2025 年 第 25 卷 第 21 期
2025, 25(21):09002-08

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2405565
引用格式:秦鹏, 陈高华, 古佳欣. 基于前景分割和多尺度特征融合的遮挡行人重识别[J]. 科学技术与工程, 2025, 25(21): 9002-9009.
Qin Peng, Chen Gaohua, Gu Jiaxin. Occluded pedestrian re-identification based on foreground segmentation and multi-scale feature fusion [J]. Science Technology and Engineering, 2025, 25(21): 9002-9009.

基于前景分割和多尺度特征融合的遮挡行人重识别

秦鹏, 陈高华*, 古佳欣
(太原科技大学电子信息工程学院, 太原 030024)

摘 要 遮挡行人重识别是一项具有挑战性的计算机视觉任务。提出了一种 FGMS-Net 网络方法,通过多个方面的改进显著提升了遮挡环境下的行人重识别能力。首先,采用改进的前景分割技术,有效地去除背景和其他杂波信息,使得特征提取更加精确。其次,针对遮挡问题,引入多尺度特征判别的方法,使得模型能够更好地捕捉局部特征,从而增强识别能力。最后,在主干网络中添加注意力机制,使得网络能够更加关注关键信息,提高整体识别性能。实验结果表明,所提方法在遮挡行人重识别任务中取得了显著的性能提升,在 Occluded-DukeMTMC 数据集上,累积匹配特征 Rank-1 和平均精度均值(mean average precision,mAP)分别达到了 71.7% 和 61.6%。

关键词 遮挡行人重识别; 前景分割; 多尺度特征; 注意力机制; 特征提取
中图法分类号 TP391.4; **文献标志码** A

Occluded Pedestrian Re-identification Based on Foreground Segmentation and Multi-scale Feature Fusion

QIN Peng, CHEN Gao-hua*, GU Jia-xin
(School of Electrical and Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China)

[**Abstract**] Occluded pedestrian re-identification is a challenging task in the field of computer vision. A method was proposed using the FGMS-Net network, which significantly enhances pedestrian re-identification in occluded environments through several improvements. Firstly, an improved foreground segmentation technique was employed to effectively remove background and other clutter information, resulting in more accurate feature extraction. Secondly, to address the occlusion issue, a multi-scale feature discrimination method was introduced, enabling the model to better capture local features and thereby enhancing identification capability. Finally, an attention mechanism was added to the backbone network, allowing the network to focus more on critical information and improve overall recognition performance. The experimental results show that method proposed has achieved significant performance improvement in the task of pedestrian re recognition with occlusion. On the Occluded-DukeMTMC dataset, the cumulative matching feature Rank-1 and mean average precision (mAP) reach 71.7% and 61.6%, respectively.

[**Keywords**] occluded pedestrian re-identification; foreground segmentation; multi-scale features; attention mechanism; feature extraction

行人重识别的基本任务是给定一组图像或视频帧,识别出在不同场景和时间内出现的同一个行人[1-2]。与传统的目标检测和图像分类任务相比,行人重识别面临着更多的挑战。首先,不同摄像头间的视角变化、光照条件差异以及行人的姿态变化都会对识别结果造成影响。其次,行人的遮挡和背景复杂性进一步增加了任务的难度。因此,如何有效地提取行人的全局和局部特征,并建立鲁棒的特征匹配算法,是行人重识别研究的核心问题。

近年来,深度学习技术的发展为行人重识别任务带来了显著的进展。基于卷积神经网络(convolutional neural network, CNN)的方法已经成为主流,通过设计精细的网络结构和损失函数,可以有效地提高识别精度。此外,度量学习(metric learning)方法以及注意力机制(attention mechanism)的引入,也在一定程度上提升了行人重识别的性能。例如,Sun 等[3]将图片进行均匀分割和独立处理人体不同部分的特征来提高识别准确性。Wei 等[4]利用人体的局部和全局线索来生成有辨别力和稳健的表示。Luo 等[5]提出了运用一些技巧用来提高准确率,包括但不

收稿日期: 2024-07-24 **修订日期**: 2025-04-11
基金项目: 山西省自然科学基金(202203021211198);太原科技大学科研启动基金(20222026)
第一作者: 秦鹏(1999—),男,汉族,山东泰安人,硕士研究生。研究方向:智能系统及软件应用技术。E-mail:qinpeng316@gmail.com。
* **通信作者**: 陈高华(1978—),女,汉族,山西原平人,博士,教授。研究方向:智能检测与信息处理。E-mail:chengaohua@tyust.edu.cn。

