SkeFi：面向无线骨架动作识别的跨模态知识迁移

Shunyu Huang、Yunjiao Zhou、Jianfei Yang，*IEEE* 高级会员

arXiv:2601.12432v1 [cs.CV] 18 Jan 2026

摘要—基于骨架的动作识别利用人体姿态关键点对人体动作进行分类，与常规端到端动作识别相比，具有更强的泛化能力和互操作性。现有方案使用 RGB摄像头标注骨骼关键点，但在黑暗环境中的性能会下降，还会引发隐私问题，限制了其在智能家居和医院中的应用。本文探索使用非侵入式无线传感器，即激光雷达和毫米波，以此作为可行的替代方案来缓解这些挑战。本文解决了两个问题：（1）无线传感器模态的数据不足，难以训练出准确的骨架估计模型；（2）无线传感器得出的骨骼关键点比 RGB 更嘈杂，给后续动作识别模型带来很大困难。我们的工作 SkeFi 通过一种新颖的跨模态知识迁移方法，利用数据丰富的 RGB模态来弥补这些不足。我们提出增强型时序相关自适应图卷积（TC-AGC），并采用帧交互增强来克服缺失帧或不连续帧带来的噪声。此外，我们的研究强调了通过双重时序卷积增强多尺度时序建模的有效性。通过将 TC-AGC 与用于跨模态迁移的时序建模相结合，我们的框架能够从含噪无线传感器数据中提取准确的姿态和动作。实验表明，SkeFi 在毫米波和激光雷达上实现了最先进的性能。代码见 https://github.com/Huang0035/Skefi。

索引词—**无线人体感知、迁移学习、动作识别。**

I. 引言

动作识别在移动系统和泛在计算中至关重要，主要利用传统摄像头拍摄的视频数据（RGB图像序列）无需专用传感器即可对人类动作进行分类，从而增强其通用性。在动作识别研究中，基于骨架的动作识别（SAR）已成为一种稳健的方案 [1]–[3]，该方案利用深度学习分析骨架关节位置及其运动，以辨识和分类动作，并且能够适应光照变化、背景复杂度及视觉异质性。因此，动作识别的重要性日益凸显。目前，基于视觉的方案是SAR中基于深度学习的方法的主流，例如OpenPose [4]、Nuitrack [5]和PoseNet [6]。这些方法已广泛应用于许多多媒体领域，包括人机交互 [7]、虚拟现实 [8]、医疗保健 [9]等。

