ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 1,Jan. 2025,pp. 217－229
DOI:10. 16337/j. 1004-9037. 2025. 01. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于可学习掩模和位置编码的遮挡行人重识别

杨真真，陈亚楠，杨永鹏，吴心怡

(南京邮电大学理学院,南京 210023)

摘　要：行人重识别虽已取得了显著进展,但在实际应用场景中,不同障碍物引起的遮挡问题仍然是一个亟待解决的挑战。为了从被遮挡行人中提取更有效的特征,提出了一种基于可学习掩模和位置编码(Learnable mask and position encoding, LMPE)的遮挡行人重识别方法。首先,引入了一种可学习的双路注意力掩模生成器(Learnable dual attention mask generator, LDAMG),生成的掩模能够适应不同遮挡模式,显著提升了对被遮挡行人的识别准确性。该模块可以使网络更灵活,能更好地适应多样性的遮挡情况,有效克服了遮挡带来的困扰。同时,该网络通过掩模学习上下文信息,进一步增强了对行人所处场景的理解力。此外,为了解决Transformer位置信息损耗问题,引入了遮挡感知位置编码融合(Occlusion aware position encoding fusion, OAPEF)模块。该模块进行不同层次位置编码融合,使网络获得更强的表达能力。通过全方位整合图像位置编码,可以更准确地理解行人间的空间关系,提高模型对遮挡情况的适应能力。最后,仿真实验表明,本文提出的LMPE在Occluded-Duke和Occluded-ReID遮挡数据集以及Market-1501和DukeMTMC-ReID无遮挡数据集上都取得了较好的效果,验证了本文方法的有效性和优越性。

关键词：行人重识别;注意力机制;掩模机制;位置编码;Transformer
中图分类号：TP391　　**文献标志码**:A

Learnable Mask and Position Encoding Based Occluded Pedestrian Re-identification

YANG Zhenzhen, CHEN Yanan, YANG Yongpeng, WU Xinyi

(College of Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Although the pedestrian re-identification task has made significant progress, the occlusion problem caused by different obstacles is still a challenge in practical application scenes. In order to extract more effective features from occluded pedestrians, a learnable mask and position encoding (LMPE) method is proposed. Firstly, a learnable dual attention mask generator (LDAMG) is introduced to adapt to different occlusion patterns, significantly improving the re-identification accuracy of occluded pedestrians. It makes the network more flexible and better adapts to diverse occlusion situations. At the same time, the network learns contextual information through the mask, which further improves the understanding of the scenes. In addition, we introduce the occlusion aware position encoding fusion (OAPEF) module to solve the problem of losing position information in Transformer. This method helps to perform the fusion of

基金项目:国家自然科学基金(62071242, 62171232);江苏省研究生科研与实践创新计划项目(KYCX22_0955, SJCX23_0251);江苏省研究生教育学改革项目(JGKT23_C019);南京邮电大学研究生教育学改革项目(JGKT23_XJ02);南京邮电大学教学改革研究项目(JG00723JX22)。
收稿日期:2024-02-14;**修订日期**:2024-04-18

different regional position encoding and allows the network to gain stronger expressive ability. The integration of position encoding in all directions enables the network to understand the spatial correlation between pedestrians more accurately, and improves the ability to adapt to the occlusion situation. Finally, simulation experiments are conducted, and results demonstrate that LMPE performs well on Occluded-Duke and Occluded-ReID occluded datasets and Market-1501 and DukeMTMC-ReID unoccluded datasets, which confirms the effectiveness and superiority of the proposed method.

Key words: pedestrian re-identification; attention mechanism; mask mechanism; position encoding; Transformer

引　言

随着计算机视觉和深度学习技术的飞速发展,行人重识别(Pedestrian re-identification, ReID)已取得了显著进展[1-3]。行人重识别的目标是通过对摄像头捕捉到的行人图像或视频进行分析,识别出在不同摄像头之间穿行的同一行人,具有广泛的应用潜力,涉及到视频监控、智能交通和城市安全等多个领域。尽管在受控环境下行人重识别取得了显著成功,然而在实际应用场景中,由于环境的复杂性和多变性,该任务仍然面临着多方面的挑战,其中遮挡问题尤为突出[4-6]。在实际场景中,行人可能会被其他行人、车辆以及建筑物等各种障碍物部分或完全遮挡,导致行人的部分信息缺失或不可见。这种遮挡现象严重影响了行人重识别系统的性能,使其在复杂环境中难以准确地区分和匹配被遮挡的行人。

现有遮挡行人重识别方法大概可分成基于卷积神经网络(Convolutional neural network, CNN)的方法和基于Transformer的方法两类。基于CNN的方法[7-11]首先对收集到的行人图像数据进行预处理,包括图像的裁剪、大小标准化、亮度和对比度调整等操作;接着选择合适的CNN模型作为基础网络结构,在此模型上对行人图像进行特征提取,得到每个行人图像的特征表示,最后进行行人特征匹配。但是这些方法大都针对可见特征进行处理,忽略了遮挡行人的不可见特征,无法获得更加有效的特征。同时,由于缺乏对行人所处场景的全面理解,这些方法也无法获取足够的上下文信息,从而限制了其在真实场景中的应用性。基于Transformer的方法[12-15]将行人图像划分为均匀大小的图像块,并使用Transformer中自注意力的强大特征处理能力进行特征提取,从而获得各图像块的区域特征进行图像匹配。但是对于图像数据,图像块之间的位置信息十分重要。上述基于Transformer的方法中大都会在训练初始添加位置信息,但经过训练后,会存在位置信息的损耗,使网络不能完整获悉图像块间的位置关系,难以获得更加有效的特征。