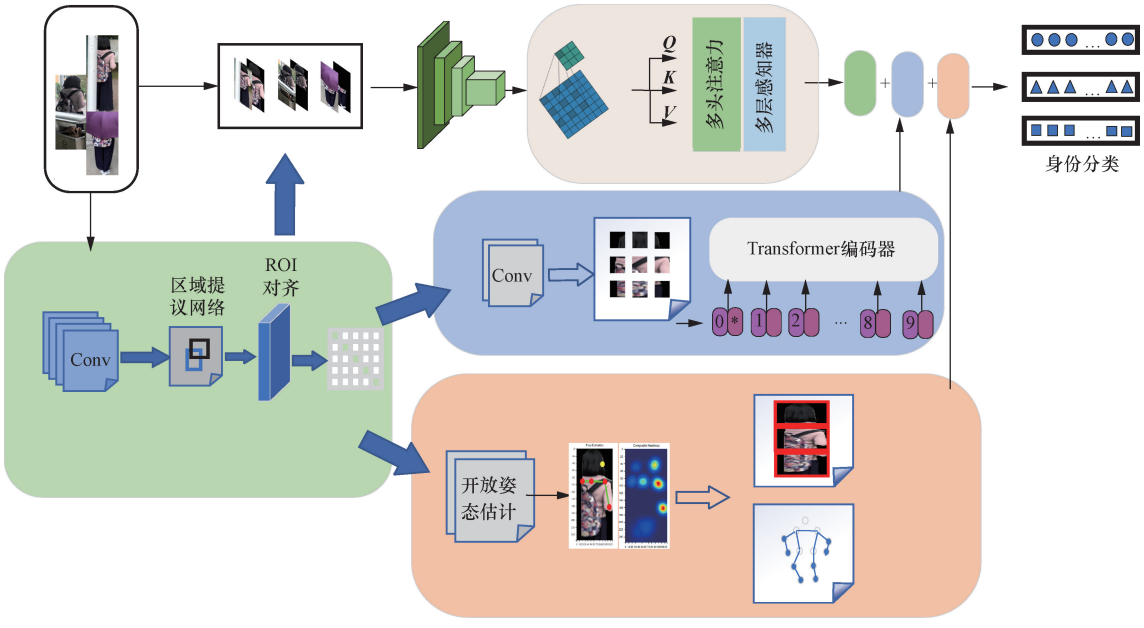
限于学习率预热、随即擦除增广、标签平滑等。这些研究大部分都是利用了卷积网络提取特征,来处理背景等杂波的影响,但网络无法较好地解决由遮挡、光照和姿态引起的模态内差异。近年来,Zhuo 等[6]提出了遮挡行人重识别。Miao 等[7]提出一种名为姿势引导特征对齐(posture guided feature alignment,PGFA)的新方法,利用姿势标志来从遮挡噪声中分离出有用信息。并且还构建了一个大规模数据集 Occluded-DukeMTMC。宋晓茹等[8]利用多个不同尺度特征进行优势互补并添加注意力机制,以此来改善模型效果。陈禹等[9]利用姿态估计进行关键点检测,生成一个基于姿态的行人特征表示,再利用 Transformer 进行训练。这些方法都提升了不少效果,但是,在特征空间中单靠单一的特征来处理行人重识别任务很难实现不同样本之间的特征对齐,特别是在遮挡的情况下模型表现很不理想。

在实际应用中,由于环境遮挡和检测器定位退化都是不可避免的,特别是在监控场景下,当发生身体遮挡时,提取的特征充满了噪声。此外,直接匹配两个图像而不考虑部件位置和可见性会导致空间错位。与以往的针对特征提取的方面相一致,但通过探索在前景分割之后,提取多种尺度的特征对行人进行表示,拉近域内距离,对遮挡下的行人重识别任务在语义信息较低的情况,特征提取如何更好地表示单个个体的特征提供了新的方向,为进一步的表征学习提供了新的研究思路。

现提出一种改进的 Mask R-CNN[10]模型,该改进模型包含两个新模块,以增强特征表达能力。第一个是像素级模块,类似于注意力机制,用于细化特征图中的信息;第二个是区域块模块,利用多层感知器(multilayer perceptron,MLP)计算区域块与区域块特征之间的关系,从而进一步提升特征表达的效果。在特征提取中,提出一种多尺度特征融合的方法,主要分为条状特征、块状特征和关键点特征。首先利用人体姿态估计模型,对输入图像中的人体姿态进行初步定位,根据人体结构的自然划分,作为其条状特征。又将前景分割后的图片输入到 ViT(vision Transformer)[11]模型中,得到块状特征。利用图神经网络,将关键点信息作为节点,通过图卷积网络生成带有姿态信息的关键点特征。最后在主干网络中融入注意力机制,以解决网络过度关注细节而忽视全局信息的问题。该方法旨在通过引导网络平衡对局部细节的关注与对全局信息的捕捉,从而提升模型的整体性能。注意力机制的加入能够自适应地调整网络对不同空间位置的关注程度,促使网络在提取特征时不仅考虑局部信息,也综合考虑全局上下文信息。

1　算法原理和网络结构

如图 1 所示整体网络并没有只采用局部特征,因为这种训练策略有可能比以前的深度特征学习有更多的判别特征,这些特征可能会过度拟合到训练集中最具判别力的部分,而忽略其他部分。在进行前景分割后,将掩膜处理后的图片与原图片结合后



