计算机科学
COMPUTER SCIENCE

基于相对位置编码转换器模块的深度步态识别网络

任禹衡, 赵云峰, 吴闯

引用本文

任禹衡, 赵云峰, 吴闯. 基于相对位置编码转换器模块的深度步态识别网络[J]. 计算机科学, 2024, 51(11A): 240400064-6.

REN Yuheng, ZHAO Yunfeng, WU Chuang. Deep Gait Recognition Network Based on Relative Position Encoding Transformer [J]. Computer Science, 2024, 51(11A): 240400064-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400064

基于相对位置编码转换器模块的深度步态识别网络

任禹衡[1]　赵云峰[2]　吴　闯[3]
1 银河水滴科技(北京)有限公司　北京 100083
2 国家管道网集团北方管道有限责任公司　北京 100084
3 武警湖南总队参谋部　长沙 410000

摘　要　步态识别是一种快速发展的远距离生物特征识别技术,在远距离、跨视角和跨着装等多种场景中具有广泛应用和优势。传统的生物特征识别技术,如指纹识别、面部识别等,往往需要近距离或在特定条件下才能有效进行,而步态识别技术则突破了这些限制,使得在更为广泛的环境下进行个体识别成为可能。以往的研究大多采用轻量级的神经网络提取步态特征,并在目前流行的跨视角和跨着装数据集上(如CASIA-B)取得了巨大的进步。然而,实验结果表明,在CASIA-B数据集上简单叠加神经网络的层数将导致识别准确率大幅度下降。基于相对位置编码转换器模块提出了一个深度步态识别网络,旨在避免陷入"局部特征关联"的陷阱,同时使网络能够持续不断地学习步态序列的时序特征。与目前主流的方法相比,所提方法在室内场景(CASIA-B,OUMVLP)和室外场景(Gait3D)步态数据集上都达到了更优的识别准确率,特别在换装任务(CL)上超出基准方法1.9%,实现了85.5%识别准确率。

关键词:步态识别;自注意力机制;相对位置建模;模式识别;深层网络
中图分类号　TP391.41

Deep Gait Recognition Network Based on Relative Position Encoding Transformer

REN Yuheng[1],ZHAO Yunfeng[2] and WU Chuang[3]
1 Watrix.AI,Beijing 100083,China
2 North Pipeline Co.,Ltd.,National Pipe Network Group,Beijing 100084,China
3 The General Staff of Hunan Provinical Corps of the Chinese People's Armed Police Force,Changsha 410000,China

Abstract　Gait recognition is a rapidly evolving long-range biometric identification technique that has wide applications and advantages in various scenarios,including long distances,non-intrusive setups,and cross-view angles. Traditional biometrics identification technique,such as fingerprint recognition and facial recognition,often require close proximity or specific conditions to be effective,while gait recognition technology breaks through these limitations,making it possible to identify individuals in a wider range of environments. Previous research predominantly employed lightweight neural networks for gait feature extraction and achieved significant progress on popular datasets like CASIA-B,which feature cross-view angles and varying attires. However,experimental results indicate a substantial decline in recognition accuracy when simply stacking neural network layers on the CASIA-B dataset. A deep gait recognition network has been proposed,incorporating the relative position encoding transformer module. This module aims to avoid the pitfall of "local feature association""and enables continuous learning of temporal features within gait sequences. Compared to current mainstream approaches,the proposed method has garnered enhanced identification precision across indoor environments,as exemplified by the CASIA-B and OUMVLP datasets,alongside outdoor settings typified by the Gait3D dataset. Especially in the task of clothes changing,wherein our method surpasses benchmark approaches by of 1.9%,achieving a recognition rate of 85.5%.

Keywords　Gait recognition mechanism,Self-attention mechanism,Relative position modeling,Pattern recognition,Deep network

1　引言

步态作为每个人独特的行走模式,具备高度的生物特征辨识性。相对于其他生物特征如面部、指纹和虹膜,步态具有一些独特的优势:难以伪装、可进行远距离非侵入性捕捉,且不需要被识别者的合作。因此,步态识别技术在公共安全中具有广泛的潜力,如嫌疑人追踪、犯罪预防和身份验证等领域[1-2]。随着深度学习模型的验证和优化,各种视觉技术的发展也得到了极大的推动,其中包括利用神经网络进行步态识别。当前主流的步态识别方法都采用浅层的神经网络来学习细粒度的步态特征,并取得了较高的识别准确率,这为步态识别技术的应用拓展和改进提供了巨大的帮助。然而,现实中的步态识别任务场景复杂多变,实际数据不仅具有各种视角和着装的变化,还受到光照、遮挡和畸变等因素的干扰。显然,浅层的步态识别网络无法学习到更复杂和更抽象的特征,也难以应对更具挑战性的实际步态识别任务。

