Modeling and Simulation 建模与仿真, 2025, 14(5), 1161-1175
Published Online May 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.145466

Hans汉斯

基于融合Capsule-Transformer的人体活动识别模型

王　星，李瑞祥*，施伟斌

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年4月30日；录用日期：2025年5月23日；发布日期：2025年5月31日

摘　要

近年来，基于可穿戴传感器的人体活动识别在智能健康监护、人机交互等场景展现出重要应用价值。传统的深度学习算法，如卷积神经网络(CNNs)和循环神经网络(RNNs)，虽然取得了一定成效，但在捕捉复杂人体活动的时序动态和空间关系方面仍存在不足。为了解决这些局限性，文章利用Transformer模型在捕获全局特征方面的优势和胶囊网络在捕获局部特征方面的优势，提出了一种结合Transformer模型与胶囊网络的新型混合架构。在两个公共数据集(UCI-HAR和WISDM)上对模型性能进行了评估。最后，该模型在UCI-HAR数据集中的总体准确率为96.0%，在WISDM数据集中的总体准确率为96.5%。实验结果表明，基于Transformer和胶囊网络的融合模型比仅基于Transformer和仅基于胶囊网络的模型表现出更好的性能。而且，该算法的性能优于近期同类研究中的其他传统深度学习算法。

关键词

人体活动识别，可穿戴传感器，Transformer，胶囊网络

Human Activity Recognition Model Based on the Fusion of Capsule-Transformer

Xing Wang, Ruixiang Li*, Weibin Shi

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 30th, 2025; accepted: May 23rd, 2025; published: May 31st, 2025

Abstract

In recent years, wearable sensor-based Human Activity Recognition (HAR) has demonstrated

*通讯作者。

文章引用: 王星, 李瑞祥, 施伟斌. 基于融合 Capsule-Transformer 的人体活动识别模型[J]. 建模与仿真, 2025, 14(5): 1161-1175. DOI: 10.12677/mos.2025.145466

significant application value in intelligent health monitoring and human-computer interaction scenarios. Although traditional deep learning algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have achieved certain results, they still exhibit limitations in capturing the temporal dynamics and spatial relationships of complex human activities. To address these limitations, this paper proposes a novel hybrid architecture combining Transformer models and Capsule Networks, leveraging the Transformer's advantages in capturing global features and Capsule Networks' strengths in extracting local features. The model performance was evaluated on two public datasets (UCI-HAR and WISDM), achieving overall accuracies of 96.0% on the UCI-HAR dataset and 96.5% on the WISDM dataset. Experimental results demonstrate that the Transformer-Capsule hybrid model outperforms both standalone Transformer-based and Capsule Network-based models. Furthermore, the proposed algorithm exhibits superior performance compared to other traditional deep learning approaches in recent related studies.

Keywords

Human Activity Recognition, Wearable Sensor, Transformer, Capsule Network

1. 引言

人体活动识别(Human Activity Recognition, HAR)是通过传感器、摄像头或可穿戴设备捕获人体运动数据，并利用算法对特定活动(如行走、跑步、跌倒、坐立等)进行自动分类和识别的技术。随着物联网、智能医疗和健康监测技术的快速发展，HAR 在多个领域展现出广阔的应用前景[1]。

随着传感器技术和普适计算技术的飞速发展，基于传感器的 HAR 技术得到了越来越广泛的应用，隐私也得到了很好的保护。研究人员探索了不同类型的传感技术在活动识别中的作用，以提高识别准确率。根据传感器在环境中的使用方式，HAR 技术可以大致分为两类：基于固定传感器的方法和基于移动传感器的方法[2]。

基于固定传感器的方法是指从安装在固定位置的传感器中获取信息，涉及声学传感器[3]、雷达[4]、静态相机[5]和其他基于环境的传感器，其中基于相机的方法是最流行的方法，通常应用背景减除法、光流法和基于能量的分割法提取特征[6]-[9]。尽管这些活动监视方法可以提供更好的识别准确性，但是它们不适合于许多室内环境，尤其是在隐私受到关注的情况下。此外，基于视觉的方法识别效果容易受到光照变化、环境遮挡和背景变化的影响，这极大地限制了其实际应用。

另一种活动识别方法是使用移动的传感器。在这些方法中，不同类型的行为信息通常是从一组专用的身体穿戴式运动传感器收集的，例如加速度计、陀螺仪和磁力计。加速度和角速度数据会随着人体运动而变化。因此，传感器的小型化和灵活性允许个人佩戴或携带嵌入有各种传感单元的移动设备。这与基于固定传感器的方法不同[10]。而且这些传感器具有低成本、低功耗、大容量、小型化、对周围环境依赖性小等特点[11]，因此基于移动传感器的活动识别因其便携性和日常生活中接受度高而受到广泛关注，相应地，已经进行了大量的研究来探索移动传感器以无处不在和普遍的方式进行活动识别的潜力。Margarito 等人[12]将加速度传感器置于受试者手腕上采集加速度数据，然后利用模板匹配算法对 8 种常见的体育运动进行分类。在文章[13]中，提出了一种针对老年人和残疾人的智能生活辅

助系统(SAIL)，Zhu 等人[13]通过多传感器融合策略的方式收集特征，实现了识别 13 种日常活动的目标。

深度神经网络已经成为当今人工智能系统的基础设施。不同类型的任务通常涉及不同类型的网络。例如，多层感知器(MLP)或全连接(FC)网络是经典类型的神经网络，它由多个线性层和非线性激活堆叠在一起[14]。卷积神经网络(CNNs)引入卷积层和池化层来处理图像等平移不变数据[15]。而递归神经网络(RNNs)利用递归单元来处理序列数据或时间序列数据[16]。Transformer 是一种新型神经网络。它主要利用自注意力机制提取内在特征[17]，并在人工智能应用中显示出广泛应用的巨大潜力。Hinton 等人[18]引入了神经网络中的胶囊概念，以利用数据的相关空间信息。一个胶囊包括一组神经层，这些神经层堆叠在一个单层中，并形成一个被称为 CapsNets 的网络[19]。与 CNNs 产生的标量值不同，胶囊中的神经元组将输出封装成信息活动向量(表示位置或大小等特征)。Sabour 等人[20]使用动态路由算法对 CapsNet 架构的设计进行了改进。该算法不需要输入活动数据，而是利用活动向量(也被称作实例化参数)来表示特征，以及向量的长度来表示特定特征的存在概率。受现有模型的启发，本文提出了一种 Transformer 模型与胶囊模型的集成融合，以提高胶囊模型用于人体活动识别的性能。

