ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 2,Mar. 2023,pp. 375－385
DOI:10. 16337/j. 1004-9037. 2023. 02. 013

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

基于改进型Transformer编码器和特征融合的行人重识别

赵　倩，薛超晨，赵　琰

（上海电力大学电子与信息工程学院，上海 201306）

摘　要：为了解决Transformer编码器在行人重识别中因图像块信息丢失以及行人局部特征表达不充分导致模型识别准确率低的问题，本文提出改进型Transformer编码器和特征融合的行人重识别算法。针对Transformer在注意力运算时会丢失行人图像块相对位置信息的问题，引入相对位置编码，促使网络关注行人图像块语义化的特征信息，以增强行人特征的提取能力。为了突出包含行人区域的显著特征，将局部patch注意力机制模块嵌入到Transformer网络中，对局部关键特征信息进行加权强化。最后，利用全局与局部信息特征融合实现特征间的优势互补，提高模型识别能力。训练阶段使用Softmax及三元组损失函数联合优化网络，本文算法在Market1501和DukeMTMC-reID两大主流数据集中评估测试，Rank-1指标分别达到97.5%和93.5%，平均精度均值（mean Average precision，mAP）分别达到92.3%和83.1%，实验结果表明改进型Transformer编码器和特征融合算法能够有效提高行人重识别的准确率。

关键词：计算机图像处理；行人重识别；局部注意力；相对位置编码；特征融合；Transformer
中图分类号：TP391　　**文献标志码：**A

Person Re-identification Method Based on Improved Transformer Encoder and Feature Fusion

ZHAO Qian，XUE Chaochen，ZHAO Yan

（School of Electronics and Information Engineering，Shanghai University of Electric Power College，Shanghai 201306，China）

Abstract：In order to solve the problem of low accuracy of Transformer encoder caused by the loss of person image blocks information and insufficient expression of person local features in person re-identification，an improved Transformer encoder and feature fusion algorithm for person re-identification is proposed. This algorithm uses relative position encoding to solve the problem that Transformer will lose the relative position information of person image blocks during attention operation so that the network can focus on the semantic feature information of person image blocks，thus enhancing the ability to extract pedestrian features. Secondly，the local patch attention module is embedded into the Transformer network to weighted strengthen the local key feature information and highlight the significant features of the person area. Finally，the fusion of global and local information features is used to achieve complementary advantages between features and improve the recognition ability of the model. In the training stage，Softmax and triple loss functions are used to jointly optimize the network. The proposed algorithm is

基金项目：国家自然科学基金（61802250）。
收稿日期：2022-04-05；**修订日期：**2022-08-27

experimentally compared and analyzed on the mainstream datasets of Market1501 and DukeMTMC-reID. The Rank-1 accuracy reaches 97.5% and 93.5% respectively, and the mean average precision (mAP) reaches 92.3% and 83.1% respectively. The experimental results show that the improved Transformer encoder and feature fusion algorithm can effectively improve the accuracy of person re-identification.

Key words: computer image processing; person re-identification; local attention; relative position coding; feature fusion; Transformer

引　言

行人重识别是一种跨视域的行人跟踪与检索技术,它是将目标行人图像与不同位置的摄像机拍摄的行人图像库或者视频序列进行匹配,从而完成跨域设备的行人检索。行人重识别技术在现实场景中广泛应用于交通、无人超市、安保和刑侦等领域,是当前计算机图像处理领域的研究热点。然而,由于拍摄场景的情况不同,会出现低分辨率、视角遮挡、姿态变化以及光照等一系列问题,给行人重识别技术的应用带来了诸多挑战。

在现有的行人重识别研究[1-2]中,传统算法主要通过特征提取以及距离度量来解决行人跨域检索的问题。首先,对行人穿衣颜色、轮廓等特征进行提取,得到特征向量;其次,通过对特征向量距离度量,使得在行人分类过程中属于同一行人类内间距离变小和不同行人类间距离变大。然而传统算法对行人特征表达不充分,难以解决行人姿态、拍摄视角、穿衣变化等问题[3]。Zheng等[4]提出以ResNet为主干、行人的ID(Identity)作为训练标签的IDE(ID-discriminative embedding)网络,成了行人重识别领域强大的基线。但该网络以提取行人的全局特征为依据,难以解决不同行人轮廓相似和外观属性相同的问题。Wang等[5]提出采用局部卷积基线(Part-based convolutional baseline,PCB),将行人特征图水平划分,从而提取分块局部特征;Szegedy等[6]提出多分支结构的多粒度网络(Multiple granularities network,MGN)以及Yan等[7]提出不同空间粒度的超图网络(Multi-granular hypergraph,MGH),通过融合全局特征与多粒度局部特征来增强算法模型对行人特征的提取能力。对于遮挡和视角问题以及背景干扰带来的负面影响[8-9],研究者提出基于注意力机制的行人重识别网络[10-12],例如CBAM、SENet和RGA网络[13-15]用注意力机制解决此类问题并提高匹配精度。此外,针对匹配效率差遏制行人重识别的可应用性问题,研究者通过对行人特征进行编码的方式提高匹配效率[16-17]。近两年来,学者们将Transformer思想运用到计算机视觉领域并取得了不错的效果。Vision Transformer(Vit)[18]、DeiT[19]等网络引入多头注意力模块解决了卷积神经网络(Convolutional neural network, CNN)在特征提取阶段卷积感受野不足的问题;He等[20]提出Transformer-baseline模型并首次融入到行人重识别领域,且引入辅助信息嵌入模块(Side information embedding,SIE)和拼图补丁模块(Jigsaw patch module,JPM)模块进一步改进构成Transreid网络得到更好的识别结果。但在已有结果中,使用Transformer作为编码器的表现并不理想,Wu等[21]认为原因是该模型在经过多头注意力运算后,采用绝对位置编码方式会丢失图像块中的相对位置信息,从而影响模型的识别效果。并且行人图像在分块处理后,如何让Transformer网络更关注包含行人特征的图像块,抑制无用的背景信息,同样对识别效果具有重要的影响作用。此外,只对编码器输出的全局信息学习与训练,忽略了关键的局部信息,从而导致行人重识别的准确性降低和识别鲁棒性差的结果。