对于步态识别任务而言,许多研究旨在获取细粒度的时空特征。其中,将步态特征横向切分为预定义数目的条状

通信作者:任禹衡(renyh16@mails.jlu.edu.cn)

特征，并分别使用损失函数进行约束已经成为了一种共识。例如，GaitSet[3]首次提出了水平金字塔映射模块（Horizontal Pyramid Mapping），将步态特征划分为 $S=4$ 个尺度，并在每个尺度下将特征图划分为2^{s-1}个横向切分的条状特征。在训练过程中对这些条状特征（共计 62 个）分别施加三元组损失[4]，而在测试过程中将这些条状特征合并为一个特征进行推理。GaitPart[5]提出了水平池化模块（Horizontal Pooling），直接将特征图划分为 $n=16$ 个横向切分的条状特征分别进行约束。GaitGL[6]提出将特征图划分为 $n=16$ 个横向切分的条状特征集，利用广义平均池化（Generalized-Mean）进一步获得特征图。3DLocal[7]提出 Localization module，动态定位身体的不同区域（如头、肩、手臂和腿部），并利用采样模块和特征提取模块选择不固定数目的帧以学习不同位置的特征。这些方法将步态看作是多个部位的运动组合，通过挖掘每个部位的时序特征，使其具有独立区分行人身份的能力。相比提取全局特征的步态识别方法，基于不同部位的步态识别方法更注重细粒度的特征，能够在一定程度上避免神经网络陷入“局部特征关联”的陷阱，并且具有更好的识别准确率。

然而，实验结果表明在 CASIA-B 数据集[8]上简单堆叠浅层的步态识别网络会导致识别准确率迅速降低。本文的研究发现，随着神经网络层数的增加，深层网络的感受野将进一步扩大，出现了“局部特征关联”的现象，进而降低了网络的识别准确率。为解决这一问题，本文提出了一种基于自注意力机制的深层步态识别网络，旨在学习更加复杂和抽象的特征。其中，相对位置编码转换器模块的引入有效地避免了网络陷入“局部特征关联”的陷阱，使网络能够持续学习步态序列的时序特征，从而提高步态识别模型的鲁棒性和准确性。

2 相关工作

2.1 步态识别

步态识别根据模型分为两类，即基于模型的方法和基于序列的方法。其中，基于模型的步态识别方法首先利用姿态估计技术建立行人模型，然后利用卷积神经网络或者时间循环神经网络提取行人显著性的特征。然而，这类方法受限于姿态估计方法或人体参数化模型[9]（Skinned Multi-Person Linear Model）的准确度。相较而言，基于序列的步态识别方法由于其结构简单且具有较高的识别精度而获得了更多研究人员的关注。基于序列的识别方法根据输入数据可以分为两类：基于模板的步态识别方法和基于集合的步态识别方法。其中，基于模板的步态识别方法的输入数据为步态能量图[10]（Gait Energy Image），通过加权平均的方法将三维的运动信息转化为二维的外形信息。尽管采用加权平均的方法能够获得更加直观的步态能量图，但是对数据直接进行加权平均将不可避免地损失一些时序信息。例如，CNN-LB[11]利用由 5 层卷积堆叠而成的孪生网络学习正负样本，其中正样本对由相同身份的步态能量图组成，负样本对由不同身份的步态能量图组成。基于集合的步态识别方法将步态序列作为输入数据，并利用神经网络提取步态的形态和时序信息。例如，GaitSet[3]假设步态序列中每帧的剪影图已经包含了位置信息，并利用 6 层卷积神经网络直接提取步态序列的时空特征。GLN[12]提出特征金字塔结构的 6 层卷积提取随机采样步态序列的时序信息。GaitPart[5]通过控制 6 层卷积的感受野，提取连续采样剪影图的局部信息。GaitGL[6]提出利用 6 层三维卷积神经网络提取连续采样步态序列的时空特征。上述方法均采用浅层的神经网络提取步态的时空特征，无法学习更复杂和更抽象的特征，也难以适应更具有挑战性的实际步态识别任务。