2. 方法

2.1. 胶囊模型

胶囊网络[20]旨在克服传统网络结构(如 CNNs)的一些限制。其整体框架可以分为两个部分：编码部分包括具有整流线性单元(ReLU)的卷积层(例如 ReLU Conv 1)、主胶囊层(即 Primary Caps)和第二个封装层(Second Caps)；解码部分包括多个具有非线性激活函数 ReLU 和 Sigmoid 的全连接层(例如 FC ReLU 和 FC Sigmoid)。

传统神经网络的神经元输入和输出只能表示提取特征的概率，而没有考虑它们的空间关系。相比之下，胶囊网络使用胶囊作为基本组件[20]，它由多个神经元组成，每个神经元由一个向量表示。值得注意的是，这些神经元的输入和输出都是向量，其中输出值表示实体存在的概率在 0 到 1 的范围内。这些向量的大小和方向相应地表示实体特征的概率和属性。

Table 1. Differences between vector neurons (VN) and scalar neurons (SN)
表 1. 矢量神经元(VN)和标量神经元(SN)之间的差异

		VN	SN
输入		u_i	x_i
操作	转换	$\hat{u}_{j\mid i}=W_{ij}u_i$	-
	加权求和	$s_j=\sum_i c_{ij}\hat{u}_{j\mid i}$	$a_j=\sum_i w_i x_i+b$
	非线性激活	$v_j=\frac{\\|s_j\\|}{1+\\|s_j\\|^2}\cdot\frac{s_j}{\\|s_j\\|}$	$h_j=g\left(a_j\right)$
输出		v_j	h_j

表 1 对比了向量神经元(Vector Neural, VN)和标量神经元(Scalar Neural, SN)之间的差异。在该表中，$x_i\left(i=1,2,\cdots,n\right)$ 表示标量神经元的输入，$w_i\left(i=1,2,\cdots,n\right)$ 表示相应的权重，b 表示偏置。变量 $u_i\left(i=1,2,\cdots,n\right)$ 表示较低级别的胶囊，而 $\hat{u}_i\left(i=1,2,\cdots,n\right)$ 表示较低级别胶囊对于较高级别胶囊的预测，$\sum$ 表示对输入的求和运算，c_{ij} 表示不同层向量元素之间的耦合系数，s_j 表示对当前层胶囊向量的输入，即预测向量的

加权和。在向量神经元的前向传播过程中，不同的胶囊之间使用动态路由算法进行交互。在标量神经元的前向传播过程中，将输入 x_i 和权重的乘积相加以形成标量 a_i，然后通过非线性函数将其转换为输出 h_j。动态路由算法旨在迭代地更新连接胶囊层的权重矩阵，以便选择表现出与初级胶囊层高度一致性的检测胶囊。该算法有助于将表示人体活动特征的初级胶囊层与表示活动类别的次级胶囊层进行匹配。计算过程概述如下：

$$\hat{u}_{j|i}=W_{ij}u_i \tag{1}$$

$$s_j=\sum_i c_{ij}\hat{u}_{j|i} \tag{2}$$

$$v_j=\frac{\left\|s_j\right\|}{1+\left\|s_j\right\|^2}\cdot\frac{s_j}{\left\|s_j\right\|} \tag{3}$$

其中 $\hat{u}_{j|i}$ 表示从向量 u_i 到向量 v_j 的预测向量，W_{ij} 表示对应的权重矩阵，v_j 表示胶囊 j 的输出向量。向量 s_j 通过非线性 squash 函数压缩和归一化后，得到单位向量 v_j。

$$c_{ij}=\frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{ik}\right)} \tag{4}$$

其中参数 b_{ij} 用于更新耦合系数，其初始值通常设为 0，具体公式如下：

$$b_{ij}\longleftarrow b_{ij}+\hat{u}_{j|i}\cdot v_j \tag{5}$$

在每次前向传播过程中，v_j 的值是基于 b_{ij} 计算得出的。通过迭代更新 b_{ij} 并对耦合系数 c_{ij} 进行后续更新，最终得到最优的耦合系数。

2.2. Transformer 模型

Transformer 最初由 Google 团队于 2017 年提出[17]，作为机器翻译的序列到序列模型。与 CNN 和 RNN 不同，它采用自注意力机制来建立全局上下文信息，并使用位置编码来表示输入数据。因此，Transformer 可以实现更多的并行计算，与传统网络结构相比，性能得到了显著改善。

Transformer 架构由多个编码器和解码器层组成。编码器和解码器都由 N 个完全相同的层组成，每个层都使用残差连接和层归一化。编码器获取输入特征并将其转换为高级嵌入，然后由解码器进行转换以生成输出。每个编码器主要由一个多头自注意力(MSA)模块和一个位置前馈网络(FFN)组成。为了实现更深层次的模型，每个模块都应用了残差连接，然后进行层归一化(LN)。相比之下，解码器在 MSA 和 FFN 模块之间包括一个额外的交叉注意(CA)模块。

在 Transformer 模型中，自注意力机制是一个重要的部分。在计算多头自注意力时，首先通过公式(6)~(8)将输入向量 x 转换为查询向量 Q、键向量 K、值向量 V：

$$Q=xW^Q x_i w_i \tag{6}$$

$$K=xW^K \tag{7}$$

$$V=xW^V \tag{8}$$

其中 W^Q、W^K、W^V 分别对应 Q,K,V 的权重矩阵。

注意机制的公式定义如下：

$$\text{Attention}\left(Q,K,V\right)=\text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{9}$$

其中 softmax 函数也被称作归一化指数函数，它将每一个元素的范围都压缩到(0, 1)之间，并且所有元素的和为 1。

MSA 机制将 Q、K 和 V 划分为 h 个头，从而实现输入 x 和其他相似输入的并行计算。然后将输出拼接起来，从而显著提高了模型的计算效率。在并行计算中，先在输入矩阵 x 和权重矩阵 W^Q、W^K、W^V 之间进行矩阵乘法，再通过获得的矩阵 Q、K、V 计算其注意力，最后将各个注意力头拼接起来乘以权重矩阵 W^O 得到输出矩阵。相关公式如下所示：

$$\mathrm{MultiHead}\left(Q,K,V\right)=\mathrm{Concat}\left(\mathrm{head}_1,\cdots,\mathrm{head}_h\right)W^O \tag{10}$$

$$\mathrm{head}_i=\mathrm{Attention}\left(Q_i,K_i,V_i\right) \tag{11}$$

其中 head_i 表示第 i 个头的注意力。

Transformer 编码器部分中的前馈层基本上由具有线性结构和卷积结构的多层感知器(MLP)组成。它利用高斯误差线性单元(GELU)和线性激活函数，并且可以从以下公式中获得，其中 x 是来自前一层的输出，W 和 b 是学习参数：

$$\mathrm{GELU}\left(x\right)=0.5x\left(1+\tanh\left(\sqrt{\frac{2}{\pi}}\left(x+0.044715x^3\right)\right)\right) \tag{12}$$

$$\mathrm{FFN}\left(x\right)=\max\left(0,xW_1+b_1\right)W_2+b_2 \tag{13}$$

2.3. 改进 Capsule-Transformer 模型

所提出的模型架构如图 1 所示，该架构由三个部分组成：门控卷积层、Transformer 层和胶囊层。前两个部分用于特征提取，而第三部分用于人体活动的分类和检测。在卷积层中，我们使用门控卷积[21]。利用该方法可以将动态特征选择机制应用于每个通道和空间位置。该方法利用三个并行的门控卷积神经网络块从输入特征中提取信息，每个并行块由三个卷积层组成。在每个块之后，沿着频率轴应用二维最大池(Max Pool)以进行降维，而时间轴保持不变以匹配目标长度。此外，在卷积层内使用相同大小的卷积核来从输入特征中提取信息。图 2 展示了传统部分卷积和门控卷积之间的比较。在部分卷积的情况下，ReLU 更新通过更新掩码来表示卷积特征。