针对以上问题,本文对Transformer-baseline网络模型进行改进,提出一种基于改进型Transformer编码器和特征融合的行人重识别算法。该方法具体如下:引入图像相对位置编码(Image relative position encoding,IRPE)学习行人图像块的相对距离,使得网络模型能够根据该距离自适应地调整注意力权值,增强对行人图像块邻近区域特征的表达能力;其次,在多头注意力机制之后,加入局部patch注意

力机制模块(Local patch attention,LPA),对包含行人区域的显著特征加权强化提取关键特征信息;最后,对编码器输出的全局与局部信息进行特征融合增强行人特征的表达能力。在Market1501和DukeMTMC-reID数据集上进行多次数据测评,实验结果表明该网络模型能够有效提高行人重识别的性能。

1　本文算法

1.1　网络框架

本文网络框架如图1所示,主要包括4个部分:行人图像分块预处理、相对位置信息编码、LPA-Transformer编码器和特征融合输出。算法具体流程为:

第1步　行人图像分块预处理。对通道数、高度和宽度分别为C、H和W的输入图像$\boldsymbol{X}$,首先用滑动窗口对其进行分块处理,得到N个图像块的序列$(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$,每个图像块$\boldsymbol{x}_i\in\mathbf{R}^{C\times p\times p}$,其中$p$表示每个图像块的维度,且$N=HW/p^2$。

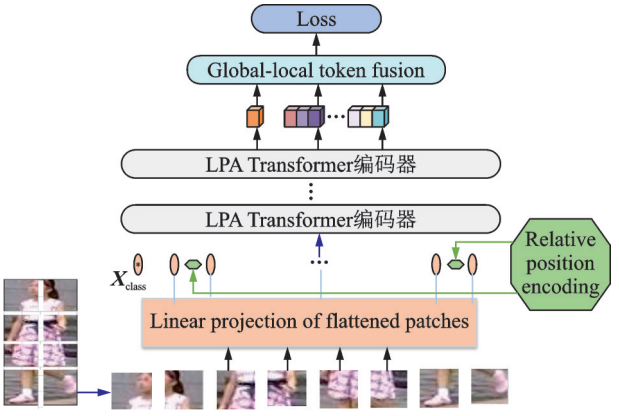

图1　改进型Transformer编码器和特征融合网络框架

Fig.1　Improved Transformer encoder and feature fusion network framework

第2步　将相对位置编码信息嵌入到图像块序列中。将此图像块序列通过一个可学习的嵌入矩阵$F\in\mathbf{R}^{(p^2\times C)\times D}$线性映射成一个$D$维向量,在$D$维向量之前添加一个可学习的分类标记$\boldsymbol{x}_{\text{class}}$,利用$\boldsymbol{f}_{\text{IRPE}}$引入图像块的相对位置信息,得到一个带有相对位置信息的图像块序列$\boldsymbol{z}_0$,表达式为

$$\boldsymbol{z}_0=(\boldsymbol{x}_{\text{class}};\boldsymbol{x}_1F;\boldsymbol{x}_2F;\cdots;\boldsymbol{x}_NF)+\boldsymbol{f}_{\text{IRPE}}\quad \boldsymbol{f}_{\text{IRPE}}\in\mathbf{R}^{(N+1)\times D}\tag{1}$$

第3步　经过嵌入表示后的图像块$\boldsymbol{z}_0$输入到LPA-Transformer编码器(Encoder)中。如图2所示,其中虚线框代表与原Transformer编码器相比融入了LPA注意力;LPA-Transformer编码器由层归一化(Layer norm, LN)、多头注意力机制(Multiheaded self-attention, MSA)、LPA注意力以及多层感知器(Multilayer perceptron, MLP)顺序处理[22],表达式为

$$\boldsymbol{z}'_l=\text{MSA}(\text{LN}(\boldsymbol{z}_{l-1}))+\boldsymbol{z}_{l-1}\quad l=1,2,\cdots,L\tag{2}$$

$$\boldsymbol{z}''_l=\text{LPA}(\text{LN}(\boldsymbol{z}'_l))+\boldsymbol{z}'_l\quad l=1,2,\cdots,L\tag{3}$$

$$\boldsymbol{z}_l=\text{MLP}(\text{LN}(\boldsymbol{z}''_l))+\boldsymbol{z}''_l\quad l=1,2,\cdots,L\tag{4}$$