针对上述问题,本文提出了一种基于可学习掩模和位置编码(Learnable mask and position encoding, LMPE)的遮挡行人重识别方法。主要工作如下:(1)引入视觉Transformer(Vision Transformer, ViT)作为主干网络。通过引入ViT,充分利用其全局上下文的建模能力,能够有效处理由遮挡引起的信息缺失,更好地处理复杂的遮挡情况,从而显著提高对遮挡行人的识别准确率;(2)引入可学习的双路注意力掩模生成器(Learnable dual attention mask generator, LDAMG)。通过采用双路注意力机制,能够更好地关注不同通道和空间位置的特征,使得生成的掩模不仅更符合被遮挡行人的实际情况,还有助于区分不同身体部位或者不同行人之间的差异,从而提取更有区分性的特征。该模块有助于提高网络对遮挡情况的适应性,并增强模型在复杂场景中的表现能力;(3)引入遮挡感知位置编码融合(Occlusion aware position encoding fusion, OAPEF)模块。该模块可以丰富位置信息,通过不同层次位置编码的融合,使网络能够灵活地表达不同位置的语义信息。同时,提高了网络对图像位置的敏感性,使网络获得对图像中细粒度特征的捕捉和理解能力。

1 相关工作

1.1 遮挡行人重识别

当前用于遮挡行人重识别方法大概可分为两类,基于CNN的和基于Transformer的方法。Wang等[16]使用CNN作为骨干网络学习特征图,使用关键点估计模型提取语义局部特征,并使用联合学习将拓扑信息嵌入到局部特征中,以减轻遮挡带来的影响。Zhou等[17]提出了一种基于嵌入图匹配网络的遮挡行人重识别方法,使用嵌入特征对齐层,直接预测两幅图像每个特征的相似度分数。同时使用节点关系学习层,用于建立个体信息与关键点信息之间的对比特征关系,以便自动更新特征之间的关系。Huang等[18]提出了一种推理与调整图注意力网络(Reasoning and tuning graph attention network, RTGAT)。该方法不仅探索了部分特征和全局特征之间的语义相关性,还利用这种相关性推理出身体部位的可见性分数。然后,将这些可见性分数作为图注意力的依据,以指导图卷积网络抑制被遮挡部分特征的噪声,并将缺失的语义信息从完整图像传播到被遮挡的图像中。这些方法都采用CNN作为骨干网络,借助其深度结构能够有效地学习和提取图像中的边缘、纹理及形状等特征。然而,当图像中存在遮挡时,CNN往往会受到影响,导致特征提取过程主要针对可见特征进行处理,忽略了不可见特征,从而导致了特征提取的不准确性。

此外,由于自注意力机制在处理不同规模的图像和复杂度的任务时展现出的出色表现,Transformer也被应用于遮挡行人重识别。He等[19]第一次将Transformer引入行人重识别的任务中,其将图像分割为一系列图像序列,建立了基于Transformer的强大基线,并通过引入拼图补丁模块(Jigsaw patch module, JPM)和边信息嵌入(Side information embeddings, SIE)模块,进一步提高基线的鲁棒性。Zhao等[20]采用可学习的张量来增强ViT提取区域特征的能力,并引入特征的切片、整合和拼接操作,以增强图像对远距离依赖的感知。Li等[21]提出了一种端到端的部件感知Transformer(Part-aware Transformer, PAT)方法,通过Transformer的编码器-解码器架构,实现遮挡行人重识别任务。Tan等[22]利用分层语义和头部丰富模块来选择可见模式空间,以应对遮挡问题,并基于先验知识,实现端到端可训练网络中的自动对齐。Wang等[23]利用姿势信息分离语义组件,并选择性地匹配非遮挡部位。同时通过对ViT提取的图像块特征进行处理,实现了对可见身体部位的强化和遮挡特征的隔离。基于Transformer的方法能够对输入序列进行全局建模,更好地捕捉图像中不同区域之间的长距离依赖关系,从而有利于全局特征的提取。但是,由于图像被划分为图像块后,各图像之间的位置信息十分重要,但Transformer对于位置信息的建模主要依赖于位置编码机制,可能会受到编码方式的限制,对某些复杂的空间关系难以准确建模。

因此,为了更好地适应多样性的遮挡情况,有效处理遮挡问题,本文引入了可学习的双路注意力掩模生成器。该生成器通过关注不同通道和空间位置的特征,增强了网络在复杂场景中的表现能力。同时,提出了遮挡感知位置编码融合模块,使网络能够灵活地表达不同位置的语义信息,并有效进行空间关系建模。

1.2 注意力机制

注意力机制与人的注意感知一致,在处理大量信息时,会优先集中在突出的部分。Li等[24]将注意力机制表示为期望最大化,并通过迭代估算出一组更为紧凑的基数,在此基础上计算注意力图。这样的计算方法对输入具有鲁棒性,并在内存和计算量方面也有优势。Pan等[25]将卷积和自注意力机制融合在一起,使模型不仅能够享受自注意力和卷积的优势,又比纯卷积或自注意力的网络计算开销小。Hu等[26]引入了紧凑的压缩和激励模块来探索通道间关系,使用全局平均池化计算通道的注意力,自适应地重新校准通道特征响应。Koohpayegani等[27]提出了用于Transformer的简单注意力模块,对查询

和键矩阵使用简单的归一化,降低计算复杂性的同时动态改变计算顺序,实现对通道数量的线性计算。

1.3 位置编码

为了克服Transformer未直接处理序列数据的位置信息这一限制,引入了位置编码,以在输入序列中引入关于元素的位置信息。位置编码使Transformer能够感知和利用序列中元素的相对位置关系,从而更好地捕捉图像和文本等数据中的空间结构。Vaswani等[28]为了让模型利用输入序列的顺序,在编码器和解码器底部的输入嵌入中添加了位置编码。该方法使用不同频率的正弦和余弦函数,让模型来关注序列位置。Shaw等[29]提出了一种扩展自注意机制,有效地考虑相对位置或序列元素之间距离的表示,是一种相对位置编码方式。Wu等[30]提出了一种专门用于二维图像的全新的相对位置编码方法,该方法考虑了方向相对距离建模以及自注意机制中查询和相对位置嵌入之间的相互作用。

