小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0082
2026 年 5 月 第 5 期
Vol.47 No.5 2026

一种基于 ViT 技术的被遮挡行人目标重识别方法

高梦兴,肖满生,许雅婷,刘振桢

(湖南工业大学 计算机学院,湖南 株洲 412007)
E-mail:2391961539@qq.com

摘 要: 行人目标重识别(ReID)是指在不同场景中匹配同一行人目标的技术.针对在有遮挡物的情况下依赖全局信息方式处理行人目标细节特征时,出现的局部信息表达能力受限问题,提出了一个基于 ViT 特征增强的 ReID 方法,主要包括:1)设计一个新型的跨尺度空洞融合模块(Dimensional Feature Reinforcement Module,CDFM),通过多维度重加权对输入特征进行优化,提升特征表达能力;2)提出一个全局与局部特征协同算法,用以提升模型的性能和鲁棒性;.该方法结合了 Transformer 模块对全局依赖的建模能力和 CNN 在捕获局部细节特征上的优势,从而增强了特征信息的流动性和表达能力;3)提出一个动态加权损失函数,通过可见区域感知对比机制明确增强可见区域特征一致性,引入动态难例采样策略缓解遮挡噪声干扰,并融合通道注意力权重优化特征对齐,进一步提升模型在遮挡场景下的判别力.实验结果表明,所提出的方法在多个主流有遮挡的 ReID 数据集上表现出更强的性能优势.

关 键 词: 行人重识别;ViT;跨尺度空洞融合;全局与局部特征协同;动态加权损失
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)05-1219-06

ViT-based Method for Occluded Pedestrian Re-identification

GAO Mengxing,XIAO Mansheng,XU Yating,LIU Zhenzhen
(School of Computer Science,Hunan University of Technology,Zhuzhou 412007,China)

Abstract: Person Re-Identification (ReID) refers to the technology of matching the same pedestrian across different scenarios. To address the limitations of local feature representation caused by relying solely on global information in handling pedestrian details under occlusion scenarios,this paper proposes a ViT-enhanced ReID method. The key contributions include:1) A novel Cross-scale Dilated Fusion Module (CDFM) that optimizes input features through multi-dimensional re-weighting and integrates multi-scale dilated convolutional branches to enhance feature discriminability;2) A Global-Local Feature Collaboration Module combining Transformer blocks and lightweight CNN layers to leverage the complementary strengths of global dependency modeling by Transformers and local detail feature extraction by CNNs,thereby improving feature fusion and robustness;3) A Dynamic Weighted Loss Function that introduces a visibility-aware contrastive learning mechanism to enforce consistency in visible regions,adopts a dynamic hard example mining strategy to mitigate occlusion-induced noise interference,and incorporates channel attention weights for refined feature alignment,significantly enhancing discriminative power in occlusion scenarios. Experimental results demonstrate that the proposed method achieves superior performance on multiple mainstream occluded ReID benchmarks,including Occluded-Duke and Occluded-REID,outperforming existing state-of-the-art methods in both Rank-1 accuracy and mAP metrics.

Keywords: Person Re-Identification(ReID);ViT;cross-scale dilated fusion;lightweight convolution;contrastive weighting loss

0 引 言

行人重识别(ReID)是指在复杂环境中,确保同一行人目标在不同的视频场景下被准确识别,从而实现行迹追踪和行为分析的技术[1].作为一个跨场景匹配任务,ReID 在安防监控、智能交通和智慧城市等领域有重要的应用价值[1,2].然而,由于实际场景中存在大量的遮挡、姿态变化和视角差异等因素,使得 ReID 任务在准确性和鲁棒性上面临很大的挑战[3],即难以实现精准的目标匹配[4].

通常情况下,遮挡可分为外部遮挡和自身遮挡,外部遮挡是行人被其他物品如车辆,植被等遮挡[5],自身遮挡是行人被自身所属物品比如背包,这些遮挡可能影响局部特征提取的准确性,从而大大增加了模型识别的难度[6].为应对这些挑战,许多方法尝试通过不同的技术手段减轻遮挡对识别性能的负面影响.Zheng 等人[7]通过多视角信息的融合,缓解了外部遮挡对行人重识别的影响.Yang 等人[8]提出了一种稳健的软匹配特征对齐方法,通过分层联合学习获取行人的局部特征.Jiang 等人[9]利用生成对抗网络恢复遮挡部分来增强局部特征的完整性.Bian 等人[10]则设计了一种基于注意力引导的特征恢复机制,通过动态感知遮挡区域并重建被遮挡的语义信息来提高特征表达的鲁棒性.

这些用于遮挡行人重识别的深度学习方法主要利用

收稿日期:2025-02-26 收修改稿日期:2025-04-16 基金项目:湖南省自然科学基金项目(2024JJ7154)资助;湖南衣朵云商品智能化项目(h2024-003)资助. 作者简介:**高梦兴**,男,2000 年生,硕士研究生,CCF 会员,研究方向为计算机视觉;**肖满生**(通信作者),男,1968 年生,博士,教授,研究方向为智能计算、智能信息处理、图像处理等;**许雅婷**,女,1997 年生,硕士研究生,CCF 会员,研究方向为计算机视觉;**刘振桢**,男,2001 年生,硕士研究生,CCF 会员,研究方向为计算机视觉.