Transformer 模型的编码器部分被添加在卷积层之后。这一增加有助于系统捕获输入信号中的全局信息。编码器结构由自注意力机制和前馈神经网络层组成，并且在自注意力层之后使用残差连接和归一化，以提高收敛速度。随后，将特征输入到前馈神经网络层，在前馈神经网络层中，使用具有 GELU 激活函数的全连接层来增强模型的泛化能力，然后在最终输出之前应用残差连接和归一化。本文所采用的 Transformer 架构如图 3 所示。

在胶囊层中，我们利用改进的胶囊网络结构，结合时间注意力层，并与第二个胶囊层并行计算输出。该层的引入有效地解决了输入数据中背景噪声导致的模型性能降低问题，特别是在复杂数据集中[22]。这些特征通过 ReLU 激活函数输入到初级胶囊层，从而使输出变成单独的时间片，这些时间片被视为后续层的单独输入。然后，时间片被传递到第二胶囊层和时间注意力(Temporal Attention, TA)层。在第二胶囊层中，动态路由算法用于训练特征并计算输出。与传统胶囊动态路由算法相比，时间注意力层受到文献[23]中概述的注意力方案的启发，对活动帧采用注意力权重，即关注重要帧，衰减无关帧，最后，将第二胶囊层和时间注意力层输出合并，得到数据特征的预测值。这些预测值可以被视为相对于从时间注意力层导出的概率分布的胶囊的预期长度。

实验结果表明，相比于原来的路由机制，使用时间注意力层能产生更好的性能。本文所使用的胶囊层的结构流程图如图 4 所示。

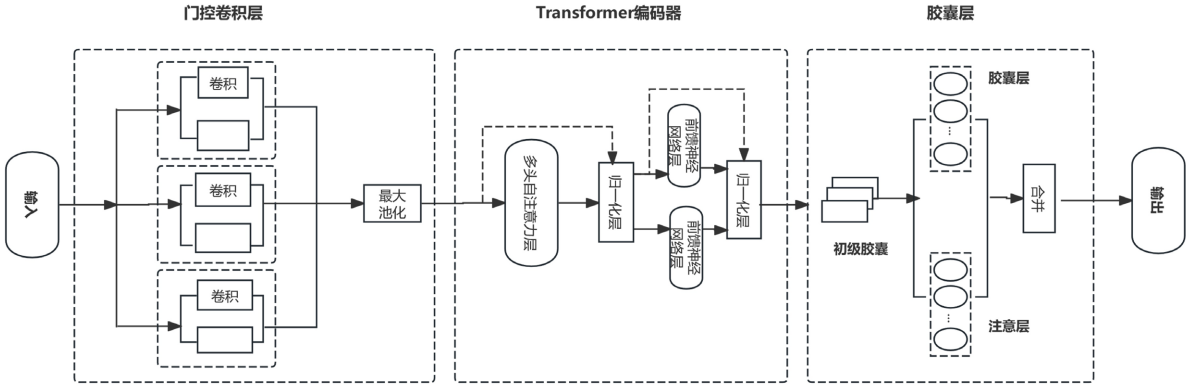

Figure 1. The neural network architecture proposed in this paper
图 1. 本文提出的神经网络架构

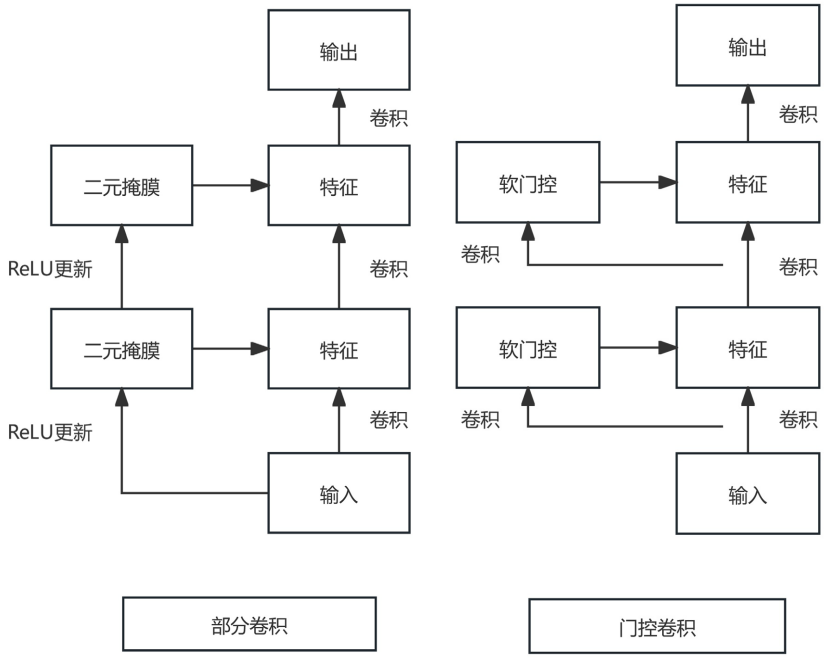

Figure 2. Illustration of the partial convolution (left) and the gated convolution (right)
图 2. 部分卷积图(左)和门控卷积图(右)

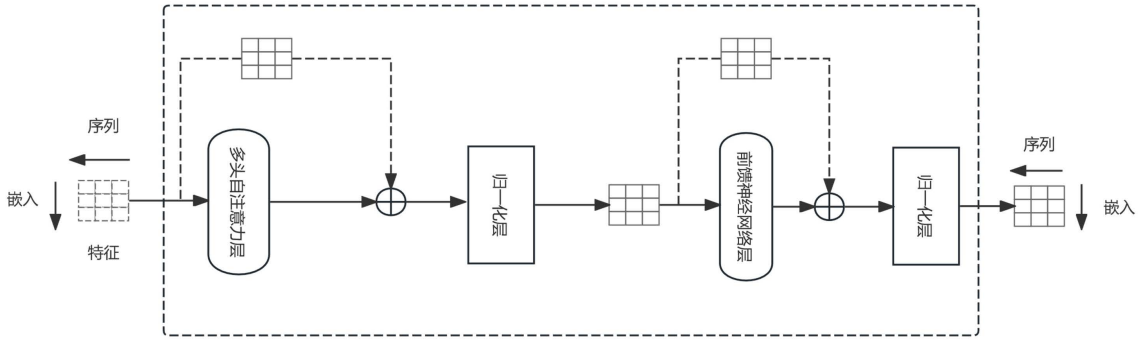

Figure 3. The network architecture of the transformer encoder
图 3. Transformer 编码器的网络结构

