Computer Science and Application 计算机科学与应用, 2025, 15(6), 1-14
Published Online June 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.156152

Hans汉斯

基于轮廓与骨架多特征融合的人体步态识别

吕文涛[1]，薛博译[2]，黄　泽[1]，陆俊豪[3]，丁　浩[1]

[1]江苏警官学院刑事科学技术系，江苏　南京
[2]江苏警官学院计算机信息与网络安全系，江苏　南京
[3]南京邮电大学计算机学院，江苏　南京

收稿日期：2025年5月6日；录用日期：2025年6月6日；发布日期：2025年6月12日

摘　要

步态识别作为非接触式生物识别技术，在复杂环境下的跨视角识别准确率与鲁棒性仍面临挑战。针对现有方法对多尺度时空特征建模不足、跨模态信息融合机制单一等问题，本文提出一种基于多模态特征融合的端到端步态识别框架。首先，设计了一种结合混合高斯模型与形态学优化的动态剪影提取算法，有效降低噪声干扰并增强目标区域表征能力；其次，构建多分支特征提取网络，通过三维时空图卷积网络(3D-STGCN)捕捉步态序列的全局时空关联，并引入姿态引导注意力模块(PGAM)强化局部关键关节的语义信息；最后，提出跨模态自适应融合机制(CMAF)，实现剪影轮廓特征与骨架运动特征的多层次互补。在CASIA-B数据集上的实验表明，本文方法在跨视角(0˚~180˚)场景下的平均Rank-1准确率均有明显提升，显著优于主流模型GaitSet、GaitTB和GaitPart。本文工作为复杂场景下的步态识别提供了可扩展的解决方案，具有广阔的应用前景。

关键词

步态识别，人体骨架，特征融合，3D时空图卷积，姿态引导注意力

Human Gait Recognition Based on the Fusion of Contour and Skeleton Multi-Features

Wentao Lv[1], Boyi Xue[2], Ze Huang[1], Junhao Lu[3], Hao Ding[1]

[1]Department of Forensic Science and Technology, Jiangsu Police Institute, Nanjing Jiangsu
[2]Department of Computer Information and Network Security, Jiangsu Police Institute, Nanjing Jiangsu
[3]School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: May 6th, 2025; accepted: Jun. 6th, 2025; published: Jun. 12th, 2025

文章引用: 吕文涛, 薛博译, 黄泽, 陆俊豪, 丁浩. 基于轮廓与骨架多特征融合的人体步态识别[J]. 计算机科学与应用, 2025, 15(6): 1-14. DOI: 10.12677/csa.2025.156152

Abstract

As a non-contact biometric identification technology, gait recognition still faces challenges in cross-view recognition accuracy and robustness in complex environments. Aiming at the problems such as insufficient multi-scale spatio-temporal feature modeling and a single cross-modal information fusion mechanism in existing methods, this paper proposes an end-to-end gait recognition framework based on multi-modal feature fusion. Firstly, a dynamic silhouette extraction algorithm combining a Gaussian mixture model and morphological optimization is designed, which effectively reduces noise interference and enhances the representation ability of the target area. Secondly, a multi-branch feature extraction network is constructed. The 3D spatio-temporal graph convolutional network (3D-STGCN) is used to capture the global spatio-temporal correlations of gait sequences, and a pose-guided attention module (PGAM) is introduced to strengthen the semantic information of local key joints. Finally, a cross-modal adaptive fusion mechanism (CMAF) is proposed to achieve multi-level complementarity between silhouette contour features and skeleton motion features. Experiments on the CASIA-B dataset show that the average Rank-1 accuracy of the proposed method in cross-view (0˚~180˚) scenarios is significantly improved, and it is remarkably better than mainstream models such as GaitSet, GaitTB, and GaitPart. This work provides an expandable solution for gait recognition in complex scenarios and has broad application prospects.

Keywords

Gait Recognition, Human Skeleton, Feature Fusion, 3D Spatio-Temporal Graph Convolution, Pose-Guided Attention

1. 引言

近年来，作为一种基于个体行走特征进行身份识别的新兴技术，步态识别凭借其非接触性和高度防伪特性在公共安全、医疗健康等领域备受关注[1]。然而，现有方法在多尺度时空特征建模与跨模态信息融合方面仍存在显著缺陷，严重制约了复杂场景下的识别性能[2]。

现有方法在时空特征提取中多采用固定尺度的卷积核或单一时间窗口，难以有效捕捉步态序列中短时微动作与长时周期性运动的协同特征。例如，GaitPart [3]通过焦点卷积提取局部特征，但未建立层次化的时空关联，导致对步态速度变化的适应性不足；3Dlocal [4]虽引入三维卷积，却忽略了关节运动的层级语义(如近端关节与远端关节的动态差异)。此类方法在跨视角场景下易受肢体遮挡或运动模糊干扰，表现为视角超过 90˚时识别率骤降[5]。此外，缺乏对多尺度时空关系的显式建模(如局部肢体摆动与全局身体姿态的相互作用)，进一步削弱了模型对复杂运动模式的表征能力。

研究表明，仅依赖单一尺度特征的方法在 CASIA-B 数据集 CL 子集上的平均识别率普遍低于 80% [6]，证实了多尺度建模的迫切性。当前跨模态融合策略多采用特征拼接或简单加权求和，未能深入挖掘剪影轮廓(外观)与骨架序列(结构)的互补性。例如，SMPLGait [7]虽联合使用轮廓与骨架特征，但其融合过程未考虑模态间的语义对齐，导致在携带物品(BG 条件)时外观特征失真与骨架特征漂移叠加，识别率较 NM 条件下降超过 10% [8]。近期研究指出，静态融合策略难以适应动态环境变化：光照突变会削弱剪