2 本文方法

2.1 总体框架

本文提出的基于可学习掩模和位置编码的遮挡行人重识别方法总体框架如图1所示。首先,将输入图像分成大小相同的小块,并通过线性投影将图像块展平为图像序列。然后,为了更好地聚合图像块信息,在图像序列中添加一个可学习的类令牌。此外,为了丰富图像结构信息,在每个序列元素以及类令牌添加位置嵌入和边信息嵌入,得到完整的图像序列。随后,图像序列经过L个Transformer块(Transformer block, TBlock)进行特征提取,得到图像序列特征$\boldsymbol{X}_i(i=1,2,\cdots,T)$。为了避免处理后出现位置信息损耗,进一步丰富图像位置信息,对输出TBlock的图像序列添加遮挡感知位置编码融合模块,通过多层次位置编码融合,建立图像块间更清晰的位置关系。同时,提取图像序列中的类令牌进行每个类的完整原型训练,以便计算三元组损失L_{tri},从而提高类内相似度,降低类间相似度。整体图像序列经归一化后,输入可学习的双路注意力掩模生成器,用于生成每个类对应的掩模。通过双路注意力,可以实现对图像各部分特征的有效关注,帮助网络构造更符合实际遮挡情况的掩模,从而使网络适应不同遮挡模式,更好地应对遮挡情况。具体来说,特征在输入LDAMG后,首先需要经过一个卷积层,再输入可学习通道注意力(Learnable channel attention, LCA)模块进行通道特征提取,经过可学习

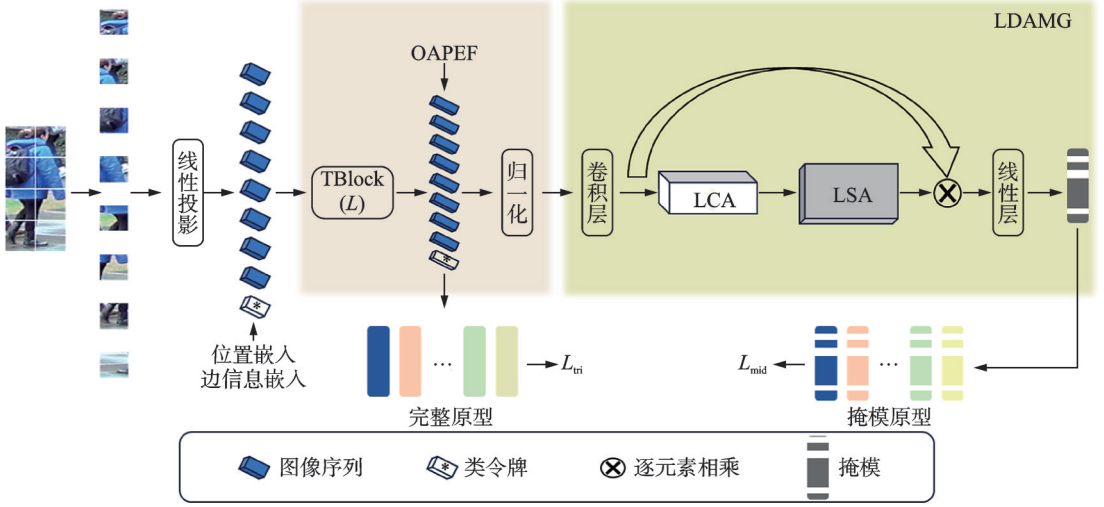

图1 LMPE总体框架

Fig.1 Overall framework of LMPE

空间注意力(Learnable spatial attention, LSA)模块进行空间特征提取,并与输入LCA的特征进行残差连接,生成最终掩模。为了平衡完整原型和掩模原型的尺度大小,同时突出训练时高质量掩模的重要性,采用掩模三元组损失L_{mid}对掩模进行优化。

2.2 遮挡感知位置编码融合模块

由于传统Transformer最初设计是用于处理文本序列数据,缺乏对元素位置信息的建模。但在图像数据中,元素的位置关系对于准确的特征提取至关重要。为了将Transformer成功应用于图像数据,研究者们普遍引入了位置编码,以赋予模型对元素位置的感知能力。然而,经过深度训练后,模型可能会损失对位置信息的敏感性,从而影响其在图像识别任务中的表现。同时遮挡行人的问题通常涉及到不同层次的特征,包括局部细节和全局上下文。为此,设计了位置编码融合模块,有助于模型更全面地捕捉不同层次信息,提高对行人多个层次特征的提取能力。位置编码融合模块格式为

$$\boldsymbol{P}_i=\sum_{j=1}^{\text{Layers}}\boldsymbol{P}_j\cdot\text{unsqueeze}(0) \tag{1}$$

式中:$\boldsymbol{P}_i(i=1,2,\cdots,T)$为融合后的位置编码,$i$表示图像序列中图像块对应的序号,$T$表示图像被分成的图像块数;$\boldsymbol{P}_j$为遮挡感知位置编码;Layers为位置编码融合的层数; unsqueeze(0)表示在张量的最前面增加一个新维度。相较于传统位置编码,遮挡感知位置编码采用基于正态分布的随机初始化,具备更强的表达能力。具体来说,首先创建了1个可学习的参数,用于表示遮挡感知的位置编码,该参数有3个维度,其中第1维表示批处理大小,第2维表示图像块数量,第3维表示每个位置的嵌入维度,决定了模型对每个位置的特征表达。对可学习参数进行正态分布初始化,这个初始值会在训练过程中通过梯度反向传播进行优化,以适应模型任务。这样获得的位置编码能够在训练过程中根据数据的特性学习适应当前任务的位置信息,从而在深度训练后仍能保持对位置关系的敏感性。