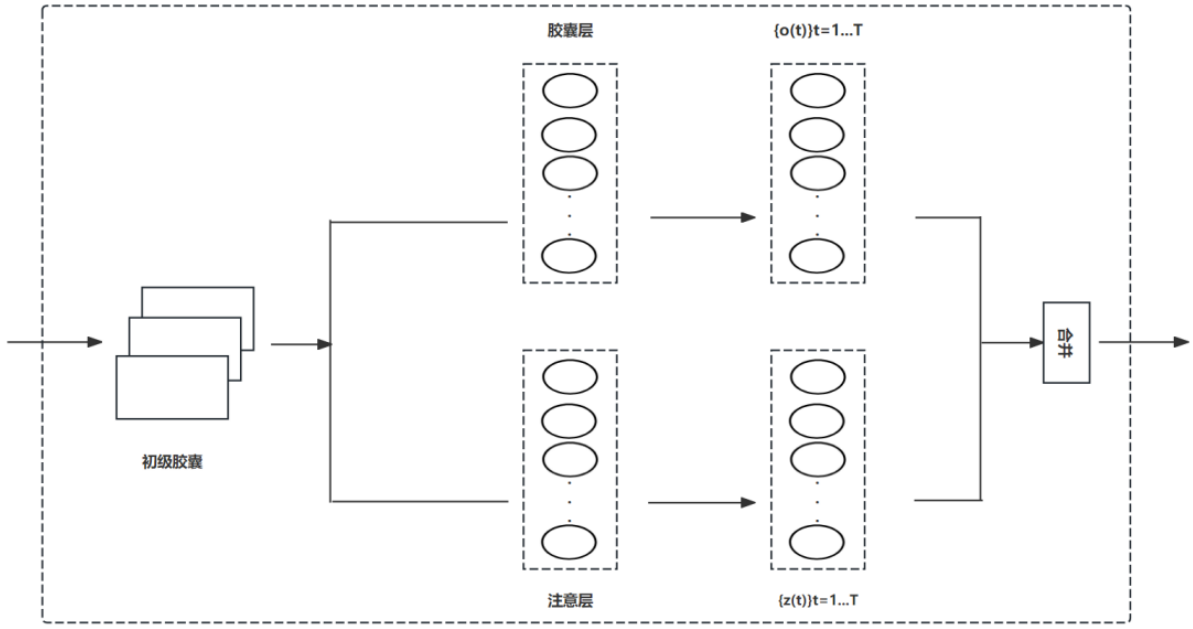

Figure 4. The network architecture of the capsule layer
图 4. 胶囊层的网络架构

3. 实验结果与分析

3.1. 数据集描述

HAR 实验使用智能手机中嵌入的加速度计传感器和陀螺仪传感器记录的数据进行。在这项工作中，本研究在两个数据集上实现并观察改进 Capsule-Transformer 模型的性能，即 UCI-HAR [24]和 WISDM [25]。

UCI-HAR 数据集[24]是根据 30 名年龄在 19~48 岁之间的受试者的记录构建的。在记录期间，所有受试者都被指示遵循活动协议。他们戴着智能手机(三星 Galaxy S II)，腰部周围嵌有惯性传感器，日常生活的六项活动分别是站立、坐下、平躺、行走、下楼和上楼。此外，该数据集还包括发生在静态姿势之间的转换：站到坐、坐到站、坐到躺、躺到坐、站到躺和躺到站。具体地说，由于姿势转换的百分比很小，本文只选择了六种基本日常活动作为输入样本。实验已经被视频记录以手动标记数据。最后，研究人员以 50 Hz 的恒定采样频率捕获了三轴加速度和三轴角速度数据。据统计，该数据集中的样本数量为 748,406，详细信息见表 2。

Table 2. Activities of UCI-HAR
表 2. UCI-HAR 活动

活动	样本数	百分比
行走	122,091	16.3%
上楼	116,707	15.6%
下楼	107,961	14.4%
坐下	126,677	16.9%
平躺	136,865	18.3%
站立	138,105	18.5%

WISDM 数据集[25]共有 1,098,209 个样本，与每种活动相关的样本占总样本的百分比如表 3 所示。可以看出，WISDM 是一个不平衡的数据集，其中走路的活动样本占比最大，达到 38.6%，而站立仅占总样本的 4.4%。其实验对象由 36 名受试者组成，这些受试者在前腿口袋里放一部 Android 手机进行某些日常活动。使用的传感器是采样频率为 20 Hz 的加速度计，它也是智能手机的内置运动传感器。该实验共记录了六种日常行为活动，分别是站立、坐、走路、上楼、下楼和慢跑。为了确保数据的质量，数据采集由专人进行监督。

Table 3. Activities of WISDM
表 3. WISDM 活动

活动	样本数	百分比
走路	424,400	38.6%
上楼	122,869	11.2%
下楼	100,427	9.1%
坐	59,939	5.5%
慢跑	342,177	31.2%
站立	48,397	4.4%

3.2. 数据预处理

为了提高模型的准确性和鲁棒性，需要对可穿戴传感器收集的原始数据进行预处理。采用噪声滤波器对传感器信号进行预处理，降噪后的信号以 2.56 s 的窗口大小采样，重叠率为 50%，每个窗口包括 128 个读数[26]。采用低通巴特沃斯滤波器分离加速度传感器信号中的物体加速度和重力加速度分量。假设重力具有低频分量，滤波器的截止频率为 0.3 Hz [26]，然后将所有特征值在[−1, 1]范围内归一化。数据集的 70%作为训练数据，30%用作测试数据。

3.3. 实验环境与评估指标

本次实验均在配备英特尔酷睿 i7-8700 处理器 8 核 3.2 GHz CPU、32 GB 内存存储和 NVIDIA GeForce GTX 2080 TI 4 GB-128 位显卡的台式机上进行。使用 Python 3.9.2 环境进行计算，在深度神经网络训练中使用 TensorFlow 库，Numpy 用于数组操作，matplotlib 用于数据可视化。以准确率、精度、召回率和 F1-Score 作为性能指标，比较不同算法的有效性，模型均采用 10 倍交叉验证。