S. Huang和Y. Zhou就职于新加坡南洋理工大学电气与电子工程学院（e220035@e.ntu.edu.sg；yunjiao001@e.ntu.edu.sg）。

J. Yang 任职于新加坡南洋理工大学机械与航空航天工程学院，并兼任电气与电子工程学院职务。J. Yang 为通讯作者（jianfei.yang@ntu.edu.sg）。

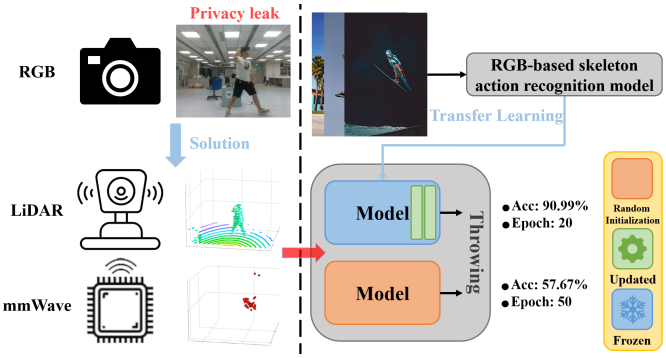


图 1. SkeFi 使用无创无线传感器通过骨骼运动识别人体动作，并通过跨模态迁移学习解决无线传感器数据不足的问题。

在城市监控、智能家居和自动驾驶汽车等应用中，使用 RGB摄像头部署动作识别系统往往会引发隐私问题，因为这可能在未经明确同意的情况下无意中收集个人数据。例如，监测行人流量时会无意捕捉到个人的面部细节；在智能家居中，居民的日常活动也可能被记录下来。如果缺乏有力的隐私保护措施，此类数据可能被滥用或泄露，从而危及用户隐私与安全。此外，光照、环境背景和视频分辨率等因素会影响关键点标注的准确性，进而影响骨架动作识别在实际场景中的应用。为应对这些挑战，近年来无线感知技术 [10]，日益受到关注，其中包括激光雷达 [11]、毫米波 [12]，和 WiFi [13]，等创新技术。这些技术利用非侵入式传感器，借助激光脉冲或电磁波的空间传播特性来检测和监测人体运动及位置，为传统的基于视觉的解决方案提供了可行替代方案，并解决了与光照条件、隐私和用户便利性相关的问题 [14]。

尽管非侵入式感知技术前景广阔，但要实现稳健的骨架动作识别（SAR），仍面临若干固有挑战：(1) 非侵入式测量设备通常比普及度高且资源丰富的 RGB 设备更昂贵。此外，这类技术的应用场景尚不广泛，导致数据稀缺。(2) 数据有限可能使复杂的深度学习模型容易过拟合。这些模型需要多样化的数据集才能进行稳健训练。然而，在数据受限的情况下，模型可能会过度重视无法代表总体趋势的非典型样本，从而削弱模型

的泛化能力。(3) 从非侵入式传感器获取的骨骼关键点比从 RGB 模态获取的噪声更大。在 RGB 领域，专用于关键点检测的深度学习工具 [15]经过精心设计，可实现高精度。相比之下，非侵入式姿态估计方法通常采用 ResNet-48 [16], 等通用深度学习架构，可能导致每一帧的结果噪声更大。此外，这些传感器容易因固有的硬件限制而发生帧丢失 [12],，从而降低基于无线感知的 SAR 任务的准确性和可靠性。

为克服上述挑战，我们提出跨模态迁移学习，以应对数据稀缺问题，并在存在缺失帧的情况下增强特征学习。迁移学习 [17], [18] 已被证明是防止在有限数据集上训练的深度学习模型过拟合的有效策略。传统上依赖 RGB 模态的骨骼运动识别领域，拥有规模庞大的数据集和丰富的现有知识。借助这些知识，我们可以将 RGB 领域的潜在模式融入模型，从而为非侵入式动作识别任务引入至关重要的先验知识。此外，为解决骨骼关键点估计中固有的噪声和帧丢失问题，我们采用了一种将空间域的结构相关性融入时间域的策略。引入时间相关性有助于捕捉跨帧的关节相关性，从而有效解决帧丢失问题。因此，我们通过引入时间相关性模块，增强了模型处理时间信息的能力，显著提升其特征学习能力。

本文提出 SkeFi，这是一种跨模态知识迁移框架，以基于骨骼的图卷积网络为骨干网络，旨在将知识从基于 RGB 的 Kinetics-Skeleton 数据集迁移至非侵入式数据集。SkeFi 适用于所有图卷积网络模型，并以 AGCN [19]为基础，融合空间建模与时序建模。为增强跨模态知识学习中非侵入式数据的特征提取能力，我们在空间建模中融入双重时序卷积，以便细致捕捉各个关键点的时间信息。此外，受 ESPNet [20]，启发，我们通过采用加法操作增强多尺度时间建模（MST），无需增加额外计算开销即可提升模型效能；我们将此方法称为 ESP-MST。SkeFi 在新型非侵入式数据集 MM-Fi 的 [21] RGB、毫米波和激光雷达模态上均表现出色，实现了有效的跨模态知识学习。SkeFi 的整体架构如图 1 所示。

