2025 年 第 25 卷 第 17 期
2025, 25(17):07276-09
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2404367
引用格式:李金成, 代雪晶, 闫睿鳌. 基于 3D-CNN 和融合 Transformer 的步态识别算法[J]. 科学技术与工程, 2025, 25(17): 7276-7284.
Li Jincheng, Dai Xuejing, Yan Ruiao. Gait recognition algorithm based on 3D-CNN and integrated Transformer[J]. Science Technology and Engineering, 2025, 25(17): 7276-7284.

基于 3D-CNN 和融合 Transformer 的步态识别算法

李金成, 代雪晶*, 闫睿鳌
(中国刑事警察学院公安信息技术与情报学院, 沈阳 100854)

摘 要 当前,步态识别的主流方法常依赖堆叠卷积层来逐步扩大感受野,以融合局部特征,这种方法大多采用浅层网络,在提取步态图像的全局特征时存在一定的局限性,并缺乏对时序周期特征信息的关注。因此提出一种融合 Transformer 和 3D 卷积的深层神经网络算法(3D convolutional gait recognition network based on adaptFormer and spect-conv,3D-ASgaitNet)。首先,初始残差卷积层将二进制轮廓数据转换为浮点编码特征图,以提供密集的低级结构特征;在此基础上,光谱层通过频域和时域的联合处理增强特征提取能力,并使用伪 3D 残差卷积模块进一步提取高级时空特征;最后融合 AdaptFormer 模块,通过轻量级的下采样-上采样网络结构,以适应不同的数据分布和任务需求,提供灵活的特征变换能力。3D-ASgaitNet 分别在 4 个公开的室内数据集(CASIA-B、OU-MVLP)、室外数据集(GREW、Gait3D)上进行,分别取得 99.84%、87.83%、45.32%、72.12% 的识别准确率。实验结果表明,所提出方法在 CASIA-B、Gait3D 数据集中的识别准确率接近 SOTA 性能。

关键词 步态识别; 融合 Transformer; 3D 残差卷积; 二进制轮廓数据
中图法分类号 TP399; **文献标志码** A

Gait Recognition Algorithm Based on 3D-CNN and Integrated Transformer

LI Jin-cheng, DAI Xue-jing*, YAN Rui-ao
(College of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang 100854, China)

[**Abstract**] Currently, mainstream gait recognition methods often rely on stacked convolutional layers to gradually expand the receptive field and integrate local features. These methods mostly use shallow networks, which have limitations in extracting global features from gait images and lack attention to temporal cycle feature information. Therefore, a deep neural network algorithm combining Transformer and 3D convolution, named 3D convolutional gait recognition network based on AdaptFormer and Spect-Conv (3D-ASgaitNet) was proposed. Firstly, the initial residual convolution layer converts the binary contour data into a floating-point encoded feature map to provide dense low-level structural features. On this basis, the spectral layer enhances the feature extraction ability through the joint processing of frequency domain and time domain, and uses the pseudo-3D residual convolution module to further extract advanced spatio-temporal features. Finally, AdaptFormer module was integrated to provide flexible feature transformation capability through lightweight down-sampling and up-sampling network structure to adapt to different data distribution and task requirements. 3D-ASgaitNet was carried out on four publicly available indoor datasets (CASIA-B, OU-MVLP) and outdoor datasets (GREW, Gait3D), and achieved recognition accuracy rates of 99.84%, 87.83%, 45.32% and 72.12%, respectively. Experimental results show that the recognition accuracy of the proposed method in CASIA-B and Gait3D data sets is close to the performance of SOTA.

[**Keywords**] gait recognition; fused Transformer; 3D residual convolution; binary silhouette data

步态识别是一种基于人的走路姿态信息进行身份认定的生物特征识别技术,与其他生物识别方式(如人脸、指纹和虹膜)相比,步态难以伪装,无需受试人员配合就可以远距离身份识别,并且步态识别对图像分辨率要求低,可以轻松地以非侵入性方式在远处识别[1]。因此,步态识别是安全应用中最重要的技术之一,在预防犯罪、法医鉴定、社会保障等方面的应用前景广阔。近年来,步态识别系统已协助多地警方快速侦破复杂案件,如广州黄埔警方利用步态识别,在案发 10 h 内成功侦破一起涉案金

收稿日期: 2024-06-12 **修订日期**: 2025-03-10
基金项目: 公安部科技强警基础工作专项(2023JC08);中央高校基本科研业务费(D2023001);中央高校基本科研业务费(D2024002)
第一作者: 李金成(2001—),男,汉族,湖北宜昌人,硕士研究生。研究方向:步态识别技术。E-mail:2023110198@cipuc.edu.cn。
* **通信作者**: 代雪晶(1970—),女,满族,辽宁凤城人,博士,教授。研究方向:声像资料技术。E-mail:1210724331@qq.com。

额达80万元的保险柜盗窃案。此外,该系统在溺水死亡、入室盗窃等案件侦破中同样发挥了重要作用,展现出在海量视频数据中高效检索、定位和追踪目标的能力,为公安侦查提供了强有力的技术支持。

现有的步态识别方法可以按照有无监督学习分成两类。在监督学习中,模型通过已标记的训练数据来学习,这些数据明确指出了每个样本的身份(或其他相关属性),该方法通常能达到较高的识别准确率,但需要大量的标记数据,且对新样本或不在训练集中的个体适应性较差。如Chao等[2]提出利用步态作为深度集,步态框架集成全局-局部融合深度网络来进行监督学习训练,该方法不受帧排列的影响,可以自然地整合在不同场景下获取的不同视频的帧。Liang等[3]提出了一种名为GaitEdge的端到端框架,能有效地阻止与步态无关的信息并释放端到端的训练算力,该方法的合成轮廓由可训练的身体边缘和固定的内部组成,以限制识别网络的接收信息。Zhang等[4]提出了一种自动编码器框架GaitNet,能准确地从RGB图像中分离出外观、姿势和典型特征,模型中LSTM模块将随时间变化的姿势特征集成为动态步态特征,而典型特征则被平均为静态步态特征。