Table 4. Model parameters (feature extraction)
表 4. 模型参数(特征提取)

	Conv1	Tf1	Conv2	Tf2	Conv3	Tf3
卷积核尺寸	64@3 × 3	-	64@3 × 3	-	64@3 × 3	-
步长	1×1	-	1×1	-	1×1	-
池化尺寸	2 × 2	-	2 × 2	-	2 × 2	-
注意力头数	-	1	-	1	-	1
丢弃率	0.2	0.3	0.2	0.3	0.2	0.3
激活函数	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU

Table 5. Model parameters (capsule layers)
表 5. 模型参数(胶囊层)

	初级胶囊层	次级胶囊层
卷积核尺寸	32@3 × 3	-
步长	1×1	-
丢弃率	0.5	-
激活函数	Squashing	Squashing
胶囊维度	8	16

表 4 和表 5 展示了模型不同阶段使用的超参数设置，其中“Tf”表示 Transformer 架构，后缀数字代表对应的 Transformer 层索引。在门控卷积网络部分，本文采用了 64 个尺寸为 3 × 3 的滤波器，池化尺寸均设置为 2 × 2。为抑制过拟合并加速收敛，在每个卷积层和初级胶囊层后均实施了批归一化操作。在 Transformer 结构中，采用的是单注意力头的编码器架构设计。

在优化策略方面，本文采用 Adam 优化器[25]作为梯度下降算法，维持 0.01 的固定学习率。动态路由的迭代次数设置为 4 次，学习率每训练 50 个周期后按 0.9 系数衰减。选用二元交叉熵作为损失函数，并据此进行梯度计算。批量大小设置为 128，每次实验均进行 300 批次训练。

在人体活动识别领域中，准确率(Accuracy)、精度(Precision)、召回率(Recall)和 F1 分数(F1-Score)是衡量分类模型性能的四个重要指标。这些指标都是基于混淆矩阵(Confusion Matrix)来计算的，混淆矩阵包括以下四种类型：真正例(True Positive, TP)，即实际为正，预测为正的样本；真负例(True Negative, TN)，即实际为负，预测为负的样本；假正例(False Positive, FP)，即实际为负，预测为正的样本；假负例(False Negative, FN)，即实际为正，预测为负的样本。

准确率(Accuracy)衡量模型整体预测正确的比例，反映模型对所有类别(正类和负类)的综合分类能力。计算公式如下：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

精度(Precision)衡量模型预测为正类的样本中实际为正类的比例，关注预测结果的可靠性。计算公式如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

召回率(Recall)衡量实际为正类的样本中被模型正确预测的比例，关注模型对正类的覆盖能力。计算公式如下：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

F1 分数(F1-Score)是精度和召回率的调和平均数，综合平衡两者性能，尤其适用于类别不平衡的数据。计算公式如下：

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

3.4. 与最新模型的比较

为了全面验证所提出模型的性能，本文使用了两个公开数据集进行测试。图 5 和图 6 分别显示了使用 UCI-HAR 和 WISDM 数据集的测试集预测模型时获得的分类混淆矩阵。以准确率、精度、召回率和 F1 分数作为性能指标，验证本文所提出算法的有效性。图 7 和图 8 分别为 UCI-HAR 和 WISDM 数据集的分类报告图表。对于 UCI-HAR 数据集，整体准确率达到 96.0%，整体精度达到 95.7%，整体召回率达到 96.0%，整体 F1 分数达到 95.9%。但在区分坐下和站立两种活动时表现相对欠佳，分类准确率仅有 89%和 91%。主要原因可能是这两种活动从运动传感器的视角来看具有相似性，仅通过加速度数据难以挖掘更深层次的差异信息。当将训练好的模型应用于不平衡的 WISDM 数据集的测试集时，整体准确率达到 96.5%，整体精度达到 96.4%，整体召回率达到 96.5%，整体 F1 分数达到 96.4%。但在区分上楼和下楼两种活动时表现相对欠佳，分类准确率仅有 89%和 96%。混合模型在多数类别上 F1 分数超过 90%，表明其泛化能力较强，适用于实际场景中的多类别分类任务。