我们的贡献总结如下：

- 我们提出了 SkeFi，这是一种新颖的骨架动作识别解决方案，利用无线传感器解决 RGB 模态带来的隐私侵犯和环境限制。据我们所知，这是首项利用非侵入式传感器和跨模态学习实现稳健骨架动作识别的工作。
- 为解决无线传感器数据不足的问题，我们提出将基于 RGB 的骨架动作识别中获得的见解迁移到非侵入式模式，以增强模型的泛化能力。
- 我们提出了 TC-AGCN 模型，用于解决非侵入式数据中的数据帧丢失和噪声问题。此外，我们改进的 ESP-MST 不仅提高了现有模型的准确率，还提升了迁移学习后的准确率。
- 我们在一个多模态骨架动作识别数据集上进行了大量实验，结果表明 SkeFi 具有稳健且出色的性能。

II. 相关工作

A. 骨架动作识别

传统的基于骨架的动作识别方法主要采用手工设计特征和关节关系。这类方法往往存在性能局限，主要是因为它们未能关注人体固有的关键语义联系。随着深度学习的发展，数据驱动方法日益受到关注，循环神经网络和卷积神经网络已成为主要的选择。然而，基于循环神经网络或卷积神经网络的方法都无法完全捕捉骨架数据的独特结构，因为骨架数据并非简单的向量序列或二维网格，而是从根本上体现为一种图结构 [1], [2], [22]。

近年来，许多图卷积网络（Graph Convolutional Networks，GCN）方法被用于从骨架数据中识别人类动作。图卷积网络擅长捕捉不规则结构数据中的嵌入特征，相较于基于循环神经网络和卷积神经网络的技术具有优势。具体而言，图卷积网络无需手动划分或制定遍历规则，因此性能优于早期方法。时空图卷积网络（ST-GCN）开创了基于图卷积网络的骨架动作识别领域 [3]。它直接以人体骨架序列为输入，并将图卷积网络扩展至时空域，从而更有效地提取动作的运动特征。此外，该方法还引入了一种新颖的权重分配策略，使模型能够对不同节点的特征进行更有区分度的学习。在 ST-GCN 的基础上，2s-AGCN 通过自注意力机制和可学习图残差掩码引入了图结构适应性 [19]。它采用双流框架利用关节特征和骨骼特征，从而提升整体性能。MS-G3D 引入多尺度卷积和时空图卷积算子，以增强特征提取能力，解决权重偏置问题，并有效捕捉复杂的时空关系 [23]。此外，MS-G3D 通过采用不同膨胀率的空洞卷积改进时间模块，扩大了感受野。CTR-GCN 引入通道拓扑细化网络，动态学习不同的拓扑结构，并有效聚合不同通道中的关节特征 [24]，，从而增强图卷积网络表征拓扑信息的能力。DG-STGCN 为骨架动作识别提出了一种新颖框架 [25]。在空间建模方面，它利用学习得到的亲和矩阵来辨别动态

图结构，摆脱传统的固定图结构。它还整合了一个用于时序建模的动态关节骨架融合模块，以实现自适应的多点合并。尽管这一方法取得了颇有前景的成果，但面向图卷积网络的技术仍存在局限，尤其是在鲁棒性、兼容性和可扩展性方面。为解决这些局限，Duan 等人提出了 PoseC3D [26],，这是一种开创性的骨架动作识别方法。PoseC3D 的独特之处在于，它采用三维热图堆叠而非图序列，作为人体骨架的基础表征。这种姿态估计方法更能抵御噪声，并在跨数据集场景中展现出更强的泛化能力。因此，PoseC3D 在主要数据集上实现了显著的性能提升 [27]–[30], ，无论是单独使用还是与 RGB 模态结合使用，均表现如此。然而，大多数基于图的骨架动作识别方法都是在干净、连续的骨架数据上开发和评估的；当关键点受到噪声干扰或帧缺失时，其假设可能不再成立，因此有必要探索超越逐帧图结构的时间相关性建模方法。

B. 无线人体感知

随着物联网（IoT）的兴起，人体感知技术引起了广泛关注，为创新生活场景铺平了道路。这些技术涵盖了诸多应用，例如智能医疗、安防监控和泛在交互。学者们深入研究了各种感知技术，以获取有关人类活动的时空信息和运动学信息 [31]–[33]。这些感知范式分为基于设备和无设备两类。以设备为中心的方法采用可穿戴传感器 [34], [35], ，确保其在各种环境中均具备多样化的功能。然而，定期佩戴这些设备的必要性可能会带来不便，或导致偶尔忘记佩戴。