无监督学习不依赖于标记的训练数据,而是直接从数据中发现模式或结构。无监督学习的步态识别方法主要包括基于生成对抗网络(generative adversarial networks,GAN)的步态识别方法、基于聚类的步态识别方法以及基于无监督域适应的步态识别方法[5],以下是针对这三类方法的具体应用实例:Talal等[6]将U-net作为图像生成器,ResNet作为身份鉴别器,并使用patchGAN作为视角分类器和真假鉴别器,利用更深层网络提取到更深层次的个体特征,达到目前利用GAN进行无监督步态识别的最高性能。Ren等[7]通过选择性集群融合策略,使用ClusterNCE Loss进行聚类约束,用K-最近邻算法在聚类中心周围选取适量样本作为聚类的支持集,来确定候选聚类。在反向传播过程中使支持集参与记忆库更新,将同一个体的正常行走和着装聚类中心拉近,以此提高着装识别率。Mu等[8]利用多个惯性测试单元采集人的行走数据,以此对标多源域适应场景。得到的数据通过MS-DANN模型训练分类,实现步态事件检测和步态模式识别,做出了无监督域在步态识别领域的早期尝试。

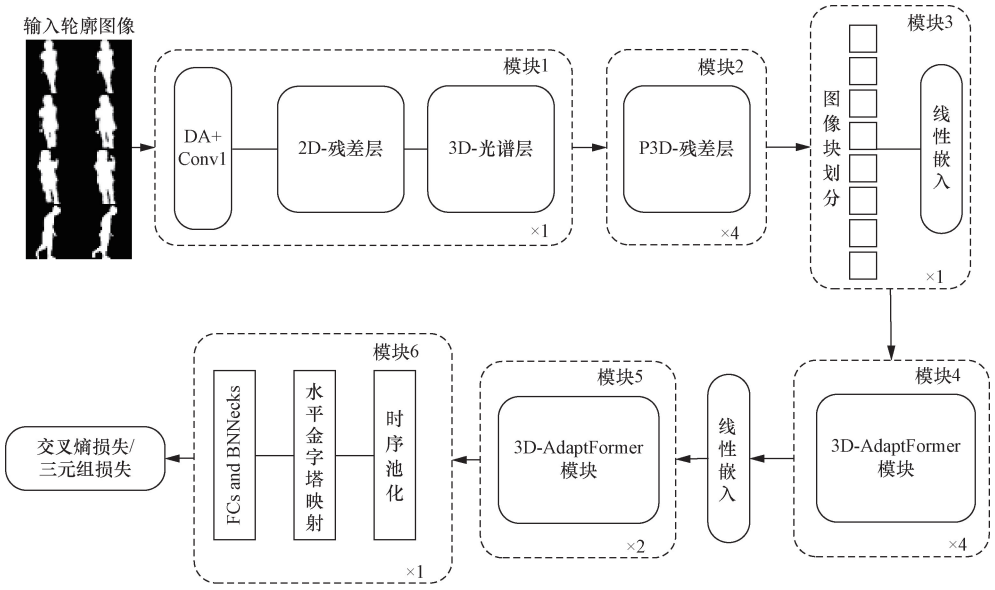
所提出方法属于监督学习中基于二进制轮廓数据的步态识别方法。由于现今步态识别方法通常过于关注每一帧图像的空间特征提取,忽略了视频帧之间的时间序列关系;或仅用卷积神经网络的算法,忽略了步态轮廓的全局特征重点信息,所以对于复杂环境的室外数据集识别准确率低下,难以保持处理室内数据集时的高性能。针对该问题,现提出一种基于3D-CNN和融合Transformer的步态识别算法(3D convolutional gait recognition network based on AdaptFormer and Spect-Conv,3D-ASgaitNet)。设计一种结合3D卷积和融合Transformer的主干网络,通过对频域和时域的联合处理增强特征提取能力,尤其加强了对连续步态帧序列的周期频率特征的提取能力;引入参数量不到视觉Transformer(vision Transformer,ViT)1.6%的AdaptMLP模块,通过轻量级的下采样-上采样网络,增强模型的适应性和全局特征感知能力,并能通过微调轻量级的适应性模块,理解步态序列的整体运动模式和长距离依赖关系,使得单个模型能够适应不同环境类型任务;在4个公开数据集上进行实验验证,并与其他模型算法进行对比,证明所提步态识别方法的有效性。

1　3D-ASgaitNet步态识别算法

针对视频中二进制轮廓图像的特点,使用3D-CNN和融合Transformer构建3D-ASgaitNet算法以解决现有步态识别算法迁移性较差以及室内外不同场景数据集识别准确率差异过大问题,主干网络如图1所示。按照数据处理的顺序主要分为6个模块:模块一(stage1)为数据增强与特征初步提取;模块二(stage2)为时空特征提取;模块三(stage3)为特征嵌入;模块四(stage4,stage5)为深度特征提取;模块五(stage6)为特征聚合与分类;模块六为损失函数(Triplet loss和Softmax loss)。

1.1　基于随机透视变换和基础轮廓转换的数据增强

在图像识别任务中,数据增强是一种基于现有的有限数据生成更多等价有效数据的技术,增加训练样本的数量以及多样性(噪声数据),提升模型鲁棒性。在步态识别领域,输入的图像通常不是传统的红绿蓝(red-green-blue,RGB)三通道彩色图像,而是去除行人外观颜色信息的二值化黑白轮廓图[9]。为了提升识别方法在行人轮廓受遮挡、环境光照变化以及视角变化等问题上的鲁棒性,提出基于随机透视变换(random perspective)和基础轮廓转换(fundamental contour)的数据增强方法,以增加受外观变化和环境变化影响的数据样本。增强后图像与原始图像如图2所示。



DA 为图像增强处理;Conv 为 Convolution 的缩写,其代表卷积层;2D、3D 为二维和三维;P3D 为伪三维;AdaptFormer 为自适应特征提取模块;FCs、BNNecks 代表 fully connected layers 和 batch normalization necks 的缩写,分别为全连接层和批归一化瓶颈层