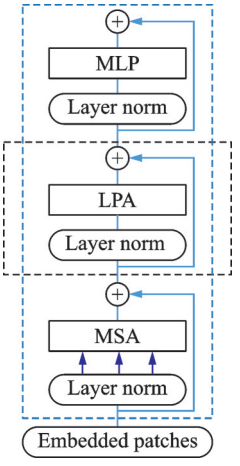

图2　LPA-Transformer编码器

Fig.2　LPA-Transformer encoder

第4步　特征融合。通过最后一层编码器得到特征尺度不变的行人图像块token$_{\boldsymbol{z}_l}$,其中包括代表行人全局特征$\boldsymbol{x}_l^0$,以及N个行人局部特征$[\boldsymbol{x}_l^1,\boldsymbol{x}_l^2,\cdots,\boldsymbol{x}_l^N]$,即$\boldsymbol{z}_l=[\boldsymbol{x}_l^0,\boldsymbol{x}_l^1,\boldsymbol{x}_l^2,\cdots,\boldsymbol{x}_l^N]$。为了避免局部信息丢失,提出全局与局部Token融合模块将局部信息融合到全局信息提高网络模型的识别能力。

1.2　图像块相对位置编码

Transreid-baseline编码器采用绝对位置编码方式,即通过一个可学习的向量对图像块按照顺序的

方式从1编码到N,这种编码方式仅标记每个图像块的位置信息,忽视了不同图像块间具有空间位置相对性。文献[21]发现不同图像块间的位置关系在图像处理领域中也有重要作用。如图3所示,以0号图像块作为参考,第1组2、4、5、7号标定的图像块与0号参考图像块的距离相同,小于第2组1、3、6、8号与0号参考图像块之间的距离,且第1组标定的图像块与0号位置的行人特征相似度要高于第2组。因此将绝对位置编码方式改进为相对位置编码,通过计算图像块相对位置距离$D(i,j)$,并对不同的距离分布施加不同程度的注意力来提高模型对行人特征的表达能力,计算过程为

$$D(i,j)=g\left(\sqrt{(\tilde{x}_i-\tilde{x}_j)^2+(\tilde{y}_i-\tilde{y}_j)^2}\right) \tag{5}$$

$$g(x)=\begin{cases}[x] & |x|\leqslant\alpha\\ \operatorname{sign}(x)\times\min\left(\beta,\dfrac{\ln(|x|/\alpha)}{\ln(|\gamma|/\alpha)}(\beta-\alpha)\right) & |x|>\alpha\end{cases} \tag{6}$$

式中:$D(i,j)$为通过计算得到两个不同图像块$(\tilde{x}_i,\tilde{y}_i)$、$(\tilde{x}_j,\tilde{y}_j)$之间的欧氏距离,为了减少位置编码带来的计算量和参数量,通过如图4所示的分段索引函数$g(x)$,将$D(i,j)$映射到$[-\beta,\beta]$范围内;x为不同图像块之间的相对距离;γ为调整对数部分的曲率;$[\cdot]$为舍入运算;sign$(\cdot)$为符号函数;α决定分段点。

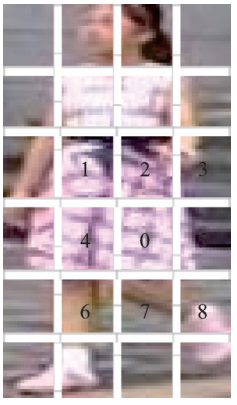

图3 图像相对位置编码示意图

Fig.3 Schematic diagram of IRPE

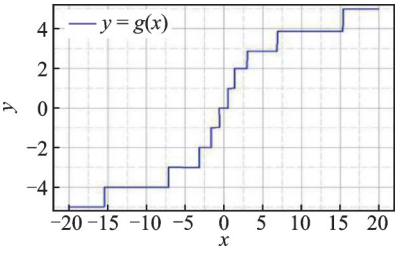

图4 分段索引函数

Fig.4 Piecewise index function

相对位置编码通过查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$进行交互计算并嵌入到多头注意力机制中,并采用乘积法的相对位置编码方式,表达式为

$$\boldsymbol{p}_{ij}=\boldsymbol{p}_{D^x(i,j),D^y(i,j)} \tag{7}$$

式中$\boldsymbol{p}_{ij}$代表的编码方式通过乘积法将与参考位置距离相同的图像块位置编码到多头注意力机制中。利用这种编码方式将输入图像块的相对位置编码到$\boldsymbol{p}_{ij}^{Q}$、$\boldsymbol{p}_{ij}^{K}$、$\boldsymbol{p}_{ij}^{V}$编码向量中,并通过多头注意力模块计算分配不同距离、不同程度的注意力来提高对行人特征的表达能力,计算过程为

$$\boldsymbol{z}_i=\sum_{j=1}^{N}\boldsymbol{\alpha}_{ij}(\boldsymbol{x}_j\boldsymbol{W}^v+\boldsymbol{p}_{ij}^{V}) \tag{8}$$

$$\boldsymbol{\alpha}_{ij}=\frac{\exp(\boldsymbol{e}_{ij})}{\sum_{k=1}^{N}\exp(\boldsymbol{e}_{ik})} \tag{9}$$

$$\boldsymbol{e}_{ij}=\frac{(\boldsymbol{x}_j\boldsymbol{W}^q+\boldsymbol{p}_{ij}^{Q})(\boldsymbol{x}_j\boldsymbol{W}^k+\boldsymbol{p}_{ij}^{K})^{\mathrm{T}}}{\sqrt{d_z}} \tag{10}$$

