。在损失函数方面，本文通过对难样本三元组的特征进行加权处理，选择性地扩大不同特征间的距离并缩小相同特征间的距离，以优化网络对特征的学习。

2 基本原理

2.1 网络结构

本文以 ResNet-50 为骨干网络，为了更多地保留行人的细粒度特征，移除 ResNet-50 网络的全局平均池化层和全连接层，同时将其最后一个卷积块的步长修改为 1，以提高骨干网络输出特征图的分辨率。

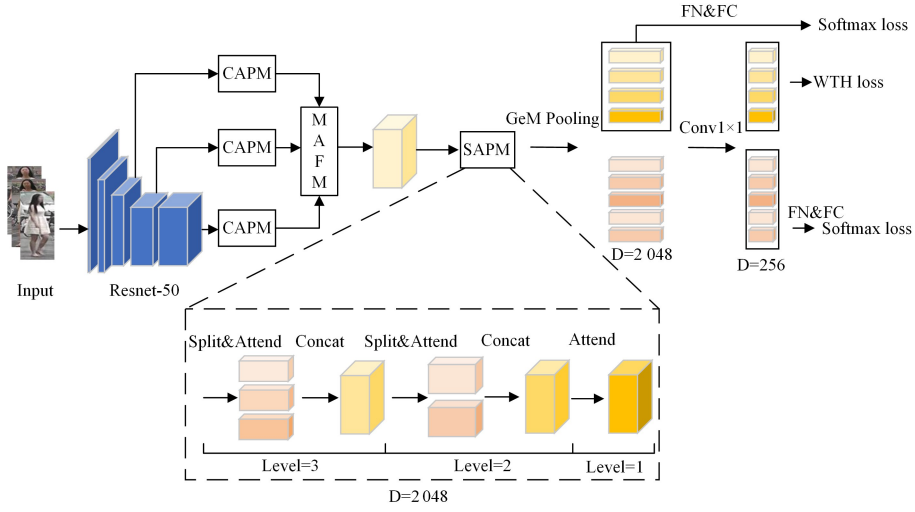


图 1 本文模型的结构图(电子版为彩图)

Fig. 1 Structure of the proposed model

ResNet-50 网络共有 5 层，每层特征输出分别用$\{F_1, F_2, F_3, F_4, F_5\}$来表示，首先将 F_3, F_4, F_5 分别送入通道注意力金字塔模块(Channel Attention Pyramid Module，CAPM)，得到 3 个新的特征图，然后通过多层级特征对齐模块(Multi-level Feature Alignment Module，MFAM)指导它们进行融合，最后将融合得到的 2048 维的全局特征 G 送入空间注意力金字塔模块(Spatial Attention Pyramid Module，SAPM)进行处理，得到多个不同切分等级下的局部特征和全局特征。如图 1 所示，SAPM 最大等级为 3 的 PAPNet 最终将得到 5 个局部特征和 4 个全局特征，使用 GeM Pooling 对它们进行池化处理后，得到 5 个 2048 维的特征向量，最后用 1×1 卷积将它们降至 256 维。训练阶段，联合加权的难样本挖掘三元组损失(Weighted TriHard loss，WTH loss)和 Softmax 损失对网络进行训练，4 个 256 维的全局特征向量计算 WTH loss，4 个 2048维的全局特征向量和 5 个 256 维的局部特征向量在进行特征归一化(Feature Normalization，FN)处理后分别送入

不共享权重的全连接层(Full Connection，FC)并计算 Softmax 损失。识别阶段，串联 9 个 256 维的特征向量作为行人图片的最终表示，使用欧氏距离度量方法与图库中的行人图片进行相似度匹配。

2.2 渐进式注意力金字塔

渐进式观察是人类仔细观察事物的一种视觉感知方式，通过由点到面，从局部到整体的逐步深入观察，将对被观察事物获得综合全面的认识。仿照这一过程，本文按照“切分-关注-合并”的原则设计了一种从部分到整体，不断让细粒度注意力指导粗粒度注意力学习渐进式注意力金字塔结构，引导网络依次发现从细到粗的不同尺度的显著信息，以充分挖掘行人的关键特征。

2.2.1 通道注意力金字塔

FPN[12]是计算机视觉领域被广泛使用的一种特征金字塔结构，它设计了一种自上而下的结构融合骨干网络不同层级的特征图，利用高层语义特征和低层细节特征的互补性来提高网络性能。受 FPN 启发，本文提出的 CAPM 也将骨干网络不同层级的特征图作为输入(由于低层提取的特征数较少，因此仅将 F_3，F_4，F_5 送入 CAPM)，与 FPN 不同的是本文并不是对各层级特征进行简单的相加融合，而是在特征的通道维度按照“切分-关注-合并”的原则，充分挖掘各层级特征在通道维度的显著信息后再送入 MFAM 进行融合。

