文章编号:1007－6735(2024)06－0589－11　DOI: 10.13255/j.cnki.jusst.20241016002

融合多尺度特征表示和注意力机制的步态识别模型

曹子康[1]，　裴颂文[1]，　黄立波[2]
（1. 上海理工大学 光电信息与计算机工程学院，上海 200093；2. 国防科技大学 计算机学院，长沙 410000）

摘要：针对步态识别模型在特征表示粒度和时空依赖建模的不足，提出了一种融合多尺度特征表示和注意力机制的步态识别模型。该模型包含两个关键模块：多尺度特征融合网络（multi-scale features fusion network，MFFN）和步态注意力融合模块（gait attention fusion module，GAFM）。其中，MFFN 通过多尺度、多粒度特征融合提高特征表示的丰富性和判别力；GAFM 通过自适应地关注步态序列中的关键帧和重要区域，从而有效地建模长期时空依赖关系。在 3 个数据集 CASIA-B，CASIA-B*和 OUMVLP 上的实验结果表明，该模型在多种复杂条件下均优于现有模型，相较于基准模型，平均识别率分别提升了 0.9%，0.3% 和 0.6%。

关键词：步态识别；多尺度特征；注意力机制；时空依赖；特征融合
中图分类号：TP 181　　**文献标志码：**A

A gait recognition model fusing multi-scale feature representation and attention mechanism

CAO Zikang[1]，　PEI Songwen[1]，　HUANG Libo[2]
(1. *School of Optical−Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *College of Computers, National University of Defense Technology, Changsha 410000*)

Abstract: To address limitations in gait recognition model regarding feature representation granularity and spatio-temporal dependency modeling, a novel model fusing multi-scale feature representation and attention mechanisms was proposed. The model consists of two key modules: multi-scale features fusion network (MFFN) structure and gait attention fusion module (GAFM). MFFN enhances the richness and discriminative power of feature representation through multi-scale and multi-granular feature fusion.

收稿日期: 2024−10−16
基金项目: 国家自然科学基金资助项目 (61975124)；四川省重点研发项目（2024YFFK0443）；上海市自然科学基金资助项目 (20ZR1438500)
第一作者: 曹子康（1999−），男，硕士研究生．研究方向：步态识别．E-mail：223330920@st.usst.edu.cn
通信作者: 裴颂文（1981−），男，教授．研究方向：智能计算、深度学习．E-mail：swpei@usst.edu.cn

引文格式: 曹子康，裴颂文，黄立波．融合多尺度特征表示和注意力机制的步态识别模型[J]．上海理工大学学报，2024，46(6)：589-599.
Citation: CAO Zikang, PEI Songwen, HUANG Libo. A gait recognition model fusing multi-scale feature representation and attention mechanism[J]. Journal of University of Shanghai for Science and Technology, 2024, 46(6): 589-599.

GAFM effectively modeled long-term spatio-temporal dependencies by adaptively focusing on key frames and important regions in gait sequences. Experimental results on CASIA-B, CASIA-B*, and OUMVLP datasets show that the model outperforms existing models under various complex conditions, with average recognition rate improved by 0.9%, 0.3% and 0.6% respectively compared to the baseline model.

Keywords: *gait recognition*; *multi-scale features*; *attention mechanism*; *spatio-temporal dependency*; *feature fusion*

生物特征识别技术是指利用人体的生理或行为特征进行识别的技术，与传统身份识别相比，其具有更高的安全性，这是因为生物特征难以被复制、盗用或遗忘。常见的生物特征包括指纹、虹膜、人脸、声纹、笔迹等。

步态识别是一种通过人体步行的特征进行身份识别的技术[1]。与其他生物特征识别方法相比，步态识别具有以下独特的优势：非接触性、隐蔽性、难以伪造性、鲁棒性等。其中：非接触性意味着步态识别可以远距离进行[2]，不需要与被识别对象直接接触；隐蔽性是指步态识别可以在不引起被识别对象注意的情况下进行；难以伪造性是因为步态是一种人体动态的特征，受个人身体结构、肌肉力量、神经控制等因素影响，每个人都有一种独属于自己的步态模式[3]。并且，步态识别对环境光照和服装变化也同时具有一定的鲁棒性，即使在户外场景中也能取得较好的识别效果。这些优势使步态识别适用于公共安全应用，例如刑事调查、嫌疑人追踪[4]和身份验证[5]。

尽管步态识别技术具有诸多优势，但现有方法仍面临两个关键挑战：首先是步态特征表示的粒度问题，现有方法往往依赖于单一尺度的特征提取，或是将全局特征和局部特征分开处理。这种处理方式难以全面捕捉步态的多尺度信息，导致一些细微但对身份识别至关重要的步态特征被忽略。例如，在处理携带物品或穿着外套等复杂场景时，由于遮挡和变形的影响，单一尺度的特征表示往往无法准确描述人体运动特征；其次是步态序列的时空依赖建模的问题，步态是一个时序运动的过程，其中包含时间和空间的依赖关系，现有方法在建模这些依赖关系时存在不足：一方面是对于长时序的建模能力有限，难以捕捉跨越多个步态周期的长期依赖关系，而另一方面，现有方法对重要性不同的区域缺乏自适应的权重分配机制，这意味着模型无法根据不同场景动态调整关注重点，进而影响识别的准确率。这些问题在实际应用中尤为突出，直接影响识别性能的稳定性和鲁棒性。为了解决上述问题，本文旨在提升步态特征的多尺度建模能力和时空依赖关系的建模效果。通过实现对不同粒度的步态信息进行有效融合，提高特征表示的丰富性和判别力，同时使用自适应的注意力机制捕捉长期时空依赖，增强模型对复杂场景的适应能力。

针对上述目标，本文提出了一种融合多尺度特征表示和注意力机制的步态识别模型（multi-scale and attention gait recognition model，MSA-Gait）。具体而言，本文的主要贡献如下：

a. 提出并实现了改进的多尺度特征融合网络（multi-scale features fusion network，MFFN）。该网络在BNNeck[6]的基础上，通过多尺度、多粒度的特征融合机制实现局部细节与全局语义特征的自适应融合。并且特征融合有效地平衡了特征的判别性和多样性，显著提升了模型捕捉细微步态信息的能力。

b. 设计并实现了步态注意力融合模块（gait attention fusion module，GAFM）。该模块通过融合不同的注意力机制，实现对步态序列中关键帧和区域的自适应关注。并通过空洞卷积增强了长期时空依赖关系的建模能力，提高了模型在复杂场景下的鲁棒性。