2.2 自注意力模型

在处理长程依赖性问题时，采用自注意力机制或者全局上下文模型是较为常见的方法。前者通过自适应地学习注意力权重，能够建模任意距离的单词之间的依赖关系，从而打破了传统固定距离依赖建模方法的局限性。后者则通过捕捉整个序列的全局上下文信息，更好地建模了长程依赖性问题。相比传统的递归神经网络和卷积神经网络等模型，这两种方法在处理序列数据时展现出了更为出色的性能。例如，Non-local 网络[13]通过建立全局性的关系来捕捉输入数据中的长程依赖关系，并将关系建模与特征提取相结合，使得模型能够同时考虑空间和通道维度上的特征。SENet 网络[14]通过 Squeeze-and-Excitation 模块来增强卷积神经网络中的特征重要性。然而，上述方法均应用在单帧图片中，不适用目前的步态识别任务。Transformer 网络[15]利用自注意力机制来建模输入序列中不同位置之间的关系，而且可以适应任意长度的输入序列，因此被广泛应用于自然语言处理和计算机视觉等领域。

2.3 位置建模

Transformer 模型[15]的位置建模是一个非常重要的研究方向，其旨在通过有效的位置编码方法来建立输入序列中不同位置之间的关系。目前的位置建模研究主要分为 3 类：直接编码位置信息建模、可学习位置信息建模和考虑上下文的位置信息建模。直接位置信息建模直接使用位置信息来编码输入序列中的位置信息，包括绝对位置编码和相对位置编码两种方法。绝对位置编码使用一个固定的位置向量来编码每个位置，例如使用正弦和余弦函数来生成位置编码向量；而相对位置编码则根据不同位置之间的相对距离来编码位置信息。可学习位置建模通过学习一个位置嵌入向量来自适应编码每个位置信息。考虑上下文的位置信息建模通过引入上下文信息来建模不同位置之间的关系，包括基于自注意力机制的上下文编码和基于卷积神经网络的上下文编码两种方法。基于自注意力机制的上下文编码通过 Transformer 模型中的自注意力机制来建立位置之间的关系；而基于卷积神经网络的上下文编码则使用更深层的卷积神经网络，扩大感受野来学习上下文信息。

3 研究方法

本章首先介绍基于深度自注意力机制的步态识别网络架构（RTBNet），然后介绍提出的基于相对位置编码的自注意力模型，最后进一步讨论步态模型的联合损失函数。

3.1 整体架构

RTBNet 的整体结构如图 1 所示，其网络参数如表 1 所列，其中 C3D 表示 3D 卷积模块，HP 表示横向切分模块，RTB 表示基于相对位置编码的转换器模块。

步态识别的核心目标是从跨视角和跨着装的场景中识别出具有相同身份的受试者。假设 $x\in R^{T\times H\times W}$ 代表剪影序列，其中 T 表示剪影序列的帧数；H 和 W 分别表示剪影的高和

宽。步态识别的特征提取过程可以表示为：

$$f=G(L(x)) \tag{1}$$

其中，f 表示步态识别网络提取每个部分的特征，G 表示全局时序特征聚合模块，而 L 表示局部时空特征提取模块。

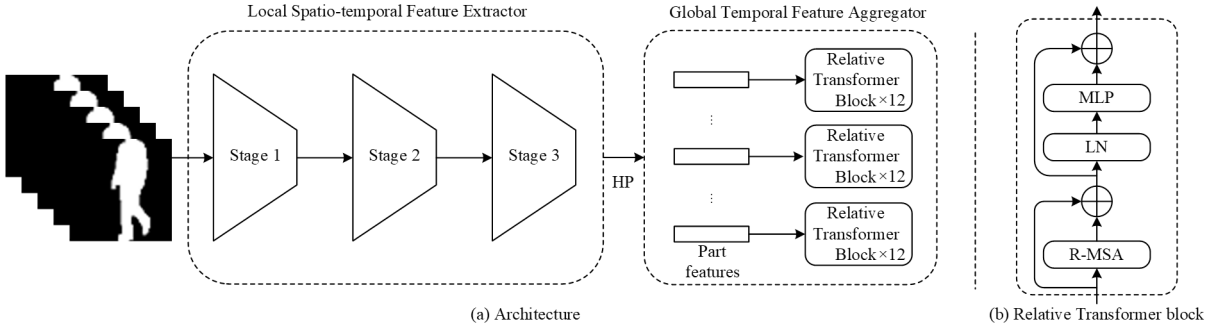

图 1　RTBNet 整体架构

Fig. 1　Overall architecture of RTBNet

表 1　基于深度自注意力机制的步态识别网络参数

Table 1　Parameters of gait recognition network based on deep self-attention mechanism