式中:$\boldsymbol{x}_i$为输入图像块;权重$\boldsymbol{\alpha}_{ij}$是对$\boldsymbol{e}_{ij}$经过Softmax函数的输出;$\boldsymbol{z}_i$为编码器输出行人序列特征;$\boldsymbol{p}_{ij}^{Q}$、$\boldsymbol{p}_{ij}^{K}$、$\boldsymbol{p}_{ij}^{V}$为相对位置编码向量;$\boldsymbol{W}^q$、$\boldsymbol{W}^k$、$\boldsymbol{W}^v$为自注意力权重矩阵。

1.3 局部patch注意力

将图像分块以后,由于每个图像块包含的行人特征信息不尽相同,甚至小部分图像块不包含行人特征信息。因此,能够重点提取包含行人信息图像块的特征,并抑制无用的图像块信息,是解决此问题的关键点,对此提出局部patch注意力,结构如图5所示。

首先将图2中MSA模块输出的维度大小为$(N+1)\times D$的patch中分离出来可学习的分类标记$\boldsymbol{x}_{\text{class}}$,如图5所示。通过维度变换将patch转换为三维图像信息$\boldsymbol{f}$,之后通过通道注意力机制对每个通道实现了差异化处理。由于平均池化能提取行人全局信息,最大池化可以抑制背景信息,故使用超参数λ和μ分别控制最大池化分支和平均池化分支的权重,将池化得到的两个特征输入到多层感知机进行特征学习之后对特征加权融合,并通过Sigmoid函数做非线性变换分配不同的权重,之后与原来的特征$\boldsymbol{f}$相乘经过维度反变换得到不同权重的patch。LPA表达式为

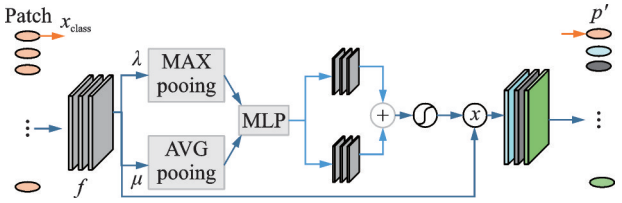

图5　局部patch注意力结构图

Fig.5　Structure diagram of local patch attention

$$\mathrm{LPA}(\boldsymbol{p})=\left[\delta(\mathrm{MLP}(\lambda\mathrm{MaxPool}(\gamma(\boldsymbol{p}))))\oplus\delta(\mathrm{MLP}(\mu\mathrm{AvgPool}(\gamma(\boldsymbol{p}))))\right] \tag{11}$$

$$\boldsymbol{p}'=\gamma'[\mathrm{LPA}(\boldsymbol{p})\otimes\boldsymbol{p}]+\boldsymbol{z}^0(\boldsymbol{x}_{\text{class}}) \tag{12}$$

式中:$\boldsymbol{p}$代表patch;γ和γ'为维度变换与反变换;δ为Sigmoid函数;λ和μ为超参数;AvgPool和MaxPool分别为平均池化和最大池化;$\boldsymbol{p}'$代表经过LPA注意力之后的行人特征。

1.4　特征融合

Transformer-baseline网络更多关注全局特征的学习,将行人全局特征$\boldsymbol{x}_l^0$输入到训练网络来约束模型,忽略关键行人局部特征信息;因此如何将更具有代表性和判别力的局部特征与全局特征融合,成为提升行人重识别性能的关键。在HPM网络[23]中,根据人体基本结构,将行人分成上、下半身2个部分提取局部特征;之后分别将行人上半身和下半身再进行2等分、3等分来提取较精细但必要的局部特征;受此启发,本文提出全局与局部token融合模块(Global-local token fusion)来分割不同的局部信息并与全局信息特征融合,如图6所示,表达式为

$$F_1^i=p[c(\boldsymbol{x}_l^0,\boldsymbol{x}_l^{64i+1},\boldsymbol{x}_l^{64i+2},\cdots,\boldsymbol{x}_l^{64i+64})]\quad i=0,1 \tag{13}$$

$$F_2^i=p[c(\boldsymbol{x}_l^0,\boldsymbol{x}_l^{32i+1},\boldsymbol{x}_l^{32i+2},\cdots,\boldsymbol{x}_l^{32i+32})]\quad i=0,1,2,3 \tag{14}$$

式中:c代表Concat操作;p代表池化;F_1^i和F_2^i分别代表支路-1、支路-2通过池化后得到与行人全局特征向量有相同尺度的融合特征,用于网络的训练。

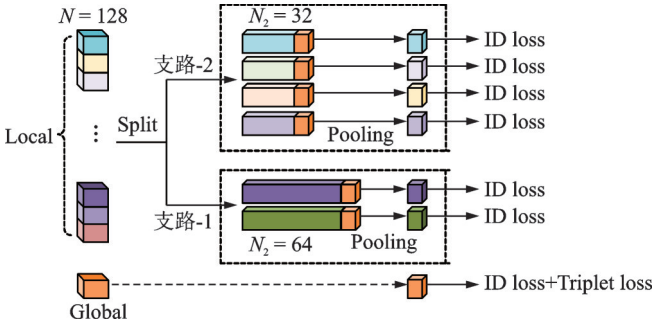

图6　全局与局部token融合模块

Fig.6　Global-local token fusion module

1.5　损失函数

为了提高行人重识别网络模型的识别准确率,采用分类损失函数和度量损失函数作为联合损失函数共同约束网络模型,使网络提取更具有判别力的行人特征。引用文献[24]所提出的损失函数作为本文分类损失函数L_{id},将网络输出行人特征$\boldsymbol{x}_i$用Softmax函数(由$p(\boldsymbol{y}_i|\boldsymbol{x}_i)$编码)计算$\boldsymbol{x}_i$被识别$\boldsymbol{y}_i$的预测概率,并通过交叉熵函数计算行人分类损失,表达式为

$$L_{\mathrm{id}}=-\frac{1}{N_1}\sum_{i=1}^{N}\log_2(p(\boldsymbol{y}_i|\boldsymbol{x}_i)) \tag{15}$$