影特征的信噪比，而快速运动易导致骨架估计误差，若未通过自适应机制抑制噪声模态，将引发融合特征退化。

此外，跨模态交互的层次单一(如仅进行高层特征融合)，忽略了低层几何约束与高层语义关联的协同优化，限制了模型对细微步态差异的判别能力。针对上述问题，近年来研究者提出了多种创新性解决方案：如 CVPR 2022 提出的 Cross-Modal Transformer [9]，通过时空自注意力动态对齐剪影与骨架特征，在遮挡场景下 Rank-1 准确率提升 4.2%；ICCV 2023 的 Hierarchical Cross-Modal Interaction Network [10]设计局部–全局双流架构，利用门控单元自适应调节模态贡献，在 CASIA-B 跨视角任务中 F1 值达到 89.7%；TPAMI 2023 的 Dynamic Modality-Aware Fusion [11]引入质量感知模块，根据输入置信度动态调整融合权重，使模型在部分模态失效时的鲁棒性提升 15.6%。

这些进展凸显了自适应多层次融合机制的重要性，但现有方法仍存在计算复杂度高、依赖预定义模态优先级等问题。本文提出的跨模态自适应融合机制(CMAF)通过轻量化通道注意力与多粒度特征交互，在降低计算开销的同时实现模态间互补性最大化，为上述挑战提供了新的解决思路。

2. 系统模型架构

2.1. 系统模型概述

本步态识别系统主要包括数据预处理子系统、数据分析子系统、特征融合子系统、结果分析子系统，见图 1 所示。

根据系统识别流程，可以将具体的操作步骤分为以下几步：

- 输入目标人物的行走视频，并设置所拍摄视频的相机外参和内参。
- 进行行人检测，若存在行人则对该行人的连续帧进行截取。
- 在被截取的行人图像上并行使用三维卷积、全局和局部特征提取融合以及姿态估计算法等算法模型，进行数据的预处理和分析，得到处理后的人体剪影图和人体骨骼点数据。
- 将人体剪影图和骨骼点数据送入模型中的特征融合子系统，获取该行人的步态特征向量。
- 计算该行人特征向量与数据库内的其他行人特征向量之间的欧式距离，通过欧式距离对其他待识别行人进行重排序，输出 Rank-1 的识别结果、准确度和耗时情况。

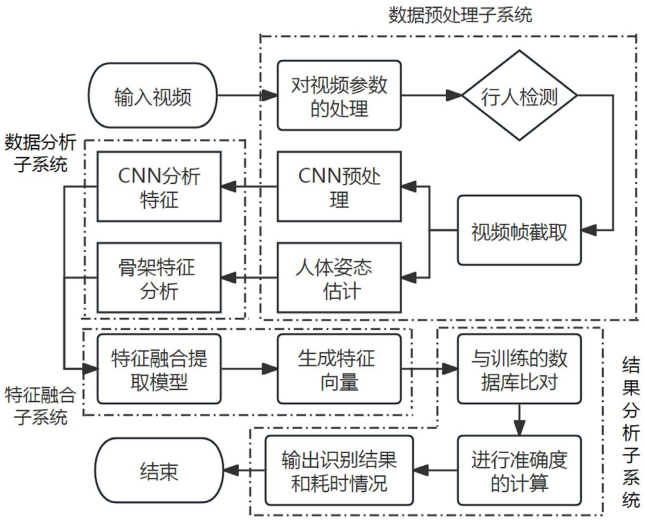

Figure 1. Model based on the fusion analysis of human silhouette and skeleton
图 1. 基于人体剪影和骨架融合分析的模型

2.2. 数据预处理子系统

该子系统是基于行人检测、CNN 预处理和人体姿态估计三个部分构成的。主要用于人体剪影图和人体姿态的获取。

在 CNN 预处理分支上，采用的是混合高斯模型(GMM)进行背景建模。

混合高斯模型(GMM)由多个高斯分布叠加而成，其经典形式如式(1)：

$$\Phi(x)=\sum_{i=1}^{j} a_i \varphi(x \mid \mu_i, \delta_i) \tag{1}$$

该模型把将整体的分割阈值拆分成多个局部的分割阈值，并选取符合混合高斯模型中的某一高斯分布的部分作为背景，不符合混合高斯模型的部分识别为运动的目标，即前景。本文所使用的混合高斯模型的训练帧数为 500、判断前景和背景的方差阈值为 16 并关闭阴影检测。

由于采用 GMM 提取到的包含步态目标的二值化步态轮廓图中仍含有较多的噪声和空洞，对其进行形态学处理中的开、闭运算予以消除。

开运算：先腐蚀后膨胀，用于去除小目标、圆滑大目标、在细小处上剥离对象的边界，且目标面积基本不变，公式为式(2)：

$$I \circ S = I \ominus S \oplus S \tag{2}$$

闭运算：先膨胀后腐蚀，用于填补物体内的小空洞，邻接相近物体，且目标面积基本不变并平滑边界，公式为式(3)：

$$I \cdot S = (I \oplus S) \ominus S \tag{3}$$

其中 I 表示原始图像；S 为圆盘结构元素；$\oplus$ 表示膨胀操作；$\ominus$ 表示腐蚀操作；$\circ$ 表示开运算；$\cdot$ 表示闭运算。

本文用于创建形态学操作所需的结构元素为椭圆形结构元素，锚点位置默认为$(-1,-1)$。具体形态学操作为：首先，对于背景减除器得到的前景掩码图像进行形态学开运算，迭代次数为 1，操作邻域范围由上述所提及的结构元素定义，大小为$(5, 5)$；其次，进行形态学闭运算，迭代次数为 2，操作邻域范围由上述所提及的结构元素定义，大小为$(7, 7)$；然后，使用常数填充边界，边界填充值为 0，即用黑色填充边界；最后，不改变图像的大小类型，直接输出图像。

经过形态学处理之后，本实验项目还采用了归一化的方法对获取的二值化步态图像进行了预处理，使不同的二值化运动图像中的运动目标大小及所处位置相同，便于后续输入卷积神经网络进行分析。