图 2 为 CAPM 结构图，输入特征图 $F\in\mathbb{R}^{C\times H\times W}$，其中 C 表示特征的通道维数，H 和 W 分别表示特征图在空间维度的高和宽。在 CAPM 的第 i 级($Level=i$)，将 $i+1$ 级的输出特征图 $F_{i+1}\in\mathbb{R}^{C\times H\times W}$ 在通道维度均匀切分得到 2^{i-1} 个子特征图 $L_{i1}\sim L_{in}\in\mathbb{R}^{C/n\times H\times W}$($n=2^{i-1}$)，分别为每个子特征图 L_{ik}($1\leqslant k\leqslant n$)引入通道注意力模块(CAM)，以提高关键通道特征的权重，帮助模型专注于更显著的特征，最后将经过 CAM 处理后的子特征图 $L^a_{i1}\sim L^a_{in}\in\mathbb{R}^{C/n\times H\times W}$($n=2^{i-1}$)合并，将得到的特征图 $F_i\in\mathbb{R}^{C\times H\times W}$ 作为 CAPM 下一等级的输入，该过程的计算式如下：

$$L_{ik}=Split_C(F_{i+1}) \tag{1}$$

$$L^a_{ik}=L_{ik}\times\sigma\{MLP[P_{Avg}(L_{ik})]+MLP[P_{Max}(L_{ik})]\} \tag{2}$$

$$F_i=Concat_C(L^a_{i1},\cdots,L^a_{in}) \tag{3}$$

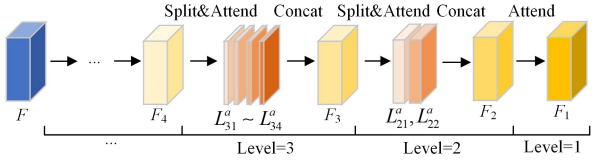


图 2 CAPM 结构图

Fig. 2 Structure diagram of CAPM

式(2)为本文使用的 CAM 公式，其结构图如图 3 所示，其中 σ 为激活函数 Sigmoid，P_{Avg} 和 P_{Max} 分别表示平均池化和最大池化，多层感知机(Multi-layer Perception，MLP)由两个 Conv 层组成，第一个 Conv 层将池化后的特征响应压缩至 C/r 维，r 表示压缩率，其后使用激活函数 ReLU 进行非线性变换，第二个 Conv 层将特征响应重新激发至 C 维，然后将两个特征响应相加，并使用 Sigmoid 激活函数进行归一化处理，得到通道注意力权重矩阵 $\boldsymbol{M}_C$，最终将输入特征图与 $\boldsymbol{M}_C$ 相乘，以实现对特征图不同通道特征的权重分配。

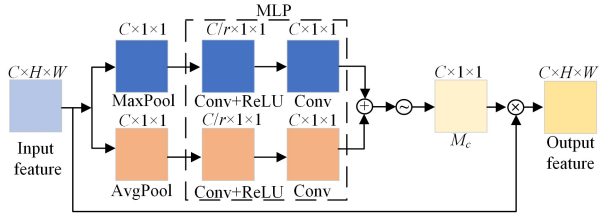


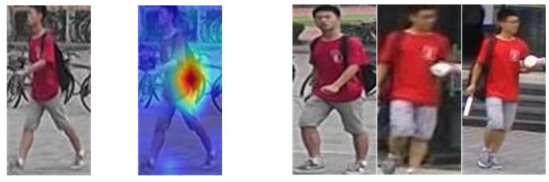
图 3 CAM 结构图

Fig. 3 Structure diagram of CAM

将特征图在通道维度均匀切分后，并行计算多个独立的通道注意力权重，通过分而治之的方式充分提取各个子特征图的关键通道特征信息，最后的合并操作相当于对多个并行注意力起到了集成的作用，由于切分带来的降维，因此多个并行注意力的计算成本与全尺寸注意力相似。利用金字塔的多级结构，在特征图的通道维度依次学习从细粒度到粗粒度的注意力，让模型从通道维度中充分挖掘不同层级特征图所包含的关键特征信息。

2.2.2 空间注意力金字塔

基于注意力机制的行人重识别算法通常仅为全局特征引入注意力，这容易让模型只关注行人最显著的特征而忽视其他细粒度特征，如图 4(b)所示，模型将注意力集中在目标行人的背包上，当行人姿态发生变化导致背包被遮挡或不可见时，模型重识别的准确度会大幅下降。