CNN 网络来提取行人的特征. CNN 以其在局部特征提取上的优异表现使其能有效捕捉图像的空间细节特征,但其在长距离依赖建模上的能力有限[11]. 随着 Transformer 架构在自然语言处理领域的突破[12],研究者开始探索其在计算机视觉任务中的应用潜力. He 等人[13]提出的 TransReID 首次将 Vision Transformer 模型应用于 ReID 任务,通过自注意力机制建模全局依赖关系,在整合全局上下文信息方面展现出显著优势. 这一突破性工作启发了后续研究,许多研究者开始探索 Vision Transformer(ViT)框架在遮挡行人重识别任务中的潜力. Wang 等[14]提出的 AA-Trans 模型,通过信息熵选择器增强细粒度特征聚合,提升了复杂遮挡场景下的识别性能. Li 等人[15]提出的层次化遮挡感知 Transformer(Occlusion-Aware Transformer,OAT)通过设计二阶注意力来捕获更全面的特征,在浅层网络捕捉局部细节特征的同时在深层网络建立跨区域全局关联,有效缓解了复杂遮挡导致的特征碎片化问题;Zhang 等人[16]开发的动态令牌选择模块(Dynamic Patch Token Selection Module,DPSM),则通过可学习的令牌重要性评分模块,实时过滤被遮挡区域的特征令牌,使模型注意力更集中于有效身份特征.

尽管 ViT 凭借其全局注意力机制展现出卓越性能,但近期研究表明其在局部特征提取方面仍存在局限. ViT 的块状注意力机制可能导致细粒度线索丢失,这一发现与早期 Sun 等[17]关于局部特征重要性的研究形成呼应. 因此,如何融合 CNN 的局部感知优势与 Transformer 的全局建模能力,已成为当前研究的重要方向.

本文通过结合 CNN 和 ViT 来提升模型,即通过以较小的开销同时捕获局部信息和全局信息,提升模型的性能. 此外,由于传统的卷积层通常具有固定的感受野,限制了其在不同尺度上捕捉特征的能力,因此在 ViT 图片分块前增加一个图片预处理的操作,使其一开始就能聚焦比较显著的特征.

1 ViT 技术框架

Vision Transformer(ViT)作为计算机视觉领域的革新性力量,近年来引发了学界与业界的广泛关注. 其核心概念是将图像数据重塑为与自然语言处理相似的序列输入模式,通过自注意力机制实现特征建模. 与传统卷积神经网络不同,ViT 完全依赖全局注意力机制来捕捉图像的上下文信息,这一创新设计使其在全局特征建模方面表现出卓越优势.

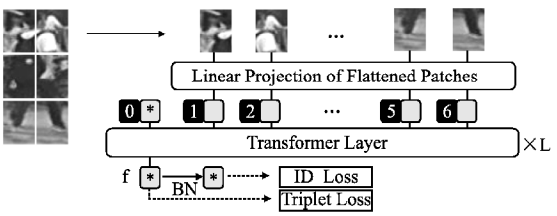

图 1 ViT 框架图

Fig. 1 ViT structure

如图 1 所示,ViT 的工作流首先对图像进行预处理,将图像切割成固定尺寸且互不重叠的小块,即 patches. 以常见的 224 × 224 像素图像为例,若将其划分为 16 × 16 像素的 patches,会得到 196 个这样的小块. 每个 patch 经过展平处理后转化为一维向量,这些向量按顺序排列形成序列输入. 随后,通过线性变换将这些向量映射到固定维度空间并引入位置编码. 这种独特方式打破了 CNN 对局部卷积的依赖,使 ViT 能够在更大的感受野中捕捉全局依赖关系.

ViT 的特征提取过程借助由堆叠的多头自注意力和前馈神经网络模块组成的 Transformer 层. 每一层中,自注意力机制会动态地对各个 patch 之间的复杂关系进行建模. 接着,前馈神经网络对自注意力机制输出的特征进行进一步加工,通过一系列线性变换和非线性激活函数,增强特征的表达能力,以便更好地应用于后续任务. 经过多层 Transformer 编码后,得到的最终特征可用于图像分类、分割等各类下游视觉任务. 在图像分类中,模型依据这些特征判断图像中物体的类别;在图像分割任务里,能够基于特征精准分割出不同物体的区域.

ViT 框架的突出优势在于其强大的全局建模能力. CNN 由于卷积操作的限制,通常只能捕捉局部信息,在识别复杂场景中的物体时,可能导致误判. 而 ViT 通过自注意力机制,能够全面关注图像中的远距离依赖关系,对图像全局信息有更精准的理解. 此外,ViT 可轻松适配不同视觉任务,在目标检测、图像生成、语义分割等领域均展现出良好的适应性. 然而,ViT 也存在一些局限性. 训练过程中,它对大规模数据有较强的依赖性,若训练数据不足,模型性能会受到较大影响. 同时,ViT 对图像遮挡较为敏感,当图像部分区域被遮挡时,难以像人类视觉系统那样依据上下文信息合理推测被遮挡部分的内容,从而限制了其在一些对数据量和遮挡情况要求较高的应用场景中的表现.