名称	输出维度	网络结构
Stage1	$T\times H\times W$	$\begin{bmatrix}C3D,[3,3,3],[1,1,1],64\\C3D,[3,3,3],[1,1,1],64\end{bmatrix}$
Stage2	$T\times H\times W$	$\begin{bmatrix}C3D,[3,3,3],[1,1,1],128\\C3D,[3,3,3],[1,1,1],128\end{bmatrix}$
Stage3	$T\times H\times W$	$\begin{bmatrix}C3D,[3,3,3],[1,1,1],256\\C3D,[3,3,3],[1,1,1],256\end{bmatrix}$
HP	$T\times P$	{MaxPool+AvgPool}
RTB	$T\times P$	{Relative Tansformer}×12

步态的数据是由二值图组成的序列，相同身份的受试者在跨视角和跨着装场景中具有显著的视觉差异，呈现出类内差远大于类间差的特点。因此，步态识别的关键在于提取每个受试者最鲁棒的运动特征。目前，现有的网络均采用浅层的卷积网络和横向切分(Horizontal Pooling，HP)模块提取步态特征。这样做的目的在于：一方面浅层的网络能够保证网络专注于提取当前位置的时空信息；另一方面将全局特征横向切分能够保证每个细粒度的特征都能够起作用。同时，增加网络层数会导致在 CASIA-B 数据集上识别准确率迅速下降，这表明：一方面网络参数增加会导致对 CASIA-B 数据集过拟合；另一方面步态神经网络由于感受野变大，陷入“局部特征关联”的陷阱中，即网络关注同一个部位的特征，停止挖掘细粒度的时空信息。为此，本文提出了一种基于相对位置编码的转换器模块的深度步态识别网络架构。首先，在局部特征提取模块中利用 3 个阶段的三维卷积网络提取步态数据的局部时空特征；然后，利用横向切分模块将全局特征切分为多个细粒度部位；最后，利用基于相对位置编码的转换器模块进一步建模步态的全局时间特征。

3.2　基于相对位置编码的自注意力模型

3.2.1　自注意力模型

步态是一种需要全身各个部位协调配合的复杂运动过程。研究表明，在跨越不同的视角和穿着不同的装备的情况下，每个人的运动模式都表现出最为稳定的特征。因此，本文提出利用自注意力模型进一步学习每个细粒度部位的时间维度上的全局运动模式。假设 $H\in R^{T\times C}$ 表示经过横向切分的特征。其中，步态时间维度的特征聚合过程可以表示为：

$$r=M(R(H)+H)+R(H)+H \tag{2}$$

其中，r 表示经过基于相对位置编码的自注意力机制模型提取的特征，M 表示前向传播模块，包含层归一化(LayerNorm)和多层感知器(MLP)；R 表示基于相对位置编码的自注意力模型。自注意力模型对步态时序特征的聚合过程可以表示为：

$$Attention(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{3}$$

其中，Q,K,V 分别表示步态特征 H 经过全连接层映射到新的特征空间后的查询(Queries)、键(Keys)和值(Values)特征。这里的注意力计算通过归一化项 $\frac{1}{\sqrt{d_k}}$ 对注意力值进行缩放，以增强稳定性。然后，通过对注意力图(Attention Map)进行归一化指数函数(softmax)操作，并与值特征相乘，最终获得步态全局时序特征。这种自注意力机制的引入，使得基于深度自注意力机制的步态识别网络(RTBNet)能够更加准确地捕捉到每个部位之间的关联信息，从而提高步态识别模型的表征能力和准确性。

3.2.2　相对位置编码

在步态的训练过程中，通常随机采样固定数目的帧进行批量训练，然而在测试过程中每个序列的帧数是变化的，也就是说，训练和测试过程中每帧数据的位置信息是未知的。为此，本文参考 RoFormer[16] 对随机采样的步态序列进行相对位置建模，并对 Q_m，K_n 分别乘以 $e^{im\theta}$，$e^{in\theta}$：

$$Q(H_m,m)=(W_qH_m)e^{im\theta} \tag{4}$$

$$K(H_n,n)=(W_kH_n)e^{in\theta} \tag{5}$$

其中，$Q(H_m,m)$ 是添加绝对位置信息 m 的 Querys，$K(H_n,n)$ 是添加绝对位置信息 n 的 Keys。对于位置为 m 的向量 $\boldsymbol{Q}$ 和位置为 n 的向量 $\boldsymbol{K}$，计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 序列的注意力关系，那么注意力关系的特征自动包含了相对位置信息。其具体形式如下：

$$(\boldsymbol{Q}_m)(\boldsymbol{K}_n)^{\mathrm{T}}=\boldsymbol{Q}\boldsymbol{R}_m\boldsymbol{R}_n^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}=\boldsymbol{Q}\boldsymbol{R}_{m-n}\boldsymbol{K}^{\mathrm{T}} \tag{6}$$