ROI(region of interest)表示感兴趣区域;***Q*** 为需要查询的信息;***K*** 为被查询的键;***V*** 为实际的值

图 1　FGMS-Net 的网络结构

Fig. 1　The structure of FGMS-Net

一同输入一个 Resnet50 网络,后串联一个注意力模块,最后与局部特征支路一同进行分类判别。

1.1 前景分割模块

在行人重识别任务中,主要关注两张图片中的个体是否为同一人,因此更注重个体本身的特征,而背景、遮挡等因素则被视为噪声。因此,忽略背景信息、仅关注行人本身是一种有效的方法。目前,有多种图像分割算法可供选择,例如 U-Net[12] 和 Mask R-CNN[10] 等。U-Net 虽然在医学图像分割中表现出色,但在处理复杂遮挡问题上不具优势。

Mask R-CNN 是目前常用的掩膜处理的手段,但是针对行人重识别任务,有很多不足,一是特征提取网络可能产生误检和漏检,这直接导致行人特征表达不足,从而影响重识别准确率;二是像素级的语义分割占感兴趣的区域较低,效率低下。针对这两个问题提出了多尺度特征增强模块。

如图 2 所示,在感兴趣区域对齐(RoI Align)之后,引入了一个并行处理模块,该模块分为两个尺度的处理。一个是局部特征细化,它通过计算每个像素点的特征值与均值之间的关系,并利用 Sigmoid 激活函数调整这些关系,以细化整张图片的特征表达。另一个是区域间特征关联,它通过注意力机制(attention mechanism)[13] 在图像的不同局部区域之间建立关联,以捕获和强调重要的局部特征。

在进行像素点的处理中,采用类似于注意力机制的方法。具体来说,首先计算每个像素与其所在通道的均值的平方差;然后归一化计算每个通道内所有像素平方差的总和,除以像素数量,并加上一个小常数;最后,使用 Sigmoid 激活函数将归一化项映射到[0,1],生成注意力权重,并与原始输入特征图逐元素相乘,可以表示为

$$\mu_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{c,i,j} \tag{1}$$

$$\alpha_c = \sigma(S_c) = \sigma\left[\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(X_{c,i,j}-\mu_c)^2+\epsilon\right] \tag{2}$$

$$\hat{X}_{c,i,j} = \alpha_c X_{c,i,j} \tag{3}$$

式中:μ_c 为通道均值;$X_{c,i,j}$ 为特征图;$\sigma(\cdot)$ 为 Sigmoid 函数;ϵ 为一个非常小的常数,用于防止除零错误,保证数值稳定性;H 和 W 分别为特征图的高度和宽度;$\hat{X}_{c,i,j}$ 为加权后的特征图。

为了进一步细化 ROI Align 特征,进行区域间的处理,首先使用提取到的 ROI Align 特征图上应用滑动窗口技术,以固定的窗口大小和步长进行扫描,从而生成一系列小块。这些小块在特征空间中保持了局部信息的完整性和互不重叠的特点。随后,将这些小块输入一个多层感知器(MLP)中,通过 MLP 的非线性变换和特征融合能力计算并分析各小块之间的关系。

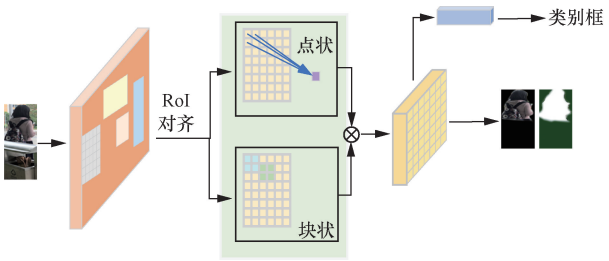
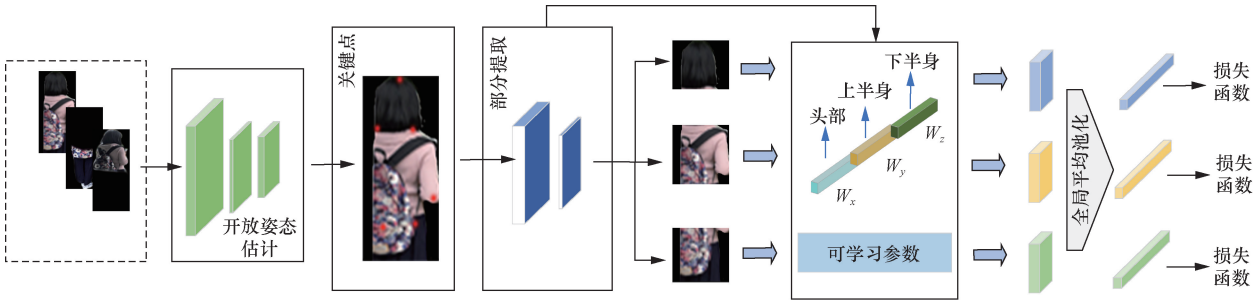