式中N_1为每批次训练样本数。

度量损失函数使用了难样本三元组损失[25]。难样本三元组损失是改进的三元组损失,即采用一种困难样本的筛选机制,以样本间的距离作为约束使网络模型在训练过程中对困难样本加以关注,从而增加网络模型的识别效果,该损失函数的表达式为

$$L_{\mathrm{tri}}=-\sum_{i=1}^{P}\sum_{a=1}^{K}\left[\alpha+\max_{p=1,\cdots,K}\left\|\boldsymbol{f}_a^{(i)}-\boldsymbol{f}_p^{(i)}\right\|_2-\min_{\substack{m=1,\cdots,K\\ j=1,\cdots,P\\ j\neq 1}}\left\|\boldsymbol{f}_a^{(i)}-\boldsymbol{f}_n^{(j)}\right\|_2\right]_+ \tag{16}$$

式中:P为数据集中行人类别的总数目;K代表每个类别行人图像的数目;α为控制类内和类内距离大小的边缘超参数;$\boldsymbol{f}_a^{(i)}$、$\boldsymbol{f}_p^{(i)}$、$\boldsymbol{f}_n^{(i)}$分别是在样本i中选定的标定样本特征、正样本特征和负样本的特征。

采取联合损失函数表达式为

$$L=\sum_{i=0}^{1}L_{\mathrm{id}}(\boldsymbol{F}_1^i)+\sum_{i=0}^{3}L_{\mathrm{id}}(\boldsymbol{F}_2^i)+L_{\mathrm{id}}(\boldsymbol{x}_l^0)+L_{\mathrm{tri}}(\boldsymbol{x}_l^0) \tag{17}$$

式中:$\boldsymbol{x}_l^0$代表行人分类的全局特征;$\boldsymbol{F}_1^i$、$\boldsymbol{F}_2^i$分别代表图6中支路-1、支路-2输出的融合行人特征。

2 实验与结果分析

2.1 数据集和评价指标

为了验证网络模型和模块的有效性,在具有属性标签行人重识别主流数据集Market-1501和DukeMTMC-reID上进行相关实验,数据集如表1所示。

表1 行人重识别数据集

Table 1 Person re-identification datasets

数据集	发布时间	摄像头/个	训练集		测试集	
			ID	数量	ID	数量
Market-1501	2015年	6	751	**12 936**	750	**19 732**
DukeMTMC-reID	2017年	8	702	**16 522**	702	**17 661**

Market-1501数据集有1 501个行人共32 668张图片,其中训练集有751个行人共12 936张图片,测试集有750个行人共19 732张图片。DukeMTMC-reID数据集包含1 404个行人共36 411张行人图片,其中训练集有702个行人的16 522张图片,测试集含702个行人共有17 661张图片。

采用累积匹配特征(Cumulative matching characteristic,CMC)中2个评价指标分别为Rank-n和平均精度均值(mean Average precision, mAP),用于评估模型性能。Rank-n即在查询集中选定一行人,计算在图库集中与之相似度最为相近且为同一行人的前n个的概率。mAP即计算平均精度AP值并对多个AP值求和加权平均得到,对模型性能具有更全面的评价,表达式为

$$\mathrm{mAP}=\frac{1}{N_2}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\mathrm{precision}(r_{ij}) \tag{18}$$

式中:N_2为查询集中图片数量;m_i为在查询集中筛选出正确样本的数量;precision(r_{ij})表示返回第j个行人正确匹配结果的平均精度。

2.2　实验设置

实验使用的计算平台是基于64位的Windows10专业版操作系统,硬件配置如下:GPU为NVIDIA GeForce GTX2080Ti,使用Pytorch作为框架模型,在训练模型时,输入行人图像的分辨率设置大小为256像素×128像素,训练批次为32,总共迭代次数为120,使用SGD优化器优化模型参数,SGD优化器的动量为0.9和权值衰减为1e−4,学习率初始化为0.008,采用余弦学习率衰减。

2.3　算法对比

将本文提出算法与其他基于CNN和Vit为骨干模型的行人重识别算法在Market-1501和DukeMTMC-reID数据集上进行实验比较,实验结果如表2、3所示。本文算法在Market-1501数据集上Rank-1和mAP分别达到了97.5%和92.3%,在DukeMTMC-reID数据集上Rank-1和mAP分别达到了93.5%和83.1%。在Market-1501数据集中,相较于以CNN为骨干模型的MGN算法,本文算法的Rank-1和mAP分别提高了1.8%和5.4%,表明以Vit为骨干模型的算法在行人重识别领域优于以CNN为骨干模型的主流算法;相较于以Vit为骨干模型的Transreid算法,本文算法的Rank-1和mAP分别提高了2.3%和3.4%,表明了Transreid算法对行人图像块的空间相对位置感知差以及对局部特征信息没有充分合理利用。在DukeMTMC-reID数据集中,本文算法相较于MGN算法Rank-1和mAP分别提高了3.8%和4.7%,相较于Transreid算法Rank-1和mAP分别提高了1.8%和2.8%。实验仿真结果表明所提出的行人重识别模型在不同数据集中均能够提高行人重识别的准确率,证明了本文算法的有效性。