进一步,对于位置编码进行融合操作。首先在$\boldsymbol{P}_j$最外层维度插入1个新维度,进行维度扩展,方便后续处理。创建1个矩阵用于存储各层次的位置编码,在前向传播中,遍历每个层次的位置编码,进行累加计算。具体来说,在融合不同层次的位置编码时,每个维度都对应输入序列的不同位置。在位置编码中,每个位置都被映射到一个特定的嵌入向量。在前向传播中,通过将这些位置编码与输入张量相加,可以根据位置编码的权重对输入序列的每个位置进行调整。于是,不同位置的元素会获得不同的重要性,从而使得模型能够更灵活地处理遮挡情况。此外,位置编码和图像特征的融合采取如下的绝对位置编码形式

$$\boldsymbol{X}_i'=\boldsymbol{X}_i+\boldsymbol{P}_i \tag{2}$$

式中:$\boldsymbol{X}_i(i=1,2,\cdots,T)$表示经TBlock模块处理后的图像序列特征;$\boldsymbol{X}_i'(i=1,2,\cdots,T)$表示添加位置编码后的特征。将添加位置编码的特征经归一化后即可馈入LDAMG模块进行进一步的特征处理,同时提取其中的类令牌用于完整原型的训练。

为了强化完整原型的特征表达,引入了三元组损失进行优化。三元组损失可有效地推动模型学习到更具区分性的特征表达。在训练过程中,通过比较同一身份的正样本和不同身份的负样本之间的相似性,模型能够更准确地区分不同的身份,从而提升了识别的准确性,三元组损失具体格式为

$$L_{\text{tri}}=\left[\|\boldsymbol{H}-\boldsymbol{H}_{\text{P}}\|_2^2-\|\boldsymbol{H}-\boldsymbol{H}_{\text{N}}\|_2^2+\gamma\right]_+ \tag{3}$$

式中:$\boldsymbol{H}$表示用类令牌训练的完整原型;$\boldsymbol{H}_{\text{P}}$和$\boldsymbol{H}_{\text{N}}$分别表示对应完整原型的正样本和负样本;$\|\cdot\|$为欧式范数;$[v]_+=\max(0,v)$表示取最大值操作;$\gamma$为超参数。

位置编码融合模块允许模型捕捉不同尺度下的位置信息,能够更好地理解输入序列的结构。同

时，还允许模型对序列中不同位置的元素赋予不同的重要性，能够更好地处理遮挡情况。此外，位置编码融合赋予了模型灵活适应不同大小行人目标的能力，不受固定位置编码长度的限制，对处理不同尺度的行人目标以及有效提取遮挡情况下的特征至关重要。

2.3 可学习的双路注意力掩模生成器

为了构建更符合遮挡情况的掩模，提出可学习的双路注意力掩模生成器，通过对空间和通道的双路特征处理，更好地聚焦图像中的显著特征。同时，行人被遮挡部分通常隐含了上下文信息，例如行人走在道路上可能被车辆遮挡，这样的上下文信息是理解整个场景的关键因素。有效的掩模生成器能够学习这些上下文信息，并帮助网络选择合适的子空间，使网络更全面、更深入地理解整个场景。为了更具体展示LDAMG模块内部结构，首先详细介绍LCA和LSA模块的完整结构，如图2所示。

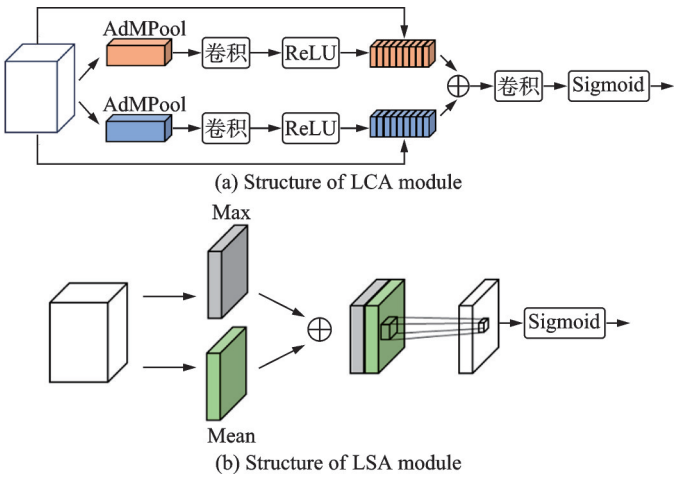

图2 LCA和LSA模块结构

Fig.2 Structures of LCA and LSA modules

由序列特征$\boldsymbol{X}'$经归一化以及卷积操作得到LCA的输入特征$\hat{\boldsymbol{X}}\in\mathbf{R}^{B\times C\times H\times W}$，其中$B$为批量大小，$C$为通道数，$H$和$W$为空间尺寸。首先要对其进行通道维度的注意力权重计算，通道注意力分数计算如下所示

$$s_{\max}=\mathrm{ReLU}(\mathrm{Conv}(\mathrm{AdMPool}(\hat{\boldsymbol{X}}))) \tag{4}$$

$$s_{\mathrm{avg}}=\mathrm{ReLU}(\mathrm{Conv}(\mathrm{AdAPool}(\hat{\boldsymbol{X}}))) \tag{5}$$

式中$s_{\max}$和s_{avg}为输入特征的通道注意力分数。在卷积操作中，每个卷积层生成的特征图由多个通道组成，每个通道代表不同的特征或特征组合。为了增强网络在学习和提取特征时的多样性，并构建更符合实际遮挡情况的掩模，采用了LCA进行特征处理，以提取不同特征之间的关系。具体而言，为了更有效地计算通道注意力，使用自适应最大池化(AdMPool)和自适应平均池化(AdAPool)对输入特征映射的空间维度进行压缩，以减少网络中需要学习的参数量。然后，通过卷积操作Conv和ReLU激活函数，获取每个通道注意力的得分。这一操作有助于网络建立不同特征之间的关系，进而提高网络在处理遮挡等复杂情况的图像时更好地捕获关键信息。将得到的通道注意力得分与原始输入特征进行加权融合，并获得最终的输出如下

$$O_{\mathrm{ch}}=\mathrm{Sigmoid}\left(\mathrm{Conv}\left(\left[\hat{\boldsymbol{X}}\otimes s_{\max},\hat{\boldsymbol{X}}\otimes s_{\mathrm{avg}}\right]\right)\right) \tag{6}$$