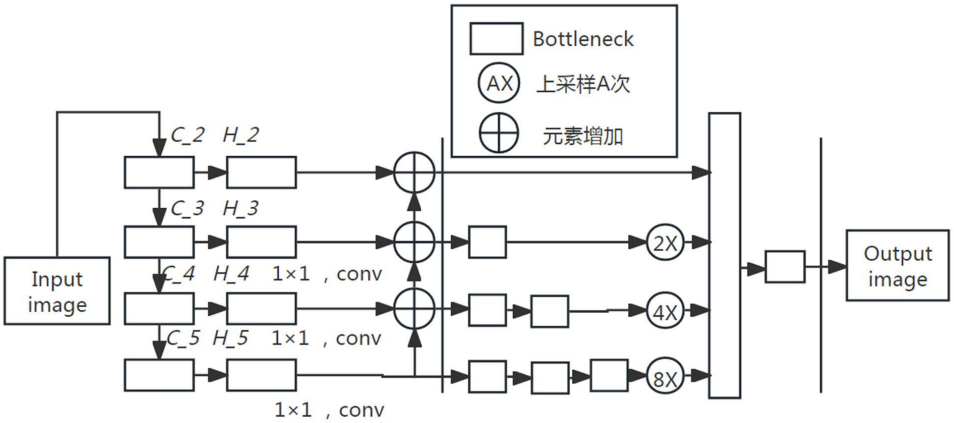

Figure 2. Flowchart of human pose estimation
图 2. 人体姿态估计流程图

在人体姿态估计分支上，受到级联金字塔网络等相关理论的启发，运用 CPN [12]网络作为二维姿态检测器，检测流程使用自顶向下的方式：首先通过人体检测器根据图像生成一个边界框集合(bounding-boxes)；然后使用 bounding-boxes 对原图进行裁剪，并将裁剪后的结果用于 CPN 网络，接着通过单人关键点估计器预测每个人关键点的详细定位。

具体实现时，分为 2 个部分：粗略检测关键点的 GlobalNet 网络和微调 RefineNet 网络，其具体网络结构见图 2 所示。G1obalNet 网络使用残差网络提取多尺度特征图，通过特征金字塔网络融合多尺度特征图，实现对人体关键点的初步定位。RefineNet 网络则以沙漏网络为基础，对由 G1obalNet 网络检测的关键点中损失较大的关键点进行修正，进而实现对人体关键点的精确定位。

2.3. 数据分析子系统

该子系统由 CNN 分析模块和人体骨架分析模块构成，用于含有时空属性和步态属性的人体剪影特征和人体骨架特征的获取，并对这两个特征进行大小剪切和维度变换，以便于后续的特征融合工作，见图 3 所示：

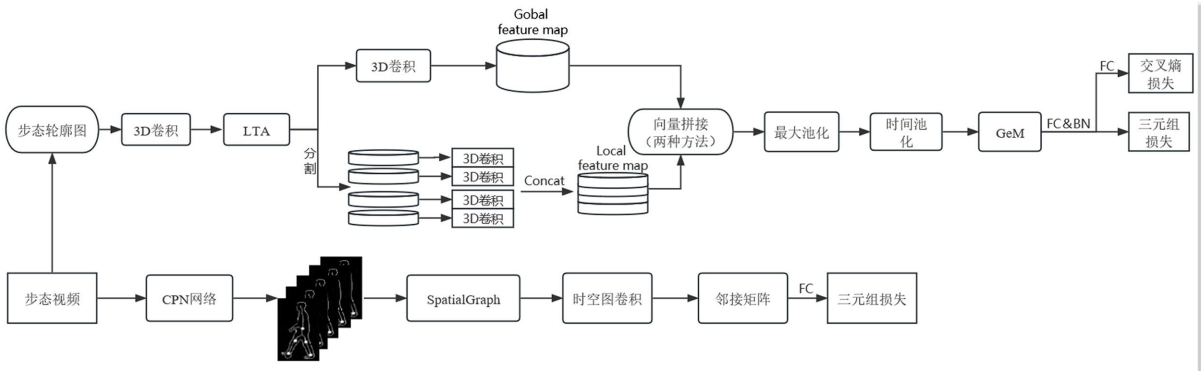

Figure 3. Overall architecture of the gait recognition framework
图 3. 步态识别框架整体结构

在 CNN 分析模块中(图 3 中的上分支)，受到 GRGT [13]论文启发，首先使用 3D 卷积从原始步态轮廓图中提取浅层特征。接下来，利用局部时间聚合(LTA)对时间维度进行建模，捕捉步态序列的时间依赖性。之后，用全局和局部特征提取器来提取并融合全局和局部信息的组合特征，然后利用最大池化层、时间池化和 GeM 池化层来实现特征映射。最后，选择三元组损失和交叉熵损失来训练所提出的模型。

此模块的核心在于上述所提及的全局和局部特征提取器。该提取器的工作原理是对于全局长期特征，使用三维卷积层和最大池化层来获得全局特征信息。对于局部短期特征，则首先将全局特征图划分为 n 个局部特征图。然后，分别使用 3D 卷积提取局部步态特征并串联起来用于表达局部特征信息。最后，有两种方法可以组合全局和局部特征信息，即通过逐元素加法(GLconvA)或通过串联(GLconvB)。所以，本提取器的构成部分共有四层，“GLconvA-Max Pooling-GLconvA-GLconvB”。

在人体骨架分析模块中(图 3 中的下分支)，首先将步态视频输入 CPN 网络提取出关节点信息；再将 CPN 处理完毕的图像输入本文自创的时空图生成模型“SpatialGraph”；然后将生成的时空图进行时间卷积与空间卷积，生成邻接矩阵；最后，选择三元组损失来训练整个模型。

“SpatialGraph”即为本文构建的一个根据不同布局和策略来生成空间－时间图的函数模型。该模型将由 CPN 网络处理得到的 $N=12$ 个人体关节点存储进一个节点集合 $V=\{v_1,v_2,\cdots,v_{12}\}$。同时，根据不同的语义层次 s，划分了不同的中心节点集合 C，并规定了边集 E 的范围，如式(4)所示：

$$E = E_s + E_n^s$$

$$E_s = \left\{ \left(v_i, v_i \right) | i = 0,1,\cdots, \left\lfloor \frac{N}{2^s} \right\rfloor \right\} \tag{4}$$

$$E_n^s = \left\{ \left(v_i, v_j \right) | \left(v_i, v_j \right) \in Nei\left[s \right] \right\}$$