图 2 Mask R-CNN 的结构

Fig. 2 The structure of Mask R-CNN

1.2 条状特征处理

如图 3 所示,在经过前景分割模块后,进行人体姿态的定位,然后进行图像分块。这里选择的人体姿态估计的模型是 OpenPose[14],它可以更好地处理遮挡和多人场景问题。因为行人重识别任务的特殊性,所提取的照片画质都特别低,大部分像素大小都为 256×128 或者更低的 128×64,这样的像素对于要呈现更细微的人体面部关键点或者其他不易区分的人体关节点是困难的,因此,对于面部关键点,只要采集到一个关键点,那么就可以认为行人的头部没有被遮挡,躯干上肩部和臀部的关键点进行



W_x、W_y、W_z 分别为头部、上半身和下半身的初值

图 3 条状特征处理

Fig. 3 Strip Feature Processing

分块时提供坐标。首先估计出人体的 17 个关键点,然后传到局部分块提取,策略如下。

(1)当检测到的关键点缺少 1、2 个时,网络继续进行。

(2)当检测到的关键点分布不均匀,那么可能的原因就是受到遮挡,此时,为了正确地分块,就要进行补充被遮挡一面的关键点坐标。

(3)如果检测到下半身,或者头部关键点为 -1,那么认为人体部位缺失,分块网络会忽略掉这些关键点。

通过 OpenPose 模型检测图像中的关键点,并获取每个关键点的置信度。然后,根据这些置信度,计算各个分块(如头部、上半身和下半身)的平均置信度,将平均置信度转化为权重,以反映各个分块的重要性。

在进行行人分类时,无法知道每个局部的重要程度,很明显的一点,如果双方头部信息均可知,那大概率只根据头部信息,就可以推断出是否属于同一身份。此外,摄像头等边缘设备大多都架设在高处,拍摄到的行人大部分都处于俯视角度,此时,下半身变得在图片中被压缩,同时在实际中大部分的遮挡都处于下半身。可学习权重可以将其初始值设置为 $\{W_x, W_y, W_z\} = \{0.43, 0.33, 0.24\}$。在网络运行时,可以根据被遮挡的程度,更新权重。可学习权重可以表示为

$$w_j = \frac{\sum_{i \in K_j} c_i}{\sum_{k=1}^{M} \sum_{i \in K_k} c_i} \tag{4}$$

式(4)中:w_j 为第 j 个分块的权重;M 为图像分块的数量;K_j 为第 j 个分块内关键点集合;c_i 为第 i 个关键点的置信度。

1.3　块状特征处理

单独使用水平划分策略会丢失一些被错误忽略的信息,导致无法揭示潜在信息之间的关系。在划分成小块后,目标是找到隐藏的空间特征及其关系,作为全局特征的补充。因为块状特征并不像条状特征一样,它具备了与其他的块状特征之间的联系,因此将其输入 ViT[11] 中进行块状特征的处理。ViT 模型将图片首先分为很多的小块,作为 Transformer 的输入,但为了保持图片小块在原先的位置,因此需要添加位置信息。ViT 模型使用多层自注意力机制和前馈神经网络来处理嵌入后的图像块,每一层的 Transformer 编码器包括多头注意力(multi-head self-attention),它通过并行的多个注意力头来捕捉不同的特征表示,从而增强模型的表达能力。

1.4　关键点特征处理

通过对水平区域的划分,进行比较简单的可见部位对齐,然而希望对于任何水平区域都关注行人的可见部分,但孤立的部分很难代表整个行人。可见部分之间的关系成为表现行人的重要因素之一。通过构建不同部分之间的关系,可以将消息传递给抽象为图节点的每个部分。这种关系提供了这些节点之间的结构理解,并且丢弃了诸如遮挡特征之类的无用特征。

图神经网络通过节点及其邻居节点之间的信息传递和特征聚合来学习图的表示。这种结构天然契合人的身体结构,利用人体姿态估计中的关键点作为图卷积网络中的节点,人体中各部分的联系作为边,以此来学习这些关键点所代表的特征之间的关系。

由于遮挡的原因,部分关键点被隐藏,所以只将关键点作为节点变得不再可靠。对围绕每个关键点的局部图像区域进行特征提取,得到特征向量,这些特征向量在空间上保留了每个关键点的位置信息,并在特征上捕捉了局部的视觉信息,将每个关键点区域的特征进行聚合,形成一个整体的表示。

将人体分为头部、四肢和躯干,按照身体结构,头部、四肢与躯干均有联系,因此可以作为拓扑结构。由于图片中的各个遮挡部分不同,但同时遮挡大部分都发生下半身,各个部分的连接信息也是不同的。将人体区域划分为 8 个部分,如图 4 所示。