其中，$\boldsymbol{R}_m$，$\boldsymbol{R}_n$ 是一个正交矩阵。对于绝对位置为 m 的向量 $\boldsymbol{Q}$，其具体形式如下：

$$\begin{pmatrix}q_0\cos m\theta_0\\q_1\cos m\theta_0\\q_2\cos m\theta_1\\q_3\cos m\theta_1\\\vdots\\q_{d-2}\cos m\theta_{d/2-1}\\q_{d-1}\cos m\theta_{d/2-1}\end{pmatrix}+\begin{pmatrix}-q_1\sin m\theta_0\\q_0\sin m\theta_0\\-q_3\sin m\theta_1\\q_2\sin m\theta_1\\\vdots\\-q_{d-1}\sin m\theta_{d/2-1}\\q_{d-2}\sin m\theta_{d/2-1}\end{pmatrix} \tag{7}$$

其中，$q_{0,1,\cdots,d-2,d-1}$表示 d 维特征所对应的值。

3.3 联合损失函数

本文在训练网络时使用一种由 Batch All 三元组损失(Triplet Loss)和交叉熵损失(Cross-entropy Loss)组成的联合损失函数。三元组损失函数的目标是优化类内和类间的距离。具体而言，它在特征空间增加类之间的距离，减少类内的距离。组合函数可以表示为：

$$L_{\text{all}}=L_{\text{cse}}+L_{\text{tri}} \tag{8}$$

其中，L_{cse}表示交叉熵损失，可以表示为：

$$L_{\text{cse}}=-\frac{1}{N}\sum_{i} y_i \log(\hat{y}_i) \tag{9}$$

其中，$\hat{y}_i$表示样本中第 i 个特征预测的类别；y_i表示其真实标签；L_{tri}表示三元组损失，它在特征空间增加类之间的距离，减少类内的距离，可以表示为：

$$L_{\text{tri}}=[D(F_i,F_k)-D(F_i,F_j)+m]_+ \tag{10}$$

其中，F_i，F_j，F_k分别表示样本 i，j，k 对应的特征，i 和 j 为同一个标签的样本，k 和 i，j 的标签不同；m 为训练过程中正负样本的间隔，在本文中 $m=0.2$。本文提出方法的网络输出为不同身体部分的特征，最后经过式(8)得到每个身体部分的损失值。

4 实验结果

本章首先介绍了 CASIA-B 数据集的基本内容，然后对基于相对位置编码的转换器模块的深度步态识别网络的训练过程进行详细的介绍，最后在 CASIA-B 数据集上与其他最先进的方法进行详细的对比。

4.1 数据集简介

CASIA-B 数据集[8]是一种广泛使用的步态数据集，由 124 名受试者组成，包含 11 个不同视角(0°～180°)，每个视角包含 10 个序列，总共涵盖 110 个序列。这些序列是在不同的行走条件下获取的，其中前 6 个序列是在正常行走条件下获得(NM)，另外 2 个序列是在背包条件下获得(BG)，最后 2 个序列是在穿着外套或夹克条件下获得(CL)。因此，每个受试者包含 110 个序列。为了保证公正性，本文采用前 74 名受试者作为训练组，剩余 50 名受试者用于测试。在测试过程中，将 NM 条件下的前 4 个序列(“NM 01－04”)作为候选集，其余 6 个序列分为 3 个子集，分别为 NM 子集(“NM 05－06”)、BG 子集(“BG 01－02”)和 CL 子集(“CL 01－02”)，用于评估步态识别性能。

OU-MVLP 数据集[17]是目前受控环境下最大的步态数据集之一，包含 10 307 名受试者。每名受试者有 2 组序列(NM-00 和 NM-01)，每组序列从 14 个角度捕获，视角均匀分布于两个区间[0°～90°]以及[180°～270°]。为了保证算法对比结果的公正性，我们遵循原论文提出的训练和测试协议[3]，采用前 5 153 名受试者作为训练集，其余的 5 154 名受试者作为测试集。测试集中将第 1 组序列(NM-00)作为 probe，第 2 组序列(NM-01)作为 gallery。