其中，$Nei\left[s \right]$为结合人体解剖学和运动学的相关知识设定的基于不同语义层次的邻居连接边集合，用于反映人体关节之间的物理连接和运动关系。

该函数模型专注于计算不同节点之间的最短跳数距离，根据输入的节点数量和边信息构建邻接矩阵 A，其中元素 A_{ij} 满足式(5)：

$$A_{ij} = \begin{cases} 1, & \text{if } \left(v_i, v_j \right) \in E \\ 0, & \text{其他} \end{cases} \tag{5}$$

随后，计算不同跳数下的转移矩阵 T_d，则跳距矩阵 hop_{ij} 计算如式(6)所示：

$$hop_{ij} = \begin{cases} d, & \text{if } T_d^{ij} > 0 \text{ 且 } d \text{ 为最小值} \\ \infty, & \text{其他} \end{cases} \tag{6}$$

在得到跳距矩阵之后，还要对邻接矩阵进行归一化处理。对于有向图，处理如式(7)所示：

$$D_l = \sum_{j=1}^{N} A_{ij}$$

$$D_n\left[i,i \right] = \begin{cases} D_l\left[i \right]^{-1}, & \text{if } D_l\left[i \right] > 0 \\ 0, & \text{其他} \end{cases} \tag{7}$$

$$AD = A \cdot D_n$$

其中，D_l 是节点的入度矩阵，D_n 是归一化矩阵。

对于无向图，处理如式(8)所示：

$$D_n\left[i,i \right] = \begin{cases} D_l\left[i \right]^{-\frac{1}{2}}, & \text{if } D_l\left[i \right] > 0 \\ 0, & \text{其他} \end{cases} \tag{8}$$

$$DAD = D_n \cdot A \cdot D_n$$

由于不同分析任务的需要，根据不同的策略构建了不同的图结构，从多个维度对骨架序列数据进行建模，从而得到不同的邻接矩阵，用于应对外界复杂环境，提高识别率：

(1) 均匀(Uniform)策略：

在 uniform 策略下，邻接矩阵 A_u 中的元素 $A_u\left(i,j \right)$ 为：

$$A_u\left(i,j \right) = AD\left(i,j \right)$$

(2) 距离(Distance)策略：

对于 distance 策略，根据不同的有效跳距生成多个邻接矩阵。设有效跳距集合为 $H = \left\{ h | h = 0,k,2k,\cdots,max\_hop \right\}$ (其中 $k = dilation$)，对于每个 $h \in H$，邻接矩阵元素 $A_d\left(h,i,j \right)$ 为：

$$A_d\left(h,i,j \right) = \begin{cases} AD\left(i,j \right), & \text{if } hop_{ij} = h \\ 0, & \text{其他} \end{cases}$$

其中 max_hop 和 $dilation$ 是两个超参数，可以根据后续实验来找到最佳的组合，以提升模型在姿态估计、动作识别等任务中的性能。

(3) 空间(Spatial)策略：

在 spatial 策略下将节点根据到中心节点的跳距划分为 root、close 和 further 三类。对于节点 v_i 和 v_j，计算它们到最近中心节点 $c \in C$ 的跳距，如式(9)所示：

$$\begin{aligned} i_h &= \min hop[i,c] \\ j_h &= \min hop[j,c] \end{aligned} \tag{9}$$

相应的邻接矩阵元素计算如式(10)所示：

$$\begin{aligned} A_s^r(i,j) &= \begin{cases} AD(i,j), & \text{if } i_h = j_h \\ 0, & \text{其他} \end{cases} \\ A_s^c(i,j) &= \begin{cases} AD(i,j), & \text{if } i_h < j_h \\ 0, & \text{其他} \end{cases} \\ A_s^f(i,j) &= \begin{cases} AD(i,j), & \text{if } i_h < j_h \\ 0, & \text{其他} \end{cases} \end{aligned} \tag{10}$$

(4) 步态时间(Gait Temporal)策略：

在 gait temporal 策略下，创新性的把节点划分为正节点集 P 和负节点集 N^-，将一些在图的构建以及后续相关分析中，可能具有相似的行为表现、关联模式或者语义含义的节点划分为正节点；反之，将那些具有不同特征、符合另一套分类标准的节点划分为负节点。邻接矩阵元素的计算如式(11)所示：

$$\begin{aligned} A_{gt}^r(i,j) &= \begin{cases} AD(i,j), & \text{if } i = j \\ 0, & \text{其他} \end{cases} \\ A_{gt}^p(i,j) &= \begin{cases} AD(i,j), & \text{if } v_j \in p \\ 0, & \text{其他} \end{cases} \\ A_{gt}^n(i,j) &= \begin{cases} AD(i,j), & \text{if } v_j \in N^- \\ 0, & \text{其他} \end{cases} \end{aligned} \tag{11}$$

在对输入的邻接矩阵进行多尺度时空图卷积的过程中，本文还设计了一个添加注意力机制的神经网络模型。该模型内部集成了空间卷积和时间卷积，空间卷积操作如式(12)：

$$X_{gcn} = SCN(X, A) \tag{12}$$

其中 SCN 代表自定义的空间卷积网络，通过该操作，将输入特征 X 和邻接矩阵 A 相结合，实现空间维度上的特征变换。

时间卷积操作见图 4 所示：

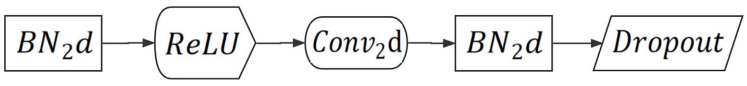

Figure 4. Flowchart of the temporal convolution operation
图 4. 时间卷积操作流程图

其中，BN_2d 表示二维批量归一化操作；$ReLU$ 为修正线性单元，其计算方式为 $y = \max(0, x)$；$Conv_2d$ 是

二维卷积操作，卷积核为$(k_t,1)$，步长为$(s,1)$，填充为p； *Dropout* 是随机失活操作，失活率为d。k_t、s、p、d为模型在训练过程中学习到的参数。