(a)People image (b)Heat map (c)People image after posture change

图 4 行人图片

Fig. 4 People images

为增强模型关注行人特征的丰富性和多样性，SAPM 同样按照“切分-关注-合并”的原则，从空间维度对行人进行从局部到整体的逐级关注，让模型捕获到不同切分等级下行人的局部细粒度特征。SAPM 的每一级首先将输入的特征图切分成相应个数的局部特征图，并单独学习每一部分的空间注意力，以发现局部关键特征，然后将局部注意力图合并，送入 SAPM 的下一级，指导粗粒度注意力的学习。图 1 中的红色框展示了最大等级为 3 的 SAPM 结构，当 $Level=i$ 时，首先将输入特征 $F_{i+1}\in\mathbb{R}^{C\times H\times W}$ 在空间维度的 H 上均匀切分为 i 份，得到 $L_{i1}\sim L_{in}\in\mathbb{R}^{C\times H/n\times W}$($n=i$)，再分别为它们引入空间注意力模块 SAM，使模型关注到各个局部特征图中的关键区域，最后合并所有局部注意力图 $L^a_{i1}\sim L^a_{in}\in\mathbb{R}^{C\times H/n\times W}$($n=i$)，得到全局特征图 $F_i\in\mathbb{R}^{C\times H\times W}$，将其作为 SAPM 下一级的输入特征图，继续挖掘更粗一级局部特征中的关键信息，该过程的具体计算式如下：

$$L_{ik}=Split_H(F_{i+1}) \tag{4}$$

$$L^a_{ik}=L_{ik}\times\sigma\{Conv\{[P_{Avg}(L_{ik});P_{Max}(L_{ik})]\}\} \tag{5}$$

$$F_i=Concat_H(L^a_{i1},\cdots,L^a_{in}) \tag{6}$$

式(5)为本文使用的 SAM 公式，其结构如图 5 所示，首先将输入特征的平均池化响应和最大池化响应在通道维度

进行拼接，然后使用 Conv 层将其通道维度降为 1，接着通过 Sigmoid 激活函数得到空间注意力权重矩阵 $\boldsymbol{M}_S$，最后将输入特征与 M_S 相乘，对关键区域和非关键区域进行编码。

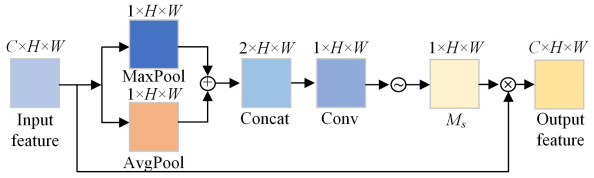


图 5 SAM 的结构图

Fig. 5 Structure diagram of SAM

2.3 多层级特征对齐

行人重识别任务的数据集是由不同摄像机拍摄而来的，摄像机拍摄角度和拍摄距离的不同会导致图片中的行人存在各种各样的形变和缩放，此外，行人姿态的变化也会对身体轮廓产生较大影响，上述现象可以被概括为空间不对齐和姿态不对齐两个问题。传统的卷积对平移、缩放、旋转等仿射变换的适应能力较差，而 Dai[13] 提出的可变形卷积(Deformable Convolution，Dconv)通过对卷积核的各个采样位置施加一个可学习的偏移量，来适应目标对象的几何形变，假设 x 为输入特征图，y 为可变形卷积处理后的输出特征图，则 y 中位置 p_0 的值可通过式(7)计算。

$$y(p_0)=\sum_{p_n\in R} w(p_n)\cdot x(p_0+p_n+\Delta p_n) \tag{7}$$

其中，R 表示卷积核的采样网格，对于 3×3 的可变形卷积 $R=\{(1,1),(0,1),\cdots,(-1,0),(-1,-1)\}$，$p_n$ 枚举 R 中的各个坐标，代表网格中不同的位置，w 为卷积核参数，Δp_n 为偏移量。

为了解决行人图片两个不对齐问题，本文基于可形变卷积设计了 MFAM，用于指导多层级特征进行融合，MFAM 的结构图如图 6 所示。首先使用 3×3 可变形卷积代替传统的卷积对各层级特征进行尺寸和维度上的变换，使网络能够自适应地调整输入特征图上的感受野，从而隐式地实现特征对齐，接着将卷积后的特征在通道维度上进行级联，以聚合多层级特征，最后再次使用 3×3 可变形卷积对级联后的特征降维处理，进一步对齐多层级特征，生成更准确的特征图。

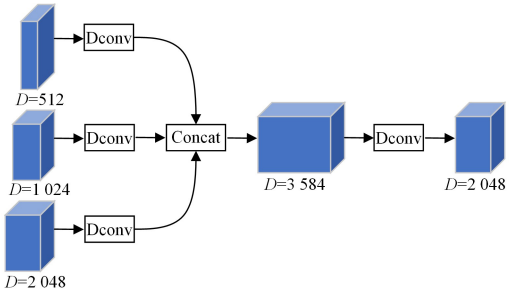


图 6 MFAM 的结构图

Fig. 6 Structure diagram of MFAM

2.4 广义均值池化

在卷积神经网络中传统的池化层有平均池化和最大池化。平均池化对每个通道特征图的像素点求均值，并将其作为该特征图的整体特征，最大池化则取每个通道特征图像素点的最大值作为整体特征。