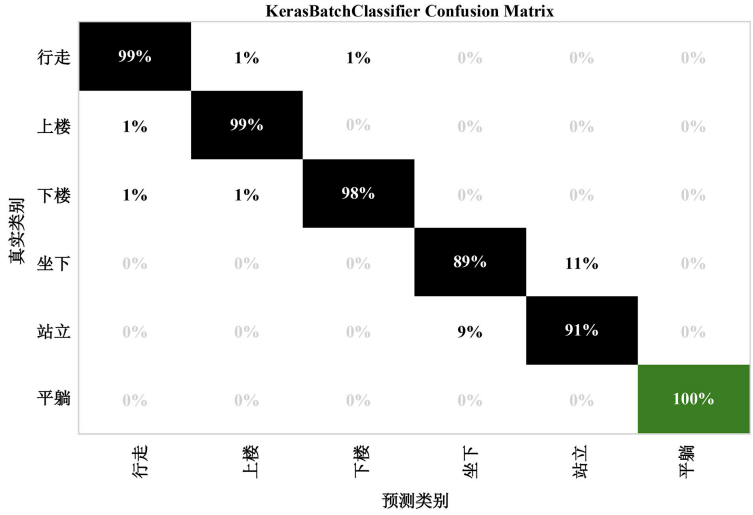

Figure 5. Classification confusion matrix on the UCI-HAR
图 5. 基于 UCI-HAR 的分类混淆矩阵

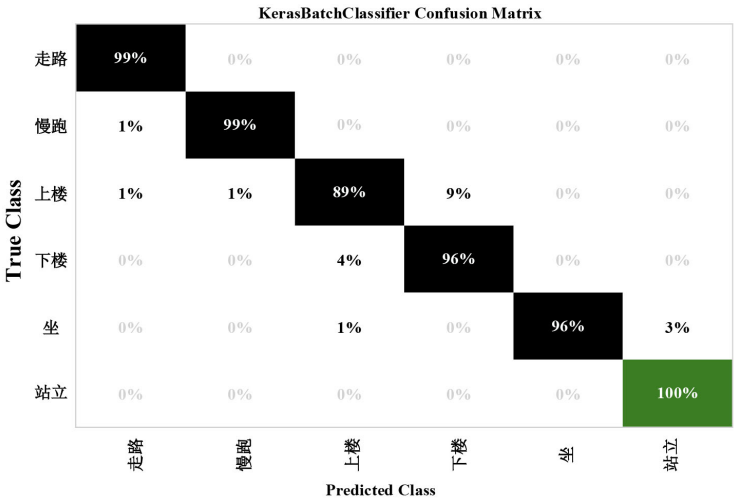

Figure 6. Classification confusion matrix on the WISDM
图 6. 基于 WISDM 的分类混淆矩阵

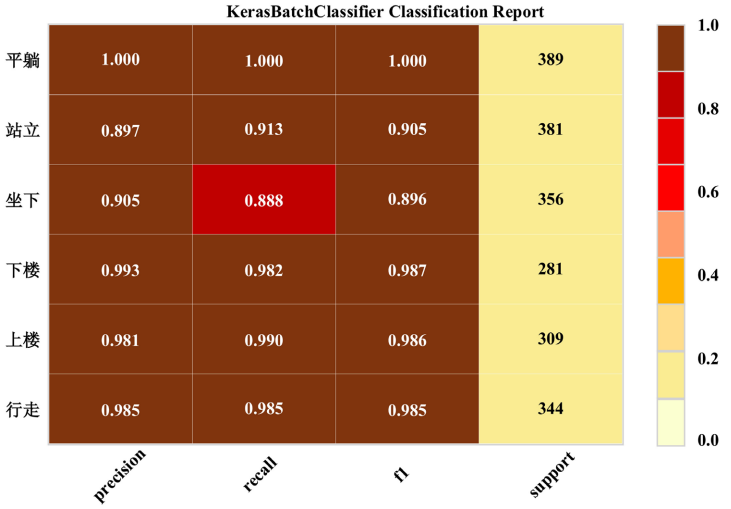

Figure 7. Classification report chart on the UCI-HAR
图 7. 基于 UCI-HAR 的分类报告图表

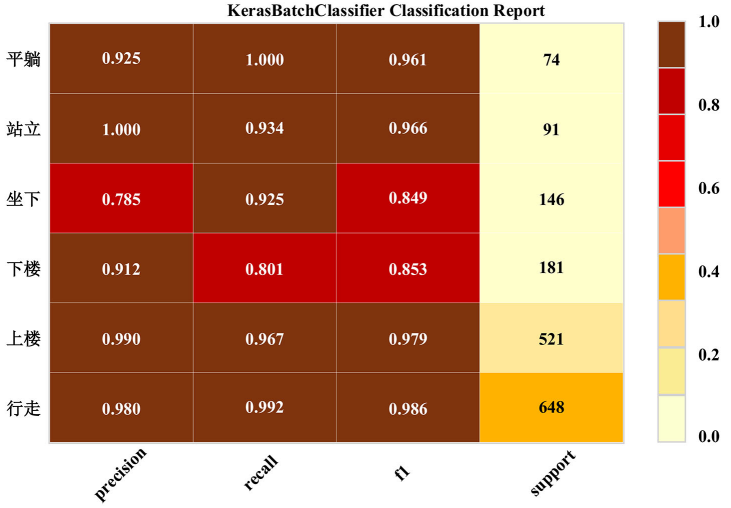

Figure 8. Classification report chart on the WISDM
图 8. 基于 WISDM 的分类报告图表

为进一步验证模型性能，在相同实验场景下将本文所提出的 Transformer-Capsule 模型与 Yang 等人提出的 CNN 模型[27]、DeepConvLSTM 模型[28]进行比较。所有结果均通过准确率、精度、召回率和 F1 分数验证，以确保后续比较的公平性与一致性。表 6 和表 7 展示了上述深度学习模型的评估结果。本文所提出的模型在 UCI-HAR 和 WISDM 数据集上的性能均超越另外两个模型。这些结果证明了所提方法在不同公共数据集上的优越性能。

Table 6. Performance of three models on UCI-HAR
表 6. 三个模型在 UCI-HAR 上的性能

UCI-HAR	准确率	精度	召回率	F1 分数
CNN	92.4%	92.0%	92.5%	92.1%
DeepConvLSTM	93.6%	93.4%	93.5%	93.2%
Transformer-Capsule	96.0%	95.7%	96.0%	95.9%

Table 7. Performance of three models on WISDM
表 7. 三个模型在 WISDM 上的性能

WISDM	准确率	精度	召回率	F1 分数
CNN	89.8%	89.6%	89.8%	89.2%
DeepConvLSTM	90.7%	90.4%	90.6%	90.3%
Transformer-Capsule	96.5%	96.4%	96.5%	96.4%

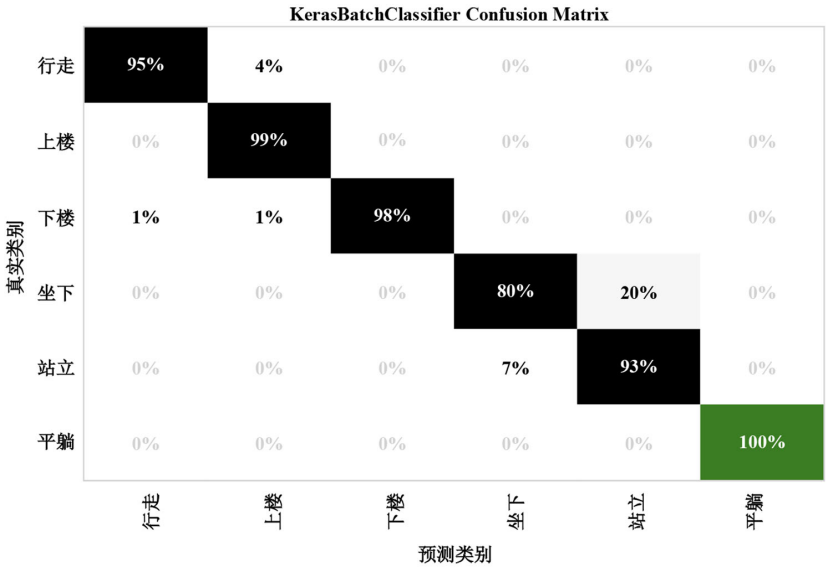

Figure 9. Classification confusion matrix of the standalone Transformer model
图 9. 独立 Transformer 模型的分类混淆矩阵

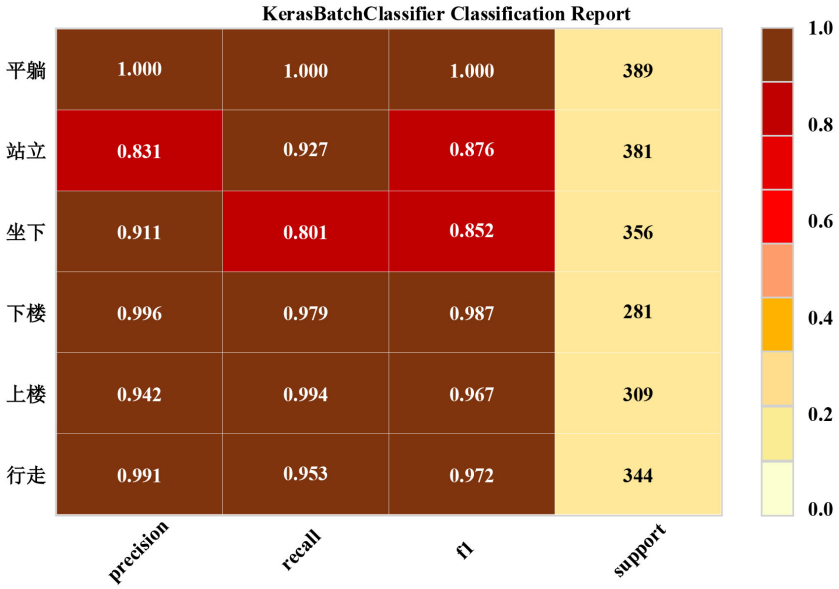

Figure 10. Classification report chart of the standalone Transformer model
图 10. 独立 Transformer 模型的分类报告图表