相比之下，无线设备人体感知（WDHS）提供了一种非侵入式替代方案，利用摄像机 [36],激光雷达、毫米波和 WiFi [13] 等工具记录人体运动。过去十年来，以图像为中心的技术取得了长足发展，并催生了众多商业项目。然而，这类技术面临着固有挑战，例如容易受到环境照明影响，以及可能侵犯隐私。因此，研究者转而关注激光雷达 [11], 毫米波 [12], [37], 和 WiFi [38],[39], 等替代传感器，它们提供了可行的解决方案 [14]。例如，Li 等人利用激光雷达在各种场景中精准获取深度数据的优势 [11],，促成了 LiDARHuman26M 数据集的诞生。该数据集开辟了新的研究方向，着重研究远距离、数据驱动的激光雷达人体动作捕捉。与此同时，毫米波技术的发展使射频（RF）成像成为克服可穿戴传感器的限制 [40]的一种有前景的方案。毫米波雷达可输出高保真三维点云，借助边缘人工智能方法进行现场处理，从而描绘人体运动 [41]–[43]。在此基础上，An 等人 [12] 推出了面向智慧医疗的毫米波辅助康复系统（MARS），能够对 19 个不同的关节和骨骼结构进行三维渲染。

基于 WiFi 的感知主要通过信道状态信息（CSI）实现，也受到关注，主要应用于手势和动作检测 [44]–[46]。Jiang 等人推出了 WiPose [47],，利用人体骨架的先验知识构建姿态，并通过循环神经网络（RNN）和损失平滑实现流畅运动，但该方法仅适用于静态位置。随后，Ren 等人开发了 Winect [48],，采用增强的二维到达角、信号分离和关键点建模来跟踪动态活动。次年，Ren 等人提出 GoPose，通过整合发射器的空间分集和 Wi-Fi OFDM 子载波的频率分集，提高二维到达角的空间分辨率 [49]。

最近，Yang 等人推出了 MM-Fi，这是一个专为无线人体感知设计的新型非侵入式多模态 4D 人体姿态数据集 [21]。该数据集涵盖了上述主要无线感知机制，是用于无线人体姿态估计的最全面基准。它在推动多模态人体姿态研究方面发挥着关键作用，并帮助学者根据具体需求合理选择传感器配置。与 RGB 相比，无线感知流程通常会产生噪声更大的骨架，且训练数据有限，这促使我们从数据丰富的 RGB 中迁移先验知识，并设计一种能够抵御噪声和帧丢失的迁移学习架构。

C. 多模态迁移学习与知识蒸馏

近年来，人体活动识别（HAR）领域的跨模态迁移研究探索了通过时间同步和表征匹配来对齐异构传感模态。C3T [50]等无监督模态适配框架利用对比目标对齐多传感器嵌入，而 MM-Fi [21] 等同步语料库则为评估模态适配提供了严格的基准。与此同时，跨模态知识蒸馏侧重于通过共享嵌入空间和特定模态的蒸馏损失，将强模态中的监督迁移至较弱模态。RadarDistill [51] 通过多级对齐，将 RGB 视频中的空间特征蒸馏至雷达点云；CRKD [52]则在雷达信号与视觉嵌入之间引入跨分辨率对比蒸馏。与这些方法相比，SkeFi 在骨架表征层面执行迁移，无需直接对齐高维雷达图或点云。SkeFi 不采用显式蒸馏损失或投影头，而是采用参数高效微调策略：冻结编码与模态无关关节结构的早期图卷积块，同时对更深层的时间模块进行适配，以应对 LiDAR 和毫米波骨架序列所特有的噪声、稀疏性和帧丢失问题。该设计无需引入额外的损失函数，即可实现轻量且稳健的跨模态迁移。