1 相关工作

1.1 基于模型的步态识别

相比于其他方法，基于模型的方法在复杂场景下面临着多重挑战。Bouchrika等[7]通过特征提取和建模实现步态分析，但在低分辨率场景下性能显著下降。SMPL(skinned multi-person linear)模

型把人体模型作为一个参数化的线性模型，但该模型在处理复杂动作和快速运动时仍存在姿态估计不准确的问题[8]。PostGait 方法利用 3D 身体姿势和先验知识来克服服装变化的影响，但其复杂的人体结构建模带来了较大的计算开销[9]。而 GaitGraph 模型虽然采用图卷积网络简化了建模过程，但在遮挡情况下，关键点定位的准确性仍然受到严重影响[10]。HMRGait(human mesh recovery gait)通过微调预训练的 HMR 网络来构建基于端对端的 SMPL 模型，但当前用于姿态特征的识别网络忽略了关节之间的结构信息[11]。SMPLGait 方法通过 SMPL 模型提取 3D 信息来增强特征学习，但对输入图像质量的要求较高，限制了其实际应用场景[12]。GPGait(generalized pose-based gait) 方法提出人体导向的姿态变换和描述器来提升骨骼特征的跨数据集泛化能力，并通过部位感知图卷积网络挖掘局部与全局关系，但在单一数据集上的识别性能略低于先前方法[13]。BiFusion(bimodal fusion) 模型提出多尺度步态图网络来集成骨架和轮廓特征，但其在衣着变化场景下的骨架估计精度仍有待提高 [14]。

1.2　基于外观的步态识别

基于外观的方法虽然避免了显式的人体建模，但在特征表示和时序建模方面仍有待改进。GEI(gait energy image) 方法把整个视频序列统一成一个包含时空信息的单个帧，但对现实场景的变化较为敏感[15]。GaitSet 模型首次将步态序列视为集合并利用最大值函数压缩帧级特征，但忽略了帧与帧之间的时序关联[16]。GaitPart 模型设计了微运动捕捉模块以提取局部动态特征，但仅关注局部特征而忽视了全局语义信息[17]。GLN(gait lateral network) 模型引入集合池化提取序列特征，但特征表示的粒度较为单一[18]。GaitGL(gait global-local) 模型提出了一种 3DCNN 网络来同时聚合局部时空信息，但其在提取全局特征时计算量较大[19]。虽然 GaitTransformer 模型引入多时间尺度变换器来建模时序依赖，但其注意力机制设计相对简单，难以充分利用跨越多个步态周期的长期依赖信息[20]。GaitSSB(gait self-supervised benchmard) 方法通过对比学习显著提升了特征的泛化能力，但未考虑不同尺度特征的融合问题[21]。DeepGaitV2模型提出了基于深层 CNN 和 Transformer 的步态识别架构，但其也有着参数量较大的问题[22]。BigGait 方法利用大视觉模型的通用知识进行步态表示学习，但其生成的步态特征缺乏直观的物理含义，且特征的鲁棒性依赖于训练数据的分布[23]。GaitPoint+方法通过将骨骼关键点序列建模为 3D 点云并结合轮廓特征来实现步态识别，但期对复杂场景的泛化能力有限[24]。

综上所述可以发现，当前步态识别模型主要存在两个局限性：在特征表示方面，比较依赖于单一尺度的特征或单独处理全局特征或局部特征，因此缺乏有效的多尺度特征融合机制；在时空依赖建模方面，现有的注意力机制设计相对简单，难以有效建模长期依赖关系。基于上述分析，本文着重探索构建融合多尺度特征表示以及高效时空依赖建模的步态识别模型。

2　MSA-Gait 模型

GaitBase 模型[25] 采用了一个类似 ResNet[26] 作为主干，通过学习步态序列的时空特征，实现了高效、准确的步态识别，在多个标准数据集上取得了优异的性能，展现出了良好的特征提取和判别能力。因此， 相比于 GaitBase 模型，本文提出了改进优化的 MSA-Gait(multi-scale and attention gait recognition model) 模型。主要包括：a. 通过融合多尺度、多粒度特征，提出并实现了 MFFN 结构，增强了捕捉微小步态信息的能力；b. 利用注意力机制自适应地关注步态序列中的关键帧和区域，设计并实现了 GAFM 模块，可以有效建模长期时空的依赖关系。MSA-Gait 的模型框架，如图 1 所示。

2.1　MSA-Gait 模型

给定输入步态序列 $\boldsymbol{I}=i_1,i_2,\ldots,i_s\in\mathbb{R}^{s\times h\times w}$，其中：$s$为序列长度；$h$和$w$分别为图像的高度和宽度。首先对输入的序列进行通道扩展和维度重排，将其转化为五维张量$\boldsymbol{X}_{\text{in}}\in\mathbb{R}^{n\times c\times s\times h\times w}$，表示为

$$\boldsymbol{X}_{\text{in}}=\mathcal{T}(\boldsymbol{I}) \tag{1}$$

式中：$\mathcal{T}$表示维度变换操作；n 为批次大小；c 为通道数。

随后 MSA-Gait 对 $\boldsymbol{X}_{\text{in}}$进行特征提取、聚合与分类，具体包含以下 3 个阶段。

2.1.1　特征提取阶段

首先对输入的特征 $\boldsymbol{X}_{\text{in}}$进行维度转换 (reshape, Res)，将五维输入转换为四维特征$\boldsymbol{X}_{\text{res}}\in\mathbb{R}^{(n\times s)\times c\times h\times w}$，表示为