每个区域的特征可以表示为

$$f_p = G\left(\sum_{i=1}^{N} f_i\right) \tag{5}$$

式(5)中:f_p 为各个部分的特征;f_i 为各个关键点;$G(\cdot)$ 为关键点聚合所代表的区域特征;N 为覆盖区域内所有关键点数量。

在图卷积中,节点和边的关系确定后,按照一定权重更新各个点的特征,这些特征既包含了自己

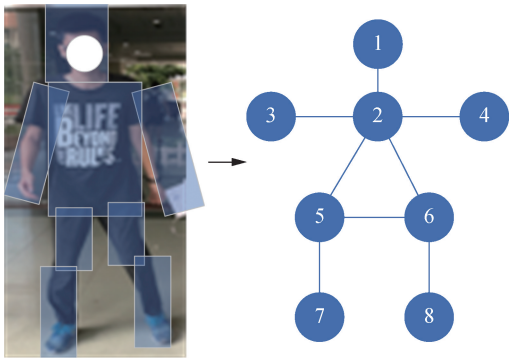


图 4　图卷积结构

Fig. 4　Graph convolutional structure

又包含了邻近关系中的特征,表示为

$$\begin{cases} f_{1,3,4} = (1-w)f_{1,3,4} + wf_2 \\ f_2 = (1-\alpha-\beta-\gamma)f_2 + \alpha f_1 + \beta f_3 + \gamma f_4 \\ f_{5,6} = (1-\delta-\varepsilon-\chi)f_{5,6} + \delta f_2 + \varepsilon f_{6,5} + \chi f_{7,8} \\ f_{7,8} = (1-\theta)f_{7,8} + \theta f_{5,6} \end{cases} \tag{6}$$

式(6)中:w、α、β、γ、δ、ε、χ、θ 均为可学习参数,使用相同的计算方法,不代表相同的值,这些参数是由聚合内关键点与所有关键点数量的比值来确定。

1.5 注意力模块

在行人重识别任务中,全局特征能够提供行人的整体外观和轮廓信息。这对于识别行人的大致形状和整体风格非常重要,例如一个人的身高、体型和整体服装款式。在处理大范围的遮挡或部分缺失时,全局特征仍然可以提供有效的信息。即使行人被部分遮挡,整体轮廓和身形仍然可以帮助识别。但单纯的卷积网络卷积神经网络(CNN)中,朴素卷积(普通卷积)尽管在提取局部特征方面表现优异,但在捕捉全局或远距离依赖的特征方面存在一定的局限性。由于朴素卷积的感受野(receptive field)主要集中在局部区域,因此对于捕捉图像中长距离的空间关系和全局上下文信息较为困难。

提出一种结合空洞卷积与多头注意力机制(multi-head deep attention,MDA),如图 5 所示。该模块首先使用空洞卷积对输入特征进行展开和重塑,通过引入不同的膨胀因子,捕捉多尺度的局部信息。多个头的注意力机制分别对这些展开的特征进行加权和聚合,动态调整对图像不同区域的关注度。最后,通过特征融合层将各头的输出整合,生成具有丰富上下文信息的全局特征表示。实验结果表明,该模块在处理复杂图像任务时,显著提升了模型对局部细节和全局结构的建模能力。

在具体实现中两个不同头的输出合并通过投影和重新组合实现,以期获得不同尺度的注意力图,

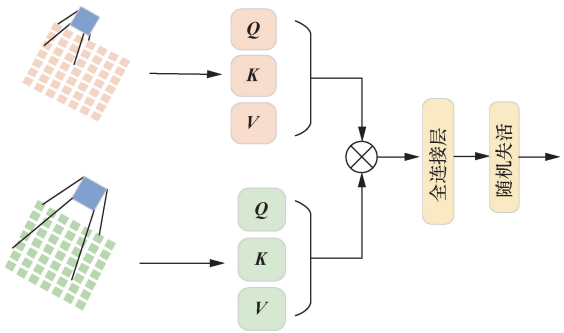


图 5 注意力机制

Fig. 5 Attention mechanism

在先行投影后添加一个 Dropout 用于防止过拟合。

2 损失函数

在整体网络中的训练,仍旧采用交叉熵损失函数和三元组损失函数,对于全局网络的学习,仍将其行人重识别任务是分类任务,交叉熵损失函数为

$$L_{\mathrm{CLS}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\ln(\hat{y}_{ic}) \tag{7}$$

式(7)中:N 为样本的数量;C 为类别的数量(即行人身份的数量);y_{ic} 为一个二值指示器,表示样本 i 是否属于类别 c(如果属于,则为 1,否则为 0);$\hat{y}_{ic}$ 为模型对样本 i 属于类别 c 的预测概率,其通过 Softmax 函数计算求得。

硬三元组损失函数(hard triplet loss)可以表示为

$$L_{\mathrm{tri}} = \sum_{i=1}^{N}\left[\max\left(0, \left\|\boldsymbol{f}_i^{\mathrm{a}} - \boldsymbol{f}_i^{\mathrm{p}}\right\|_2^2 - \left\|\boldsymbol{f}_i^{\mathrm{a}} - \boldsymbol{f}_i^{\mathrm{n}}\right\|_2^2 + \alpha\right)\right] \tag{8}$$