基于这些见解，第 III 节将介绍 SkeFi 如何结合经 RGB 预训练的图骨干网络与自适应时间模块，实现跨无线骨架模态的稳健迁移。

III. SkeFi框架

A. 概述

SkeFi 是一种采用图卷积网络进行骨架动作识别的迁移学习框架。该框架包含空间和时间卷积模块，用于捕捉骨骼数据中固有的结构信息。本节将介绍相关符号，定义标准图卷积，然后按模块对当前的图卷积网络算法进行分类和讨论。最后，我们将详细说明改进后的模型，并介绍 SkeFi 框架中迁移学习的具体实现。

B. 问题表述

标准图卷积使用权重 W 进行特征变换。它通过 v_{tj} via a_{ij} 汇聚邻接顶点的表示，以更新节点 v_{ti} 的相应表示 $f_{out}(v_{ti})$，其中 $\{v_{ti} \mid t, i \in \mathbb{Z}, 1 \le t \le T, 1 \le i \le N\}$ 表示由 T 个时间步组成的序列中骨架的 N 个人体关节。此过程可形式化为：

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in N(v_{ti})} a_{ij} f_{in}(v_{ti}) \mathrm{W} \tag{1}$$

其中，$f_{in}(v_{ti})$ 表示节点 v_{ti} 的输入特征，a_{ij} 是邻接矩阵 A 中的一个元素，$\in \mathbb{R}^{N \times N}$ 用于描述图的结构。$a_{ij} = 1$ 表示在时间步 t 节点 v_{ti} 与 v_{tj} 之间存在连接，否则为 $a_{ij} = 0$。

对于静态方法 [3]，GCN 的拓扑结构在推理过程中保持不变，且通常由人工定义。相反，动态方法允许在推理过程中动态推断 GCN 的拓扑结构 [19], [53]，其中 a_{ij} 通常由模型根据输入样本确定。

C. 空间建模

ST-GCN [3]。在空间建模模块中，具体的输入特征图是一个 $C \times T \times N$ 张量，其中，C 表示通道数，T 表示时间长度，N表示顶点数。为实现 ST-GCN，我们将公式 1 转换如下：

$$\mathrm{f}_{out} = \sum_{k}^{K_v} \mathrm{W}_k (\mathrm{f}_{in} \mathrm{A}_k) \odot \mathrm{M}_k \tag{2}$$

其中，K_v 表示空间维度的卷积核大小。根据 ST-GCN 设计的分区策略，K_v 设为 3。$\bar{\Lambda}_k^{-\frac{1}{2}} \bar{\mathrm{A}}_k \bar{\Lambda}_k^{-\frac{1}{2}} = \mathrm{I}$，其中 A_k 类似于一个 $N \times N$ 邻接矩阵，其中元素 $\bar{\mathrm{A}}_k^{ij}$ 表示顶点 v_j是否属于子集 N_i。它从 f_{in} 中提取特定子集内相互连接的顶点，并与相应的权重向量对应。$\bar{\Lambda}_k^{ii} = \sum_j^{N} \bar{\mathrm{A}}_k^{ij} + \varepsilon$ 是归一化对角矩阵，其中 $\varepsilon = 0.001$ 用于避免出现空行。$\mathrm{M}_k \in \mathbb{R}^{N \times N}$表示可学习注意力图，用于突出每个邻接矩阵中的元素。$\mathrm{W}_k \in \mathbb{R}^{C_{out} \times C_{in} \times 1 \times 1}$ 表示权重矩阵，$\odot$表示点积。

2s-AGCN [19]。图卷积网络结构具有层次性，能够在各层中封装多层语义信息。在相比之下，ST-GCN 采用静态方法，在所有层中固定图的拓扑结构，从而降低了其灵活性，也削弱了对各层固有多层语义信息的建模能力。为解决这一局限，2s-AGCN 引入了自适应图卷积层。该方法在端到端学习框架中优化图的拓扑结构和网络的其他参数。为使图结构具备自适应性，2s-AGCN 将公式 2 修改如下：

$$\mathrm{f}_{out} = \sum_{k}^{K_v} \mathrm{W}_k \mathrm{f}_{in} (\mathrm{A}_k + \mathrm{B}_k + \mathrm{C}_k) \tag{3}$$