Gait3D 数据集[18]是一个近期提出的在实际场景中采集的步态数据集，共计包含了 4 000 名受试者和 25 309 个序列。这些序列是在现实场景中受试者无配合的条件下获取的，一共包含了 37 种视角，每个序列的长度也不固定。为了保证算法对比结果的公正性，我们采用官方的划分方法，将前 3 000 名受试者作为训练集，剩余的 1 000 名受试者作为测试集。在测试集中选择 10 00 个序列作为 probe，剩余的 5 369 个序列作为 gallery。

4.2 训练细节

在训练过程中，本文参考 GaitSet 的方法提取输入序列的中心，然后重新归一化到统一的尺寸大小(64×64)。在采样过程中，设置输入批量大小的策略为(p，k)，其中 p 表示一个 batch 中采样的人数，k 表示每个人的采样数量。值得注意的是，本文训练过程中采样的数目为统一值，在测试过程中对整个序列提取一个完整的步态特征。

4.3 实验结果

表 2 列出了在 CASIA-B 数据集上本文方法与目前流行的方法进行了充分的对比，分别在各个着装和视角的条件下测试了识别准确率。1)相比基于模版的方法(CNN-LB[11])，本文提出的方法能够有效的提取步态的时空信息并在各个视角和着装的情况下识别准确率均大幅领先。从输入数据方面进行对比，CNN-LB 输入的数据是 GEI，直接对每个序列进行加权平均损失了大量的时空信息。而本文采用的输入数据是随机采样的连续序列，通过 3 维卷积进行局部时空特征提取和 RTB 全局时间维度特征建模，能够获得稳定的运动模式。从网络架构方面进行对比，CNN-LB 采用孪生网络提取输入样本对的特征，这种方式需要对所有数据进行组对。而本文采取度量学习的方法，对每个序列提取完整的特征，在测试过程中直接在特征空间计算相似度。从实验结果方面进行对比，CNN-LB 方法在各个角度和着装的状态下都低于本文采用的方法，具体而言，在 NM 状态下识别准确率领先 8%，在 BG 状态下识别准确率领先 22.3%，在 CL 状态下识别准确率领先 31.5%。2)相比基于 C2D 的方法(GaitSet[3]，GaitPart[5]，GLN[12])，本文提出的方法能够有效地获得时空特征，并在 CL 状态下明显优于其他方法。从输入数据方面对比，基于 2D 卷积的方法和深度自注意力机制的步态识别网络架构(RTBNet)均采用随机采样的方法输入序列，其中 GaitSet 和 GLN 采用随机采样不连续帧作为输入，而 GaitPart 和 RTBNet 均采用随机采样连续帧作为输入。相较而言，采样连续帧能够进一步获得步态的时序特征。从网络架构方面对比，基于 2D 卷积的方法均采用数层 2 维卷积作为 backbone，然后对全局特征进行横向切分，并对每个特征单独约束。而 RTBNet 采用 3 维卷积作为 backbone 提取特征，并利用 12 层的 RTB 模块建模时间维度的特征。从实验结果方面进行对比，RTBNet 的方法在各个角度和着装状态下都优于其他方法。3)相比基于 C3D 的方法(MT3D[19]，GaitGL[6])，本文提出的方法能够专注于时间维度的全局特征提取。从输入数据方面进行对比，基于 3D 卷积的方法与 RTBNet 均采用采样连续帧作为输入。相较而言，本文选择不固定数目的连续采样方式。从网络架构方面进行对比，基于 3D 卷积的方法均采用数层 3 维卷积作为 backbone，然后对全局特征进行横向切分并单独约束。而 RTBNet 在 backbone 中对时间维度的特征不进行降维，经过横向切分模块后，利用 RTB 模块为输入序列增加相对位置信息，并使用深度网络建模步态运动过程中的时序信息。

表 2　在 CASIA-B 数据集上的识别准确率(不包括相同视角的情况)

Table 2　Accuracy on CASIA-B dataset(excluding the same viewpoint)

(%)