式(8)中:N 为三元组的数量;$\boldsymbol{f}_i^{\mathrm{a}}$ 为第 i 个三元组中 Anchor 样本的特征向量;$\boldsymbol{f}_i^{\mathrm{p}}$ 为第 i 个三元组中 Positive 样本的特征向量;$\boldsymbol{f}_i^{\mathrm{n}}$ 为第 i 个三元组中 Negative 样本的特征向量;α 为边际(margin),用于控制正样本与负样本之间的最小距离差异。

对于在第二个局部特征模块中,不使用三元组损失函数,以避免在因为错位或者其他的原因破坏模型,而是使用多个 Softmax 损失函数,模型能够学习到不同任务的特定特征。这些特征可以在共享的底层特征基础上进一步优化,使模型在每个任务上的表现更加优秀。

3 实验与分析

3.1 实验数据集

Market1501[15]数据集包括由 6 个摄像头(其中 5 个高清摄像头和 1 个低清摄像头)拍摄到的 1 501 个行人、32 668 个检测到的行人矩形框。每个行人至少由 2 个摄像头捕获到,并且在一个摄像头中可能具有多张图像。训练集有 751 人,包含 12 936 张图像,平均每个人有 17.2 张训练数据;测试集有 750 人,包含 19 732 张图像,平均每个人有 26.3 张测试数据。3 368 张查询图像的行人检测矩形框是人工绘制的,而 gallery 中的行人检测矩形框则是使用 DPM 检测器检测得到的。

Occluded-DukeMTMC[16]遮挡行人数据集是从 DukeMTMC-reID 数据集人工挑选出来的,训练集中有 12 927 张图像(665 个身份),用于查询的 2 163 张图像(634 个身份),图库集中有 9 053 张图像。

DukeMTMC-reID[16-17]数据集在杜克大学内采集,图像来自 8 个不同摄像头。该数据集提供训练集和测试集。训练集包含 16 522 张图像,测试集包含 17 661 张图像。训练数据中一共有 702 人,平均每类(每个人)有 23.5 张训练数据。是目前最大的行人重识别数据集,并且提供了行人属性的标注。

Occluded-REID[6]数据集图像是由校园内的移动摄像设备捕获的,包括属于 200 个身份的 2 000 张带注释的图像。在数据集中,每个人由 5 张全身人物图像和 5 张具有各种遮挡的遮挡人物图像组成。

Partial-REID[18]数据集包括 60 名行人的 900 张图像。每个人有 5 张全身人物图像、5 张遮挡人物图像和 5 张从遮挡图像中手动裁剪的部分人物图像。

3.2　评价指标

选用的测试指标是累积匹配特征(cumulative matching characteristic,CMC)和平均精度均值(mean average precision,mAP)。

累积匹配特征用于评估在前 k 个候选中找到正确匹配的概率,主要有 Rank-1、Rank-5 和 Rank-10。对于查询集 Q 中的所有查询图像,CMC 曲线在排名 k 处的值定义为

$$\mathrm{CMC}(k) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\delta(r_i \leqslant k) \tag{9}$$

式(9)中:$|Q|$ 为查询图像的数量;$\delta(r_i \leqslant k)$ 为指示函数,如果在第 k 个位置是正确匹配,则为 1,否则为 0。

平均精度均值(mAP)对所有平均精度取均值。对于查询图像 q 和对应的检索列表 L,平均精度计算公式为

$$\mathrm{AP}(q) = \frac{1}{m_q}\sum_{k=1}^{|L|}P(k)\,\mathrm{rel}(k) \times 100\% \tag{10}$$

式(10)中:m_q 为查询 q 的正确匹配数量;$P(k)$ 为在第 k 个位置上的精确值;$\mathrm{rel}(k)$ 为指示函数,如果在第 k 个位置是正确匹配,则为 1,否则为 0。

$$\mathrm{mAP} = \frac{1}{|Q|}\sum_{q\in Q}\mathrm{AP}(q) \times 100\% \tag{11}$$

式(11)中:Q 为所有查询的集合;$|Q|$ 为查询的总数量。mAP 是评估模型在检索任务中整体性能的重要指标,它考虑了检索结果的排序和准确性。

3.3　实验设置

使用 ImageNet 预训练的参数对模型进行初始化,输入图像的分辨率为 256×128,采用水平翻转、填充、随机裁剪和随机擦除等操作实现数据增强。通过 SGD 优化器端到端的进行训练,动量为 0.9,权重衰减 1×10^{-4},学习率初始为余弦学习率衰减将学习率初始化为 0.008,模型批次为 64,因为需要进行 OpenPose 人体姿态估计,所以占用比较高,该模型采用在 COCO 数据集上进行预训练的模型,生成 17 个关键点和置信度来预测人体姿态。前景分割中的 Mask R-CNN 同样采用在 COCO 数据集上的预训练模型。总体网络训练了 180 轮,网络在 ubuntu 系统上的 Pytorch 平台下利用两张 RTX 4090 进行训练。