式中：O_{ch}表示可学习通道注意力模块的输出；$\otimes$表示逐元素相乘；$[\cdot]$表示对两个特征进行拼接操作。具体来说，输入特征跟通道注意力得分融合后，对自适应最大池化和自适应平均池化两个分支进行特征拼接，并再一次进行卷积。最后经过Sigmoid激活函数，获得可学习通道注意力模块的输出。经过权重融合的通道注意力，更加关注了对特定任务或特征更为关键的通道，明确了每个通道在局部区域的重要性。同时通过动态地调整每个通道的权重，使网络更容易适应输入数据的复杂特征分布，从而提高了网络的泛化性能。

对于可学习双路注意力掩模生成器中的LSA，首先分别对输入特征计算最大值和平均值，然后对结果进行拼接，进而获得最终输出结果，其格式如下

$$O_{\mathrm{sp}}=\mathrm{Sigmoid}\left(\mathrm{Conv}\left(\left[\max\left(O_{\mathrm{ch}}\right),\mathrm{mean}\left(O_{\mathrm{ch}}\right)\right]\right)\right) \tag{7}$$

式中:O_{sp}表示可学习空间注意力模块的输出;$\max(O_{\mathrm{ch}})$表示对特征进行最大池化处理;$\mathrm{mean}(O_{\mathrm{ch}})$表示对特征进行平均池化处理。不同于LCA,LSA关注的是特征在特征图中的位置和排列,可以明确强调或抑制某些位置上的特征。该模块通过计算输入特征图在每个空间位置的最大值和平均值,能够捕捉到图像中不同位置的重要信息,使网络更加关注包含重要特征的空间位置,特别是对于被遮挡行人的关键部位。同时,该模块还能够使网络自适应地调整对每个位置的关注程度,更好地适应不同遮挡情况。

对于整体的可学习双路注意力掩模生成器来说,首先需要对输入的特征图进行特征重塑,然后进行卷积操作,进一步提取和整合特征。使特征图分别经过可学习通道注意力和可学习空间注意力进行各自维度的加权,最后将最终的输出与输入张量进行残差连接,生成最终的掩模,其具体格式为

$$O=\mathrm{Linear}\left(\hat{\boldsymbol{X}}\otimes O_{\mathrm{sp}}\right) \tag{8}$$

输入特征经过残差连接后,再统一经过线性层,获得最终的掩模,通过通道和空间的双路加权,掩模可以更好地关注和利用不同通道和空间位置的特征,更加符合遮挡行人的被遮挡情况。

为了对完整分支和掩模分支进行有效约束,同时强调在训练阶段学习高质量掩模的重要性,采用交叉熵损失进行优化,其具体格式为

$$L_{\mathrm{mid}}=-\frac{1}{n}\sum_{t=1}^{n}\ln\left(p\left(O_t|O\right)\right) \tag{9}$$

式中:n表示每批内训练样本个数;O为生成的掩模;O_t为类标签;$p(O_t|O)$表示O被识别为类别标签O_t的预测概率。

3 实验结果与分析

3.1 数据集和评价指标

为了证明提出方法的有效性和优越性,在两个遮挡数据集Occluded-Duke[31]和Occluded-ReID[32],以及两个完整行人数据集Market-1501[33]和DukeMTMC-ReID[34]上分别进行了实验。

采用Rank-k[35]和平均精度均值(mean Average precision, mAP)[36]来定量评估模型性能。其中Rank-k用于衡量模型在前k个排名中正确识别的比例,mAP综合考虑了精确度和召回率,体现了模型的平均准确率。

3.2 参数设置

本文实验选择PyTorch作为实现框架,并利用强大的RTX 3090 GPU进行模型训练和推理。为了提高模型的学习能力,选用了在ImageNet21K[37]上预训练的ViT作为骨干网络。为了适应行人重识别任务需求,将输入图像的大小调整为256像素×128像素。此外,还采用多种数据增强技术,包括随机翻转、随机裁剪和随机擦除[38],以增加训练数据的多样性,提高模型的泛化能力。在优化器方面,采用随机梯度下降(Stochastic gradient descent, SGD)[39]作为优化器,并将学习率初始化为0.008。为了更好地调整学习率,使用余弦学习率衰减策略,有助于模型更快地收敛到最优解。对于遮挡感知位置编码融合模块,通过实验对层数取不同值进行了实验,在Occluded-Duke数据集上的实验结果如图3所示。由图3可知,当层数取值在1~3时,Rank-1和mAP的值整体呈轻微下

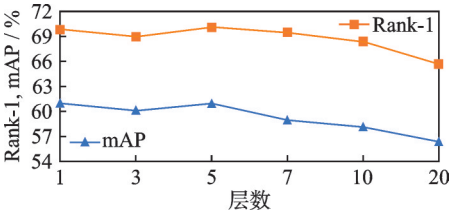

图3 层数大小的影响

Fig.3 Effect of layer size

降趋势;在取值为3~5时,呈上升趋势;在层数取值大于5时,开始出现大幅下降趋势。说明当层数取值过大时,模型在训练数据集上过度拟合,导致性能下降。在取值3~5时,模型性能有明显提升,同时在层数为5时,Rank-1的值也高于层数为1时的值,体现了位置编码融合的有效性,因此选取层数为5作为最终取值。