$$\boldsymbol{X}_{\text{res}}=\text{Res}(\boldsymbol{X}_{\text{in}}) \tag{2}$$

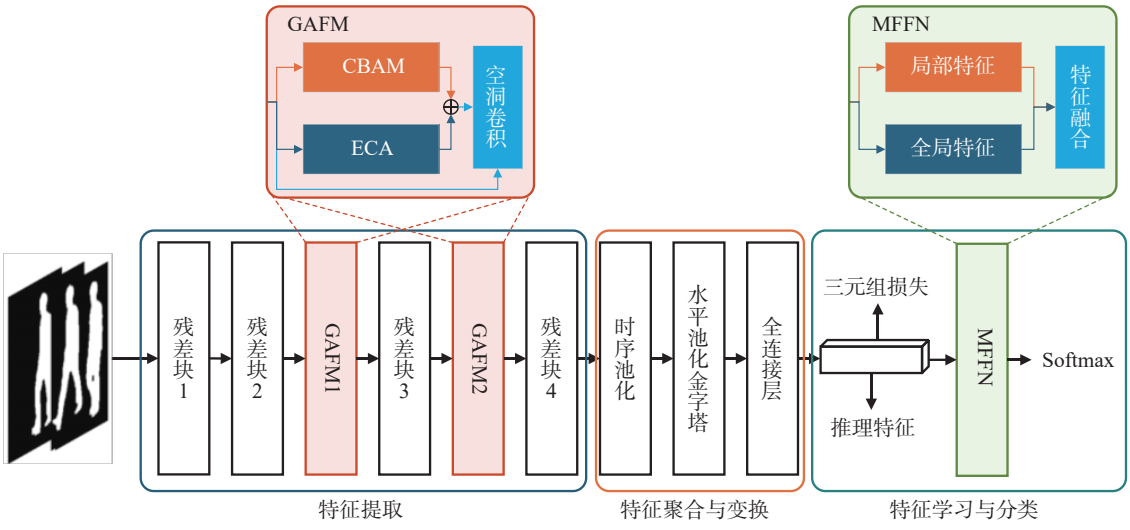

图 1　MSA-Gait 模型架构图

Fig.1　The architecture figure of MSA-Gait

通过残差块和 GAFM 模块对$\boldsymbol{X}_{\text{res}}$交替进行如下步态特征提取：

$$\boldsymbol{Y}_1 = \text{ResBlock}_1(\boldsymbol{X}_{\text{res}}) \tag{3}$$

$$\boldsymbol{Y}_2 = \text{ResBlock}_2(\boldsymbol{Y}_1) \tag{4}$$

$$\boldsymbol{Y}_3 = \text{GAFM}_1(\boldsymbol{Y}_2) \tag{5}$$

$$\boldsymbol{Y}_4 = \text{ResBlock}_3(\boldsymbol{Y}_3) \tag{6}$$

$$\boldsymbol{Y}_5 = \text{GAFM}_2(\boldsymbol{Y}_4) \tag{7}$$

$$\boldsymbol{Y}_6 = \text{ResBlock}_4(\boldsymbol{Y}_5) \tag{8}$$

随后对$\boldsymbol{Y}_6$进行维度转换，将四维特征转化为五维形式$\boldsymbol{Y}_{\text{res}} \in \mathbb{R}^{n\times c\times s\times h\times w}$以进行特征聚合，表示为

$$\boldsymbol{Y}_{\text{res}} = \text{Res}(\boldsymbol{Y}_6) \tag{9}$$

2.1.2　特征聚合与变换阶段

经过特征提取后的特征$\boldsymbol{Y}_{\text{res}}$首先在序列维度$s$上通过时序池化（temporal pooling，TP），将步态序列中所有帧的特征进行最大值聚合，得到一个统一的特征$\boldsymbol{Z}_1 \in \mathbb{R}^{n\times c\times h\times w}$，表示为

$$\boldsymbol{Z}_1 = \text{TP}(\boldsymbol{Y}_{\text{res}}) \tag{10}$$

随后通过水平池化金字塔（horizontal pooling pyramid，HPP）对特征图在水平方向上进行分段池化，提取不同尺度下的局部特征$\boldsymbol{Z}_2 \in \mathbb{R}^{n\times c\times p}$，其中$p$代表水平分段的总数，表示为

$$\boldsymbol{Z}_2 = \text{HPP}(\boldsymbol{Z}_1) \tag{11}$$

最后使用全连接层（fully connected layer，FC）对每个局部特征进行独立的特征映射，得到增强判别特征$\boldsymbol{Z}_3$，表示为

$$\boldsymbol{Z}_3 = \text{FC}(\boldsymbol{Z}_2) \tag{12}$$

2.1.3　特征学习与分类阶段

经过特征聚合与变化后的特征$\boldsymbol{Z}_3$通过 MFFN 模块进行多尺度特征融合，得到分类预测分数$\boldsymbol{L} \in \mathbb{R}^{n\times C\times p}$，其中$C$表示类别数，表示为

$$\boldsymbol{L} = \text{MFFN}(\boldsymbol{Z}_3) \tag{13}$$

随后预测分数$\boldsymbol{L}$应用 Softmax 函数，获得最终的类别概率分布$\boldsymbol{P}$，表示为

$$\boldsymbol{P} = \text{Softmax}(\boldsymbol{L}) \tag{14}$$

模型针对训练和推理阶段组织不同的输出形式。在训练阶段，输出包含用于度量学习的嵌入特征和用于分类学习的预测分数，其中嵌入特征由全连接层的输出$\boldsymbol{Z}_3$直接得到，预测分数则是由 MFFN 输出的$\boldsymbol{L}$。在推理阶段，模型仅输出用于步态识别的嵌入特征，最终通过优化三元组损失和分类损失的组合来指导模型的训练过程。

2.2　GAFM

在步态识别中，不同的注意力机制往往关注特征的不同方面。为了全面捕捉步态序列中的关键信息，本文提出了 GAFM 模块，通过融合互补的注意力机制来增强特征表示。如图 2 所示，该模块主要包含 ECA(efficient channel attention) [27] 和 CBAM(convolutional block attention module)[28] 两个分支，以及特征融合部分。在 MSA-Gait 的特征提取阶段中，GAFM 模块分别作用于不同层级的中间特征以增强其表示能力。