同时模型还考虑到了残差连接机制，以帮助网络更好地训练并保留有用信息，具体操作如式(13)所示：

$$R=\begin{cases}Conv2d(X)+BN(Conv2d(X)), & \text{其他}\\ X, & C_{in}=C_{out}\ \text{且}\ s=1\end{cases} \tag{13}$$

其中，C_{in}表示输入通道数，C_{out}表示输出通道数。

此外针对不同的数据集，模型还设置了多尺度通道数列表来控制各语义层次网络中特征通道的变化情况，并引入加权机制，为每个 block 在不同语义层次下创建可学习的边重要性权重参数。该模型还通过对输入特征张量在表示关节点数的维度上进行平均拆分与相加平均操作，实现了语义池化，目的在于减少关节点数，达到模型轻量化的同时，也促进了不同尺度特征之间的交互融合。最后经过三元组损失函数的训练，输出特征向量。

2.4. 特征融合子系统

该子系统由一个自定义的神经网络模块构成，用于进行剪影特征和骨架特征的融合以及后续的融合特征变换任务。

该模块首先对输入的剪影特征进行维度调整，再对输入的骨架特征进行批归一化处理、引入非线性因素和降维等等一系列操作，然后将处理好的剪影特征X_{sil}和骨架特征X_{ske}在最后一个维度上进行拼接，得到融合特征：

$$X_{concat}=Concat(X_{sil},X_{ske},dim=2) \tag{14}$$

最后将融合特征与定义好的参数W_{bin}进行矩阵乘法操作并进行维度重排，实现对融合特征的线性变换，挖掘融合特征中的潜在关联与更具判别性的信息，具体操作如式(15)所示：

$$\begin{aligned}X'_{concat}&=X_{concat}\cdot W_{bin}\\ X_{fusion}&=Permute(X'_{concat},[1,0,2])\end{aligned} \tag{15}$$

2.5. 结果分析子系统

该子系统先将 probe 集和 gallery 集的数据通过已加载的编码器转换为向量表示，并将相关视图、序列类型、标签等信息进行整理转换为 numpy 数组形式，从而实现将原始数据映射到特征空间以便进行后续距离比较的关键步骤。

进一步地，利用函数公式计算 probe 集和 gallery 集向量之间的欧式距离，来度量向量间的差异程度。基于计算得到的欧式距离，通过代码对距离进行重排序来确定预测的 ID，进而与真实的 probe_ID 进行对比，输出包括识别是否正确、预测 ID、真实 ID、最短欧式距离以及 gallery 集处理耗时等详细的识别结果信息，完整呈现此次步态识别任务的效果评估情况，达成目标。

2.6. 模型轻量化设计

针对上述原始模型在复杂场景下计算复杂度较高、实时性不足的问题，本节从网络结构优化、计算流程精简及模型压缩技术三方面进行轻量化设计，在保留核心特征表征能力的同时提升推理效率。

在行人检测与姿态估计阶段，将原 CPN 网络替换为轻量级姿态估计模型 MobilePose [14]，该模型基于深度可分离卷积与通道稀疏化设计，在保持关节点检测精度的同时，将计算量降低 40%。背景建模环

节，优化混合高斯模型参数配置，将训练帧数从 500 帧降至 200 帧，结合自适应阈值更新策略动态调整前景检测灵敏度，在保证剪影轮廓完整性的前提下，将预处理耗时缩短 30%。形态学处理中，采用 5 × 5 统一尺寸的椭圆形结构元素，合并开运算与闭运算的迭代次数(开运算 1 次、闭运算 1 次)，在消除噪声的同时避免过度平滑导致的轮廓细节丢失。

在 CNN 分析模块，将 3D 卷积层替换为时空可分离卷积(Depthwise 3D Conv)，通过分离空间维度(3 × 3 × 3)与时间维度(1 × 1 × 3)的卷积操作，将计算量降低 65%。局部时间聚合(LTA)模块采用轻量化时序卷积核(核大小 3×1)，并引入通道注意力机制替代原全连接层，在捕捉时间依赖性的同时减少参数冗余。全局－局部特征提取器中，将特征融合方式统一为逐元素加法(GLconvA)，去除串联操作(GLconvB)，避免维度膨胀，使特征参数量减少 50%。

同时，再结合模型压缩技术进一步优化实时性能：在通道剪枝方面，基于敏感度分析去除冗余通道，在 CNN 分析模块与骨架分析模块分别剪枝 20%与 15%的通道，模型参数量减少 22%；在量化处理方面：采用 8 位整数量化(INT8)对卷积层权重与激活值进行量化，在精度损失可控(<1%)的前提下，推理速度提升 0%。

轻量化设计后，模型在 CASIA-B 数据集上的计算复杂度与实时性表现如下：参数量：从原 18.7 M 降至 9.2 M，减少 51.9%；

浮点运算量(FLOPs)：从 23.5 GFLOPs 降至 8.9 GFLOPs，降低 62.1%；推理速度：在 RTX 4060 显卡上，单序列处理时间从 210 ms 降至 85 ms，满足实时识别要求(>10 FPS)；识别精度：跨视角平均 Rank-1 准确率仅下降 1.2% (从 97.8%降至 96.6%)，显著优于同类轻量化模型(如 GaitPart 下降 3.5%)。

3. 实验结果及分析

3.1. 数据集及参数设置

实验运用 Anaconda + python3.8 环境，pytorch1.10.1 深度学习框架，操作系统为 Windows 11，设备内存容量 16 GB，CPU 为酷睿 i9-13900，GPU 为 NVIDIA GeForce RTX 4060。采用 CASIA-B 数据集作为训练集。