ViT 自诞生以来,不断推动着计算机视觉领域的发展. 从最初在简单图像分类任务上崭露头角,到如今逐步拓展至复杂场景理解、视频分析等多个领域,研究人员也在不断探索改进方法,以克服其现有局限,进一步挖掘其潜力.

2 改进的被遮挡行人 REID 技术

前面阐述了 ViT 框架的基本概念与原理,本文在 ViT 基础上,结合复杂场景中的行人目标,提出了一种结合卷积神经网络(CNN)和 Vision Transformer 的图像特征增强网络结构(如图 2 所示),以解决行人重识别(ReID)任务中面临的复杂问题. 该网络结构主要分为 3 个部分:图像特征预增强部分、全局与局部特征协同部分,以及损失计算部分.

图像特征预增强部分先将图像进行预处理,增加模型的感受野并对其特征进行增强. 通过这种特征预处理方式,可以有效简化后续模块的任务复杂度,同时确保输入数据的多样性和模型的收敛性. 全局与局部特征协同部分由 Transformer 和 CNN 模块组成,Transformer 从全局视角提取图像的语义信息,CNN 则聚焦于细粒度特征的捕获. 能够实现全局与局部特征的互补,通过结合两者可以增强对复杂场景的适应能力. 损失计算部分通过融合 3 种损失使模型在遮挡场景下更专注于判别性区域的学习.

这一网络架构的提出,是基于对行人重识别任务中存在的典型挑战的深入分析. 在遮挡场景中,行人的部分特征信息可能丢失,通过并行使用 Transformer 和 CNN 模块,不仅实现

了全局与局部特征的互补,还增强了模型对复杂场景的适应能力. 此外,这种结合方式能够有效利用两种网络在不同规模数据集上的表现特性,解决 Transformer 在小规模数据集上泛化能力不足的问题,同时保持 CNN 在局部特征提取方面的高效性.

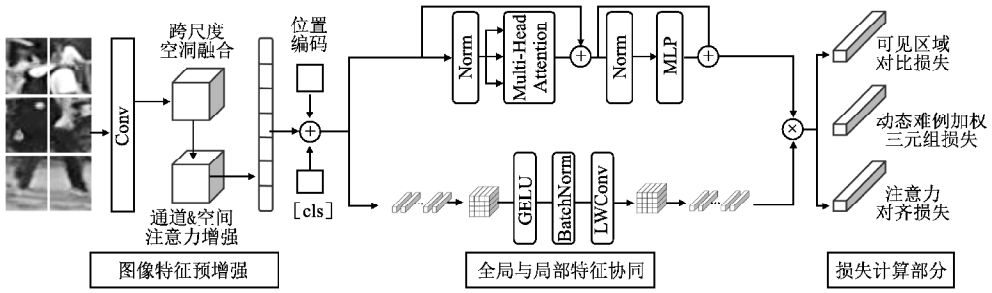

图 2　整体框架图

Fig. 2　Overall framework diagram

2.1　图像特征预增强部分

为了同时捕获图像中的细节信息和广泛上下文信息,本文设计了一种跨尺度空洞融合模块(Cross-ScaleDilated Fusion Module,CDFM)来进行图像特征预增强. 如图 3 所示该模块结合空洞卷积和通道空间注意力对图像进行预处理. 整合和增强相关特征,显著提升了特征表达的判别力和鲁棒性,尤其在处理遮挡、姿态变化以及复杂背景干扰等行人重识别的关键问题时表现出卓越的能力.

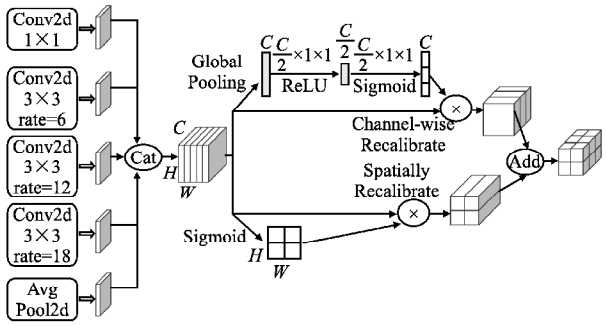

图 3　CDFM 模块框架图

Fig. 3　CDFM structure

CDFM 模块主要由多尺度空洞卷积部分以及通道和空间注意力的融合机制所构成. 在特征提取环节,通过引入跨尺度空洞卷积这一核心技术得以在多样的空间尺度上对图像特征展开全面提取. 传统的卷积网络在尝试增加感受野时,往往采用堆叠多层卷积层的方式. 然而,这种方法虽然在一定程度上能够扩大感受野范围,但却不可避免地带来了计算量呈指数级增长的弊端,同时还可能致使图像信息在多层处理过程中出现稀释现象,导致关键信息的丢失. 与之形成鲜明对比的是,空洞卷积技术,特别是运用不同空洞率的策略,能够在完美保留图像解析度的前提下,显著地扩大感受野范围. 这一特性对于精准捕捉图像中广泛的上下文信息而言,具有举足轻重的意义. 例如,在一幅行人图像中,更广泛的感受野能够帮助模型获取行人周围的环境信息,包括背景中的场景元素、其他物体的位置等,这些上下文信息对于准确识别行人身份至关重要. 在图像预处理阶段,CDFM 模块借助并行设置的多个不同空洞率,使得网络能够同时捕捉到图像中不同尺度的特征. 这一独特优势有助于大幅提高模型对图像中各类尺度特征的适应性与识别能力,无论是微小的细节特征,如行人衣服上的纹理图案,还是较大尺度的整体特征,如行人的整体姿态轮廓,都能被模型有效地感知和处理.