其中，A_k 与公式 2 中原始归一化的 $N \times N$邻接矩阵 A_k 对齐，表示人体的物理结构。第二个分量 B_k同样是一个 $N \times N$ 邻接矩阵。与 A_k不同，B_k 中的元素是参数化的，并在训练过程中与其他网络参数同时优化。由于 B_k 的取值不受约束，最终得到的表示图完全由训练数据学习得出。C_k表示一种数据依赖图，它会为每个样本学习一个不同的图。为确定两个顶点之间是否存在连接以及连接强度，2s-AGCN 使用归一化的嵌入高斯函数来计算它们之间的相似度：

$$f(v_i, v_j) = \frac{e^{\theta(v_i)^T \phi(v_j)}}{\sum_{j=1}^{N} e^{\theta(v_i)^T \phi(v_j)}} \tag{4}$$

其中，N 是顶点总数。在具体实现中，2s-AGCN 使用点积来衡量嵌入空间中两个顶点之间的相似度。给定一个维度为 $C_{in} \times T \times N$ 的输入特征图 f_{in}，通过两个卷积层（分别记为 θ 和 ϕ）将其嵌入到维度为 $C_e \times T \times N$ 的空间中。随后，将得到的嵌入特征图重新排列并整形为 $N \times C_e T$ 和 $C_e T \times N$ 矩阵。两者相乘得到一个相似度矩阵 C_k，其中每个元素 C_k^{ij} 反映顶点 v_i 与顶点 v_j 之间的相似度。然后将该矩阵的值归一化到范围 [0, 1]。C_k 可由公式 4 推导如下：

$$\mathrm{C}_k = softmax(\mathrm{f}_{in}{}^T \mathrm{W}_{\theta k}^T \mathrm{W}_{\phi k} \mathrm{f}_{in}) \tag{5}$$

为了保持原始模型的性能，2s-AGCN 不会用 B_k 或 C_k 替换 A_k。相反，它会初始化两者参数，并在训练过程中采用加法，以增强模型的适应性。

D. 跨模型迁移的空间建模改进

为缓解非侵入式数据经关键点估计后产生的噪声和帧信息丢失问题，我们的方法引入结构相关性原理来改进模型。结构相关性为理解图的拓扑结构提供了重要见解。然而，在本文提出的跨模型知识迁移方法中，现有模型忽略了源数据的时间相关性，这可能导致潜在模式丢失。因此，我们引入两个时间卷积，以增强缺失帧与正常帧之间的交互，改善序列的整体