其中使用$\boldsymbol{U}$作为模块的局部输入符号，来描述 GAFM 的具体结构，给定输入特征图$\boldsymbol{U} \in$

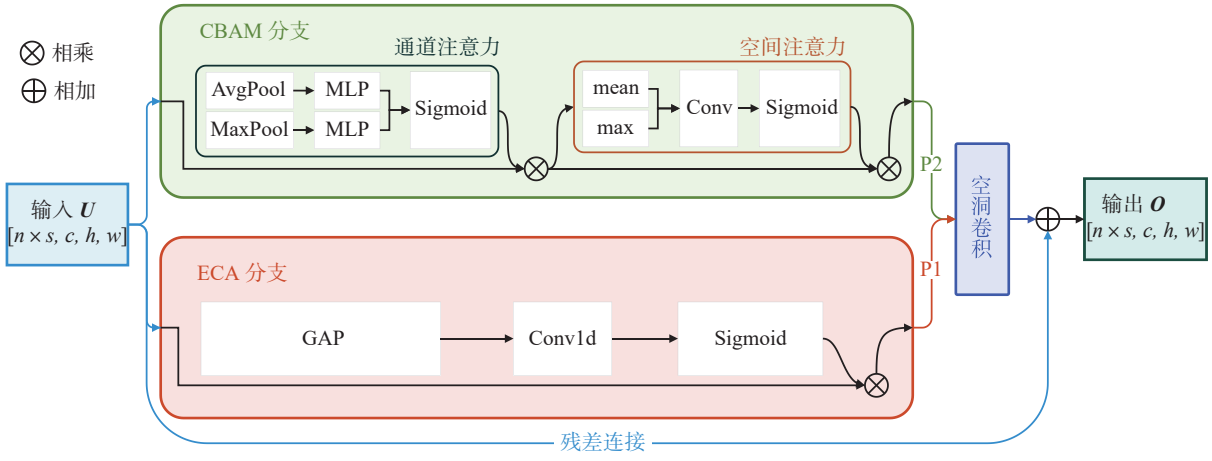

图 2　GAFM 结构图

Fig.2　The structure picture of GAFM

$\mathbb{R}^{(n\times s)\times c\times h\times w}$。

2.2.1　ECA 分支

ECA 通过轻量级的一维卷积实现高效的通道注意力计算。ECA 首先通过全局平均池化 (global average pooling，GAP) 对空间维度进行压缩得到 $\boldsymbol{y}\in\mathbb{R}^{(n\times s)\times c\times 1\times 1}$，表示为

$$\boldsymbol{y}=\mathrm{GAP}(\boldsymbol{U}) \tag{15}$$

将特征重排后通过一维卷积捕捉通道间的局部依赖得到通道注意力权重 $\boldsymbol{w}\in\mathbb{R}^{(n\times s)\times c\times 1\times 1}$，表示为

$$\boldsymbol{w}=\mathrm{Sigmoid}(\mathrm{Conv1d}(\boldsymbol{y})) \tag{16}$$

其中 Conv1d 使用卷积核大小为 7 的一维卷积，最终得到 ECA 注意力分支的输出 $\boldsymbol{P}_1\in\mathbb{R}^{(n\times s)\times c\times h\times w}$，表示为

$$\boldsymbol{P}_1=\boldsymbol{U}\odot\boldsymbol{w} \tag{17}$$

2.2.2　CBAM 分支

CBAM 串联了通道注意力和空间注意力。通道注意力模块首先通过平均池化 (AvgPool) 和最大池化 (MaxPool) 获取两个通道特征 $\boldsymbol{f}^{\mathrm{c}}_{\mathrm{avg}}$ 和 $\boldsymbol{f}^{\mathrm{c}}_{max}$，表示为

$$\boldsymbol{f}^{\mathrm{c}}_{\mathrm{avg}}=\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{U})) \tag{18}$$

$$\boldsymbol{f}^{\mathrm{c}}_{\mathrm{max}}=\mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{U})) \tag{19}$$

式中，上标 c 表示通道。通过 Sigmoid 融合两个特征可得到通道注意力权重 $\boldsymbol{M}_{\mathrm{c}}$，表示为

$$\boldsymbol{M}_{\mathrm{c}}=\mathrm{Sigmoid}(\boldsymbol{f}^{\mathrm{c}}_{\mathrm{avg}}+\boldsymbol{f}^{\mathrm{c}}_{\mathrm{max}}) \tag{20}$$

空间注意力模块则聚合通道维度的平均值和最大值 $\boldsymbol{f}^{\mathrm{s}}_{\mathrm{avg}}$ 和 $\boldsymbol{f}^{\mathrm{s}}_{\mathrm{max}}$，表示为

$$\boldsymbol{f}^{\mathrm{s}}_{\mathrm{avg}}=\mathrm{mean}(\boldsymbol{U}) \tag{21}$$

$$\boldsymbol{f}^{\mathrm{s}}_{\mathrm{max}}=\max(\boldsymbol{U}) \tag{22}$$

式中，上标 s 表示空间。接着把通道维度上的平均值和最大值在通道维度上进行拼接，随后通过 7×7 的卷积层处理拼接后的特征得到空间注意力权重 $\boldsymbol{M}_{\mathrm{s}}$，表示为

$$\boldsymbol{M}_{\mathrm{s}}=\mathrm{Sigmoid}(\mathrm{Conv}(\boldsymbol{f}^{\mathrm{s}}_{\mathrm{avg}},\boldsymbol{f}^{\mathrm{s}}_{\mathrm{max}})) \tag{23}$$

通过将通道注意力和空间注意力的输出依次作用于输入特征，得到 CBAM 注意力分支的输出 $\boldsymbol{P}_2\in\mathbb{R}^{(n\times s)\times c\times h\times w}$，表示为

$$\boldsymbol{P}_2=\boldsymbol{U}\odot\boldsymbol{M}_{\mathrm{c}}\odot\boldsymbol{M}_{\mathrm{s}} \tag{24}$$

2.2.3　特征融合

为了有效的融合两种注意力机制的输出，本文首先将两个注意力分支的输出 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 相加，随后通过空洞卷积[29]得到融合特征 $\boldsymbol{F}$，表示为

$$\boldsymbol{F}=\mathrm{Conv}(\boldsymbol{P}_1+\boldsymbol{P}_2) \tag{25}$$

其中空洞卷积层卷积核大小为 1，空洞率为 2，使用空洞卷积可以在不增加参数量的情况下扩大感受野，最后通过残差连接得到增强融合特征 $\boldsymbol{O}$，表示为

$$\boldsymbol{O}=\boldsymbol{F}+\boldsymbol{U} \tag{26}$$

通过以上设计，GAFM 模块能够通过融合不同注意力机制提取的互补特征，增强对步态序列关键信息的捕捉能力。

2.3　MFFN

本文在 BNNeck 的基础上改进提出了 MFFN 结构，通过多尺度、多粒度特征融合来提升步态特征的表示能力，其结构如图 3 所示。该模块包含 3 个关键分支：局部特征分支、全局特征分支和特征融合分支，每个分支针对不同尺度的特征