图 1 3D-ASgaitNet 框架图

Fig. 1 3D-ASgaitNet structure diagram

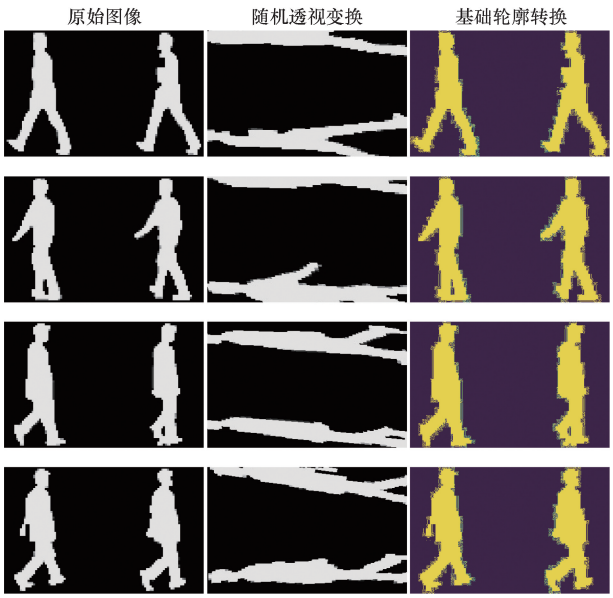


图 2 数据增强对比图

Fig. 2 Data augmentation comparison chart

随机透视变换是指利用透视中心、像点、目标点三点共线的条件,按透视旋转定律使承影面(透视面)绕迹线(透视轴)旋转某一角度,破坏原有的投影光线束,仍能保持承影面上投影几何图形不变的变换。$[x \quad y \quad w]$ 为原始图像中任意一点,$w=1$,透视变换用式(1)表示为

$$[x' \ y' \ w'] = [u \ v \ w]\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{1}$$

式(1)中:$[x' \quad y' \quad w']$ 为经过矩阵变换后映射到的新位置,w' 代表归一化因子;$[u \ v \ w]$ 为原始图像的齐次坐标表示;式子最右侧为一个 3×3 的变换矩阵 $\boldsymbol{H}$,该矩阵可以分成 4 个部分:第一部分 $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ 用于图像缩放、旋转,第二部分 $[a_{31} \quad a_{32}]$ 用于平移,第三部分 $[a_{13} \quad a_{23}]$ 用于透视变换,第四部分 $[a_{33}]$ 中 $a_{33}=1$。

实际应用中,为得到归一化坐标,经过透视变换后,图像的具体坐标公式表示为

$$x'' = \frac{x'}{w'} = \frac{a_{11}u + a_{21}v + a_{31}w}{a_{13}u + a_{23}v + w^2} \tag{2}$$

$$y'' = \frac{y'}{w'} = \frac{a_{12}u + a_{22}v + a_{32}w}{a_{13}u + a_{23}v + w^2} \tag{3}$$

基础轮廓转换通过对图像中的行人轮廓进行几何和结构上的调整,模拟真实世界中可能遇到的各种变化情况(服装变化、携带物品、部分遮挡等)。以形态学上的变换为例,使用形态学操作,如腐蚀和膨胀来模拟服装的松紧变化,用式(4)和式(5)表示为

$$\mathrm{dst}(x,y) = \min_{(x',y'):\mathrm{element}(x',y')\neq 0} \mathrm{src}(x+x', y+y') \tag{4}$$

$$\mathrm{dst}(x,y) = \max_{(x',y'):\mathrm{element}(x',y')\neq 0} \mathrm{src}(x+x', y+y') \tag{5}$$

式中:$\mathrm{dst}(x,y)$ 为目标图像位置 (x,y) 的像素值;$\mathrm{src}(x+x', y+y')$ 为原图像中相对于位置 (x,y) 偏

移(x',y')的像素值,其中(x',y')是结构元素中非零元素的位置;element(x',y') ≠ 0 表示结构元素中的非零位置。

综上所述,随机透视变换通过对图像的顶点进行随机的位置扰动,得到不同视角的视图效果,从而增强数据集的多样性和模型的适应性。基础轮廓转换提升了模型对实际环境中行人外观变化的适应能力。通过这种数据增强技术,模型能够有效地处理服装变化、部分遮挡、携带物品以及体态和姿势的差异,从而在各种复杂场景下实现更准确的步态识别。

1.2　基于 Spect-Conv 的时序特征提取模块

针对室内外数据集,图像的边缘、纹理结构特性(如光照变化、阴影等)在空间域中处理非常复杂,但在频域中可以通过简单的线性变换得到有效处理。同时,步态具有周期性特征,而周期性变化在频域中表现为明显的频率峰值,所以频域分析适合识别和处理步态中的周期性变化[10]。因此采用 Spect-Conv 模块,结合光谱注意力机制和卷积操作,在一定时间窗口内对多帧数据进行集合的频域变换来识别动态特征。Spect-Conv 模块如图 3 所示。

图 3 是一种傅里叶处理结构,通过快速傅里叶变换(fast Fourier transform,FFT)得到的频谱层来捕获架构初始层中的相关特征,将图像空间域转换到频域,用式(6)表示为

$$F(\boldsymbol{u},\boldsymbol{v}) = \sum_{\boldsymbol{x}=0}^{M-1}\sum_{\boldsymbol{y}=0}^{N-1} f(\boldsymbol{x},\boldsymbol{y})\,\mathrm{e}^{-\mathrm{j}2\pi\left(\frac{\boldsymbol{ux}}{M}+\frac{\boldsymbol{vy}}{N}\right)} \tag{6}$$

式中:$f(\boldsymbol{x},\boldsymbol{y})$ 为原始图像中位置 $(\boldsymbol{x},\boldsymbol{y})$ 的像素值;$F(\boldsymbol{u},\boldsymbol{v})$ 为转换到频域后在 $(\boldsymbol{u},\boldsymbol{v})$ 处的复数值;M、N 为图像的尺寸,M 为行数,N 为列数;j 为虚数单位;$\mathrm{e}^{-\mathrm{j}2\pi\left(\frac{\boldsymbol{ux}}{M}+\frac{\boldsymbol{vy}}{N}\right)}$ 代表一个旋转和缩放的复平面波形,其中 $2\pi\left(\frac{\boldsymbol{ux}}{M}+\frac{\boldsymbol{vy}}{N}\right)$ 为相位项。

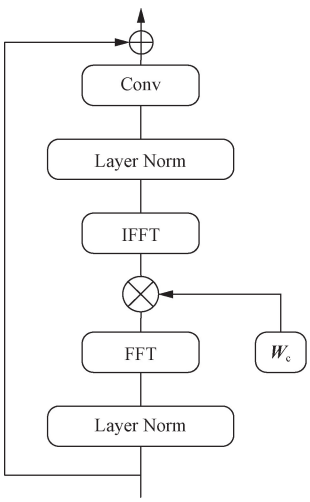


图 3　Spect-Conv 框架图

Fig. 3　Spect-Conv structure diagram

转换到频域后,对频域数据应用一个权重参数矩阵 $\boldsymbol{W}_c$,来调整不同频率成分的强度。在频域中,图像的每个频域成分 $F(\boldsymbol{u},\boldsymbol{v})$ 会与权重矩阵 $\boldsymbol{W}_c(\boldsymbol{u},\boldsymbol{v})$ 中相对应的元素相乘,以实现频域成分的调整,所以在连续时间帧的频域变换可以间接实现对视频数据时间序列特征的处理。然后进行快速傅里叶逆变换(inverse fast Fourier transform,IFFT)以重新获取新的空间信号,表达式为

$$F'(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{MN}\sum_{\boldsymbol{u}=0}^{M-1}\sum_{\boldsymbol{v}=0}^{N-1} G(\boldsymbol{u},\boldsymbol{v})\,\mathrm{e}^{-\mathrm{j}2\pi\left(\frac{\boldsymbol{ux}}{M}+\frac{\boldsymbol{vy}}{N}\right)} \tag{7}$$