3.4　对比试验

表 1 所示为与当前常用模型的对比结果,在与之对比的模型包括了解决整体行人重识别的方法(Aligned,PCB),基于部分匹配的方法(DSR),基于外部语义信息的方法(PVPM,PGFA,HOReID)和基于 Transfomer 的方法(PAT,TransReID,DPM)。由结果可以看出,利用前景分割消除背景等杂波,再利用人体姿态作为定位重要局部特征是可行的,在 3 个数据集上的结果都达到领先的水平。

如表 2 所示,在 Market-1501 数据集和 DukeMTMC 数据集上,本文方法在 Rank-1 和 mAP 指标上分别达到了 95.3%,91.2% 和 88.9%,82.4%。这些

表 1　遮挡数据集对比结果

Table 1　Comparison results of occlusion datasets

模型	O-REID		P-REID		O-DukeMTMC	
	Rank-1	mAP/%	Rank-1	mAP/%	Rank-1	mAP/%
PCB[3]	41.3	38.9	66.3	63.8	42.6	33.7
Aligned[19]	—	—	—	—	28.8	20.2
DSR[20]	72.8	62.8	73.7	68.07	40.8	30.4
PGFA[7]	—	—	69.0	61.5	51.4	37.3
PVPM[21]	70.4	61.2	—	—	47.0	37.7
HOReID[22]	80.3	70.2	85.3	—	55.1	43.8
OAMN[23]	—	—	86.0	—	62.6	46.1
PAT*[24]	81.6	72.1	88.0	—	64.5	53.6
TransReID[25]	70.2	67.3	71.3	68.6	64.2	55.7
DPM[26]	85.5	79.7	—	—	71.4	61.8
FGMS 本文	86.3	80.1	87.3	77.4	71.7	61.6

结果表明,提出的方法不仅在遮挡场景下表现出色,还在全身行人重识别任务中优于大多数现有方法。尽管网络主要针对遮挡场景设计,但其鲁棒性和泛化性使其在无遮挡场景中同样具有出色的性能。这归功于本文方法不仅保留了全局特征,还综合利用了多粒度的局部特征,从而在不同场景下都能进行有效识别。

为了进一步检验每个模块在整体网络中是否有效,设计了消融实验,将一个 ResNet50 作为基线模型,然后依次添加各模块,以探究它们的真实作用。

实验设置每 10 轮保存一次最佳模型文件,总计训练 180 轮。表 3 中可以看出,设计的各个模块均有效。在没有引入人体姿态估计模块,仅依靠 PCB 方法对前景分割后的图像进行 3 块区域的分割,效果有着少许提升。引入块状特征的处理后,ViT 模型依靠自身网络结构的特性,使得模型准确度大幅度上升。关键点特征的引入使得网络同时关注了图片中行人的拓扑信息,规避了被遮挡特征引起的信息不足。

表 2 全身数据集对比结果

Table 2 Comparison results of full-body datasets

模型	Market1501		DukeMTMC	
	Rank-1	mAP/%	Rank-1	mAP/%
PCB[3]	93.8	81.6	81.8	66.1
MGN[27]	95.7	86.9	88.7	78.4
ISP[28]	95.3	88.6	89.6	80.0
Trans ReID[25]	95.2	88.9	90.7	82.0
FPR[29]	95.4	86.6	88.6	78.4
PGFA[7]	91.2	76.8	82.6	65.5
HOReID[22]	94.2	84.9	86.9	75.6
PAT *[24]	95.4	88.0	88.8	78.2
DPM[26]	95.5	89.7	91.0	82.6
FGMS 本文	95.3	88.9	91.2	82.4

表 3 消融实验对比结果

Table 3 Comparison results of ablation studies

方法	Rank-1	mAP/%
ResNet50	37.4	30.6
ResNet50 + Mask R-CNN	39.8	34.8
ResNet50 + stipe	48.1	40.6
ResNet50 + patch	60.7	55.3
ResNet50 + keypoint	71.7	61.6

4 结论

遮挡的行人重识别任务是一个细粒度问题,由于现有的网络在实际中会将关注点投入更具判别的信息中,但是在遮挡的图片中更重要信息占语义信息较低。提出了一种结合前景分割和多尺度特征结合的模型,首先,将图片输入改进的 Mask R-CNN 进行行人的实例分割,排除掉背景,然后将之输入 3 个尺度的网络中,以获得更具有判别力的特征,在条状特征中,引入一个可学习的参数调整不同的占比,图卷积模块进一步利用人体姿态信息中的关键点处理被遮挡部分,加入扩充感受野的注意力机制使得网络可以更好地利用全局特征。在公开数据集上进行验证,结果表明本文模型表现良好,在遮挡场景下具有较强的竞争力。