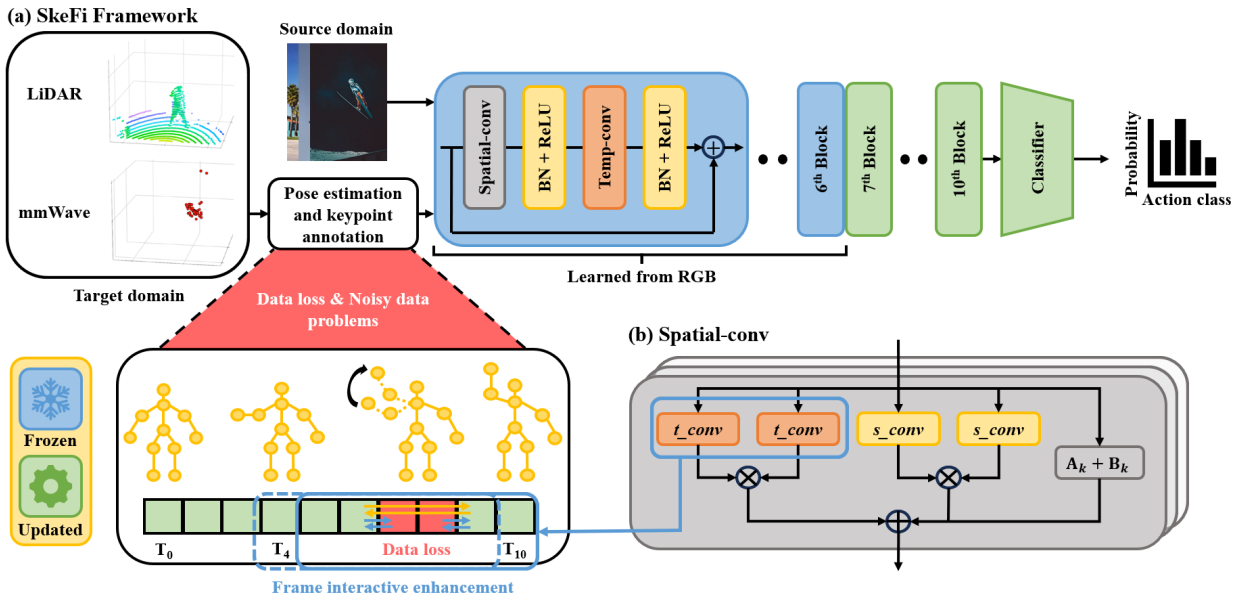


图2。（a）SkeFi，一种迁移学习框架。该框架使用源域中的RGB模态训练模型，并通过冻结参数将先验知识迁移到目标域。绿色和蓝色分别表示可训练参数和冻结参数。（b）TC-AGC。我们提出了一个时间相关性模块（如蓝色方框所示），通过促进缺失帧与正常帧之间更好的交互，增强帧序列的连通性。$\oplus$ 表示逐元素求和。$\otimes$ 表示矩阵乘法。

连通性。时间相关性的自适应空间建模（TC-AGC）如图2(b) 所示。图中，目标域在骨架估计后，T_7 和 T_8 处的帧丢失导致举手动作不完整。关键动作特征的此类丢失可能造成分类错误。因此，我们引入时间相关模块，以维持动作过程的连贯性并避免动作分类错误。

为将时间信息融入自适应空间图，我们使用公式4中的高斯函数构建D_k 。D_k 在形式上与C_k （公式5）相似——二者都采用双线性相似度，随后进行逐行Softmax——但 θ 和 ϕ的具体实现有所不同：D_k 使用沿时间轴采用相同填充的时间卷积替代 1×1 9×1 ，以保留序列长度T，并在亲和度估计之前注入局部时间归纳偏置。因此，公式3中的输出特征变为：

$$\mathrm{f}_{out} = \sum_{k}^{K_v} \mathrm{W}_k \mathrm{f}_{in}(\mathrm{A}_k + \mathrm{B}_k + \mathrm{C}_k + \mathrm{D}_k) \qquad (6)$$

E. 时序建模

为捕捉骨架序列中的时间动态并考量动作中的特定运动，必须将节点特征信息传播到序列中的其他骨架。现有研究可分为单尺度时间建模和多尺度时间建模。

单尺度时间建模 (SST)。在单尺度时间建模中，由于每个顶点都有固定数量的邻居（2 个，对应连续两帧中的关节），因此可以像传统卷积操作一样使用图卷积。具体而言，在由空间建模得到的特征图上执行 a$K_t \times 1$ 卷积

其中，K_t 表示时间维度上的卷积核大小，通常设为 9。

多尺度时间建模（MST）。多尺度时间建模采用多个并行卷积块来刻画输入骨架中关节之间的时间关系。该结构包含四个分支，每个分支都整合了一个 1×1 卷积以降低通道维度。前两个分支各包含两个具有不同膨胀率的时间卷积，且卷积核大小固定为 3×1。第三个分支采用最大池化操作，第四个分支则采用 1×1 卷积。将四个分支的结果拼接，得到输出。该方法利用不同的膨胀率，无需使用更大的卷积核即可扩大感受野。然而，引入空洞卷积可能会产生网格效应，破坏骨架信息的连续性。受ESP模块启发 [20]，我们在最终连接之前，将前两个采用不同膨胀率的分支所产生的时间卷积结果进行整合。改进后的方法不会增加多尺度时间建模的复杂度，并能有效缓解网格效应。我们将改进后的多尺度时间建模简称为ESP-MST。