式(7)中:$G(\boldsymbol{u},\boldsymbol{v})$ 为在频率 $(\boldsymbol{u},\boldsymbol{v})$ 处的幅度和相位的复数表示;$F'(\boldsymbol{u},\boldsymbol{v})$ 为在时空域中位置 $(\boldsymbol{x},\boldsymbol{y})$ 的像素值或信号强度。

同时为保证数据处理的稳定性和效率,在 FFT 和 IFFT 前后应用层归一化,可以表示为

$$\mathrm{LN}(\boldsymbol{x}) = \frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{8}$$

式(8)中:$\boldsymbol{\mu}$ 和 $\boldsymbol{\sigma}$ 分别为层内所有元素的平均值和标准差。

1.3　伪 3D 残差卷积时空特征提取模块

为提取步态视频的时空特征和减少模型的训练参数,在步态识别模型中使用伪 3D 残差卷积网络(P3D),将 2D 卷积模型的训练参数作为 3D 模型的空间卷积层参数的初始化[11]。整体来说,P3D 卷积将 3D 卷积解耦成 2D 的空间卷积和 1D 的时间卷积,如图 4 所示。

在步态识别视频帧序列中,空间特征和时间特征之间存在直接和间接的影响,并且都直接和间接地影响最终输出,所以设计出图 4 的残差连接方式。图 4 的伪 3D 残差卷积框架图可以表示为

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \boldsymbol{x}_t + \boldsymbol{S}(\boldsymbol{x}_t) + \boldsymbol{T}[\boldsymbol{S}(\boldsymbol{x}_t)]\\ &= (\boldsymbol{I}+\boldsymbol{S}+\boldsymbol{TS})\,\boldsymbol{x}_t\end{aligned} \tag{9}$$

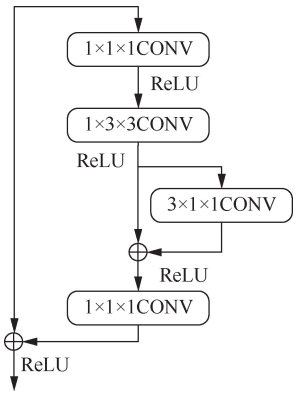


图 4　伪 3D 残差卷积框架图

Fig. 4　Pseudo-3D residual convolution framework diagram

式(9)中:$\boldsymbol{S}$ 为空间维度上的二维滤波器;$\boldsymbol{T}$ 为时间域的一维滤波器;$\boldsymbol{x}_t$ 和 $\boldsymbol{x}_{t+1}$ 为第 t 个残差单元的输入和输出。

1.4 融合 Swin Transformer 和 AdaptFormer 的深度特征提取模块

在使用 ViT 步态识别网络对二值化轮廓数据进行训练时,由于数据中存在大量无信息(哑)斑块,自注意力机制的计算量会成倍增加,生成无用梯度的风险也会相应提高。针对这一问题,融合 Swin Transformer 和 AdaptFormer 时,采用了 Swin Transformer 的基于移位窗口的多头自注意力模块,使算法具有类似 CNN 滑动窗口机制和分层架构的特性[12]。融合 Transformer 是由基于移位窗口的多头自注意力模块、AdaptMLP 模块、层归一化 Layer Norm 等组成,整体架构如图 5 所示。

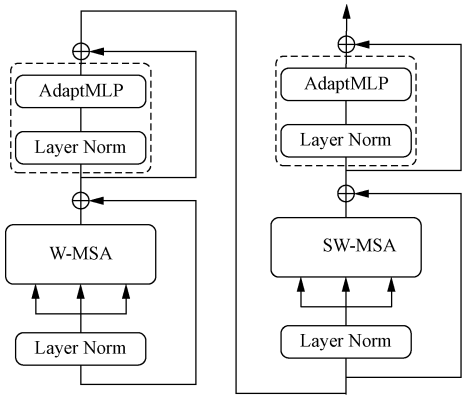


图 5 Swin Transformer 框架图

Fig. 5 Swin Transformer structure diagram

图5 中,窗口自注意力机制(window-based multi-head self-attention,W-MSA)是在固定大小的局部窗口内进行的多头自注意力计算,滑动窗口自注意力机制(sliding window multi-scale attention,SW-MSA)是在 W-MSA 的基础上,通过对窗口进行移位操作来捕获跨窗口的全局信息。融合 Transformer 通过多头注意力模块,实现对特征矩阵进行更深层次的特征提取。多头注意力模块的计算过程如式(10)~式(12)所示。

$$\boldsymbol{x}_1' = \mathrm{Atten}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{10}$$

$$\boldsymbol{h}_i = \mathrm{Atten}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{11}$$

$$\mathrm{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{concat}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i) \tag{12}$$

式中:$\boldsymbol{x}_1'$ 为输入特征;$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} \in \mathbf{R}^{N^2\times d}$ 分别为查询、键和值,N^2 为进行 self-attention 的 token 数目,d 为 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 的维度;$\sqrt{d_k}$ 为对注意力值进行归一化;$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$ 为模块内权重矩阵;$\boldsymbol{h}_i$ 为注意力机制的头数,$i = 1,2,\cdots,l$。

在此基础上,对该自注意力机制计算相似度过程中对每个 head 加入相对位置偏置 $\boldsymbol{B} \in \mathbf{R}^{M^2\times M^2}$,表达式为