CDFM 架构通过 5 个并行的卷积分支来实现不同尺度的特征提取,每个分支配置有不同的空洞率,以扩大感受野并捕捉不同范围的空间信息. 第 1 分支使用 1 × 1 卷积核,不改变空间尺度,直接提取特征. 第 2 分支使用 3 × 3 卷积核,空洞率为 6,适中地扩大感受野. 第 3 分支使用 3 × 3 卷积核,空洞率为 12,进一步扩大感受野以捕获更广泛的上下文信息. 第 4 分支使用 3 × 3 卷积核,空洞率为 18,提供最广的感受野. 第 5 分支是用的一个额外的全局平均池化分支被用来提取全局上下文特征,增强模型对于整体布局的理解能力. 5 个不同分支提取不同尺度的特征,然后在通道维度上进行拼接,合成一个综合的特征图. 合并后的特征图并行经过两种注意力机制的校准.

对综合特征图进行全局平均池化(Global Pooling),得到每个通道的全局特征. 然后,通过两个全连接层(分别使用 ReLU 和 Sigmoid 激活函数)学习每个通道的重要性权重. 接着,将这些权重与原始特征图逐通道相乘,实现通道加权. 加权后对综合特征图在通道维度上进行全局池化,得到空间特征图. 通过一个 1 × 1 卷积和 Sigmoid 激活函数,学习每个空间位置的重要性权重. 将这些权重与原始特征图逐元素相乘,实现空间加权. 最后通道注意力和空间注意力机制的输出特征图通过元素加法(Add)操作进行融合,得到增强后的特征图. 此时分别经过通道和空间校准的特征图与原始合并特征图进行元素加和操作,以整合和增强相关特征. 增强后的特征图最后通过一个 1 × 1 卷积层进行降维和整合,生成最终的输出特征图.

2.2　全局与局部特征协同部分

为了同时捕获全局依赖关系和局部细节特征,本文提出了一种全局与局部特征协同模块,本模块基于 Transformer Block-CNN 并行架构的模型设计. 在 Vision Transformer (ViT)的基础上,引入了一种轻量化卷积模块(Lightweight Convolution Module,LWC Module),以弥补 ViT 在局部特征捕获方面的不足. 该架构结合了 Transformer 和卷积网络(CNN)的互补特性,既能建模图像的全局依赖关系,又能捕获图像的局部细节特征,从而在提升特征提取能力的同时,有效缓解了 ViT 模型在小规模数据集上收敛速度慢和泛化能力不足的问题.

本文的核心思想是通过并行使用 Transformer 模块和

LWC 模块实现全局和局部特征的建模. Transformer 模块利用多头自注意力机制(Multi-Head Self-Attention,MHSA)和前馈神经网络(Feed-Forward Network,FFN),动态建模特征之间的长距离依赖关系,提取全局语义信息. 输入特征首先经过归一化层(LayerNorm),以增强数值的稳定性,接着通过 MHSA 机制计算输入特征的相似性并分配权重,动态聚合上下文信息,最后通过 FFN 增强特征的非线性表达能力. 这一设计能够捕获图像的全局依赖关系,为模型的全局特征建模提供了有力支持. 然而,由于 Transformer 模块缺乏局部归纳偏差的建模能力,容易忽略像素间的局部相关性,特别是在小规模数据集或复杂场景中表现欠佳. 因此,设计了 LWC 模块以弥补这一不足.

LWC 模块作为一种轻量化卷积增强模块,专注于捕获图像的局部细节特征. 其处理流程包括以下几个步骤:首先,将输入的一维补丁标记(1D patch tokens)重新塑造成二维特征图(2D feature maps),以适应卷积操作;然后,对重塑后的特征图进行批量归一化(Batch Normalization)和 GELU 激活处理,以增强数值的稳定性和非线性特征提取能力;接着,通过轻量化卷积操作(Lightweight Convolution)捕获局部区域的上下文信息;最后,将卷积处理后的二维特征图再次重塑为一维补丁标记形式,并与 Transformer 模块的输出特征逐元素相加,形成最终的增强特征. 值得注意的是,LWC 模块仅对图像补丁标记进行操作,类标记(class token)直接跳过 LWC 模块的处理,从而确保全局语义特征的完整性.