卷积神经网络在进行特征提取时出现的误差主要来自两个方面：1)邻域大小受限造成估计值方差增大；2)卷积层参数误差造成估计均值偏移。平均池化可以让前一个误差减小，有利于保留行人图像的整体信息，最大池化可以让后一个误差减小，有利于突出行人图像的轮廓和细节纹理信息。为利用平均池化和最大池化的互补性，在保留行人图片信息完整性的同时突出行人细节特征，本文引入可学习的池化层 GeM Pooling，其计算式如下：

$$f=[f_1\cdots f_k\cdots f_K]^{\mathrm{T}},\ f_k=\left(\frac{1}{|X_k|}\sum_{x_i\in X_k}x_i^{p_k}\right)^{\frac{1}{p_k}} \tag{8}$$

其中，f_k 代表特征图，K 为通道数，X_k 是特征映射 $k\in\{1,2,\cdots,K\}$ 的 $W\times H$ 激活集，P_k 为池化超参数。池化超参数 P_k 在反向传播的过程中自动学习到最优解，由式(8)可得当 P_k 趋近于无穷时 GeM Pooling 近似于最大池化，当 $P_k=1$ 时 GeM Pooling 则为平均池化，因此引入的 GeM Pooling 介于平均池化和最大池化之间，同时具备平局池化和最大池化的优势。

2.5 损失函数

传统三元组损失[14]的三元组是从数据集中随机抽取的，很容易被区分开，这不利于网络的训练。针对这一弊端，难样本挖掘三元组损失[15](TriHard loss)选择最难的正样本和最难的负样本，与固定样本一起构成三元组，其计算式如下：

$$L_{TH}=\sum_{a,p_h,n_h}[d(a,p_h)-d(a,n_h)+\alpha]_+ \tag{9}$$

其中，α 为边际值，$d(a,p_h)$ 表示固定样本与最难正样本之间的欧氏距离，$d(a,n_h)$ 表示固定样本与最难负样本之间的欧氏距离。TriHard loss 未考虑到最难负样本与固定样本之间存在共同特征，最困难正样本与固定样本之间存在的不同特征，导致在训练过程中某些共同特征被拉开距离而某些不同特征却相互靠近，这会影响网络对特征的学习，因此本文通过加权的方式区别对待样本的不同特征向量，具体表达式如下：

$$L_{WTH}=\sum_{a,p_h,n_h}[d_T(a,p_h)-d_T(a,n_h)+\alpha]_+ \tag{10}$$

$$d_T(x,y)=d(T_{a,n_h}\times x,T_{a,n_h}\times y) \tag{11}$$

$$T^k_{a,n_h}=\begin{cases}\dfrac{W^k_{ij}}{\max(W_{ij})}+b, & \dfrac{W^k_{ij}}{\max(W_{ij})}\geqslant t\\ 0, & \dfrac{W^k_{ij}}{\max(W_{ij})}<t\end{cases} \tag{12}$$

$$W_{ij}=|W^i_c-W^j_c| \tag{13}$$

其中，T_{a,n_h} 为权重，b 为可学习的参数，t 为 0 到 1 之间的常数，$k\in[1,q]$，q 为特征向量的维数，样本 i 和样本 j 在网络中 FC 层权重 W_c 之差的绝对值 W_{ij} 表示它们在特征向量的每个元素上的差异程度，WTH loss 通过加权的方式扩大样本间差异较大的特征间的距离，缩小相似特征间的距离。

除了加权的难样本挖掘三元组损失外，本文还使用标签平滑的 Softmax 损失[16]对网络进行训练，其计算式如下：

$$L_{\mathrm{softmax}}=-\sum_{i=1}^{N}q_i\log(p_i) \tag{14}$$

$$q_i=\begin{cases}1-\dfrac{N-1}{N}\varepsilon, & i=y\\ \dfrac{\varepsilon}{N}, & \text{otherwise}\end{cases} \tag{15}$$

其中，N 表示数据集中行人个数，p_i 表示输出的行人身份的预测概率，y 为行人身份标签，ε 为超参数，本文将 ε 设为 0.1。

由于 Softmax 损失用于优化余弦距离，而三元组损失是通过计算欧氏距离得到的，如果在训练过程中直接结合这两种损失函数，就会产生冲突，导致损失难以收敛，影响模型