本研究提出了一种基于 Transformer 与 Capsule 网络的混合架构。为深入探究模型的有效性，我们在

统一实验环境下设计了消融实验：通过固定门控卷积以及模块参数配置，分别使用独立 Transformer 以及独立 Capsule 网络在 UCI-HAR 数据集上进行基准测试。实验结果显示如图 9~12，独立 Transformer 分支取得 94.17%的准确率，Capsule 网络分支达到 93.67%的准确率，而混合模型则展现出显著优势，准确率跃升至 96.0%。表 8 展示了消融实验的评估结果。在综合评估指标方面，混合模型在精度、召回率和 F1 分数上均相较于单一 Transformer 模型和单一 Capsule 模型有所提升。这些实验数据充分表明，通过有效整合 Transformer 模型在全局特征以及 Capsule 网络在局部特征上的各自优势，混合架构能够产生显著的协同效应，为神经网络结构设计提供了新的优化路径。

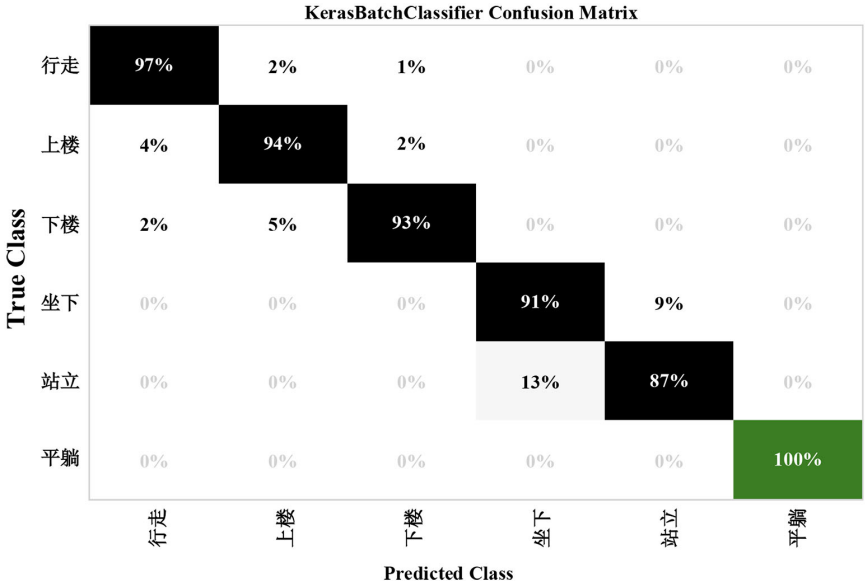

Figure 11. Classification confusion matrix of the standalone Capsule model
图 11. 独立 Capsule 模型的分类混淆矩阵

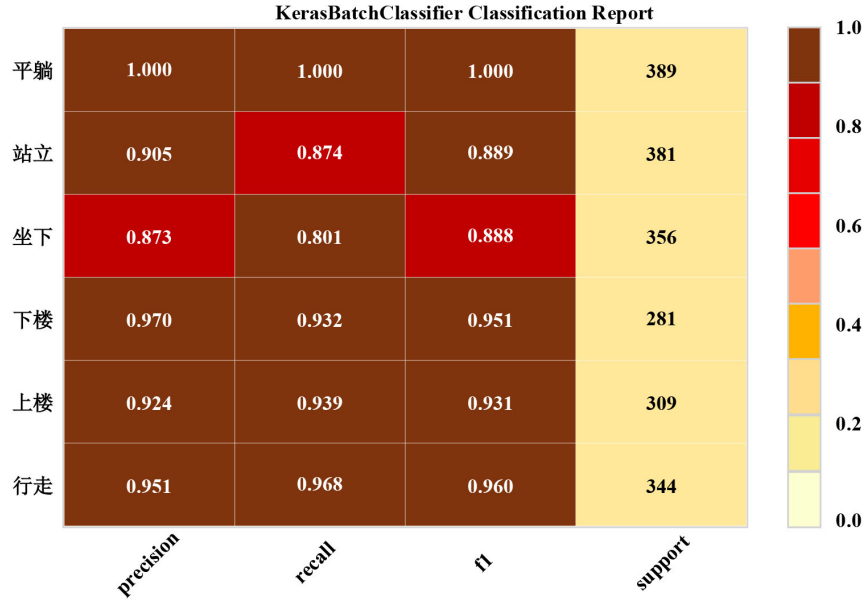

Figure 12. Classification report chart of the standalone Capsule model
图 12. 独立 Capsule 模型的分类报告图表

Table 8. Performance of three models on UCI-HAR
表 8. 三个模型在 UCI-HAR 上的性能

UCI-HAR	准确率	精度	召回率	F1 分数
独立 Transformer	94.1%	94.1%	94.1%	94.1%
独立 Capsule	93.6%	93.4%	93.6%	93.3%
Transformer-Capsule	96.0%	95.7%	96.0%	95.9%

4. 结论

本研究针对传统深度学习算法在人体活动识别任务中时序动态建模与空间关系捕捉的局限性，创新性地在人体活动识别中提出了一种基于 Transformer 与胶囊网络的混合架构。通过 Transformer 的多头注意力机制有效捕获传感器数据的全局时序依赖特征，结合胶囊网络的动态路由机制精准提取局部空间关联特性，实现了对人体复杂活动模式的多层次特征表征。实验验证表明，本模型在 UCI-HAR 和 WISDM 数据集上分别取得 96.0%和 96.5%的识别准确率，相较单一 Transformer 架构和独立胶囊网络模型在 UCI-HAR 数据集上分别提升 1.9%和 2.4%的识别准确率，同时显著优于 CNN、LSTM 等传统深度学习模型。实验结果充分验证了所提模型在时空特征协同学习方面的优越性，其通过异构网络的优势互补机制，在保证模型轻量化的基础上实现了端到端特征学习，避免了传统方法中复杂的特征工程需求。与近期同类研究相比，本方法在识别精度与模型鲁棒性方面均展现出竞争优势，为可穿戴设备在智能健康监护等实际应用场景提供了更可靠的技术支撑。未来研究将重点优化模型计算效率，并拓展其在多模态传感器数据融合中的应用。

参考文献

[1] 陈金瑶, 李瑞祥, 王星, 等. 基于 DWT-VMD 混合信号分解技术的人体活动识别[J]. 数据采集与处理, 2024, 39(3): 736-749.