在整个架构中,每一个 Transformer 块都配备了一个 LWC 模块. 这样的设计既可以确保全局特征和局部特征的融合,又能够在前向传播过程中,绕过 Transformer 模块的全局建模逻辑,专注于捕获局部细节信息,并将这些信息补充到 Transformer 的输出中. 这种并行架构不仅显著提升了特征表达的丰富性,还解决了 ViT 模型在局部特征提取能力上的不足. 通过将全局特征和局部特征进行互补融合,该模型在遮挡、姿态变化和复杂背景等场景下展现了优异的鲁棒性. 此外,LWC 模块的设计还引入了卷积操作的局部归纳偏差,使得模型在小规模数据集上能够更快收敛,同时增强了泛化能力.

该模块通过 ViT 模块建模全局依赖关系,LWC 模块捕获局部细节特征,再结合两者进行特征融合,显著提升了模型的性能和鲁棒性. 这一设计特别适用于复杂场景下的特征提取任务,为行人重识别等应用提供了更强大的支持.

2.3 损失计算部分

在行人重识别任务中,传统的损失函数设计通常采用交叉熵损失(Cross-Entropy Loss)与三元组损失(Triplet Loss)相结合的方式. 然而,这种组合方式存在两个显著缺陷:首先,三元组损失对样本难例挖掘策略高度敏感,在遮挡场景下易受噪声干扰导致收敛不稳定;其次,全局特征对齐方式忽略了遮挡区域与可见区域的差异性权重分配,导致模型对关键可见区域的特征判别力不足.

针对上述问题,本文引入基于对比学习的动态加权损失函数(Contrastive Weighting Loss,CWL). 该损失函数通过以下 3 个核心机制实现改进:1)构建可见区域感知的对比学习机制,增强局部特征区分度;2)设计动态难例加权策略,降低遮挡噪声样本的负面影响;3)融合通道注意力权重实现特征空间的精细化对齐.

改进后的损失函数由 3 部分组成:

1)可见区域对比损失:对于输入图像 I,通过特征图 $F \in R^{H\times W\times C}$ 提取空间注意力权重 $A_s \in R^{H\times W}$,计算可见区域特征对比:$\mathcal{L}_{vc} = -\log \dfrac{\sum_{i\in\Omega_v} \exp\left(f_i \cdot \frac{f_i^+}{\tau}\right)}{\sum_{i\in\Omega_v} \exp\left(f_i \cdot \frac{f_i^+}{\tau}\right) + \sum_{j\in\Omega_o} \exp\left(f_j \cdot \frac{f_j^-}{\tau}\right)}$.

其中 Ω_v 和 Ω_o 分别表示可见区域与遮挡区域索引集合,τ 为温度系数.

2)动态难例加权三元组损失:引入自适应权重系数 α 调整难例样本贡献度:$\mathcal{L}_{tw} = \alpha[\,\|f_a - f_p\|^2 - \|f_a - f_n\|^2 + \gamma]_+$,$\alpha = \sigma\left(\dfrac{\|f_a - f_p\|}{\|f_a - f_n\|}\right)$ 通过 Sigmoid 函数动态调节难例权重,γ 为边界超参数.

3)通道注意力对齐损失:利用 CDFM 模块输出的通道权重 w_c 优化特征分布:$\mathcal{L}_{ca} = \dfrac{1}{C}\sum_{c=1}^{C} w_c \cdot \|f_c - f_c^{id}\|_2$. 最终联合损失函数为:$\mathcal{L}_{total} = \lambda_1 L_{vc} + \lambda_2 L_{tw} + \lambda_3 L_{ca}$.

其中 $\lambda_1, \lambda_2, \lambda_3$ 为平衡系数,通过网格搜索分别设置为 0.7,0.2,0.1.

3 实验结果与分析

3.1 数据集和评价指标

Occluded-Duke 数据集是行人重识别(ReID)领域中一个用于评估遮挡场景下性能的基准数据集. 它是从 DukeMTMC-reID 数据集派生而来的,专门针对行人部分被遮挡的挑战进行了扩展和标注[18]. Occluded-Duke 数据集包含 15,618 张训练图像,17,661 张查询图像以及 176,661 张库(gallery)图像. 所有图像来自 DukeMTMC 数据集,部分添加了遮挡标注[19].

Occluded-REID 是行人重识别(ReID)领域另一个专门针对遮挡场景的基准数据集[20]. 它由传统的行人重识别数据集扩展而来,旨在评估和提升遮挡条件下 ReID 模型的表现. Occluded-REID 数据集包含 200 个行人,每个行人有 5 张遮挡图像以及 5 张完整图像. 总计约 2000 张图像,但数据量相对较小,适合快速验证模型性能或作为其他数据集的补充.

P-DukeMTMC-reID 数据集是从经典的 DukeMTMC-reID 数据集派生而来的一个用于研究部分遮挡场景下行人重识别(ReID)的基准数据集[21]. 它被设计为一个测试遮挡问题的高质量数据集,适用于验证和提升 ReID 模型在真实场景中的鲁棒性. P-DukeMTMC-reID 的数据规模与 DukeMTMC-reID 基本一致. 包含 702 名行人用于训练,702 名行人用于测试. 提供的图像包括遮挡图像和完整图像,以支持遮挡与非遮挡场景下模型的评估.