3.3 可视化实验

为了定性评估提出的LMPE模型的性能,分别展示了Occluded-Duke和Market-1501数据集上的可视化结果。

在Occluded-Duke遮挡数据集上展示提出的LMPE的识别性能,其可视化结果如图4所示,其中第1列查询图像是在该数据集上随机选取的,检索得到的正确和错误结果分别用绿色和红色序列号显示。对比方法是基于CNN的HoReID以及基于Transformer的TransReID。从图4中可以看出,HoReID在姿态信息的帮助下,检索到了一系列与查询图像行人姿态相似的图片,但忽略了行人的其他相关特征,导致检索结果并不理想。TransReID检测到了5个正确结果,展现出了基于Transformer的方法在遮挡行人重识别方向的优秀特征处理能力。提出的LMPE检测到了8个正确结果,识别准确率得到有效提升,表明了LMPE在处理遮挡问题的有效性和优越性。

图4 Occluded-Duke数据集的可视化结果

Fig.4 Visualization results of Occluded-Duke dataset

此外在Market-1501无遮挡数据集上展示LMPE和其他两种不同对比方法的可视化结果,其可视化结果如图5所示,其中第1列查询图像是在该数据集上随机选择的,检索得到的正确和错误结果分别用绿色和红色序列号显示。从图5可以看出,HoReID检测到了4个正确结果;TransReID检测到8个正确结果;提出的LMPE检测结果全部正确,达到了最高的检测精度。与TransReID相比,提出的LMPE具有更高的鲁棒性,不容易检索到异常值。

图5 Market-1501数据集的可视化结果

Fig.5 Visualization results of Market-1501 dataset

总的来说,无论是遮挡情况还是非遮挡情况,LMPE都表现出最出色的性能。

3.4 对比实验

为了证明本文方法的有效性,在两个公开的遮挡数据集上定量比较了提出的LMPE和一些先进的行人重识别方法。用于对比的方法包括基于CNN的方法[8,9,16,18]和基于Transformer的方法[12,19,21-23]。其中ISP、HoReID、TransReID、DPM以及PFD是在相同服务器环境下根据文献提供的代码进行实验得到的数据,其他结果来自于提出该方法的文献。具体实验结果如表1所示。

表1 遮挡数据集上的实验性能

Table 1 Performance comparison on occluded datasets %

模型	Occluded-Duke		Occluded-ReID	
	Rank-1	mAP	Rank-1	mAP
ISP[8]	61.2	49.8	—	—
PVPM[9]	47.0	37.7	70.4	61.2
HoReID[16]	54.9	43.6	79.3	70.1
RTGAT[18]	61.0	50.1	71.8	51.0
FED[12]	68.1	56.4	**86.3**	79.3
TransReID[19]	67.1	59.6	81.5	76.2
PAT[21]	64.5	53.6	81.6	72.1
DPM[22]	67.6	58.6	84.2	76.9
PFD[23]	66.3	59.1	80.5	76.7
LMPE	**71.8**	**61.4**	84.6	**79.4**

基于CNN的方法通过多层卷积和池化操作,有助于从图像中学习抽象特征。但是在面对大范围遮挡时,由于卷积操作的局部感受野,遮挡可能导致部分特征信息的丢失,影响模型性能。ISP和HoReID通过辅助网络提取行人姿态,弥补了基于CNN网络因遮挡带来特征信息丢失的问题,ISP在Occluded-Duke数据集上Rank-1可达到61.2%。基于Transformer方法能够对整个输入序列进行全局关系建模,而不受卷积操作的感受野限制。同时自注意力机制和多头注意力机制使其在处理不同尺度和复杂关系时更具有可扩展性,有助于适应遮挡场景的多样性。基于Transformer方法在遮挡行人重识别领域都获得了优异的性能,例如,PFD的Rank-1和mAP达到66.3%和59.1%。与PFD相比,提出的LMPE的Rank-1和mAP分别提高了5.5%和2.3%,达到了71.8%和61.4%,在遮挡行人重识别任务上展现出了优异的性

能。在Occluded-Duke数据集上，LMPE的Rank-1和mAP均高于FED。在Occluded-ReID数据集上，LMPE的Rank-1比FED低了1.7%，mAP高出0.1%。FED采用的遮挡增强策略，主要针对常见遮挡发生的4个位置（即上、下、左、右各方向上$\frac{1}{4}\sim\frac{1}{2}$区域）进行进一步处理，符合实际情况，不受数据集大小限制。因此无论是主流Occluded-Duke数据集还是小型Occluded-ReID数据集都有良好的表现。提出的LMPE需要在训练过程中收集大量数据来不断进行优化掩模的生成，以便网络能适应多种情况，因此可能在小型数据集上性能略差于FED，但在主流数据集上性能好于FED。

为了验证提出方法的通用性，在完整行人重识别数据集Market-1501和DukeMTMC-ReID上与一些先进的行人重识别方法进行定量比较。用于对比的方法包括基于CNN的方法[8,10,16,18,31]，和基于Transformer的方法[12,14,19,23]。其中ISP、PCB、HoReID、NFormer、TransReID以及PFD是在相同服务器环境下进行实验得到的数据，其他结果来自于提出该方法的作者。具体实验结果如表2所示。从表2可以看出，提出的LMPE在Market-1501中Rank-1达到95.7%，mAP达到89.6%，展现出了Transformer的强大特征处理能力。与TransReID相比，提出方法的Rank-1和mAP分别提高了2.5%和2.7%。在DukeMTMC-ReID数据集上，与先进的方法相比，提出的LMPE依旧体现出了出色的性能，Rank-1达到90.9%，mAP达到83.1%。面对完整行人图像，提出的LDAMG可以使模型在训练过程中更好地处理图像噪声，提高了模型的鲁棒性。同时，还能通过学习上下文信息，建立行人与周围环境的关系，提高场景的理解力。即使提出的LMPE是专门为遮挡行人重识别任务设计的，但是在完整行人图像上，依然展现出优秀性能，体现了提出方法的普适性。