表2　在Market-1501上不同算法结果比较

Table 2　Comparison of different algorithms on Market1-501　%

主干模型	算法	Market-1501			
		mAP	Rank-1	Rank-5	Rank-10
CNN	PCB+RPP[5]	81.6	93.8	97.5	98.5
	OSNet[26]	84.9	94.8	—	—
	DG-Net[27]	86.0	94.8	—	—
	HSP[28]	83.4	93.7	97.6	98.4
	MGN[6]	86.9	95.7	98.3	99.0
Vit	VIT-BoT[18]	86.8	94.7	—	—
	DeiT-BoT[19]	86.6	94.4	—	—
	OH-Former[29]	88.0	94.9	—	—
	AAformer[30]	87.7	95.4	—	—
	Transreid[20]	88.9	95.2	—	—
	本文算法	**92.3**	**97.5**	**98.9**	**99.5**

2.4　消融实验

为了验证IRPE模块、LPA模块和特征融合(G-LTF)对实验的影响,在Transreid-baseline网络添加不同模块,其他保持不变,分别在Market-1501和DukeMTMC-reID数据集进行实验分析,实验结果如表4所示。对于Market-1501数据集,单独融入IRPE模块、LPA模块或G-LTF均对模型效果有所提升;将IRPE模块、LPA模块和G-LTF同时融入到Transreid-baseline网络中,Rank-1和mAP得到了进一步的提升。上述实验表明,

表3　在DukeMTMC-reID上不同算法结果比较

Table 3　Comparison of different algorithms on DukeMTMC-reID　%

主干模型	算法	DukeMTMC-reID			
		mAP	Rank-1	Rank-5	Rank-10
CNN	PCB+RPP[5]	69.2	83.3	90.5	92.5
	HSP[28]	73.3	86.0	93.0	94.5
	VPM[31]	72.6	83.6	91.7	94.2
	OSNet[26]	73.5	88.6	—	—
	DG-Net[27]	74.8	86.6	—	—
	MGN[6]	78.4	88.7	—	—
Vit	VIT-BoT[18]	79.3	88.8	—	—
	DeiT-BoT[19]	78.9	89.3	—	—
	Transreid[20]	82.0	90.7	—	—
	AAformer[30]	80.0	90.1	—	—
	OH-Former[29]	82.8	91.0	—	—
	本文算法	**83.1**	**93.5**	**95.7**	**97.8**

表4 不同模块在Market-1501和DukeMTMC-reID数据集上消融实验对比

Table 4 Ablation comparison of different modules on Market-1501 and DukeMTMC-reID datasets %

方法	Market-1501				DukeMTMC-reID			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	86.6	94.7	—	—	79.3	88.8	—	—
Baseline+LPA	87.7	95.3	98.1	99.2	80.1	89.4	95.1	97.3
Baseline+IRPE	87.4	95.2	97.8	99.0	79.8	89.1	95.3	97.1
Baseline+GLTF	90.5	96.9	98.5	**99.6**	81.8	91.8	**95.7**	97.3
Baseline+IRPE+LPA+GLTF	**92.3**	**97.5**	**98.9**	99.4	**83.1**	**93.5**	**95.7**	**97.8**

IRPE模块能弥补Transreid-baseline网络对行人图像块在空间位置相对关系的缺陷，LPA模块让网络更加关注具备行人信息的图像块，G-LTF融合局部与全局行人信息，使网络可以提取到更加具有鲁棒性的行人特征，从而提高行人重识别的准确度和精度。

在LPA注意力中，为了确定不同权重的池化对行人图像块特征提取的影响，引进2个超参数λ(Max pooling)和μ(Avg pooling)进行实验分析。表5给出了不同的λ、μ取值时在Market-1501数据集上的实验结果，表明当λ=0.4、μ=0.6时能使网络模型取得较好的结果。

在IRPE编码中，测试了不同的相对编码方式对整体模型性能的影响。IRPE-$\boldsymbol{K}$、RPE-$\boldsymbol{QK}$和IRPE-$\boldsymbol{QKV}$分别代表作用于向量$\boldsymbol{K}$、向量$\boldsymbol{Q}$和$\boldsymbol{K}$、向量$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的编码方式。将这3种编码方式应用于本文算法在Market-1501数据集上进行测试，结果如表6所示。可以看出，当分别采用3种不同的编码方式，模型在Market-1501数据集的性能均有提升；当采用IRPE-$\boldsymbol{QKV}$时，行人识别效果更好。这是因为相对位置编码通过$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$向量在自注意力机制计算时，IRPE能辨别更多行人图像块间位置关系，从而验证了模型的有效性。

表5 λ和μ在Market-1501数据集上的实验比较

Table 5 Experimental comparison of λ and μ on Market-1501 dataset %

λ	μ	mAP	Rank-1
0.2	0.7	92.0	97.1
0.3	0.9	91.7	97.0
0.4	0.6	**92.3**	**97.5**
0.5	0.5	91.6	97.3
0.6	0.4	92.1	96.9
0.8	0.3	91.9	97.1
1.0	1.0	91.6	97.0