选用的两个评估指标分别为:累计匹配特性(Cumulative Matching Characteristics,CMC)曲线和平均准确率均值(mean Average Precision,mAP). CMC 聚焦于前 K 幅图像匹配成功的概率,该方法主要分析 Rank-1 即分别为前 1 幅图像匹配成功的概率[22]. 此外,还计算了每个查询图像的平均准确率,并据此汇总得出所有查询图像平均准确率的均值,即 mAP 值.

3.2 实验参数设置和实验环境

在本次针对 Vision Transformer(ViT)的深入研究实验

中,本文构建了严谨的实验体系并进行了多维度的技术优化,搭建了实验环境并细致设置了各类参数,以确保实验的高效性与准确性. 实验整体依托于 Pytorch1.7.0 框架开展,该框架凭借其强大的张量计算功能、动态计算图特性以及丰富的预训练模型库,为本文的研究提供了坚实且灵活的开发基础. 同时,本文选用英伟达 RTX4090 显卡作为核心计算硬件. 在训练参数设置方面,批处理大小(Batch Size)被设定为 64,总训练周期为 120 个 epoch,前 10 个 epoch 为线性学习率预热阶段.

为有效防止过拟合现象的出现,本文在模型架构内部进行了针对性设置. 在 Transformer 模块中,应用了 Dropout 率为 0.1. Dropout 作为一种常用的正则化手段,能够在训练过程中随机将部分神经元的输出置零,以此迫使模型学习到更加鲁棒的特征表示,避免神经元之间过度协同适应,从而提升模型的泛化能力. 此外,在 CNN 分支的卷积层后添加了 Layer Normalization. Layer Normalization 能够对每一层的输入数据进行归一化处理,使数据分布更加稳定,加速模型收敛,同时在一定程度上缓解梯度消失或梯度爆炸问题,增强模型训练的稳定性.

在损失计算部分,温度系数 $\tau=0.07$,特征嵌入维度为 512,边界参数 $\alpha=0.3$,难例挖掘比例设为前 20%,加权系数 $\lambda=0.5$.

3.3　实验结果

本文通过 3 个数据集来验证本文的模型在目前主流方法上的效果. 根据表 1 的结果显示,本文的方法在 Occluded-Duke,Occluded-REID,P-DukeMTMC 这 3 个数据集上的表现都比较好. 本文的骨干网络是由 Transformer 分支和 CNN 分支组成,CNN 分支即本文设计的 LWC 模块. 两者的结合使得模型能很好的捕获全局和局部的信息. 而本文在预处理阶段进行的图像预增强也是模型性能提升的一个不可缺少的部分. 与之前的 PAT 方法相比,本文的方法 mAP 平均提高了 11.2%,Rank-1 平均提高了 3.5%. 同时与其他的遮挡行人重识别方法相比,本文提出的方法也有不少的提升. 以上的实验

表 1　3 个遮挡数据集在各个方法上的实验结果

Table 1　Experimental results of three occlusion datasets across various methods

方法	Occluded-Duke		Occluded-REID		P-DukeMTMC	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PVPM[23]	37.7	47.0	37.7	66.8	69.9	85.1
PAT[24]	53.6	64.5	72.1	**81.6**	–	–
TransReID[13]	59.2	66.4	78.5	79.8	79.6	82.5
PFD[25]	61.8	69.5	83.0	81.5	–	–
Ours	**63.8**	**71.9**	**84.3**	81.2	**85.2**	**84.3**

表明,ViT 与卷积网络相结合能更好的解决行人重识别的遮挡问题. 作为第 1 个将 ViT 引入行人再识别领域的方法,TransReID 刷新了众多行人再识别的最佳表现,在遮挡行人再识别领域也是如此. 本文复现了 TransReID,并在 Occluded-Duke,Occluded-REID 和 P-DukeMTMC 3 个数据集上进行测试. 根据表 1 所示的实验结果,本文所提出的方法相较于 TransReID,其 Rank-1 和 mAP 在 Occluded-Duke 数据集上分别增加了 5.5% 和 4.6%,在 Occluded-REID 数据集上分别增加了 5.8% 和 1.7%,在 P-DukeMTMC 数据集上分别增加了 5.6% 和 1.8%. 这证明了在加入卷积的局部特征感受野后,ViT 的性能得到了进一步的提高.

本文方法在 Occluded-REID 数据集的 Rank-1 指标上较 PAT 方法存在微小差距. 现有方法(如 PAT)通过预定义部位分块和显式局部对齐机制,在遮挡集中于稳定区域(如躯干下半部)的特定场景下具有局部优势. 相比之下,本文方法采用的动态加权策略更关注遮挡程度自适应性,虽牺牲了部分固定遮挡模式的定位精度,但在多样化的复杂遮挡场景(如多目标交叉遮挡)下表现出更强的鲁棒性,整体上展现出普适性优势.

3.4　消融实验

为了验证模型中各模块对整体性能的贡献,本文在 Occluded-Duke、Occluded-REID 和 P-DukeMTMC 数据集上进行了详细的消融实验,结果如表 2 所示.