	Methods	Probe View										Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	CNN-LB[11]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GaitSet[3]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart[5]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	MT3D[19]	95.7	98.2	99.0	97.5	95.1	93.9	96.1	98.6	99.2	98.2	92.0	96.7
	GaitGL[6]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GaitBase[20]	95.3	99.5	100.0	99.1	97.0	95.6	98.0	99.4	99.9	99.2	93.8	97.6
	RTBNet(ours)	96.5	98.9	99.3	98.6	96.8	96.1	97.8	99.2	99.5	99.2	95.8	97.9
BG	CNN-LB[11]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitSet[3]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart[5]	89.1	94.8	96.7	95.1	88.3	94.9	89.0	93.5	96.1	93.8	85.8	91.5
	MT3D[19]	91.0	95.4	97.5	94.2	92.3	86.9	91.2	95.6	97.3	96.4	86.6	93.0
	GaitGL[6]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GaitBase[20]	92.3	95.5	96.3	96.0	91.7	90.5	92.3	96.1	97.4	95.8	88.6	94.0
	RTBNet(ours)	93.4	97.2	98.0	97.0	92.2	89.3	92.3	96.5	97.5	96.6	91.5	94.7
CL	CNN-LB[11]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitSet[3]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart[5]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	MT3D[19]	76.0	87.6	89.8	85.0	81.2	75.7	81.0	84.5	85.4	82.2	68.1	81.5
	GaitGL[6]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitBase[20]	69.4	80.2	82.7	81.3	76.7	75.0	76.1	78.9	81.0	78.9	66.9	77.4
	RTBNet(ours)	75.5	89.9	92.1	89.6	84.3	81.6	85.8	89.4	92.2	87.4	72.6	85.5

表 3 列出了在 OU-MVLP 数据集上本文方法与目前流行方法的识别准确率对比。RTBNet 在很大程度上超过了之前的方法,尤其是在信息较少的 0°和 180°视角下,能够更加准确地捕捉每个部位之间的关联信息,从而提高步态识别模型的表征能力,使得性能得到显著的提升。

表 4 列出了在 Gait3D 数据集上本文方法与目前流行方法的识别准确率对比。1)RTBNet 相比基于 2D 卷积的方法(GaitSet[3],GaitPart[5],GLN[12])能够有效地建模时空特征,实验结果证实了本文方法的有效性。2)相比基于 3D 卷积的方法(GaitGL[6]),RTBNet 利用相对位置编码对步态序列进行编码,并采用 12 层的 RTB 模块建模时间维度的特征。3)相比基于 SMPL 的方法(SMPLGait[17]),RTBNet 只采用剪影图作为输入,网络结构相对而言更加简单有效,并获得了更高的识别准确率。

表 3　在 OU-MVLP 数据集上的识别准确率(不包括相同视角的情况)

Table 3　Accuracy on OU-MVLP dataset(excluding the same viewpoint)

(%)

Methods	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet[3]	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitPart[5]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN[12]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
GaitGL[6]	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
GaitMPL[21]	83.9	90.1	91.3	91.5	91.2	90.6	90.1	85.3	89.3	90.7	90.7	90.7	89.8	88.9	89.6
Lagrange[22]	85.9	90.6	91.3	91.5	91.2	91.0	90.6	88.9	89.2	90.5	90.6	89.9	89.8	89.2	90.0
GaitBase[20]	87.8	91.4	91.5	91.8	91.5	91.3	91.0	89.3	90.7	91.2	91.4	90.9	90.7	90.3	90.8
RTBNet(ours)	89.2	91.8	91.8	92.1	92.1	91.7	91.5	91.2	91.1	91.1	91.3	91.1	90.9	90.6	91.2

表 4　在 Gait3D 数据集上与目前流行方法的识别准确率对比

Table 4　Recognition accuracy comparison with current popular methods on Gait3D dataset

Methods	Publication	R-1/%	R-5/%
GaitSet[3]	AAAI 2019	36.7	58.3
GaitPart[5]	CVPR 2020	28.2	47.6
GLN[12]	ECCV 2020	31.4	52.9
GaitGL[6]	ICCV 2020	29.7	48.5
SMPLGait[17]	CVPR 2022	46.3	64.5
RTBNet(ours)	—	47.6	67.3

为了量化所提出的 RTB 模块对整个 RTBNet 模型性能的贡献,在 CASIA-B 数据集上对其进行消融实验,结果如表 5 所列。实验结果充分证明了 RTB 模块对整体模型性能的积极影响。具体而言,NM 的识别准确率从 96.1%提升至 97.9%,BG 的准确率从 91.4%提高到 94.7%,CL 的识别准确率也得到了 8.8%的绝对增长,表明了 RTB 模块的有效性和必要性。

表 5　RTB 模块的消融实验

Table 5　Ablation experiments of RTB

(%)