提出的改进前景分割模块显著提高了 Mask R-CNN 在目标检测与分割任务中的表现,尤其在复杂场景下能够更好地平衡对局部细节与全局信息的关注,从而为深度学习在视觉任务中的应用提供了新的思路和方法。例如在智慧城市中的安防与监控,利用改进的目标检测算法,结合摄像头和传感器数据,开发智能安防系统可以实时监测公共区域中的异常行为,如人群聚集、非法入侵等。在智能驾驶领域,结合激光雷达(lightlaser detection and ranging,LiDAR)和摄像头数据,能够实时识别和分割行人、车辆、障碍物等。特别是在复杂的城市道路环境中,提出的多尺度特征融合方法和注意力机制能够帮助车辆更好地理解复杂的场景,提高安全性和驾驶决策能力。在医学影像分析领域,特别是在肺部计算机断层(computed tomography,CT)图像分析中,通过多尺度特征融合的方法,能够同时捕捉到肺结节的细节特征以及全局信息,提高早期肺癌诊断的准确性。研究内容和方法都可以提供一定的借鉴,而不单单仅在行人重识别领域,充分体现了本文模型和方法的泛用性。

参 考 文 献

[1] Zheng Z, Zheng L, Yang Y. A discriminatively learned CNN embedding for person reidentification[J]. ACM Transactions on Multimedia Computing Communications and Applications, 2018, 14(1): 1-20.

[2] 罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展[J]. 自动化学报, 2019, 45(11): 2032-2049.
Luo Hao, Jiang Wei, Fan Xing, et al. A survey on deep learning based person re-identification[J]. Acta Automatica Sinisa, 2019, 45(11): 2032-2049.

[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. New York: IEEE, 2018: 480-496.

[4] Wei L, Zhang S, Yao H, et al. GLAD: global-local-alignment descriptor for scalable person re-identification[J]. IEEE Transactions on Multimedia, 2019, 21(4): 986-999.

[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. Long Beach, CA: IEEE, 2019: 1487-1495.

[6] Zhuo J, Chen Z, Lai J, et al. Occluded person re-identification [C]//IEEE International Conference on Multimedia and Expo. San Diego: IEEE, 2018: 1-6.

[7] Miao J, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification[C]//Proceedings of the IEEE/ CVF International Conference on Computer Vision. New York: IEEE, 2019: 542-551.

[8] 宋晓茹, 杨佳, 高嵩, 等. 基于注意力机制与多尺度特征融合的行人重识别方法[J]. 科学技术与工程, 2022, 22(4): 1526-1533.
Song Xiaoru, Yang Jia, Gao Song, et al. Person re-identification method based on attention mechanism and multi-scale feature fusion [J]. Science Technology and Engineering, 2022, 22(4): 1526-1533.

[9] 陈禹, 刘慧, 梁东升, 等. 基于姿态估计和 Transformer 模型的遮挡行人重识别[J]. 科学技术与工程, 2024, 24(12): 5051-5058.
Chen Yu, Liu Hui, Liang Dongsheng, et al. Occluded person re-identification based on pose estimation and Transformer model [J]. Science Technology and Engineering, 2024, 24(12): 5051-5058.

[10] He K, Gkioxari G, Piotr D, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 42(2): 386-397.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. ArXiv, 2021: 2010.11929.

[12] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]. Cham: Springer International Publishing, 2015.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[14] Cao Z, Simon T, Wei S E, et al. OpenPose: real-time multi-person 2D pose estimation using part affinity fields[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Onlin: IEEE, 2017: 172-186.

[15] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: a Benchmark[C]//2015 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2015: 1079-1087.

[16] 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. New York: IEEE, 2017: 3754-3762.

[17] Lee B, Erdenee E, Jin S, et al. Computer Vision-ECCV 2016 workshops[M]. Cham: Springer International Publishing, 2016.

[18] Zheng W S, Li X, Yang T, et al. Partial person re-identification [C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2015: 4678-4686.

[19] Zhao L, Li X, Zhuang Y, et al. Deeply-learned part-aligned representations for person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2017: 3219-3228.

[20] He L, Liang J, Li H, et al. Deep spatial feature reconstruction for partial person re-identification: alignment-free approach[C]// Proceedings of the IEEE Conference on Computer vision and Pattern Recognition. New York: IEEE, 2018: 7073-7082.

[21] Gao S, Wang J, Lu H, et al. Pose-guided visible part matching for occluded person ReID[C]//Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Amsterdam: IEEE, 2020: 11744-11752.

[22] Wang G, Yang S, Liu H, et al. High-order information matters: Learning relation and topology for occluded person re-identification [C//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 6449-6458.

[23] Chen P, Liu W, Dai P, et al. Occlude them all: occlusion-aware attention network for occluded person ReID[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 11833-11842.

[24] Li Y, He J, Zhang T, et al. Diverse part discovery: occluded person re-identification with part-aware transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Huston: IEEE, 2021: 2898-2907.

[25] He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 15013-15022.

[26] Tan L, Dai P, Ji R, et al. Dynamic prototype mask for occluded person re-identification [C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 531-540.

[27] 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. New York: ACM, 2018: 274-282.

[28] Zhu K, Guo H, Liu Z, et al. Identity-guided human semantic parsing for person re-identification[C]//Computer Vision-ECCV 2020: 16th European Conference. Glasgow: IEEE, 2020: 346-363.

[29] He L, Wang Y, Liu W, et al. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 8450-8459.