[2] Cornacchia, M., Ozcan, K., Zheng, Y. and Velipasalar, S. (2017) A Survey on Activity Detection and Classification Using Wearable Sensors. *IEEE Sensors Journal*, **17**, 386-403. https://doi.org/10.1109/jsen.2016.2628346

[3] Yatani, K. and Truong, K.N. (2012) BodyScope: A Wearable Acoustic Sensor for Activity Recognition. *Proceedings of the* 2012 *ACM Conference on Ubiquitous Computing*, Pittsburgh, 5-8 September 2012, 341-350. https://doi.org/10.1145/2370216.2370269

[4] Cagliyan, B., Karabacak, C. and Gurbuz, S.Z. (2014) Human Activity Recognition Using a Low Cost, COTS Radar Network. 2014 *IEEE Radar Conference*, Cincinnati, 19-23 May 2014, 1223-1228. https://doi.org/10.1109/radar.2014.6875784

[5] Yang, X. and Tian, Y. (2017) Super Normal Vector for Human Activity Recognition with Depth Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1028-1039. https://doi.org/10.1109/tpami.2016.2565479

[6] Liu, J., Shahroudy, A., Xu, D., Kot, A.C. and Wang, G. (2018) Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 3007-3021. https://doi.org/10.1109/tpami.2017.2771306

[7] Kitani, K.M., Okabe, T., Sato, Y. and Sugimoto, A. (2011) Fast Unsupervised Ego-Action Learning for First-Person Sports Videos. *CVPR* 2011, Colorado, 20-25 June 2011, 3241-3248. https://doi.org/10.1109/cvpr.2011.5995406

[8] Amer, M.R. and Todorovic, S. (2016) Sum Product Networks for Activity Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 800-813. https://doi.org/10.1109/tpami.2015.2465955

[9] Lin, W., Xing, S., Nan, J., Wenyuan, L. and Binbin, L. (2019) Concurrent Recognition of Cross-Scale Activities via Sensorless Sensing. *IEEE Sensors Journal*, **19**, 658-669. https://doi.org/10.1109/jsen.2018.2874506

[10] Lopez-Nava, I.H. and Munoz-Melendez, A. (2016) Wearable Inertial Sensors for Human Motion Analysis: A Review. *IEEE Sensors Journal*, **16**, 7821-7834. https://doi.org/10.1109/jsen.2016.2609392

[11] Chen, L., Hoey, J., Nugent, C.D., Cook, D.J. and Yu, Z. (2012) Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C* (*Applications and Reviews*), **42**, 790-808.

https://doi.org/10.1109/tsmcc.2012.2198883

[12] Margarito, J., Helaoui, R. and Bianchi, A.M. (2016) User-Independent Recognition of Sports Activities from a Single Wrist-Worn Accelerometer: A Template-Matching-Based Approach. *IEEE Transactions on Biomedical Engineering*, **63**, 788-796. https://doi.org/10.1109/TBME.2015.2471094

[13] Zhu, C. and Sheng, W. (2011) Wearable Sensor-Based Hand Gesture and Daily Activity Recognition for Robot-Assisted Living. *IEEE Transactions on Systems, Man, and Cybernetics—Part A*: *Systems and Humans*, **41**, 569-573. https://doi.org/10.1109/tsmca.2010.2093883

[14] Orbach, J. (1962) Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms. *Archives of General Psychiatry*, **7**, 218-219. https://doi.org/10.1001/archpsyc.1962.01720030064010

[15] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[16] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[18] Hinton, G.E., Krizhevsky, A. and Wang, S.D. (2011) Transforming Auto-Encoders. *Artificial Neural Networks and Machine Learning—ICANN* 2011, Espoo, 14-17 June 2011, 44-51. https://doi.org/10.1007/978-3-642-21735-7_6

[19] Kwabena Patrick, M., Felix Adekoya, A., Abra Mighty, A. and Edward, B.Y. (2022) Capsule Networks—A Survey. *Journal of King Saud University-Computer and Information Sciences*, **34**, 1295-1310. https://doi.org/10.1016/j.jksuci.2019.09.014

[20] Sabour, S., Frosst, N. and Hinton, G.E. (2017) Dynamic Routing between Capsules. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3859-3869.

[21] Liu, G., Reda, F.A., Shih, K.J., Wang, T., Tao, A. and Catanzaro, B. (2018) Image Inpainting for Irregular Holes Using Partial Convolutions. *Computer Vision—ECCV* 2018, Munich, 8-14 September 2018, 89-105. https://doi.org/10.1007/978-3-030-01252-6_6

[22] Iqbal, T., Xu, Y., Kong, Q. and Wang, W. (2018) Capsule Routing for Sound Event Detection. 2018 26*th European Signal Processing Conference* (*EUSIPCO*), Rome, 3-7 September 2018, 2255-2259. https://doi.org/10.23919/eusipco.2018.8553198

[23] Dauphin, Y.N., Fan, A., Auli, M. and Grangier, D. (2017) Language Modeling with Gated Convolutional Networks. *Proceedings of the* 34*th International Conference on Machine Learning* (*ICML*), Sydney, 6-11 August 2017, 933-941.

[24] Reyes-Ortiz, J., Oneto, L., Samà, A., Parra, X. and Anguita, D. (2016) Transition-Aware Human Activity Recognition Using Smartphones. *Neurocomputing*, **171**, 754-767. https://doi.org/10.1016/j.neucom.2015.07.085

[25] Kwapisz, J.R., Weiss, G.M. and Moore, S.A. (2011) Activity Recognition Using Cell Phone Accelerometers. *ACM SIGKDD Explorations Newsletter*, **12**, 74-82. https://doi.org/10.1145/1964897.1964918

[26] Anguita, D., Ghio, A., Oneto, L., Parra, X. and Reyes-Ortiz, J.L. (2013) A Public Domain Dataset for Human Activity Recognition Using Smartphones. *ESANN* 2013 *Proceedings*, *European Symposium on Artificial Neural Networks*, *Computational Intelligence and Machine Learning*, Bruges, 24-26 April 2013, 437-442.

[27] Yang, J., Nguyen, M.N., Li, X.L. and San, P.P. (2015) Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. *Proceedings of the* 24*th International Conference on Artificial Intelligence*, Buenos, 25-31 July 2015, 3995-4001.

[28] Ordóñez, F. and Roggen, D. (2016) Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, **16**, Article 115. https://doi.org/10.3390/s16010115