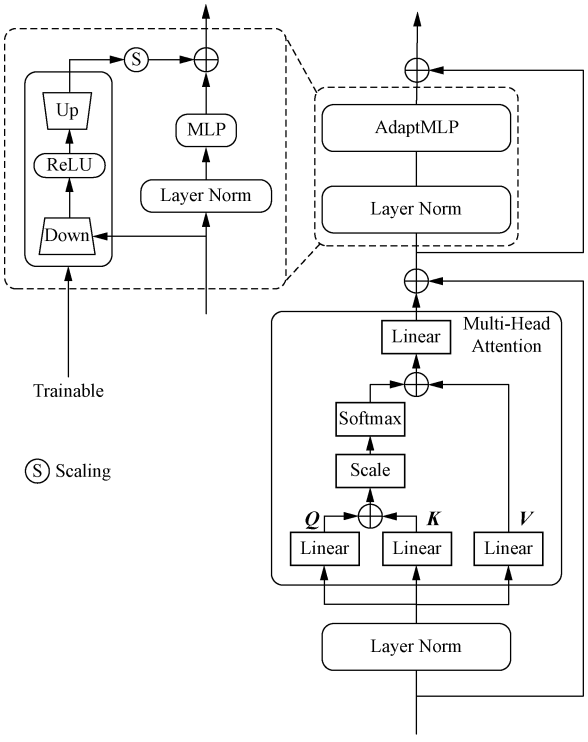
$$\mathrm{Atten}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{SoftMax}(\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}/\sqrt{d_k} + \boldsymbol{B})\boldsymbol{V} \tag{13}$$

式(13)中:M^2 为(局部)窗口内的 patches 数;因为沿各轴的相对位置均处于 $[-M+1,M-1]$ 范围内,所以参数化一个更小尺寸的偏置矩阵 $\hat{\boldsymbol{B}} \in \mathbf{R}^{(2M-1)\times(2M-1)}$,$\boldsymbol{B}$ 中的值均取自 $\hat{\boldsymbol{B}}$。

所以通过移位窗口来进行特征提取计算,移位窗口方案通过将自注意力计算限制在不重叠的局部窗口,同时允许跨窗口连接来提高效率,这种分层架构具有在各种尺度上灵活建模的特性,使得处理图像数据时有着线性计算的复杂度。

引入 AdaptMLP 模块来调整轻量级模块,以生成适应不同场景(如室内和室外)的特征,能够以较小的计算量提高融合 Transformer 的可迁移性,使得该模型在室内和室外的步态数据集上都能够以较小的计算成本获得较高的识别准确率[13]。AdaptFormer 整体架构如图 6 所示。

图 6 中左分支被设计为限制参数数量的瓶颈结构,生成的适应性特征 $\hat{\boldsymbol{x}}_1$,其表达形式为



Trainable 代表该虚线内模块可训练变化;Scaling 为可缩放因子;Multi-Head Attention 为多头注意力机制;Linear 为线性层即全连接层

图 6 AdaptFormer 框架图

Fig. 6 AdaptFormer structure diagram

$$\hat{\boldsymbol{x}}_1 = \mathrm{ReLU}[\mathrm{LN}(\boldsymbol{x}_1')\boldsymbol{W}_{\mathrm{down}})\boldsymbol{W}_{\mathrm{up}} \tag{14}$$

式(14)中:$\boldsymbol{x}_1'$为特定的输入特征;$\boldsymbol{W}_{\mathrm{down}} \in \mathbf{R}^{d\times\hat{d}}$为下投射层的一个参数;$\boldsymbol{W}_{\mathrm{up}} \in \mathbf{R}^{\hat{d}\times d}$为上投射层的一个参数;$\hat{d}$为瓶颈的中间维度,满足$\hat{d} \ll d$。

此外,由于非线性性质,在这些投射层之间存在一个 ReLU 层。该瓶颈模块通过一个比例因子 s 和原始的 MLP 网络(右分支)残差连接,然后特征$\hat{\boldsymbol{x}}_1$和$\boldsymbol{x}_1'$通过残差连接与$\boldsymbol{x}_1$融合,其表达式为

$$\boldsymbol{x}_1 = \mathrm{MLP}[\mathrm{LN}(\boldsymbol{x}_1')] + s\hat{\boldsymbol{x}}_1 + \boldsymbol{x}_1' \tag{15}$$

1.5　特征聚合模块

特征聚合模块将输入序列中的特征信息进行整合,以便于后续的分类或识别任务。在步态识别任务中,输入通常是一个连续帧序列,通过特征提取网络(如 CNN 或 Transformer)得到的特征是一个四维张量,包含了时间、空间和通道信息。特征聚合模块的作用是对这些高维特征进行聚合,得到一个固定长度的特征表示。

1.5.1　时序池化

时序池化(temporal pooling)是指在时间序列数据上进行池化操作,其主要目的是通过在时间维度上聚合信息,提取序列中的关键特征,简化后续的处理流程。因为步态序列通常由连续的帧组成,不同的序列长度可能不同,所以通过时序池化可以将变长的序列转换为固定长度的特征向量,从而减少计算量,并保留关键信息。在本文模型中对时间维度进行最大池化,表达式为

$$\boldsymbol{Y}_{bchw} = \mathrm{Max}_{d=1}^{D}\boldsymbol{X}_{bcdhw} \tag{16}$$

式(16)中:$\boldsymbol{X}_{bcdhw}$为输入的特征矩阵,表示第“b”个样本,第“d”个时间步,通道数、高度、宽度分别为c、h、w;$\boldsymbol{Y}_{bchw}$为池化后的输出矩阵的值。

1.5.2　水平金字塔映射

水平金字塔映射(horizontal pooling matching,HPM)[2]将特征图在水平方向上进行不同尺度的重复分割,再进行池化操作后将其依次堆叠,形成金字塔状的特征图,HPM 结构如图 7 所示。

利用平均池化和最大池化进行相加,从而同时保留全局、局部信息和最具辨别力的特征表示,如式(17)所示。该模块通过关注不同尺度的特征,使网络聚焦在行走过程中不同部位的信息。

$$\boldsymbol{f}' = \mathrm{AvgPool}(\boldsymbol{F}_{i,j}) + \mathrm{MaxPool}(\boldsymbol{F}_{i,j}) \tag{17}$$