表 2　模型在 3 个遮挡数据集上的消融研究

Table 2　Ablation study of the model on three occlusion datasets

方法	Occluded-Duke		Occluded-REID		P-DukeMTMC	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
C	47.8	58.4	49.6	52.6	77.0	89.6
T	57.1	65.7	76.2	80.7	80.9	92.7
C-T	59.6	67.0	74.4	76.6	84.8	94.2
C-T-CDFM	61.9	69.6	76.1	**79.6**	86.0	94.7
C-T-NL	61.0	68.8	75.6	78.2	85.5	93.9
C-T-CDFM-NL	**62.3**	**70.2**	**76.8**	78.8	**86.6**	**95.8**

实验中依次移除模型中的关键组件,包括 CNN 模块、Transformer 模块以及跨尺度空洞融合模块(CDFM),并记录模型在 mAP 和 Rank-1 准确率上的变化. 消融实验的结果表明:

在所有数据集上,仅使用 CNN 的模型表现出较低的 mAP 和 Rank-1、准确率,说明 CNN 虽然在提取局部特征上表现良好,但缺乏全局特征的建模能力.

仅使用 Transformer 的模型在全局特征建模方面表现更优,但其对局部细节特征的捕获能力有限,尤其在小规模数据集上易受到遮挡的影响.

当模型同时使用 CNN 和 Transformer 模块时,性能显著提升. Transformer 提供了强大的全局上下文建模能力,而 CNN 补充了细粒度的局部特征信息,两者的结合展现出较好的互补性.

在包含 CDFM 模块的完整模型中,mAP 和 Rank-1 准确率进一步提高,尤其是在遮挡严重的 Occluded-Duke 和 Occluded-REID 数据集上,性能提升尤为显著. 这表明 CDFM 模块能够有效强化特征表达,通过结合通道注意力和空间注意力机制增强模型对细粒度信息的感知能力,从而提升了在遮挡场景下的鲁棒性.

在 Occluded-REID 数据集中,完整模型的 Rank-1 表现相较于其不采用本文损失函数的结果略差. 多次分析可知有限的训练样本量放大了噪声鲁棒性损失(公式(2))的收敛方差,轻微抑制了关键局部特征的聚焦能力. 且本文损失函数的特征对齐优化方向(公式(3)通道注意力对齐)与 Occluded-REID 的评估目标存在目标函数偏差. 模型容量有限性下,多目标优化导致特征判别力的边际效益递减. 微小性能波动反映了动态加权机制对遮挡模式分布差异的敏感性,但这种局部适应性代价与全局创新收益的权衡符合复杂视觉任务的优化规律,不影响方法论对行人重识别领域的突破.

其中,C 代表单独以 CNN 的方法来完成遮挡的行人重识别任务,T 代表单独以 ViT 的方法来实现. 而 C-T 则表示由 CNN 和 ViT 的混合网络实现,即加上本文的全局和局部特征模块来实现. C-T-CDFM 则是在混合网络的基础上只加上本文的跨尺度空洞融合模块(其他按照主流方法),C-T-NL 是只加上本文改进的损失函数(其他按照主流方法),最后的 C-T-CDFM-NL 即以本文完整的网络来实现有遮挡的行人重识别任务.

由上述表格可以看出,本文提出的网络架构逐步引入全局与局部特征模块,跨尺度空洞融合模块和改进的损失函数后,在遮挡数据集上的性能得到显著提升,表明这些设计对解决遮挡问题的有效性.

4 结 论

本文设计了一种基于 ViT 技术的遮挡行人重识别方法,解决了遮挡场景下模型性能不足的问题,并在 Occluded-Duke、Occluded-REID 和 P-DukeMTMC 数据集上的实验结果验证了该方法的有效性. 其中,Transformer 模型擅长全局依赖关系建模,CNN 模型则在捕获细粒度的局部特征方面具有优势,两者的结合在遮挡场景中展现出良好的互补性. 同时,通过设计跨尺度空洞融合模块(CDFM),增强了通道和空间的特征,使模型能够更全面、深入地挖掘行人的特征信息,从而显著提升了模型在遮挡行人重识别任务中的表现.

尽管改进后的模型在遮挡行人重识别领域取得了令人瞩目的进展,但不可否认的是,模型仍存在进一步优化的空间. 在复杂背景下的行人重识别场景中,模型面临着诸多严峻挑战. 背景噪声的干扰以及光线变化的影响,常常致使模型的识别精度出现明显下降. 具体而言,在复杂背景中,局部特征极易被背景中相似物体的特征所淹没,导致模型难以准确提取有效的局部信息;与此同时,全局特征也因背景的复杂性而无法有效地区分不同行人,使得模型在识别过程中频繁出现误判.

深入探究其原因,当前训练数据集中复杂背景样本的匮乏是一个不容忽视的关键因素. 由于训练数据无法充分涵盖现实场景中复杂多样的背景情况,模型在面对真实复杂背景时,泛化能力受到极大限制,难以准确适应并作出正确判断. 因此,针对复杂背景下的干扰问题以及训练数据集中复杂背景样本不足的问题展开深入研究,将成为未来进一步提升模型性能、拓展模型应用范围的重要研究方向. 通过优化模型结构以更好地应对背景噪声和光线变化,以及扩充和优化训练数据集,增加复杂背景样本的多样性和数量,有望使模型在复杂背景下的行人重识别任务中取得更为出色的表现.