性能，对特征向量进行归一化处理可以平衡这种冲突，归一化之前的特征计算三元组损失，归一化之后的特征计算 Softmax 损失，这种做法可以有效解决上述问题。本文使用的特征归一化的计算公式如下：

$$\gamma=\frac{1}{m}\sum_{i=1}^{m}f_i \tag{16}$$

$$\sigma^2=\frac{1}{m}\sum_{i=1}^{m}(f_i-\gamma)^2 \tag{17}$$

$$\hat{f}=\frac{f-\gamma}{\sqrt{\sigma^2+\varepsilon}} \tag{18}$$

其中，f_i 表示归一化前的特征向量，m 表示 f_i 的维数，$\hat{f}$ 表示归一化后的特征向量。经过归一化处理之后，分布异常的特征向量将得到平衡，损失更容易收敛。

4 实验结果与分析

4.1 实验环境及数据集

本文在 Ubuntu16.04 的操作系统下，使用 NVIDIA GTX1080Ti GPU，基于 Pytorch-1.8.1 框架进行实验，在 Market-1501[17] 和 DukeMTMC-reID[17] 这两个公开数据集上，对本文所提出的基于渐进式注意力金字塔的行人重识别方法进行一系列实验，以验证其有效性，并将其与近期提出的行人重识别算法进行对比。

Market-1501 是当今最流行的基于图像的 Person-ReID 数据集，它所包含的 32668 张图像是由 6 个非重叠摄像头从不同角度拍摄的，对应 1501 个行人身份。整个数据集被分为两部分，分别进行训练和测试，训练集包含 12936 张图像，对应 751 个行人，测试集包含 3368 张查询图像和 19732 张图库图像，对应 750 个行人。

DukeMTMC-reID 是 DukeMTMC 数据集的子集，它包含了从 8 个高分辨率摄像头收集到的 1812 个行人的 36411 张图像，其中 1404 个行人的图像是通过 2 个以上的摄像头收集的。DukeMTMC-reID 也被划分为两部分，训练集包含 16522 张图像，对应 702 个行人，测试集包含 2228 张查询图像和 17661 张图库图像，共对应 702 个行人。

4.2 实验参数设置及评价指标

对模型进行训练时，输入的行人图像大小为 384×128，对训练数据使用水平翻转和随机擦除[19]进行增强，训练批次大小设置为 64(16 个行人，每个行人 4 张图像)，三元组损失的阈值参数 α 设置为 0.3，采用随机梯度下降法(SGD)对本文模型进行优化，动量设置为 0.9，权重衰减因子设置为 0.0005。模型共训练 120 个周期(epoch)，前 10 个周期学习率从 3.5×10^{-6} 线性增加到 3.5×10^{-4}，保持学习率为 3.5×10^{-4}，训练 50 个周期，然后在 $epoch=40$，$epoch=80$ 时让学习率分别衰减为 3.5×10^{-5} 和 3.5×10^{-6}。本文使用首位命中率(Rank-1)和平均精度均值(mean Average Precision，mAP)作为评估指标，它们是评价行人重识别精度的常用指标，Rank-1 表示搜索结果的第 1 张图像就是正确结果的概率，而 mAP 则通过计算 PR 曲线下的面积得到。

4.3 渐进式注意力金字塔实验结果

本文在 Market-1501 数据集上分别对 CAPM 和 SAPM 的最大等级进行了实验，其中最大等级为 0 表示不引入注意力机制，实验时未引入 MFAM，多层级特征采用直接相加的方式进行融合。此外，为了验证渐进式注意力金字塔结构的有效性，本文在进行最大等级实验时增设了非金字塔结构的对照组，非金字塔结构也将特征图进行多等级切分，并对切分后的特征单独引入注意力机制，但没有采用 PAPNet 逐级关注的模式，当最大等级为 0 和 1 时，非金字塔结构和金字塔结构对特征图的操作相同，因此精度一致。

4.3.1 CAPM 实验结果

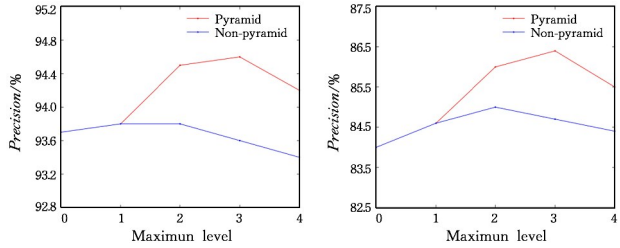
CAPM 最大等级实验结果如表 1 所列，图 7 给出了根据表 1 数据绘制的精度曲线，实验时没有引入 SAPM，模型直接学习融合后的全局特征 G。当最大等级为 0 时，即采用直接相加融合的方式，模型的重识别精度最低；为多层级特征引入注意力机制后再相加，模型的精度小幅提升；当最大等级为 2 和 3 时，通过逐级关注的方式充分挖掘关键通道，模型的精度有了大幅提升，相比直接相加融合的方式，mAP 分别提高了 2%和 2.4%，Rank-1 分别提高了 0.8%和 0.9%；当最大等级为 4 时模型的 mAP 和 Rank-1 有所下降，因为在 $Level=4$ 时，特征在通道维度会被平均分成 8 份后再单独引入注意力，过多的切分可能增加模型对一些非关键通道的关注，从而导致模型精度降低。另外，本文提出的金字塔结构在最大等级为 2，3，4 时模型的重识别精度均优于非金字塔结构，最大等级为 3 时，两者的 mAP 相差 1.7%，Rank-1 相差 1%，实验结果证明了注意力金字塔结构逐级关注模式的有效性，CAPM 通过细粒度注意力指导粗粒度注意力学习的方式使得模型精度取得了显著提升。