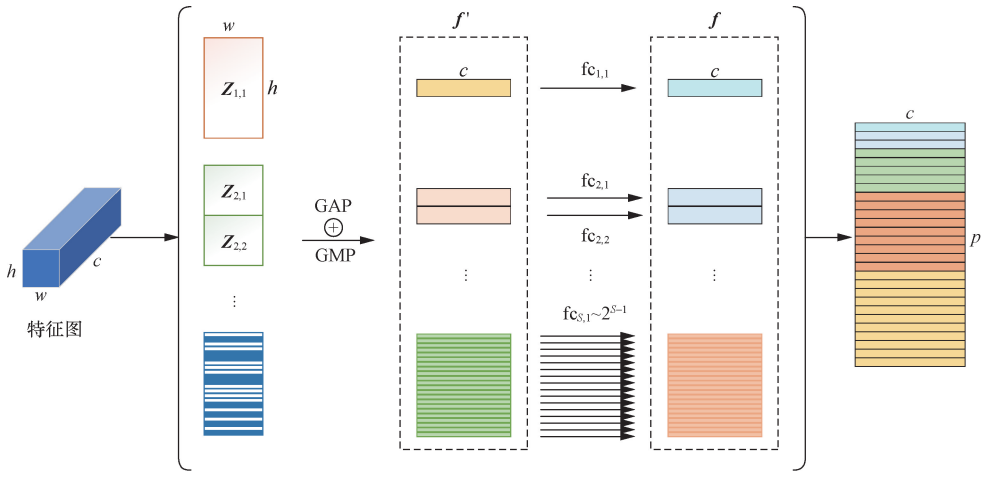
式(17)中:$\boldsymbol{F}_{i,j}$为将输入特征$\boldsymbol{F}_i$水平分割的第j部分;AvgPool 为平均池化;MaxPool 为最大池化。

2　实验结果与分析

2.1　实验设置

实验所涉代码复现以及实验均在服务器(2×NVIDIA GeForce RTX 3090,Inter(R)Xeon(R)Platinum 8350C CPU@2.60GHz,84RAM);操作系统为 Ubuntu 20.04 系统;PyTorch2.0 深度学习框架;Python3.10.2 版本;所使用的测试平台为 OpenGait[14]。采用的数据集如表 1 所示。

实验的 4 个数据集(CASIA-B、OU-MVLP、GREW、Gait3D)的参数设置如表 2 所示。4 个数据集的输入分辨率都设置为 64×44;评估器均启用浮点 16 位计算以便优化内存和速度,距离度量使用欧几里得距离(euc)来评估样本间相似性;使用三元组



h、w 分别为特征图的高与宽;c 为通道数;p 为水平特征向量的数量;S 为 HPM 的尺度数目;$\boldsymbol{Z}_{s,t}$ 为特征块在尺度 S 下的索引,其中 $t \in 1,2,\cdots,2^{S-1}$;“⊕”表示对应元素相加;GAP 与 GMP 分别为全局平均池化和全局最大池化;fc 为独立的全连接操作;$\boldsymbol{f}'$ 与 $\boldsymbol{f}$ 分别为全连接操作前后的特征向量

图 7　水平金字塔映射结构

Fig. 7　Horizontal pyramid mapping structure

表 1 数据集

Table 1 Data set

数据集名称	训练集		测试集		采集环境
	Id	Seq	Id	Seq	
CASIA-B	74	8 140	50	5 500	室内
OU-MVLP	5 153	144 284	5 154	144 412	室内
Gait3D	3 000	18 940	1 000	6 369	室外
GREW	20 000	102 887	6 000	24 000	室外

注:数据集 Id 为每个数据项的唯一标识符;Seq 为序列(sequence)。

表 2 数据集训练参数

Table 2 Dataset training parameters

数据集名称	批次大小	优化器	学习率调整阶段	迭代次数
CASIA-B	(4,8)	SGD	(20k,40k,60k)	80k
OU-MVLP	(32,8)	SGD	(60k,80k,100k)	120k
Gait3D	(32,4)	AdamW	I_{max} =60k	80k
GREW	(32,4)	AdamW	I_{max} =150k	200k

注:I_{max} 为学习率设置的最大值;k 为 1 000,用来对比不同数据集的学习率调参大小。

损失和交叉熵损失,均设置权重为 1.0,三元组损失的边距设为 0.2,交叉熵损失的放大比例设置为 16;SGD 优化器的初始学习率和权重衰减分别设为 0.1 和 0.000 5;AdamW 优化器学习率设置为 0.000 3,权重衰减设置为 0.02;输入帧数不固定;此外,采用随机透视变换和基础轮廓转换为空间数据增强技术。

2.2 实验结果与分析

将 3D-ASgaitNet 模型的测试结果与近年来提出的 SMPLGait、GaitSet 等 6 种先进的模型在 4 个步态数据集 CASIA-B、OU-MVLP、GREW 和 Gait3D 上的测试数据进行对比(使用 Rank-1 准确率作为评价指标)。对比结果如表 3 所示。

多个数据集上的实验结果表明,在CASIA-B数据集的 NM、BG 和 CL 子集上,本文模型的 Rank-1 准确率分别为 98.8%、95.4% 和 88.7%,优于对比模型,这是由于 Spect-Conv 模块和伪 3D 卷积模块的时序特征提取能力,有效捕捉了行走节奏和模式。在 OU-MVLP 数据集上,模型的 Rank-1 准确率为 78.2%,与其他 SOTA 模型有一定的差距,这是由于 3D-ASgaitNet 网络层数较深,同时 OU-MVLP 数据集数据量大且缺少多样性,所以本文模型在该数据集中容易出现过拟合,导致性能下降。在更复杂的 GREW 数据集上,Rank-1 准确率为 36.8%,低于 SOTA 模型,GREW 数据集相对于 Gait3D 数据集环境更加复杂,且 GREW 数据集根据年龄的不同进行了分类采集,本文模型在 GREW 数据集中过于关注视频帧序列的时序特征,对空间深度特征的采集不足,导致该数据集 Rank-1 准确率相对其他 SOTA 模型较低。值得注意的是,在 Gait3D 数据集上,模型表现出较高的 Rank-1 准确率(63.6%),说明本文模型在训练 Gait3D 数据集时,有效地提取了时空特征。同时 CSTL 在 Gait3D 数据集的 Rank-1 准确率相对室内数据集过于低下,说明仅使用浅层网络在面临复杂数据集时无法准确提取步态特征,所以所提的深层网络在处理复杂数据集时占到了一定的优势。总体来说,本文模型在CASIA-B、Gait3D 上的性能表现优异。

在训练阶段,模型的 softmax 损失和准确率变化如图 8 所示。模型在 CASIA-B 和 OU-MVLP 数据集上的准确率迅速上升并趋于稳定,表明模型在两者数据集上能够有效学习并泛化。在 GREW 和 Gait3D 数据集上,识别准确率上升较慢且波动较大,表明模型在处理复杂和多样化的数据集时仍面临挑战。损失值曲线显示了模型的收敛过程。在 CASIA-B、OU-MVLP 以及 Gait3D 数据集上,损失值迅速下降并趋于稳定,反映出模型的收敛性较好。