表2 完整数据集上的实验性能

Table 2 Performance comparison on holistic datasets %

模型	Market-1501		DukeMTMC-ReID	
	Rank-1	mAP	Rank-1	mAP
ISP[8]	95.0	88.1	89.5	80.1
PCB[10]	92.1	76.6	80.7	65.3
HoReID[16]	93.2	81.8	85.1	72.5
RTGAT[18]	95.3	88.2	89.1	80.2
PGFA[31]	91.2	76.8	82.6	65.5
FED[12]	95.0	86.3	89.4	78.0
NFormer[14]	93.2	83.7	90.3	82.1
TransReID[19]	93.2	86.9	90.2	81.0
PFD[23]	94.9	88.1	90.4	82.2
LMPE	**95.7**	**89.6**	**90.9**	**83.1**

3.5 消融实验

本文提出的方法主要由LDAMG和OAPEF组成，其中LDAMG又包含LCA和LSA两个子模块。为了验证每个模块的有效性，在遮挡数据集Occluded-Duke上进行了消融实验，实验结果如表3所示。实验选取DPM网络作为基线模型，基线结果根据文献[22]提供的代码得到。在基线上引入LCA后，Rank-1和mAP分别提高了1.2%和1.6%，表明该模块在性能上具有一定的促进作用。但是Rank-5和Rank-10存在轻微下降，表明LCA会导致模型更关注对特定行人特征的通道，忽视了其他通道的特征，这种权衡会影响到在Rank-5和Rank-10这些更宽松的匹配条件下的识别性能。在基线上引入LSA后，虽然Rank-1和mAP的提升略少于LCA，但LSA能够更有效地建模图像中不同区域之间的空间关系，有助于在更宽松的匹配条件下提升识别准确度。在基线上添加LDAMG后，Rank-1和mAP分别提高了2.5%和2.7%。这表明提出的LDAMG在LCA和LSA的共同作用下，其能够更精细地关注图像中的重要区域，更好地学习局部和全局信

表3 消融实验效果

Table 3 Results of ablation experiment %

模型	Occluded-Duke			
	Rank-1	Rank-5	Rank-10	mAP
Base	67.6	81.1	84.7	58.6
Base+LCA	68.8	80.4	84.6	60.2
Base+LSA	68.6	81.2	85.1	59.9
Base+LDAMG	69.3	80.5	84.8	60.7
Base+OAPEF	70.1	82.6	86.4	61.0
Base+LDAMG+OAPEF	**71.8**	**83.9**	**87.6**	**61.4**

息。因此,生成的掩模能够适应不同的遮挡模式,使网络更加灵活。在基线上添加OAPEF后,Rank-1达到70.1%,mAP达到61.0%,相比基线分别提高了2.5%和2.4%。这说明了OAPEF的有效性,它获得丰富的位置信息,并建立了各图像块之间的关联。在基线上同时加入LDAMG和OAPEF后,Rank-1和mAP都展现出更加出色的效果,分别达到71.8%和61.4%。实验结果证实了LDAMG和OAPEF对提高模型性能均有积极的促进作用。

4 结束语

为了从被遮挡行人中提取更有效的特征,提出了基于可学习掩模和位置编码的遮挡行人重识别方法。一方面,为了使网络更适应遮挡情况,提出了可学习的双路注意力掩模生成器,生成的掩模可以选择合适的子空间,使模型在处理遮挡问题时更加灵活。另一方面,由于传统Transformer缺少位置信息,还提出遮挡感知位置编码融合模块,通过融合不同层次位置信息,使模型更准确地理解行人与场景之间关系,提高了遮挡行人重识别任务的识别准确率。实验表明,提出的LMPE能够显著提升行人重识别效果。未来的研究中,将进一步深入研究遮挡问题,考虑对遮挡信息与掩模信息进行有效融合,通过生成的掩模,引导网络获得准确定位遮挡位置,进一步提高识别准确率。

参考文献:

[1] YE Meng, CHEN Shuoyi, LI Chenyue, et al. Transformer for object re-identification: A survey[EB/OL]. (2024-01-13)[2024-03-30]. https://arxiv.org/pdf/2401.06960.

[2] 夏道勋, 郭方, 刘浩杰,等. 开放式行人再识别研究进展综述[J]. 数据采集与处理, 2021, 36(3): 449-467.
XIA Daoxun, GUO Fang, LIU Haojie, et al. Review on research progress of open-world person re-identification[J]. Journal of Data Acquisition and Processing, 2021, 36(3): 449-467.

[3] ZAHRA A, PERWAIZ N, SHAHZAD M, et al. Person re-identification: A retrospective on domain specific open challenges and future trends[J]. Pattern Recognition, 2023, 142: 109669.

[4] PENG Yunjie, HOU Saihui, CAO Chunshui, et al. Deep learning based occluded person re-identification: A survey[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(3): 1-27.

[5] NING Enhao, WANG Changshuo, ZHANG Huang, et al. Occluded person re-identification with deep learning: A survey and perspectives[J]. Expert Systems with Applications, 2023, 239: 122419.

[6] NGUYEN V D, KHALDI K, NGUYEN D, et al. Contrastive viewpoint-aware shape learning for long-term person re-identification[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA: IEEE, 2024: 1041-1049.

[7] NING Enhao, WANG Yangfan, WANG Changshuo, et al. Enhancement, integration, expansion: Activating representation of detailed features for occluded person re-identification[J]. Neural Networks, 2024, 169: 532-541.

[8] ZHU Kuan, GUO Haiyun, LIU Zhiwei, et al. Identity-guided human semantic parsing for person re-identification[C]//Proceedings of Computer Vision ECCV 2020: 16th European Conference. Glasgow, UK: Springer International Publishing, 2020: 346-363.

[9] GAO Shang, WANG Jingya, LU Huchuan, et al. Pose-guided visible part matching for occluded person reid[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2020: 11744-11752.