表 1 CAPM 实验结果

Table 1 CAPM experiment results

(单位：%)

Maximum level	Non-pyramid Structure		Pyramid Structure	
	Rank-1	*mAP*	*Rank*-1	*mAP*
0	93.7	84.0	93.7	84.0
1	93.8	84.6	93.8	84.6
2	93.8	85.0	94.5	86.0
3	93.6	84.7	94.6	86.4
4	93.4	84.4	94.2	85.5



(a)Rank-1 (b)mAP

图 7 CAPM 的实验结果

Fig. 7 CAPM experiment results

4.3.2 SAPM 实验结果

在 CAPM 最大等级为 3 的基础上对 SAPM 最大等级进行实验，实验结果如表 2 所列，图 8 给出了根据表 2 数据绘制的精度曲线。当最大等级为 0 时，由于没有对特征进行任何操作，因此模型精度和表 1 中 CAPM 最大等级为 3 时的精度相同；最大等级为 1 时，由于全局注意力的引入使得模型的关注点单一，降低了模型的鲁棒性，从而导致模型精度下降；在最大等级提高至 2 和 3 时，通过对局部细粒度特征的学习，模型精度逐级大幅提升，mAP 相比上一等级依次提升了 2%和 1%，Rank-1 依次提升了 1.3%和 0.4%；当最大等级提高至

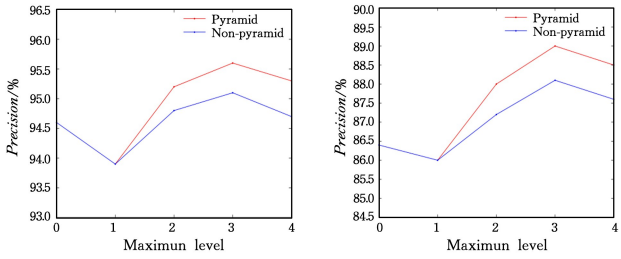
4 时，模型精度有所下降，因为在空间维度对特征图进行过多的切分，会破坏特征图的语义信息，而且切分越多模型精度就越容易受到行人不对齐问题的影响。另外，通过与非金字塔结构进行对比，使用 SAPM 在空间维度对行人进行逐级关注的方式同样使模型取得了更高的重识别精度，再一次验证了本文提出的渐进式注意力金字塔结构的有效性。

表 2 SAPM 实验结果

Table 2 SAPM experiment results

（单位：%）

Maximum level	Non-pyramid Structure		Pyramid Structure	
	Rank-1	*mAP*	*Rank*-1	*mAP*
0	94.6	86.4	94.6	86.4
1	93.9	86.0	93.9	86.0
2	94.8	87.2	95.2	88.0
3	95.1	88.1	95.6	89.0
4	94.7	87.6	95.3	88.5



(a)Rank-1 (b)mAP

图 8 SAPM 实验结果

Fig. 8 SAPM experiment results

4.4 MFAM 实验结果

本文在 Market-1501 与 DukeMTMC-reID 这两个数据集上对比普通卷积相加融合、可变形卷积相加融合以及可变形卷积级联融合后再次进行可变形卷积（MFAM）4 种多层级特征融合结构，实验时模型的 CAPM 和 SAPM 的最大等级都设为 3，最终实验结果如表 3 所列。结果表明，采用 MFAM 结构进行多层级特征融合的模型在两个数据集上均取得了最高的重识别精度，可变形卷积和级联结构可以增强模型对行人图片仿射变换的建模能力。

表 3 MFAM 实验结果

Table 3 MFAM experiment results

（单位：%）

Model	Market-1501		DukeMTMC-reID	
	Rank-1	*mAP*	*Rank*-1	*mAP*
Conv+Add	95.6	89.0	89.2	78.5
Dconv+Add	95.7	89.3	89.6	79.1
Dconv+Concat	95.7	89.4	89.8	79.4
Dconv+Concat+Dconv	95.9	89.7	90.1	79.8