RTB	NM	BG	CL
—	96.1	91.4	76.7
√	97.9	94.7	85.5

结束语　本文介绍了一种基于相对位置编码转换器的深度步态识别的模型。该模型利用步态剪影数据,通过建模时空特征来判断其身份或状态。为了解除浅层的步态识别网络难以学习更复杂和更抽象的特征的限制,该模型采用了基于自注意力机制的深层步态识别网络,旨在学习更加复杂和抽象

的特征，使得步态识别模型具有更强的鲁棒性。最后，该模型在CASIA-B数据集和Gait3D上进行了实验验证，并与其他先进的步态识别模型进行了比较。实验结果表明，该模型在准确率和鲁棒性方面均优于其他模型，具有很高的实用价值。

综上所述，本文提出的基于相对位置编码转换器的深度步态识别模型在人身份识别和状态判断等领域具有很高的实用价值。通过对局部时空特征和全局时间特征建模，该模型能够准确地识别行人的步态特征，同时提高了模型的鲁棒性和泛化能力。本文的研究为步态识别任务提供了新的思路和方法，在提高人身份识别和状态判断的准确性和效率方面具有重要的实用价值。在未来的工作中，我们将进一步深入探究基于相对位置编码转换器在步态识别任务中的应用，进一步提升模型的性能。

参 考 文 献

[1] LARSEN P, SIMONSEN E, LYNNERUP N. Gait analysis in forensic medicine[J]. Journal of Forensic Sciences, 2008, 53: 1149-1153.

[2] BOUCHRIKA I, GOFFREDO M, CARTER J, et al. On using gait in forensic biometrics [J]. Journal of Forensic Sciences, 2011, 56: 882-889.

[3] CHAO H Q, HE Y W, ZHANG J P, et al. Gaitset: Regarding gait as a set for cross-view gait recognition [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 8126-8133.

[4] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification [J]. arXiv: 1703. 07737, 2017.

[5] FAN C, PENG Y J, CAO C S, et al. Gaitpart: Temporal part-based model for gait recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 14225-14233.

[6] LIN B B, ZHANG S L, YU X. Gait recognition via effective global-local feature representation and local temporal aggregation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2021: 14648-14656.

[7] HUANG Z, XUE D X, SHEN X, et al. 3D local convolutional neural networks for gait recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 14920-14929.

[8] YU S Q, TAN D L, TAN T N. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition[C]// The International Conference on Pattern Recognition. 2006: 441-444.

[9] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: A skinned multi-person linear model [J]. ACM Transactions on Graphics, 2015: 1-16.

[10] SHIRAGA K, MAKIHARA Y, MURAMATSU D, et al. Geinet: Viewinvariant gait recognition using a convolutional neural network [C]// International Conference on Biometrics. 2016: 1-8.

[11] WU Z F, HUANG Y Z, WANG L, et al. A comprehensive study on crossview gait based human identification with deep cnns [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39: 209-226.

[12] HOU S H, CAO C S, LIU X, et al. Gait lateral network: Learning discriminative and compact representations for gait recognition [C]// Proceedings of the European Conference on Computer Vision. 2020: 382-398.

[13] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

[14] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on in Neural Information Processing Systems. 2017: 6000-6010.

[16] SU J L, LU Y, PAN S F, et al. Roformer: Enhanced transformer with rotary position embedding [J]. arXiv: 2104. 09864, 2021.

[17] TAKEMURA N, MAKIHARA Y, MURAMATSU D, et al. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition[J]. IPSJ transactions on Computer Vision and Applications, 2018, 10: 1-14.

[18] ZHENG J K, LIU X C, LIU W, et al. Gait recognition in the wild with dense 3d representations and a benchmark [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 20228-20237.

[19] LIN B B, ZHANG S L, BAO F. Gait recognition with multiple-temporal-scale 3d convolutional neural net-work [C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020: 3054-3062.

[20] FAN C, LIANG J H, SHEN C F, et al. OpenGait: Revisiting Gait Recognition Toward Better Practicality [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 9707-9716.

[21] DOU H Z, ZHANG P Y, ZHAO Y H, et al. Gaitmpl: Gait recognition with memory-augmented progressive learning[J]. IEEE Transactions on Image Processing, 2022, 33: 1464-1475.

[22] CHAI T R, LI A N, ZHANG S X, et al. Lagrange motion analysis and view embeddings for improved gait recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 20249-20258.

REN Yuheng, born in 1993, postgraduate. His main research interests include computer vision and deep learning.