在训练阶段,本文模型在不同数据集上的三元组损失样本数量和三元组平均距离的变化情况如图 9 所示。在三元组损失样本数量图中,所有数据集的损失样本数量在训练初期迅速下降,并趋于稳定,表明模型能够逐渐识别并优化三元组样本。在三元组平均距离图中,随着训练迭代次数的增加,所有数据集的平均距离逐渐增大并趋于稳定,反映模型在

表 3 不同方法 Rank-1 准确率对比

Table 3 Comparison of Rank-1 accuracy of different methods

模型	期刊信息	准确率/%					
		CASIA-B			OU-MVLP	GREW	Gait3D
		NM	BG	CL			
本文模型	—	98.8	95.4	88.7	78.2	36.8	63.6
GaitSet[2]	AAAI 2019	95.8	90.0	75.4	87.1	48.4	36.7
GaitPart[15]	CVPR 2020	96.1	90.7	78.7	88.7	47.6	28.2
GaitGL[16]	ICCV 2021	97.4	94.5	83.8	89.7	47.3	29.7
CSTL[17]	ICCV 2021	98.0	95.4	87.0	90.2	50.6	11.7
3DLocal[18]	ICCV 2021	98.3	95.5	84.5	90.9	—	—
SMPLGait[19]	CVPR 2022	—	—	—	—	—	46.3

注:NM、BG 以及 CL 分别为 CASIA-B 数据集中测试集包含的正常、携包以及穿着外套的步态序列。

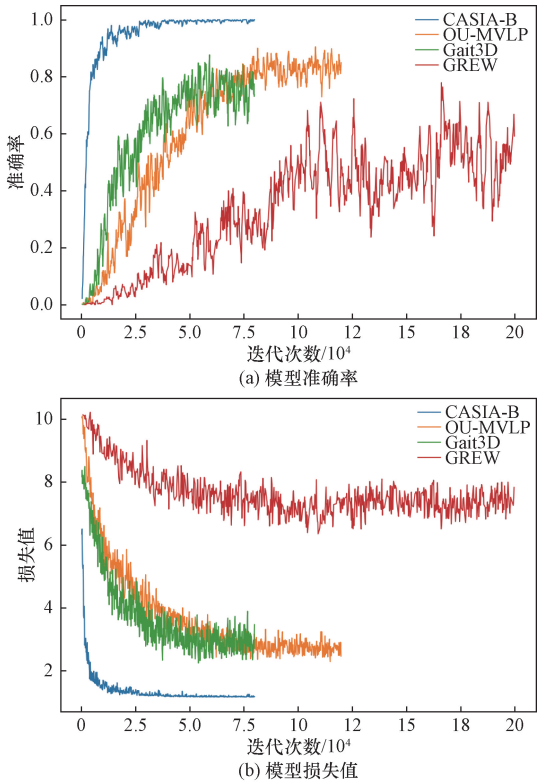


图 8　4 种数据集上迭代的准确率和损失值

Fig. 8　Accuracy and loss values of iterations on four datasets

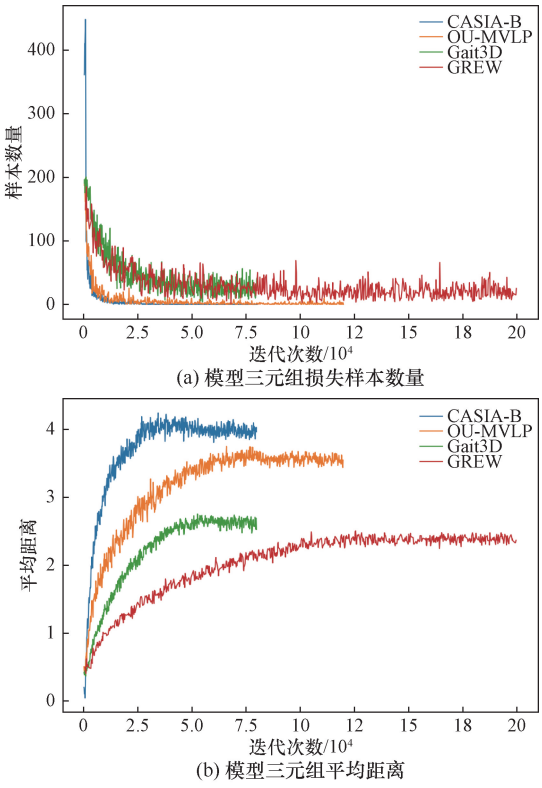


图 9　4 种数据集上迭代的三元组损失样本数量和平均距离变化曲线

Fig. 9　Iterative triplet loss sample size and average distance variation curves on four datasets

逐渐拉开正负样本距离,提高识别效果,CASIA-B 和 OU-MVLP 数据集的曲线表现出较高的稳定性和较大的平均距离,说明模型在室内数据集上表现较好;而 GREW 和 Gait3D 数据集的曲线波动较大,且最终距离较小,表明在复杂数据集上的识别效果还有提升空间。

2.3　消融实验

为探究模型各个模块对步态识别的作用,在 Gait3D 数据集上进行消融实验,其他实验设置均相同,具体结果如表 4 所示。

表 4　Gait3D 数据集上的消融实验结果

Table 4　Results of ablation experiments on the Gait3D dataset

实验序号	Spect-Conv	3D-CNN	AdaptFormer	识别准确率/%
0	√	—	—	32.64
1	—	√	—	61.23
2	—	—	√	68.23
3	√	√	—	62.11
4	√	—	√	67.22
5	—	√	√	70.63
6	√	√	√	72.12

注:√表示添加了该模块;—表示未添加该模块。

如表 4 所示,实验 0 仅使用 Spect-Conv 模块,得到了较低的识别准确率 32.64%,这表明 Spect-Conv 单独使用时特征提取能力有限。实验 1 中单独使用 3D-CNN,识别准确率提升至 61.23%,显示出 3D-CNN 在处理时空数据上的强大能力。实验 2 中单独使用 AdaptFormer 模块,识别准确率达到了 68.23%,表现出该模块能够准确地提取全局关键特征信息。实验 3 结合 Spect-Conv 和 3D-CNN,准确率较单独使用 Spect-Conv 模块提升 29.47%,说明 Spect-Conv 与 3D-CNN 的耦合性较好。实验 4 ~ 实验 6 中都引入 AdaptFormer 模块,准确率逐步提高,表现出 AdaptFormer 在多模块融合时,其提高步态识别效果上的显著作用。这些实验表明,虽然单一模块有其局限,但多模块联合使用能有效提升模型的整体性能,尤其是 AdaptFormer 的加入,为模型带来了显著的性能提升。