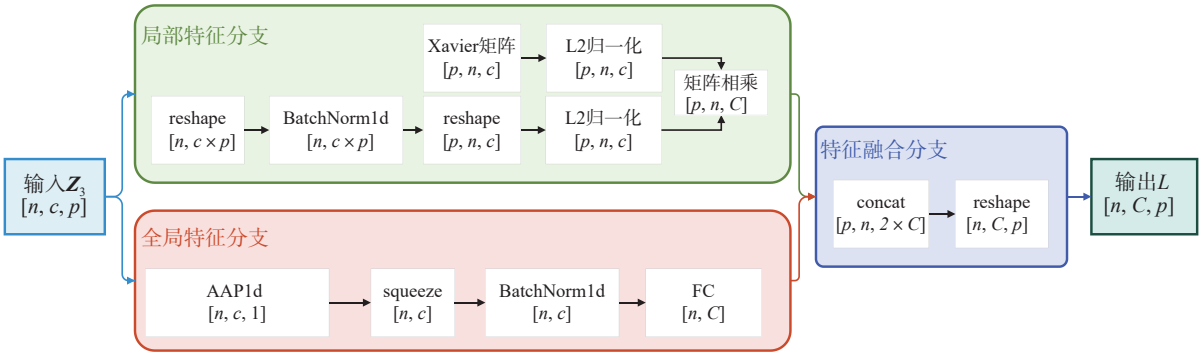

图 3 MFFN 结构图

Fig.3 The structure picture of MFFN

表示进行优化设计。其中 MFFN 模块输出的$\boldsymbol{L}$用于训练阶段的分类损失计算，以指导模型学习更具判别性的特征表示。

2.3.1 局部特征分支

给定输入特征图$\boldsymbol{Z}_3 \in \mathbb{R}^{n\times c\times p}$，其中$p$表示水平池化金字塔的总分段数。局部特征分支首先进行特征重排得到$\widetilde{\boldsymbol{M}} \in \mathbb{R}^{n\times(c\times p)}$，表示为

$$\widetilde{\boldsymbol{M}} = \mathrm{Res}(\boldsymbol{Z}_3) \tag{27}$$

使用重排操作将特征展平为二维形式，使 BatchNorm1d 能够对所有维度的特征分布进行标准化处理。接着对重排后的特征$\widetilde{\boldsymbol{M}}$进行归一化，表示为

$$\widehat{\boldsymbol{M}} = \mathrm{BatchNorm1d}\left(\widetilde{\boldsymbol{M}}\right) \tag{28}$$

其中 BatchNorm1d 操作能够标准化特征分布，接着对归一化后的特征$\widehat{\boldsymbol{M}}$重新调整维度得到$\widehat{\boldsymbol{M}}_{\mathrm{local}} \in \mathbb{R}^{p\times n\times c}$，表示为

$$\widehat{\boldsymbol{M}}_{\mathrm{local}} = \mathrm{Res}\left(\widehat{\boldsymbol{M}}\right)$$

并进行 L2 归一化得到单位向量$\widehat{\boldsymbol{L}}$，表示为

$$\widehat{\boldsymbol{L}} = \frac{\widehat{\boldsymbol{M}}_{\mathrm{local}}}{\|\widehat{\boldsymbol{M}}_{\mathrm{local}}\|_2} \tag{29}$$

最后通过 L2 归一化之后的特征$\widehat{\boldsymbol{L}}$与 Xavier[30] 初始化得到的权重矩阵$\boldsymbol{W}_{\mathrm{local}}$进行矩阵相乘，得到局部特征的预测分数$\boldsymbol{L}_{\mathrm{l}} \in \mathbb{R}^{n\times C\times p}$，表示为

$$\boldsymbol{L}_{\mathrm{l}} = \widehat{\boldsymbol{L}} \cdot \frac{\boldsymbol{W}_{\mathrm{local}}}{\|\boldsymbol{W}_{\mathrm{local}}\|_2} \tag{30}$$

对权重矩阵进行 L2 归一化可以约束特征分布、增强判别性，同时有助于提高训练过程的稳定性。

2.3.2 全局特征分支

全局特征分支首先通过自适应平均池化 (adaptive average pool，AAP) 压缩时序维度信息，提取序列的全局统计信息$\widetilde{\boldsymbol{G}} \in \mathbb{R}^{n\times c\times 1}$，表示为

$$\widetilde{\boldsymbol{G}} = \mathrm{AAP1d}(\boldsymbol{Z}_3) \tag{31}$$

随后对池化后的特征$\widetilde{\boldsymbol{G}}$进行归一化，得到标准化全局特征$\widehat{\boldsymbol{G}} \in \mathbb{R}^{n\times c}$，表示为

$$\widehat{\boldsymbol{G}} = \mathrm{BatchNorm1d}\left(\widetilde{\boldsymbol{G}}\right) \tag{32}$$

最后通过全连接层将标准化特征映射到类别空间，得到全局预测分数$\boldsymbol{L}_{\mathrm{g}} \in \mathbb{R}^{n\times C}$，表示为

$$\boldsymbol{L}_{\mathrm{g}} = \mathrm{FC}\left(\widehat{\boldsymbol{G}}\right) \tag{33}$$

2.3.3 特征融合分支

为了实现局部和全局特征的有效融合，将局部分支对应$\boldsymbol{L}_{\mathrm{l}}$和全局分支对应的$\boldsymbol{L}_{\mathrm{g}}$在类别维度上进行拼接，并进行维度调整，得到最终预测分数$\boldsymbol{L}_{\mathrm{e}} \in \mathbb{R}^{n\times C\times p}$，表示为

$$\boldsymbol{L}_{\mathrm{e}} = \mathrm{Res}\left(\mathrm{concat}(\boldsymbol{L}_{\mathrm{l}}, \boldsymbol{L}_{\mathrm{g}})\right) \tag{34}$$

通过以上设计，MFFN 的局部特征分支保留细粒度的时序信息，全局特征分支提供了整体的语义信息，特征融合分支则实现了多尺度特征的自适应融合。

3 实验结果分析

本文中所有实验均重复进行 3 次，并报告其平均值和标准差。具体实验设置如下：操作系统为 Ubuntu 22.04.4，处理器为 Intel Xeon Gold 5220 CPU，图形处理器为 NVIDIA A10，基 Pytorch 深度学习框架实现，其中实验的训练配置如下：优化器采用 SGD，初始学习率为 0.1，动量系数为 0.9，权重衰减为 0.0005，采用混合精度训练和同步批归一化提高训练效率。

3.1 实验数据集

本文选取 CASIA-B[31]、CASIA-B*[32] 和

OUMVLP[33]3 个数据集进行实验。CASIA-B 数据集包含 124 名受试者在 11 个视角下的 RGB 图像和轮廓图。每位受试者有 3 种行走状态：正常行走（NM）、背包（BG）和外套（CL）。标准测试协议通常选取前 74 名受试者的步态序列作为训练集，第 75 至 124 名用于测试。测试时，Gallery 集合选取 NM 场景中的 4 段序列（NM#1-4），剩下 6 段作为 Probe 集合，包括 NM#5-6、BG#1-2 和 CL#1-2。测试过程中，Probe 集合样本与 Gallery 集合中除自身视角外的所有视角进行匹配。

CASIA-B*数据集是 CASIA-B 的重标注版本，通过对剪影图像的精细化处理，减少了原始数据中的噪声。数据集保持了与 CASIA-B 相同的受试者数量和采集条件，但对图像质量进行了显著提升，使其更适合于步态识别研究。数据集采用与 CASIA-B 相同的训练和测试划分方案。