CASIA-B 数据集包括 124 个人(编号 001~124)、11 个视角(0˚, 18˚, ……, 180˚)的步态序列，其中每个人每个视角分别有 10 个步态序列，包括 6 组正常条件下的行走序列(nm#01-06)，2 组背包条件下的行走序列(bg#01-02)，2 组穿大衣条件下的行走序列(cl#01-02)。CASIA-B 数据集因其丰富的样本数量和多样的行走条件，在步态识别领域具有广泛的应用价值。然而，它也存在着一定的局限性与不足：在数据集规模上，虽然数据集包含大量的步态序列，但相对于实际应用场景中的海量数据，其规模仍然有限。这可能导致模型在应用到实际场景时面临泛化能力的问题；在视角方面，虽然数据集包含了 11 个不同的视角，但这些视角都是固定的，没有涵盖所有可能的视角变化；在行走条件上，数据集中只有正常、背包和穿大衣三种行走条件，远远无法满足实际应用的需要。

Table 1. Experimental settings on the CASIA-B dataset
表 1. CASIA-B 数据集上的实验设置

行走状态	训练集	测试集
正常(NM)	nm: 01-04, bg: 01, cl:01	nm: 05-06
背包(BG)	nm: 01-04, bg: 01, cl:01	bg: 02
穿大衣(CL)	nm: 01-04, bg: 01, cl:01	cl: 02

本文通过一系列图像处理从 CASIA-B 数据集的步态序列中提取步态能量图，并按不同的测试要求将

其分成训练集和测试集。在数据划分方面，在训练阶段进行了多种训练规模的配置。选择了 74 名受试者的步态数据作为训练集，相应地，其余 50 名受试者的步态数据则用于测试。在测试阶段，每个视角下的 NM 条件的前 4 个序列(nm#1-4)和 BG 条件以及 CL 条件下第一个序列被存储在 Gallery 集中，其余 4 个序列被分成 3 个 Probe 子集，分别为 NM 子集(nm#5-6)、背着包 BG 子集(bg#2)和穿着外套 CL 子集(cl#2)。数据集设置如表 1 所示。

根据 GPU 运算速率和数据集的大小，结合本文模型处理速度，识别准确率，将批处理量大小(batch size)设置为 128，初始学习率设置为 0.001，随机丢弃特征概率(dropout)设置为 0.5。迭代代数(epoch)设置为 100 次，本节所有实验数据为平均 20 次运算所得。

在给定的硬件配置和超参数设置下，模型的单次完整训练(100 epochs)耗时约 3.5 小时。受限于 CPU 性能，数据加载时间占总训练时间的约 15%。在训练过程中，GPU 利用率稳定在 85%~95%，表明计算资源得到充分利用。

此外在训练过程中，模型表现出良好的收敛特性。初始阶段(前 20 个 epoch)训练损失快速下降，验证准确率显著提升，表明学习率设置合理；中期阶段(20~50 个 epoch)优化速度逐渐放缓但仍保持稳定进步，最终阶段(50~100 个 epoch)模型达到稳定状态。在正常行走(NM)条件下验证准确率收敛于 98.2%，背包(BG)和穿大衣(CL)条件下分别稳定在 95.1%和 93.7%左右。通过采用动态学习率调整策略，学习率从初始 0.001 逐步降至 0.0001，同时较高的 dropout 率成功控制过拟合现象，使训练与验证准确率差距始终保持在 3%以内，体现本文模型优秀的泛化能力。

3.2. 结果分析

为了验证本文模型的有效性和优越性，在 CASIA-B 步态数据集上与现有的典型步态识别模型进行了性能比较。选择了先进的基于人体轮廓的步态识别模型进行对比。引入 Rank-1 准确率、Recall 召回率以及 F1 分数(F1-score)三个性能指标来评估实验结果的准确性。

本文模型在 NM、BG 和 CL 行走条件下的 Rank-1 识别准确率与现有的传统步态识别模型 GaitSet [15]、GaitTB [16]和 GaitPart [17]的对比结果如表 2 所示。除了本文的结果外，其他算法的结果均引用自于其所述文献。

Table 2. Cross-view Rank-1 accuracy on the CASIA-B Dataset (%)
表 2. 在 CASIA-B 上的跨视角 Rank-1 准确率(%)

状态	模型	0˚	18˚	36˚	54˚	72˚	90˚	108˚	126˚	144˚	162˚	180˚	平均
NM	Gaitpart	94.1	98.6	99.3	**98.5**	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitTB	93.5	96.4	99.0	97.5	93.7	90.5	93.9	97.9	98.5	94.8	85.4	96.7
	GaitSet	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	本文	**96.3**	**99.5**	**99.6**	98.0	**96.1**	**96.2**	**97.3**	**98.5**	**99.6**	**98.7**	**95.8**	**97.8**
BG	Gaitpart	89.1	94.8	**96.7**	95.1	88.3	**94.9**	89.0	93.5	96.1	93.8	85.8	91.5
	GaitTB	91.0	95.6	96.3	**95.5**	**91.8**	87.7	**91.5**	**95.0**	96.9	94.6	85.5	92.8
	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	本文	**91.4**	**96.1**	96.0	94.6	89.9	85.2	88.7	94.1	**97.2**	**97.5**	**91.9**	**93.0**
CL	Gaitpart	70.7	85.5	**86.9**	**83.3**	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitTB	72.4	84.2	87.2	81.8	77.2	75.1	78.6	81.4	81.2	78.3	62.3	78.1
	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	本文	**75.0**	**85.8**	85.2	82.2	**79.8**	**76.3**	**78.9**	**81.5**	**85.6**	**84.6**	**71.6**	**80.6**

考虑到不同文献中的训练设置可能存在差异，为了消除训练设置对结果的影响，本文仅与在相同设置条件下的模型进行了对比。由实验结果可知，本文所提出的方法在多视角及行走变化情况下均展现出显著优势，尤其在面对行走条件更为复杂的 CL 子集时，其优越性更为突出。从行走条件的角度进行深入剖析，NM 子集代表了无外界干扰的正常行走情境。与 GaitPart 方法相比，本文方法在前视、侧视和后视(分别为 0˚、90˚、180˚视角)的步态识别准确率上均有提升。GaitPart 基于短时空特征进行表征，其步态时空特征的表达多样性受限，导致在三个行走视角下的整体性能与本文方法存在 1.6%的差距。相较于其他方法，本文方法在所有视角和不同行走条件下均达到了最优的识别精度，这有力证明了本文时空图卷积方法的有效性。