4.5 可视化实验结果

图 9 是模型在 Market-1501 数据集上训练结束后，使用 Grad-CAM[20] 技术得到的类激活热力图，图中颜色高亮的部分就是训练过程中模型的关注点。4 组图片均按照输入图片、基准网络（Baseline）热力图、加入 CAPM 后的热力图和再次加入 SAPM 后的 PAPNet 热力图的顺序排列，Baseline 选用未引入 CAPM 和 SAPM 的模型。通过对比 4 组热力图可以发现，Baseline 对行人的关注区域较小，加入 CAPM 后模型的关注区域得到了扩大，但仍较多地集中在行人的上半身，加入 SAPM 后模型倾向于对行人的全身进行关注，实验结果表明本文方法通过挖掘多层级通道维度和空间维度的显著信息，让模型对行人获得了更综合全面的感知。

图 9 Grad-CAM 热力图

Fig. 9 Grad-CAM heat map

图 10 给出了 Baseline 和 PAPNet 在 Market-1501 数据集上的部分测试结果，绿色框表示识别正确，红色框表示识别错误。对比第一组图片可以发现，PAPNet 在行人姿态发生变化的场景下识别的正确率高于 Baseline。在第二组图片中目标行人被自行车遮挡，Baseline 出现了 5 张错误检测结果，而 PAPNet 仍保持着较高的正确率，说明 PAPNet 在遮挡场景下具有较高的鲁棒性。第三组图片的实验结果表明，Baseline 对行人的下半身关注度不足，导致识别结果中出现了两张上半身很相似的错误图片，而 PAPNet 的识别结果全部正确，因为通过在通道和空间两个维度上充分挖掘行人的特征信息，模型对行人获得了更全面的感知。

图 10 在 Market-1501 数据集上的部分测试结果（电子版为彩图）

Fig. 10 Part of test results on Market-1501 dataset

4.6 消融实验结果

为评估每个模块对模型精度的影响，本文在 Market-1501 和 DukeMTMC-reID 两个数据集上对模型进行了消融实验，实验结果如表 4 所列。本次实验的 Baseline 是以骨干

网络最后一层输出的全局特征为特征向量，用平均池化对特征进行池化处理，联合难样本挖掘三元组损失和 Softmax 损失进行训练的网络，在 Market-1501 数据集上的实验结果表明 WTH loss 相比 TriHard loss 让模型的 Rank-1 和 mAP 分别提升了 0.5% 和 0.7%，GeM Pooling 让模型的 Rank-1 和 mAP 分别提升了 0.7% 和 1.2%，加入最大等级为 3 的 CAPM 后，模型的 Rank-1 和 mAP 又分别提升了 0.7% 和 1.6%，加入最大等级的 3 的 SAPM 后模型的 Rank-1 提升至 95.9%，mAP 提升至 89.0%，加入 MFAM 后，模型精度最高，最终模型的 Rank-1 达到了 95.9%，mAP 达到了 89.7%。在 DukeMTMC-reID 数据集上的消融实验结果也表明，各个模块的加入让模型重识别精度取得了不同程度的提高。

表 4　消融实验结果

Table 4　Ablation experiment results

（单位：%）

Model	Market-1501		DukeMTMC-reID	
	Rank-1	*mAP*	*Rank*-1	*mAP*
Bseline	92.7	82.9	84.3	73.2
＋Weighted TriHard loss	93.2	83.6	84.9	73.7
＋GeM Pooling	93.9	84.8	85.7	74.6
＋CAPM	94.6	86.4	86.9	75.7
＋SAPM	95.6	89.0	89.2	78.5
＋MFAM	95.9	89.7	90.1	79.8

4.7　与其他方法对比

本文在 Market-1501 与 DukeMTMC-reID 这两个数据集上将本文提出的 PAPNet 与近几年提出的行人重识别方法进行对比，对比结果如表 5 所列。为证明本文所提出的渐进式注意力金字塔结构的有效性，对比方法中包括使用均匀分块思想获取行人局部特征的 PCB＋RPP 和 MSMGN[21] 等算法，以及基于注意力机制的 HACNN，IANe[22]，MHOA[23] 和 AMFC 算法。其中 MSMGN 利用 FPN 的方式融合多层级特征后进行多粒度分块，AMFC 利用注意力机制对多层级特征进行融合，与 APANet 的部分思想一致。另外，本文还对比了利用人体属性信息辅助网络进行训练的 AANet[24]，以及通过特征丢弃强化网络对局部特征学习的 BDB[25] 等算法。为保证公平比较，本文实验所有方法均没有采用重排序（Re-ranking[26]），最终的实验结果表明本文方法取得了不错的效果，利用注意力金字塔结构在通道和空间两个维度上对行人特征进行充分挖掘，让模型取得了更高的重识别精度。