[10] SUN Yifan, ZHENG Liang, YANG Yi, 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. Cham: Springer, 2018: 480-496.

[11] CHENG Xinhua, JIA Mengxi, WANG Qian, et al. More is better: Multi-source dynamic parsing attention for occluded person re-identification[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM, 2022: 6840-6849.

[12] WANG Zhikang, ZHU Feng, TANG Shixiang, et al. Feature erasing and diffusion network for occluded person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2022: 4754-4763.

[13] HONG Xing, ZHANG Langwen, YU Xiaoyuan, et al. MBA-Net: Multi-branch attention network for occluded person re-identification[J]. Multimedia Tools and Applications, 2024, 83(2): 6393-6412.

[14] WANG Haochen, SHEN Jiayi, LIU Yongtuo, et al. NFormer: Robust person re-identification with neighbor transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2022: 7297-7307.

[15] ZHANG Xin, FU Keren, ZHAO Qijun. Dynamic patch-aware enrichment transformer for occluded person re-identification [EB/OL]. (2024-02-16)[2024-03-30]. https://arxiv.org/html/2402.10435v1.

[16] WANG Guanan, YANG Shuo, LIU Huanyu, 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. Seattle: IEEE Computer Society, 2020: 6449-6458.

[17] ZHOU Shuren, ZHANG Mengsi. Occluded person re-identification based on embedded graph matching network for contrastive feature relation[J]. Pattern Analysis and Applications, 2023, 26(2): 487-503.

[18] HUANG Meiyan, HOU Chunping, YANG Qingyuan, et al. Reasoning and tuning: Graph attention network for occluded person re-identification[J]. IEEE Transactions on Image Processing, 2023, 32: 1568-1582.

[19] HE Shuting, LUO Hao, WANG Pichao, et al. TransReID: Transformer-based object re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 15013-15022.

[20] ZHAO Yunbin, ZHU Songhao, WANG Dongsheng, et al. Short range correlation transformer for occluded person re-identification[J]. Neural Computing and Applications, 2022, 34(20): 17633-17645.

[21] LI Yulin, HE Jianfeng, ZHANG Tianzhu, 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. Seattle: IEEE Computer Society, 2021: 2898-2907.

[22] TAN Lei, DAI Pingyang, JI Rongrong, et al. Dynamic prototype mask for occluded person re-identification[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM, 2022: 531-540.

[23] WANG Tao, LIU Hong, SONG Pinhao, et al. Pose-guided feature disentangling for occluded person re-identification based on Transformer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022, 36(3): 2540-2549.

[24] LI Xia, ZHONG Zhisheng, WU Jianlong, et al. Expectation-maximization attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korean: IEEE, 2019: 9167-9176.

[25] PAN Xuran, GE Chunjiang, LU Rui, et al. On the integration of self-attention and convolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2022: 815-825.

[26] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7132-7141.

[27] KOOHPAYEGANI S A, PIRSIVASH H. SimA: Simple softmax-free attention for vision transformers[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA: IEEE, 2024: 2607-2617.

[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[29] SHAW P, USKOREIT J, VASWANI A. Self-attention with relative position representations[EB/OL]. (2018-04-12)[2024-03-30]. https://arxiv.org/pdf/1803.02155.

[30] WU Kan, PENG Houwen, CHEN Minghao, et al. Rethinking and improving relative position encoding for vision transformer [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 10033-10041.

[31] MIAO Jiaxu, WU Yu, LIU Ping, et al. Pose-guided feature alignment for occluded person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korean: IEEE, 2019: 542-551.

[32] ZHUO Jiaxuan, CHEN Zeyu, LAI Jianhuang, et al. Occluded person re-identification[C]//2018 IEEE International Conference on Multimedia and Expo. San Diego, CA, United States: IEEE, 2018: 1-6.

[33] ZHENG Liang, SHEN Liyue, TIAN Lu, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Computer Society, 2015: 1116-1124.

[34] ZHENG Zhedong, ZHENG Liang, YANG Yi. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3754-3762.

[35] DENG Yichuan, LI Zhihang, SONG Zhao. An improved sample complexity for rank-1 matrix sensing[EB/OL]. (2023-03-13)[2024-03-30]. https://arxiv.org/pdf/2303.06895.

[36] 孙明浩,王洪元,吴琳钰,等. 基于特征金字塔分支和非局部关注的行人重识别[J]. 数据采集与处理, 2023,38(1): 121-131.
SUN Minghao, WANG Hongyuan, WU Linyu, et al. Person re-identification based on feature Pyramid branch and non-local attention[J]. Journal of Data Acquisition and Processing, 2023, 38(1): 121-131.

[37] PINTOR M, ANGIONI D, SOTGIU A, et al. ImageNet-Patch: A dataset for benchmarking machine learning robustness against adversarial patches[J]. Pattern Recognition, 2023, 134: 109064.

[38] GARCEA F, SERRA A, LAMBERTI F, et al. Data augmentation for medical imaging: A systematic literature review[J]. Computers in Biology and Medicine, 2023, 152: 106391.

[39] BEZNOSIKOV A, GORBUNOV E, BERARD H, et al. Stochastic gradient descent-ascent: Unified theory and new efficient methods[C]//Proceedings of International Conference on Artificial Intelligence and Statistics. Valencia, Spain: IEEE, 2023: 172-235.

作者简介:

杨真真(1984-),**通信作者**,女,副教授,研究方向:深度学习、计算机视觉,E-mail: yangzz@njupt.edu.cn。

陈亚楠(2000-),女,硕士研究生,研究方向:深度学习、计算机视觉,E-mail: 1373792125@qq.com。

杨永鹏(1984-),男,博士研究生,研究方向:深度学习、计算机视觉,E-mail: yangyp@njcit.cn。

吴心怡(1999-),女,硕士研究生,研究方向:深度学习、计算机视觉,E-mail: 1323746365@qq.com。

(编辑:张黄群)