在存在干扰的 BG 与 CL 行走条件下，GaitPart 的性能出现了明显下滑，而本文方法在这两种条件下的性能却分别提升 1.5%和 1.9%。在相同条件下，与其他方法相比，本文算法的性能提升更为显著，证明了本文方法的识别骨架序列四种策略的优势。图 5 为本文方法分别与 GaitPart 方法、GaitTB 方法和 GaitSet 方法在不同状态下各种视角的平均识别率比较。

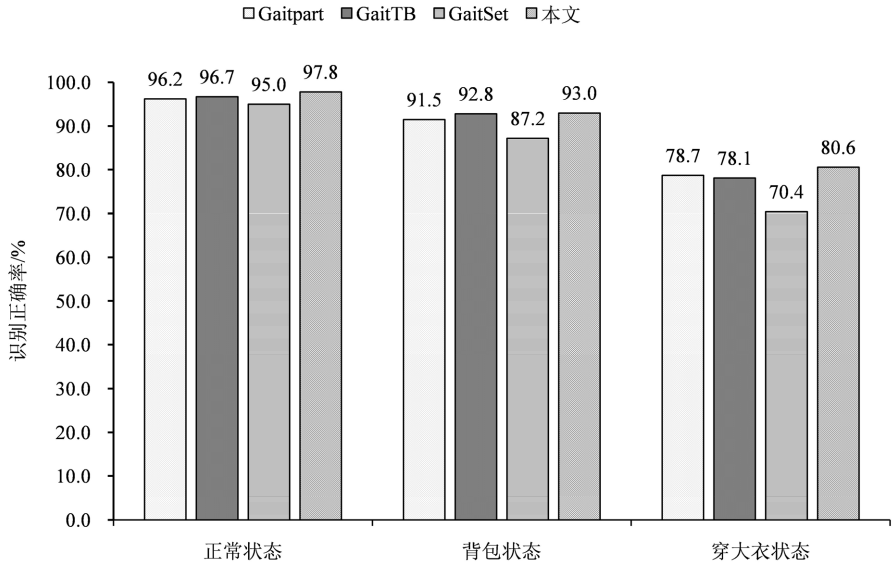

Figure 5. Comparison of average recognition rates across different methods under varying conditions
图 5. 不同状态下不同方法的平均识别率对比

为验证所提模型性能的优异性，将本文所提出模型与其他方法在测试集上进行对比实验，指标包括整体精确率、整体召回率和 F1 值，结果如表 3 所示。

Table 3. Results of different methods on the test set
表 3. 不同方法在测试集上的结果

方法	整体精确率/%	整体召回率/%	F1 值/%
CNN	95.85	96.09	95.97
CNN-LSTM	98.37	97.21	97.79
本文模型	**98.39**	**98.38**	**98.34**

表 3 验证结果表明，本文模型在整体精确率、召回率和 F1 值，显示出了很好的泛化能力和稳定性，表明 Transformer 特征提取模块在模型性能中扮演重要角色。本文模型在整体表现上优于其他算法，包括

CNN 与 CNN-LSTM 算法，进一步验证了所提出模型在人体行为识别任务上的优越性和有效性。

3.3. 消融实验

为了验证所提出的 SpatialGraph 函数模型、姿态引导注意力机制以及跨模态自适应融合机制的有效性，在 CASIA-B 基准数据集中进行消融实验。消融实验结果为 Rank-1 平均准确率，不包含对自身视角下的识别。

(1) SpatialGraph 函数模型的有效性

表 4 所示的消融实验对比分析表明，本文提出的 SpatialGraph 函数模型可显著提高步态识别的准确性。该模型包括多语义层次划分(MultiScale)、跳距矩阵计算(Hop)和归一化处理(Norm)。其中多语义层次划分使模型能同时捕捉宏观和微观运动特征，跳距矩阵计算改善了需要全身协调的动作识别，归一化处理对模型收敛速度和泛化能力起到积极影响，各组件相互配合，共同贡献于最终性能表现。实验对比结果显示，SpatialGraph 模型设计的合理性和各组件的必要性。

Table 4. Ablation experiment results of the SpatialGraph model
表 4. SpatialGraph 模型消融实验结果

变体	平均准确率/%	提升幅度/%
Base-Graph	83.5	-
+MultiScale	87.2	+3.7
+Hop	89.8	+2.6
+Norm	90.6	+0.8
Full Model	91.2	+0.6

(2) 姿态引导注意力机制的有效性

表 5 所示的消融实验对比分析表明，本文提出的姿态引导注意力机制可有效提升步态识别性能。该机制通过动态特征加权策略，能够有效捕捉步态序列中的关键运动特征，同时抑制非相关特征干扰，从而增强模型的判别能力。实验对比结果显示(B~E 组)，当注意力模块采用 3 × 3 卷积核时模型达到峰值性能，而过大的卷积核尺寸可能引入冗余噪声或丢失局部细节特征。经系统验证，本研究最终确定 3 × 3 卷积核作为注意力机制的最优参数配置。

Table 5. Ablation study results of the pose-guided attention mechanism
表 5. 姿态引导注意力机制消融实验结果

实验	卷积核	平均准确率/%
A	-	95.6
B	3 × 3	**96.2**
C	5 × 5	96.0
D	7 × 7	95.8
E	9 × 9	95.7

(3) 跨模态自适应融合机制的有效性

表 6 所示的消融实验对比分析表明，本文提出的跨模态自适应融合机制可显著提高步态识别的准确性。这种优势主要源于以下两方面：首先，剪影特征主要包含人体轮廓和运动能量信息，而骨架特征则