References:

[1] Ye Mang, Shen Jianbing, Lin Gaojie, et al. Deep learning for person re-identification: a survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.

[2] Zhang Guoqing, Yang Shan, Wang Hairui, et al. Multi-modal person re-identification based on deep learning: a review[J]. Journal of Nanjing University of Information Science & Technology, 2024, 81(1): 25-51.

[3] Wang Jiahe, Gao Xizhan, Zhu Fa, et al. Exploring frontier technologies in video-based person re-identification: a survey on deep learning approach[J]. Computers, Materials & Continua, 2024, 79(3): 4123-4145.

[4] Sun Yifan, Zheng Liang, Yang Yi, et al. Beyond part models: person retrieval with refined part pooling[C]//European Conference on Computer Vision, Munich: Springer, 2018: 480-496.

[5] Zhou Yu, Zhao Xiaofeng, Wang Yi, et al. Multi-scale occluded person re-identification guided by key fine-grained information[J]. Journal of Electronics & Information Technology, 2024, 46(6): 2578-2586.

[6] Liu Zhigang, Wang Qi, Zhao Yijun, et al. Occluded person re-identification with pose estimation correction and feature reconstruction [J]. IEEE Access, 2023, 11(2): 14906-14914.

[7] Zheng Liang, Shen Liyue, Tian Lu, et al. Scalable person re-identification: a benchmark[C]//IEEE International Conference on Computer Vision, 2015: 1116-1124.

[8] Yang Zhenzhen, Chen Yanan, Yang Yongpeng, et al. Robust feature mining transformer for occluded person re-identification[J]. Digital Signal Processing, 2023, 141(10): 104166, doi: 10.1016/j.dsp.2023.104166.

[9] Jiang Yi, Xu Jiajie, Yang Baoqing, et al. Image inpainting based on generative adversarial networks[J]. IEEE Access, 2020, 8(1): 22884-22892.

[10] Bian Yuan, Liu Min, Wang Xueping, et al. Occlusion-aware feature recover model for occluded person re-identification [J]. IEEE Transactions on Multimedia, 2024, 26(11): 5284-5295.

[11] Zhang Wenfeng, Huang Lei, Wei Zhiqiang, et al. Appearance feature enhancement for person re-identification[J]. Expert Systems with Applications, 2021, 163: 113771, doi: 10.1016/j.eswa.2020.113771.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems, Long Beach: Curran Associates, 2017: 5998-6008.

[13] Luo Hao, He Shuting, Wang Pichao, et al. TransReID: transformer-based object re-identification[C]//IEEE International Conference on Computer Vision, 2021: 14993-15002.

[14] Wang Qi, Wang Jianjun, Deng Hongyu, et al. AA-trans: core attention aggregating transformer with information entropy selector for fine-grained visual classification [J]. Pattern Recognition, 2023, 140(8): 109547, doi: 10.1016/j.patcog.2023.109547.

[15] Li Yanping, Liu Yizhang, Zhang Hongyun, et al. Occlusion-aware transformer with second-order attention for person re-identification [C]//IEEE Transactions on Image Processing, 2024: 3200-3211.

[16] Zhang Xin, Fu Keren, Zhao Qijun. Dynamic patch-aware enrichment transformer for occluded person re-identification [J]. arXiv preprint arXiv, 2024: 2402.10435.

[17] Sun Yifan, Zheng Liang, Yang Yi, et al. Beyond part models: person retrieval with refined part pooling[C]//European Conference on Computer Vision, 2018: 501-518.

[18] Zheng Zhedong, Zheng Liang, Yang Yi, et al. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro [C]//IEEE International Conference on Computer, 2017: 3774-3782.

[19] Ergys Ristani, Francesco Solera, Roger Zou, et al. Performance measures and a data set for multi-target, multi-camera tracking [C]//European Conferenceon Computer Vision, 2016: 17-35.

[20] Zhuo Jiaxuan, Chen Zeyu, Lai Jianhuang, et al. Occluded person re-identification[C]//IEEE International Conference on Multimedia and Expo, 2018: 1-6.

[21] He Linxiao, Wang Yinggang, Liu Wu, et al. Foreground-aware pyramid reconstruction for occluded person re-identification [C]// IEEE International Conference on Computer Vision, 2019: 8449-8458.

[22] Luo Hao, Jiang Wei, Gu Youzhi, et al. A strong baseline and batch normalization neck for deep person re-identification [J]. IEEE Transactions on Image Processing, 2020, 29(12): 4022-4035.

[23] Gao Shang, Wang Jingya, Lu Huchuan, et al. Pose-guided visible part matching for occluded person reid[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11744-11752.

[24] Li Yulin, He Jianfeng, Zhang Tianzhu, et al. Diverse part dis-covery: occluded person re-identification with part-aware transformer [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2021: 2897-2906.

[25] Wang Tao, Liu Hong, Song Pinhao, et al. Pose-guided feature disentangling for occluded person re-iden-tification based on transformer [C]//AAAI Conference on ArtificialIntelligence, 2022: 2540-2549.