2.4　跨数据集可迁移性实验

为了验证 3D-ASgaitNet 添加了 AdaptMLP 模块后模型的可迁移性,采用了室外数据集 Gait3D 和室内数据集 CASIA-B 进行迁移性实验,观察迁移训练后的识别准确率是否显著降低。实验设计为先后在两个不同环境的数据集上进行训练。由于引入了 AdaptMLP 模块,相比原始模型,后者的训练时间显然会缩短,参数数量也更少。从表 5 的验证结果可以看出,尽管训练时间减少,但模型的识别准确率却得到了维持。

表 5 跨数据集实验结果

Table 5 Cross dataset experimental results

测试标准	Gait3D	CASIA-B	跨数据集再训练	
			Gait3D→CASIA-B	CASIA-B→Gait3D
识别准确率/%	72.12	99.84	99.98	68.89

3 结论

为解决现有步态识别算法在室内外场景中识别准确率较低的问题,基于 3D-CNN 与融合 Transformer,提出了一种高精度的步态识别算法 3D-ASgaitNet。

(1) 3D-ASgaitNet 算法通过融合 Transformer 和 3D-CNN,显著提高了步态识别的性能,尤其在动态特征提取和跨数据集适应性方面表现突出。

(2) 3D-CNN 有效地减少了视频数据的维度和复杂性,为 Transformer 提供了结构化的底层信息,从而提升了模型的效率和精度。

(3) 融合 Transformer 的全局建模能力和动态注意力机制弥补了 CNN 在全局建模方面的不足,增强了步态识别的精度。

(4) 轻量级 AdaptMLP 模块的引入,提高了模型在不同数据集中的适应性,并能够自适应调整参数,精确捕捉行走动作的特定特征。

在 CASIA-B、Gait3D 数据集上的其他步态识别算法对比,3D-ASgaitNet 的识别效果优于同等规模的卷积神经网络,并且能够有效提升步态识别动态特征提取能力,同时网络收敛速度也有较大的提升。然而对于训练完成后的模型,在跨数据集上测试模型性能时,其识别准确率有所下降,所以未来的研究将重点解决这一问题,进一步提高模型在跨数据集测试时的鲁棒性和准确率。

参 考 文 献

[1] 王紫薇. 基于深度学习的步态目标检测与识别方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2023.
Wang Ziwei. Research on gait object detection and recognition method based on deep learning [D]. Harbin: Harbin Engineering University, 2023.

[2] Chao H, He Y, Zhang J, et al. GaitSet: regarding gait as a set for cross-view gait recognition[J]. ArXiv, 2018: 1811.06186.

[3] Liang J, Fan C, Hou S, et al. Gaitedge: beyond plain end-to-end gait recognition for better practicality[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 375-390.

[4] Zhang Z, Tran L, Liu F, et al. On learning disentangled representations for gait recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 345-360.

[5] 陈福仕, 沈尧, 周池春, 等. 无监督学习步态识别综述[J]. 计算机科学与探索, 2024, 18(8): 2014-2033.
Chen Fushi, Shen Yao, Zhou Chichun, et al. Review of unsupervised learning gait recognition[J]. Computer Science and Exploration, 2024, 18(8): 2014-2033.

[6] Talal E B, Oraibi Z A, Wali A. Gait recognition using deep residual networks and conditional generative adversarial networks[C]//2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). Torino: IEEE, 2023: 1179-1185.

[7] Ren X, Hou S, Cao C, et al. Unsupervised gait recognition with selective fusion[J]. ArXiv Preprint ArXiv, 2023: 2303.10772.

[8] Mu F, Gu X, Guo Y, et al. Unsupervised domain adaptation for position-independent IMU based gait analysis[C]//IEEE Sensors Conference. Rotterdam: IEEE, 2020: 1-4.

[9] 高毅, 王彪, 王梦阳, 等. 基于反向传播神经网络的三维足迹触觉步态特征识别[J]. 科学技术与工程, 2022, 22(24): 10646-10653.
Gao Yi, Wang Biao, Wang Mengyang, et al. 3D recognition of footstep tactile gait characteristics based on back propagation neural network[J]. Science Technology and Engineering, 2022, 22(24): 10646-10653.

[10] 闫锋, 苏忠允. 基于时频域融合和 ECA-1DCNN 的航空串联故障电弧检测[J]. 科学技术与工程, 2024, 24(5): 1937-1945.
Yan Feng, Su Zhongyun. Aviation series arc fault detection method based on time frequency domain fusion and ECA-1DCNN [J]. Science Technology and Engineering, 2024, 24(5): 1937-1945.

[11] 廖金雷, 张磊, 周湘山, 等. 融合植被指数的 3D-2D-CNN 高光谱图像植被分类方法[J]. 科学技术与工程, 2021, 21(27): 11656-11662.
Liao Jinlei, Zhang Lei, Zhou Xiangshan, et al. 3D-2D-CNN hyperspectral image vegetation classification method fusion with vegetation index[J]. Science Technology and Engineering, 2021, 21(27): 11656-11662.

[12] Fan C, Hou S, Huang Y, et al. Exploring deep models for practical gait recognition[J]. ArXiv Preprint ArXiv, 2023: 2303.03301.

[13] Chen S, Ge C, Tong Z, et al. AdaptFormer: adapting vision transformers for scalable visual recognition[J]. Advances in Neural Information Processing Systems, 2022, 35: 16664-16678.

[14] Fan C, Liang J, Shen C, et al. OpenGait: revisiting gait recognition towards better practicality[J]. ArXiv, 2023: 2211.06597.

[15] Fan C, Peng Y, Cao C, et al. Gaitpart: temporal part-based model for gait recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 14225-14233.

[16] Lin B, Zhang S, Wang M, et al. GaitGL: Learning discriminative global-local feature representations for gait recognition[J]. ArXiv Preprint ArXiv, 2022: 2208.01380.

[17] Huang X, Zhu D, Wang H, et al. Context-sensitive temporal feature learning for gait recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 12909-12918.

[18] Huang Z, Xue D, Shen X, et al. 3D local convolutional neural networks for gait recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 14920-14929.

[19] Zheng J, Liu X, Liu W, et al. Gait recognition in the wild with dense 3D representations and a benchmark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Louisiana: IEEE, 2022: 20228-20237.