OUMVLP 是由大阪大学开发的大规模多视角步态数据集，包含 10 307 个受试者，每个受试者从 14 个不同视角（0°～270°）采集步态序列。该数据集在室内受控环境下采集，所有视频序列都被处理成对齐的二值化剪影序列。数据集将 5 153 个受试者的数据用于训练，剩余 5 154 个受试者的数据用于测试，每个受试者的第一个序列作为 Gallery，其余序列作为 Probe。

3.2　实验结果分析

图 4 展示了本文提出的 MSA-Gait 与 GaitSet，GaitPart 等主流步态识别模型在 3 个数据集上的平均识别率对比。从图中可以看出，在 CASIA-B 数据集上，MSA-Gait 达到 90.9% 的平均识别率，相比 GaitSet 和 GaitPart 分别提升了 6.7% 和 2.1%，相对于基线模型提升了 0.9%；在 CASIA-B*数据集上，MSA-Gait 取得 88.7% 的识别率，较 GaitSet 和 GaitPart 分别提升了 4.8% 和 4.0%，相对基线模型提升了 0.3%；在 OUMVLP 大规模数据集上，MSA-Gait 实现了 88.9% 的平均识别率，相比 GaitSet 提升了 1.8%，相比基线模型提升了 0.6%。实验结果表明，MSA-Gait 通过融合多尺度特征表示和注意力机制，在不同规模和场景的数据集上都实现了性能提升。下面将对各数据集上的实验结果进行详细分析。

3.2.1　CASIA-B

不同步态识别模型在 CASIA-B 数据集上的识别准确率如表 1 所示，其中加粗的数值表示在各条件下的最佳性能。CASIA-B 作为经典的步态数据集，其实验结果在正常行走、背包和外套 3 种条件下充分体现了模型的鲁棒性。在正常行走条件下，MSA-Gait 和 GTIEN 都表现出了优异的性能，MSA-Gait 达到了 98.0% 的识别准确率，而 GTIEN 达到了 97.9% 的识别准确率。MSA-Gait 相较于 GaitSet 提升了 3.0%，相较于 GaitPart 提升了 1.8%。这一性能提升主要得益于本文提出的多尺度特征融合策略，有效增强了特征的判别能力。特别是在处理 CASIA-B 数据集中 11 个不同视角的步态序列时，多尺度特征的融合帮助模型捕捉了更丰富的视角不变特征。

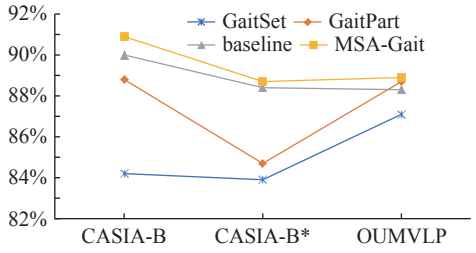

图 4　不同步态识别模型的平均准确率对比图

Fig.4　Comparison of the average recognition accuracy of different gait recognition models

表 1　不同步态识别模型在 CASIA-B 数据集上的识别准确率对比

Tab.1　Comparison of the recognition accuracy of different gait recognition models on the CASIA-B dataset

模型	正常/%	背包/%	外套/%
GaitSet	95.0	87.2	70.4
GaitPart	96.2	91.5	78.7
GaitGraph	87.7	74.8	66.3
FR-GCN[34]	91.4	80.0	75.7
PGOFI[35]	95.4	91.6	79.0
GaitMGL[36]	85.8	75.9	70.6
GTIEN[37]	97.9	**94.3**	79.0
Gaitbase(baseline)	97.5±0.3	93.7±0.2	78.8±0.2
MSA-Gait	**98.0±0.3**	94.0±0.3	**80.6±0.2**

在背包条件下，GTIEN 和 MSA-Gait 展现出很强的鲁棒性，分别达到 94.3% 和 94.0% 的识别率。其中 MSA-Gait 相比 GaitSet 的 87.2% 提升了 6.8%，相比 GaitPart 的 91.5% 提升了 2.5%。这表明本文提出的 MSA-Gait 在处理 CASIA-B 数据集中的步态条件出现形态变化时具有较强的适应能力。GTIEN 通过对时序特征的精细建模取得了最佳性能，而 MSA-Gait 则通过 MFFN 模块平衡了局

部细节和全局语义信息。

在最复杂的外套条件下，MSA-Gait以80.6%的识别率领先于所有对比方法，相较于GaitSet提升了10.2%，相较于GaitPart提升了1.9%。相比FR-GCN虽引入骨架信息来增强特征表示，但在外套遮挡情况下仅达到75.7%的识别率，本文方法提升了4.9%。这一显著提升归功于GAFM模块融合的ECA和CBAM两种注意力机制，它们能够自适应地关注未被外套遮挡的显著区域。PGOFI虽采用基于先验运动信息的部分表示，在外套条件下达到79.0%的准确率，但其复杂的特征提取过程限制了实用性，而MSA-Gait通过有效的特征融合和注意力机制，实现了1.6%的性能提升。

最后，MSA-Gait在所有条件下都优于基线模型。在正常行走、携带背包和穿着外套3种条件下分别提升了0.5%、0.3%和1.8%，这证实了所提出改进的有效性。尤其在外套条件下的显著提升，表明本文设计的多尺度特征融合和注意力机制能够更好地处理CASIA-B数据集中的复杂遮挡场景。

3.2.2 CASIA-B*

表2对比了不同步态识别模型在CASIA-B*数据集上的准确率。在CASIA-B*数据集上，本文方法在正常行走和背包条件下略低于基线模型(分别降低0.3%和0.2%)，但在最复杂的外套条件下实现了1.2%的性能提升。与GaitSet和GaitPart等方法相比，MSA-Gait在正常行走条件下分别提升了3.4%和2.6%的识别率，在背包条件下分别提升了5.4%和5.5%，在外套条件下分别提升了5.4%和3.7%，显示出其显著的性能优势。

表2 不同步态识别模型在CASIA-B*数据集上的识别准确率对比

Tab.2 Comparison of the recognition accuracy of different gait recognition models on the CASIA-B* dataset

模型	正常/%	背包/%	外套/%
GaitSet	92.3	86.1	73.4
GaitPart	93.1	86.0	75.1
Gaitbase(baseline)	**96.0±0.3**	**91.7±0.2**	77.6±0.2
MSA-Gait	95.7±0.2	91.5±0.3	**78.8±0.2**