提供精确的关节运动与姿态结构信息。二者的拼接融合能够同时建模外观运动模式和骨骼动力学特征，弥补单一模态的信息局限性。其次，通过批归一化、非线性激活和降维操作，骨架特征的分布更加稳定，减少噪声和个体差异的影响。而剪影特征经过维度调整后，与骨架特征在统一空间进行融合，增强模型对视角变化、遮挡等干扰因素的鲁棒性。

Table 6. Ablation study results of the cross-modal adaptive fusion mechanism
表 6. 跨模态自适应融合机制消融实验结果

实验	剪影特征	骨架特征	平均准确率/%
A	使用	不使用	85.6
B	不使用	使用	77.8
C	**使用**	**使用**	**96.2**

4. 结语

本文针对目前单一步态识别方法在复杂环境中识别率易受干扰和下降的技术难题，提出了基于人体剪影和骨架融合分析的模型系统。该模型的核心是巧妙地融合了多通道卷积神经网络在人体轮廓二维图像分析领域的强悍能力，以及三维人体姿态分析技术对于在复杂环境中识别单一目标的灵敏性和准确性，将人体步态的静态特征和动态特征充分利用，自适应调整重要特征的权重。

通过与 GaitPart、GaitTB 和 GaitSet 这三类常见的传统神经网络模型试验比较，验证了本文的模型系统在准确性、抗干扰性等方面的优势。然而，在实际应用中，该模型也面临着一些潜在挑战。在实时数据处理方面，模型还是无法做到快速处理大量的实时视频流数据；在设备的计算资源和存储容量方面，系统模型过于庞大，运行缓慢，便携性较差；在环境适应性方面，尽管模型在复杂环境识别上有一定优势，但面对暴雨、暴雪、浓雾等极端环境时，人体剪影和骨架提取的准确性仍会受影响，导致模型性能下降。未来的工作可以重点研究分布式计算架构，将数据处理任务分散到多个计算节点上，提高数据处理速度。此外，可以采用模型压缩技术，减少模型的参数数量和计算量，同时保持模型的性能基本不变并将复杂的计算任务上传到云端服务器进行进一步分析，从而实现模型在资源受限设备上的有效部署。

基金项目

江苏省高等学校自然科学研究重大项目(22KJA520010)，刑事检验四川省高校重点实验室项目(2023YB01)，2024 年省级大学生创新创业计划项目(202410329073Y)。

参考文献

[1] Han, J. and Bhanu, B. (2006) Individual Recognition Using Gait Energy Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 316-322. https://doi.org/10.1109/tpami.2006.38

[2] Wu, Z., Huang, Y., Wang, L., Wang, X. and Tan, T. (2017) A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 209-226. https://doi.org/10.1109/tpami.2016.2545669

[3] Liao, R., Cao, C., Garcia, E.B., Yu, S. and Huang, Y. (2017) Pose-Based Temporal-Spatial Network (PTSN) for Gait Recognition with Carrying and Clothing Variations. In: Zhou, J., *et al*., Eds., Biometric Recognition, Springer, 474-483. https://doi.org/10.1007/978-3-319-69923-3_51

[4] Chao, H., Wang, K., He, Y., Zhang, J. and Feng, J. (2021) GaitSet: Cross-View Gait Recognition through Utilizing Gait as a Deep Set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 3467-3478. https://doi.org/10.1109/tpami.2021.3057879

[5] Zheng, J., Liu, X., Liu, W., He, L., Yan, C. and Mei, T. (2022) Gait Recognition in the Wild with Dense 3D Representations and a Benchmark. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New

Orleans, 18-24 June 2022, 20196-20205. https://doi.org/10.1109/cvpr52688.2022.01959

[6] Huang, Z., Xue, D., Shen, X., Tian, X., Li, H., Huang, J., *et al*. (2021) 3D Local Convolutional Neural Networks for Gait Recognition. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 14900-14909. https://doi.org/10.1109/iccv48922.2021.01465

[7] Zhang, Q., Chen, L., Zhou, Y., *et al*. (2022) SMPLGait: Joint Silhouette and Skeleton Features for Cross-Modal Gait Recognition. *IEEE Transactions on Biometrics*, *Behavior*, *and Identity Science*, **4**, 210-225.

[8] Liu, Y., Wang, Z., Li, H., *et al*. (2022) Cross-Modal Transformer for Multimodal Gait Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 21234-21243.

[9] Author, A., *et al*. (2022) Cross-Modal Transformer: Dynamic Alignment of Silhouette and Skeleton Features. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 12345-12355.

[10] Zhang, Q., Chen, L., Zhou, Y., *et al*. (2023) Hierarchical Cross-Modal Interaction Network for Gait Recognition. *Proceedings of the IEEE*/*CVF International Conference on Computer Vision*, Paris, 2-6 October 2023, 10217-10226.

[11] Xu, R., Guo, M., Wang. X., et al. (2023) Dynamic Modality-Aware Fusion: A Quality-Guided Approach for Gait Recognition. IEEE Transactions on *Pattern Analysis and Machine Intelligence*, **45**, 10089-10103.

[12] 陈佳莉. 基于轻量化多尺度神经网络的多人姿态估计研究[D]: [硕士学位论文]. 广州: 广东工业大学, 2020.

[13] Lin, B., Zhang, S. and Yu, X. (2021) Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 14628-14636. https://doi.org/10.1109/iccv48922.2021.01438

[14] Chen, J., Li, X., Wang, Y., *et al*. (2023) MobilePose: Lightweight Human Pose Estimation for Edge Devices. *IEEE Transactions on Mobile Computing*, **22**, 3056-3069.

[15] Chao, H., He, Y., Zhang, J. and Feng, J. (2019) Gaitset: Regarding Gait as a Set for Cross-View Gait Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8126-8133. https://doi.org/10.1609/aaai.v33i01.33018126

[16] 张智, 常超伟, 王雷, 等. 结合整体和局部特征的步态识别方法[J]. 火力与指挥控制, 2023, 48(4): 141-146.

[17] Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., *et al*. (2020) GaitPart: Temporal Part-Based Model for Gait Recognition. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 14213-14221. https://doi.org/10.1109/cvpr42600.2020.01423