表6 不同编码方式的实验比较

Table 6 Experimental comparison of different coding methods %

方法	mAP	Rank-1
Backbone+IRPE-$\boldsymbol{K}$	92.1	97.1
Backbone+IRPE-$\boldsymbol{QK}$	92.0	**97.5**
Backbone+IRPE-$\boldsymbol{QKV}$	**92.3**	97.3

2.5 可视化效果

为了进一步验证算法的有效性，本文采用行人图像排序可视化的方法，图7给出了数据可视化的结果。图7中第1列query为待查询行人图像，图像1～10为按照余弦相似度排序前10名的行人图像结果。图7(a)为Transreid-baseline查询结果，可以发现当行人轮廓相似、行人图像分辨率低时，易出现检测错误的现象。图7(b,c)为加入LPA与IRPE模块的查询结果，可以看出查询准确性有所提高。图7(d)为加入LPA、IRPE和GLTF模块的查询结果，可以看出，即使在图像低分辨率背景模糊且存在不属于行人特征的干扰因素(如背包)的情况下，该网络仍然可以从图库集中找出正确的行人图像。这是由于引入的改进型Transformer和特征融合算法能进一步提高检索排序结果，提高准确性，便于解决现实复杂场景下行人识别的问题。

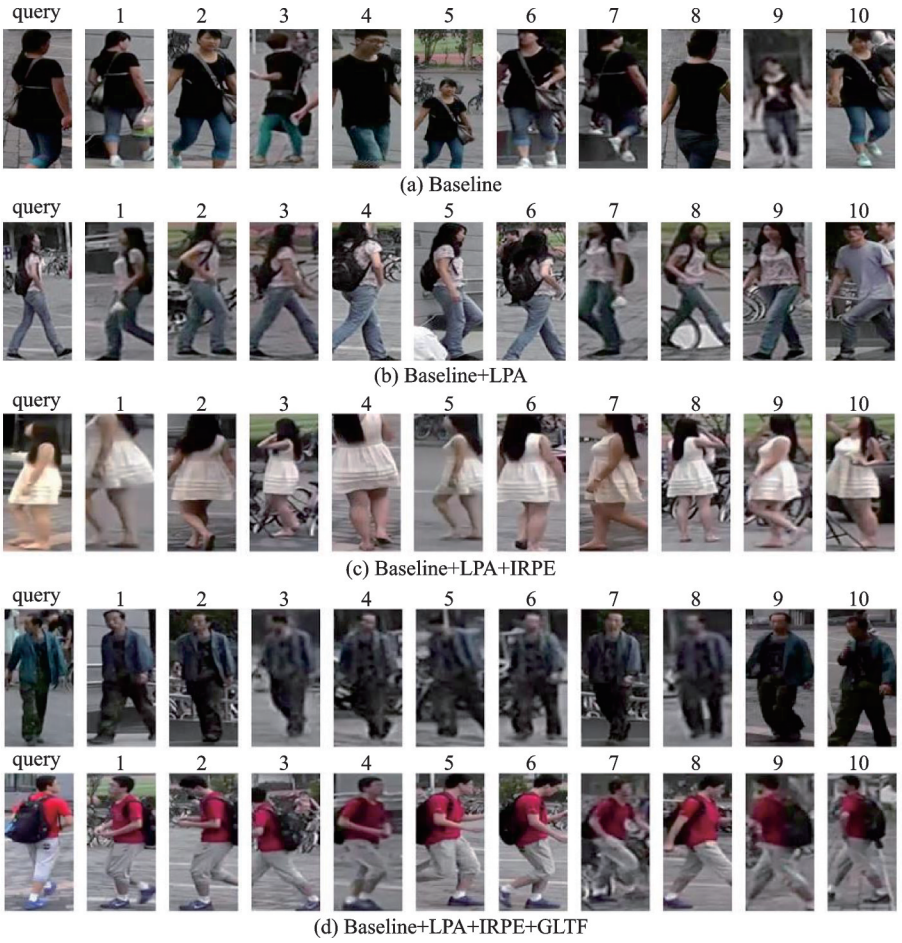

图7　在Market-1501数据集上的查询可视化结果

Fig.7　Query visualization results on Market-1501 dataset

3　结束语

本文提出一种改进型Transformer编码器和特征融合的行人重识别算法,将原网络模型中使用的绝对位置编码改进为相对位置编码。在对行人图像块相对位置编码的过程中,根据不同的距离分布施加不同程度的注意力来提高模型对行人特征的表达;融合局部patch注意力机制模块使网络对划分的不同patch具有不同的权重分配,从而选择性的挖掘更多行人图像块特征抑制无用背景信息。此外,将行人局部特征与全局特征融合使网络模型在训练中关注更具有判别性质的行人特征提高模型鲁棒性。实验结果证明本文算法比目前主流算法获得更高的准确性。在未来的研究中,将尝试对模型进行轻量化,保证准确度的同时提高模型在设备方面部署的速度。

参考文献:

[1]　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]. Journal of Automation, 2019, 45 (11): 2032-2049.

[2]　李梦静,吉根林, 赵斌. 基于步行周期聚类的视频行人重识别关键帧提取算法[J]. 南京航空航天大学学报, 2021, 53(5):

780-788.