这种性能权衡反映了模型在不同场景下的特性：在简单场景(正常行走、背包)中，步态特征相对显著，基线模型就足以获得较好性能，这是因为CASIA-B*数据集通过精细化处理提高了剪影图像的质量，降低了数据中的噪声，使得基本的特征提取方法就能获得良好的表示。但本文提出的多尺度特征融合和注意力机制在这些简单场景下引入了一定的冗余信息，导致性能略有下降。而在复杂场景(外套)下，由于衣着遮挡导致步态特征不易被提取，本文提出的MFFN通过多尺度特征融合增强了对局部细节的捕捉能力，同时GAFM模块的注意力机制帮助模型聚焦于未被遮挡的区域，因此在复杂场景下MSA-Gait表现出更好的识别性能。

这些实验结果表明，尽管本文提出的方法在简单场景下可能不会带来显著提升，但在处理复杂的场景时具有明显的优势。这种特点使得MSA-Gait适合于实际应用场景，在现实环境中，外套等遮挡情况是步态识别面临的主要挑战之一。

3.2.3 OUMVLP

表3对比了不同步态识别模型在大规模数据集OUMVLP上的准确率。GTIEN和MSA-Gait都达到了88.9%的准确率，相比于基线模型的88.3%准确率，MSA-Gait获得了0.6%的性能提升，与其他方法相比，MSA-Gait较GaitSet提升了1.8%，较GaitPart提升了0.2%，较PGOFI提升了1.7%，展现出了MSA-Gait在大规模数据集上的优越性。

表3 不同步态识别网络在OUMVLP数据集上的识别准确率对比

Tab.3 Comparison of the recognition accuracy of different gait recognition networks on the OUMVLP dataset

模型	准确率/%
GaitSet	87.1
GaitPart	88.7
PGOFI	87.2
GTIEN	**88.9**
Gaitbase(baseline)	88.3±0.2
MSA-Gait	**88.9±0.2**

模型性能的提升主要得益于以下两点：一是MFFN模块通过多尺度特征融合提高了特征的判别能力，使模型能够更好地区分大规模的行人样本，在OUMVLP数据集包含的10307个受试者中，不同行人之间的步态特征差异可能非常细微，而多尺度特征融合策略能够同时捕捉全局运动模式和局部细节特征，从而提升模型的识别精度。二是GAFM模块的注意力机制有效捕捉了步

态序列中的关键帧和重要区域，增强了模型对长期时空依赖关系的建模能力，这一特性在处理 OUMVLP 数据集中来自 14 个不同视角的步态序列时尤为重要，这是因为 GAFM 能够自适应地关注不同视角下的显著特征。以上结果验证了 MSA-Gait 在大规模应用场景中的可靠性和实用性，为实际部署提供了有力支持。

3.3　消融实验

为了全面评估本文提出的 MFFN 和 GAFM 模块的贡献，本文在 CASIA-B 数据集上进行了详细的消融实验。随后评估了每个模块在单独使用以及组合使用下的性能表现，实验结果如表 4 所示。

表 4　CASIA-B 数据集的消融实验

Tab.4　Ablation experiments on the CASIA-B dataset

模型	正常/%	背包/%	外套/%
baseline	97.5±0.3	93.7±0.2	78.8±0.2
MFFN	97.7±0.2	93.7±0.3	79.2±0.2
GAFM	97.8±0.3	**94.0±0.3**	79.3±0.3
MFFN+GAFM	**98.0±0.3**	**94.0±0.3**	**80.6±0.2**

表 4 的消融实验结果不仅验证了各个模块的有效性，同时也展示了它们之间的协同效果。在正常行走条件下，单独使用 MFFN 和 GAFM 模块分别将识别率从 97.5% 提升至 97.7% 和 97.8%，表明两个模块均能有效增强模型的特征学习能力。在背包条件下，MFFN 保持了与基线模型相当的性能（93.7%），而 GAFM 提升至 94.0%，显示出注意力机制在处理携带物品时导致的局部形变时具有优势。在最复杂的外套条件下，MFFN 和 GAFM 分别实现了 0.4% 和 0.5% 的性能提升，将识别率从 78.8% 提高到 79.2% 和 79.3%。

当组合使用两个模块时，模型在各种条件下都取得了最佳性能：正常行走条件下达到 98.0%，背包条件下达到 94.0%，外套条件下达到 80.6%。这种性能提升揭示了 GAFM 和 MFFN 之间存在协同作用：GAFM 通过自适应加权机制首先增强了特征的表达能力，为后续的多尺度特征融合提供了基础；而 MFFN 通过多尺度特征融合进一步提升了特征的判别性，使得模型能够更好地捕捉步态的细节特征。特别是在外套条件下，两个模块的组合带来了 1.8% 的提升，这表明多尺度特征融合和注意力机制的结合能够更好地应对复杂的场景。这些实验结果有力地支持了本文提出的模型设计的合理性和有效性。

为了深入分析模块融合带来的计算开销，本文对模型的参数量、推理时间和计算量进行了详细统计，参数量对比图如图 5 所示。

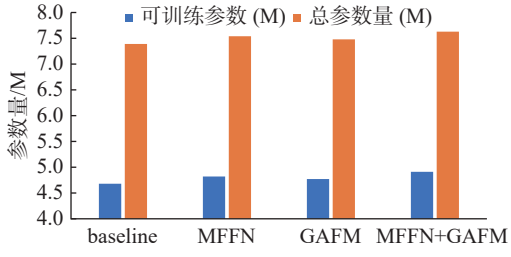

图 5　参数量对比图

Fig.5　Parameter comparison diagram

在参数量方面，基线模型的可训练参数为 4.68 M，总参数量为 7.39 M。而完整模型的可训练参数和总参数量分别为 4.91 M 和 7.63 M，相比基线模型分别增加了 4.9% 和 3.3%。其中计算量和推理时间对比图如图 6 所示。

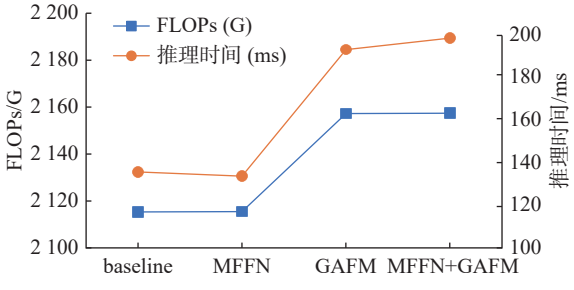

图 6　计算量与推理时间对比图

Fig.6　Comparison diagram of computational cost and inference time

在模型计算量方面，基线模型的 FLOPs 为 2 115.35 G，而融合 MFFN 和 GAFM 后的完整模型为 2 157.41 G，仅增加了约 2% 的计算量。在推理时间方面，基线模型的平均推理时间为 135.60±1.56 ms，完整模型为 198.38±1.96 ms。