表 5　对比实验结果

Table 5　Results of comparative experiments

（单位：%）

Model		Market-1501		DukeMTMC-reID	
		Rank-1	*mAP*	*Rank*-1	*mAP*
Local Features	PCB＋RPP	93.8	81.6	83.3	69.2
	MGN	95.7	86.9	88.7	78.4
	MSMGN	95.1	86.3	86.9	76.4
	RCMGN[27]	95.9	88.7	89.5	78.9
Attention Mechanism	HACNN	91.2	75.7	80.5	63.9
	IANet	94.4	83.1	87.1	73.4
	MHOA	95.1	85.0	89.1	77.2
	AMFC	93.8	84.1	86.2	73.9
Others	AANet	93.9	82.5	86.4	72.6
	BDB	94.2	84.3	86.8	72.1
	BagTricks	94.5	85.9	86.4	76.4
	MltB[28]	94.7	84.5	85.8	72.9
	CtF[29]	93.7	85.4	87.7	75.7
Ours	APANet	95.9	89.7	90.1	79.6

结束语　本文提出了基于渐进式注意力金字塔的行人重识别算法，通过金字塔的多级结构让细粒度注意力不断指导粗粒度注意力的学习，让模型对行人进行从局部到整体的逐级关注，以获取行人更丰富的细粒度特征；利用可变形卷积和级联结构融合多层级特征，增强了网络的鲁棒性；对难样本三元组特征进行加权处理也优化了网络对特征的学习。实验证明，本文方法可以让模型关注到更丰富的行人特征，行人重识别精度与现有方法相比有明显的提升。本文使用的注意力机制较为简单，下一步将对注意力机制进行研究，让模型更好地获取行人的重要特征。

参 考 文 献

[1] YE M，SHEN J，LING，et al. Deep learning for person re-identification：A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，44(6)：2872-2893.

[2] DALAL N，TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'05). Ieee，2005，1：886-893.

[3] LOWE D G. Object recognition from local scale-invariant features[C]// Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE，1999，2：1150-1157.

[4] 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.

[5] LUO H，GU Y，LIAO X，et al. Bag of tricks and a strong baseline for deep person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019：4321-4329.

[6] 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(ECCV). 2018：480-496.

[7] FAN X，LUO H，ZHANG X，et al. Scpnet：Spatial-channel parallelism network for joint holistic and partial person re-identification[C]// Asian Conference on Computer Vision. Cham：Springer，2018：19-34.

[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] ZHANG Z，LAN C，ZENG W，et al. Relation-aware global attention for person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：3186-3195.

[10] 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.

[11] ZHANG Z Y，DING J W，WEI H W，et al. Cascaded Multi-level Features Learning For Attention Based Person Re-Identification[J]. Laser & Optoelectronics Progress，2021，58(22)：2215003.

[12] LIN T Y，DOLLÁR P，GIRSHICK R，et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.
[13] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 764-773.
[14] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]// Proceedings of the IEEE Conference on Computer Vision Pnd Rattern recognition. 2015: 815-823.
[15] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv: 1703. 07737, 2017.
[16] 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.
[17] 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.
[18] RISTANI E, SOLERA F, ZOUR, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]// European Conference on Computer Vision. Cham: Springer, 2016: 17-35.
[19] ZHONG Z, ZHENG L, KANG G, et al. Random erasing data augmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 13001-13008.
[20] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.
[21] WANG D, ZHOU D K, HUANG Y D, et al. Multi-scale Multigranularity Feature for Pedestrian Re-identification[J]. Computer Science, 2021, 48(7): 238-244
[22] HOU R, MA B, CHANG H, et al. Interaction-and-aggregation network for person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9317-932.
[23] CHEN B, DENG W, HU J. Mixed high-order attention network for person re-identification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 371-381.
[24] TAY C P, ROY S, YAP K H. Aanet: Attribute attention network for person re-identifications[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7134-7143.
[25] DAI Z, CHEN M, GU X, et al. Batch dropblock network for person re-identification and beyond[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3691-3701.
[26] 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.
[27] DONG H S, ZHONG S, YANG Y F, et al. Person Re-identification by Region Correlated Deep Feature Learning with Multiple Granularities[J]. Computer Science, 2021, 48(12): 269-277.
[28] 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/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1389-1398.
[29] WANG G, GONG S, CHENG J, et al. Faster person re-identification[C]// European Conference on Computer Vision. Cham: Springer, 2020: 275-292.

ZHANG Shuaiyu, born in 1998, postgraduate. His main research interests include computer vision and person re-identification.

PENG Li, born in 1967, Ph. D, professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include visual Internet of things and deep learning.