LI Mengjing, JI Genlin, ZHAO Bin. Key frame extraction algorithm for video-based person re-identification based on walking cycle clustering[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(5): 780-788.

[3] LI Rui, ZHANG Baopeng, ZHU Teng, et al. A divide-and-unite deep network for person re-identification[J]. Applied Intelligence, 2021, 51(7): 1479-1491.

[4] ZHENG Liang, YI Yang, HAUPTMANN A G. Person re-identification: Past, present and future[EB/OL]. (2016-10-10)[2022-03-15]. https://arxiv.org/abs/1610.02984v1.

[5] WANG Guanshuo, YUAN Yufeng, XIONG Chen, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proceedings of the 26th ACM International Conference on Multimedia. Amsterdam, Netherlands: ACM, 2018: 274-282.

[6] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016: 2818-2826.

[7] YAN Yichao, QIN Jie, CHEN Jiaxin, et al. Learning multi-granular hypergraphs for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 2899-2908.

[8] LI Jiabin, LI Xuewei, LIU Hongzhe, et al. Person re-identification based on local feature association and global attention mechanism[EB/OL]. (2021-01-29)[2022-03-15]. https://doi.org/10.19678/j.issn.1000-3428.0059940.

[9] CHEN Tianlong, DING Shaojin, XIE Jingyi, et al. ABD-Net: Attentive but diverse person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. San Diego, CA, USA: IEEE, 2019: 8351-8361.

[10] LIU Zhigang, DU Juan, WANG Mei, et al. ADCM: Attention dropout convolutional module[J]. Neurocomputing, 2020, 394: 95-104.

[11] HOU Ruibing, MA Bingpeng, CHANG Hong, et al. Interaction-and-aggregation network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, California, USA: IEEE, 2019: 9317-9326.

[12] YAN Yichao, QIN Jie, NI Bingbing, et al. Learning multi-attention context graph for group-based re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. DOI: 10.1109/TPAMI.2020.3032542.

[13] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 3-19.

[14] 王立辉, 杨贤昭, 刘惠康, 等. 基于GhostNet与注意力机制的行人检测跟踪算法[J]. 数据采集与处理, 2022, 37(1): 108-121.
WANG Lihui, YANG Xianzhao, LIU Huikang, et al. Pedestrian detection and tracking algorithm based on GhostNet and attention mechanism[J]. Journal of Data Acquisition and Processing, 2022, 37(1): 108-121.

[15] ZHANG Zhizheng, LAN Cuiling, ZENG Wenjun, et al. Relation-aware global attention for person re-identification[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 3186-3195.

[16] LIU Zheng, QIN Jie, LI Annan, et al. Adversarial binary coding for efficient person re-identification[C]//Proceedings of 2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE, 2019: 700-705.

[17] CHEN Jiaxin, QIN Jie, YAN Yichao, et al. Deep local binary coding for person re-identification by delving into the details [C]//Proceedings of the 28th ACM International Conference on Multimedia. [S.l.]: ACM, 2020: 3034-3043.

[18] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL].(2020-10-22)[2022-03-15]. https://arxiv.org/abs/2010.11929.

[19] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 10347-10357.

[20] HE Shuting, LUO Hao, WANG Pingchao, et al. Transreid: Transformer-based object re-identification[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021.

[21] 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.

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

[23] YANG FU, WEI Yunchao, ZHOU Yuqing, et al. Horizontal Pyramid matching for person re-identification[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. New York: AAAI, 2019: 8295-8302.

[24] YE Mang, SHEN Jianbing, LIN Gaojie, et al. Deep learning for person re-identification: A survey and outlook[EB/OL]. (2021-01-06)[2022-03-15]. https://arxiv.org/abs/2001.04193v2.

[25] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[EB/OL]. (2017-11-21)[2022-03-15]. http://arxiv.org/abs/1703.07737v2.

[26] ZHOU Kaiyang, YANG Yongxin, CAVALLARO A, et al. Omni-scale feature learning for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 3702-3712.

[27] ZHENG Zhedong, YANG Xiaodong, YU Zhiding, et al. Joint discriminative and generative learning for person re-identification [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, California, USA: IEEE, 2019: 2138-2147.

[28] KALAYEH M M, BASARAN E, GÖKMEN M, et al. Human semantic parsing for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 1062-1071.

[29] CHEN Xiayu, XU Chunlin, CAO Qiong, et al. OH-Former: Omni-relational high-order transformer for person re-identification [EB/OL]. (2021-05-20)[2022-03-15]. http://arxiv.org/abs/2109.11159.

[30] ZHU Kuan, GUO Haiyun, ZHANG Shiling, et al. AAformer: Auto-aligned transformer for person re-identification[EB/OL]. (2021-08-13)[2022-03-15]. http://arxiv.org/abs/2104.00921v2.

[31] SUN Yifan, XU Qin, LI Yali, et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, California, USA: IEEE, 2019: 393-402.

作者简介:

赵倩(1969-),女,博士,副教授,研究方向:视频图像处理、深度学习,E-mail:zhaoqian@shiep.edu.cn。

薛超晨(1996-),**通信作者**,男,硕士研究生,研究方向:图像处理、行人重识别,E-mail:xbcxcc163@163.com。

赵琰(1979-),女,博士,副教授,研究方向:数字图像处理、信息安全。

(编辑:张黄群)