进一步分析各个模块的计算开销发现：MFFN 模块在保持接近基线模型推理速度（133.63±39 ms）的同时，仅增加了 0.13 G 的 FLOPs 和 0.15 M 的总参数量；GAFM 模块虽然引入了 41.93 G 的额外 FLOPs 和 0.09 M 的参数量，但为模型带来了较为显著的性能提升。尤其在外套条件下，GAFM 单模块实现了 0.5% 的准确率提升。考虑到在实际应用中，准确率的微小提升对系统的可靠性具有重要影响，且硬件平台的计算能力在不断提升，本文认为这种程度的计算开销增加是可以接受的。

4 结 论

针对步态识别中存在的特征表示粒度不足和时空依赖建模不充分等问题，提出了一种改进的步态识别模型。通过引入MFFN结构实现多尺度特征融合，利用GAFM模块建模时空依赖关系，本文的模型在3个公开数据集上取得了优于现有方法的识别性能，尤其在复杂场景下展现出了良好的鲁棒性和泛化能力。

现有的步态识别方法大多依赖于精心设计的深度网络结构，而对于数据的利用和增强还不够充分。未来可以探索更有效的数据增强策略，以进一步提高模型的鲁棒性和泛化能力。其次，当前的步态识别研究主要集中在封闭场景下，而在开放场景中，如户外、夜间、拥挤等环境，有待提高模型的适应性。

参考文献：

[1] WANG L, TAN T N, NING H Z, et al. Silhouette analysis-based gait recognition for human identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(12): 1505−1518.

[2] SEPAS-MOGHADDAM A, ETEMAD A. Deep gait recognition: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 264−284.

[3] PATAKY T C, MU T T, BOSCH K, et al. Gait recognition: highly unique dynamic plantar pressure patterns among 104 individuals[J]. Journal of the Royal Society Interface, 2012, 9(69): 790−800.

[4] WU Z F, HUANG Y Z, WANG L, et al. A comprehensive study on cross-view gait based human identification with deep cnns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(2): 209−226.

[5] NIXON M S, CARTER J N. Automatic recognition by gait[J]. Proceedings of the IEEE, 2006, 94(11): 2013−2024.

[6] LUO H, JIANG W, GU Y Z, et al. A strong baseline and batch normalization neck for deep person re-identification[J]. IEEE Transactions on Multimedia, 2020, 22(10): 2597−2609.

[7] BOUCHRIKA I, NIXON M S. Model-based feature extraction for gait analysis and recognition[C]//Proceedings of the Third International Conference on Computer Vision/Computer Graphics Collaboration Techniques. Rocquencourt, France: Springer, 2007: 150–160.

[8] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: a skinned multi-person linear model[J]. Seminal Graphics Papers: Pushing the Boundaries, 2023, 2: 88.

[9] LIAO R J, YU S Q, AN W Z, et al. A model-based gait recognition method with body pose and human prior knowledge[J]. Pattern Recognition, 2020, 98: 107069.

[10] TEEPE T, KHAN A, GILG J, et al. Gaitgraph: graph convolutional network for skeleton-based gait recognition[C]//Proceedings of 2021 IEEE International Conference on Image Processing. Anchorage: IEEE, 2021: 2314–2318.

[11] LI X, MAKIHARA Y, XU C, et al. End-to-end model-based gait recognition using synchronized multi-view pose constraint[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021.

[12] 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/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 20228–20237.

[13] FU Y, MENG S B, HOU S H, et al. GPGait: generalized pose-based gait recognition[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 19595–19604.

[14] PENG Y J, MA K, ZHANG Y, et al. Learning rich features for gait recognition by integrating skeletons and silhouettes[J]. Multimedia Tools and Applications, 2024, 83(3): 7273−7294.

[15] HAN J, BHANU B. Individual recognition using gait energy image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(2): 316−322.

[16] 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 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 8126–8133.

[17] FAN C, PENG Y J, CAO C S, et al. GaitPart: temporal part-based model for gait recognition[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 14225–14233.

[18] 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 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 382–398.

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

[20] CUI Y F, KANG Y M. GaitTransformer: multiple-temporal-scale transformer for cross-view gait recognition[C]//Proceedings of 2022 IEEE International Conference on Multimedia and Expo. Taipei, China: IEEE, 2022: 1–6.

[21] FAN C, HOU S H, WANG J L, et al. Learning gait representation from massive unlabelled walking videos: a benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14920−14937.

[22] FAN C, HOU S H, HUANG Y Z, et al. Exploring deep models for practical gait recognition[DB/OL]. [2024-01-10]. https://arxiv.org/abs/2303.03301.

[23] YE D Q, FAN C, MA J Z, et al. BigGait: learning gait representation you want by large vision models[C]// Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 200–210.

[24] REN H, CHEN J J, VELIPASALAR S. GaitPoint+: a gait recognition network incorporating point cloud analysis and recycling[DB/OJ]. [2024-04-16]. https://arxiv.org/abs/2404.10213.

[25] FAN C, LIANG J H, SHEN C F, et al. OpenGait: revisiting gait recognition toward better practicality[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 9707–9716.

[26] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770–778.

[27] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11534–11542.

[28] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 3–19.

[29] [29] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. [2016-04-30]. https://arxiv.org/abs/1511.07122

[30] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Sardinia: JMLR. org, 2010: 249–256.

[31] 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]//Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong, China: IEEE, 2006: 441–444.

[32] LIANG J H, FAN C, HOU S H, et al. GaiteDge: beyond plain end-to-end gait recognition for better practicality[C]//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 375–390.

[33] 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): 4.

[34] WANG L K, CHEN J Y, LIU Y X. Frame-level refinement networks for skeleton-based gait recognition[J]. Computer Vision and Image Understanding, 2022, 222: 103500.

[35] XU J, LI H, HOU S J. Attention-based gait recognition network with novel partial representation PGOFI based on prior motion information[J]. Digital Signal Processing, 2023, 133: 103845.

[36] ZHANG Z P, WEI S W, XI L Y, et al. GaitMGL: multi-scale temporal dimension and global–local feature fusion for gait recognition[J]. Electronics, 2024, 13(2): 257.

[37] CHEN Y F, LI X L. Gait feature learning via spatio-temporal two-branch networks[J]. Pattern Recognition, 2024, 147: 110090.

（编辑：董　伟）