F. 跨模型知识迁移框架：*SkeFi*

在我们的方法中，每个关节的邻域覆盖整个人体骨架图，这遵循了先前的研究；相关研究证明，全身感受野对动作建模十分有效 [19]。完整的网络由十个图卷积块组成，后接全局平均池化和一个 Softmax 分类器，用于最终的动作预测。

对于跨模态迁移，我们采用 Kinetics-Skeleton [27] 作为源域，并采用 MM-Fi [21]——包括其 RGB、

激光雷达和毫米波模态——作为目标域。这一选择基于三项共同特性：(1) 两个数据集使用完全相同的骨架关键点集合，因此无需进行特征重映射；(2) 两者都存在帧丢失和传感器噪声，因此从 RGB 骨架中学到的鲁健时空模式能够惠及无线模态；(3) Kinetics-Skeleton 提供覆盖 400 个类别的大规模监督，使模型能够学习可迁移至非侵入式感知任务的模态无关运动先验。

SkeFi 的核心组件之一是分块式层冻结策略，旨在保留可迁移的空间表征，同时使模型能够适配目标域的噪声特征。使用在 Kinetics-Skeleton 上预训练的权重初始化网络后，冻结前六个图卷积块——这些模块负责与模态无关的结构编码。仅对更深层的时序块和分类器在 MM-Fi 上进行微调。这种选择性适配可防止模型对稀疏且含噪的激光雷达/毫米波骨架过拟合，同时保持较高的参数效率。

我们采用包含交叉熵损失 [3], [19] 的两阶段训练方案。在第二阶段，将全连接分类头替换为与目标域类别数量相匹配的分类头。完整的迁移工作流程总结于算法 1 中，适配过程如图2(a)所示。

算法 1跨模态知识迁移

训练（Kinetics-Skeleton 特征 f_{kin}）
初始化空间划分权重 $\{\mathrm{W}_k\}$ 和骨干网络（TC-AGC →时间）；
将分类器设为 N_s 类；将所有参数设为可训练的；
对于 *epoch*，**执行**
$p_{kin} \leftarrow$前向传播(f_{kin});
计算交叉熵损失 $\mathcal{L}_{ce}$;
更新所有图卷积网络/骨干网络和分类器参数;
结束循环
保存预训练权重 $\theta^{\star}$;

SkeFi框架(f_{mm}, $\{\mathrm{W}_k\}$)
加载 $\theta \leftarrow \theta^{\star}$；将分类器替换为 N_t 个类别;
冻结浅层（早期模块）；初始化/解冻深层和分类器;
对于 *epoch* **执行**
$x \leftarrow$ 对齐帧(f_{mm})（例如，填充/采样至固定的 T）；
$p_{mm} \leftarrow$前向传播(x);
计算交叉熵损失 $\mathcal{L}_{ce}$;
仅更新未冻结的图卷积网络/骨干网络层和分类器;
结束循环
返回验证集上表现最佳的检查点;

函数 FORWARD(x)：
对每个模块执行 $x \leftarrow$ TC-AGC(x)； $x \leftarrow$ TEMPORAL(x)**结束循环**；**返回** SOFTMAX(FC(GAP(x)));

IV. 实验

我们首先使用标准训练模型对 MM-Fi 数据集进行了基准评估。随后，我们使用 Kinetics-Skeleton 数据集评估性能，接着采用 SkeFi框架开展从 Kinetics-Skeleton 到 MM-Fi 数据集的迁移学习实验。我们深入探讨了影响迁移学习结果的各项因素。

A. 数据集

MM-Fi [21]: MM-Fi 是一个新型的非侵入式多模态四维人体姿态数据集，面向无线人体感知。MM-Fi 整合了 1,080 个连续序列，包含超过 32 万个同步帧，覆盖五种感知模态：RGB图像、深度图像、激光雷达点云、毫米波雷达点云和 WiFi CSI数据。凭借这些特性，MM-Fi 成为目前可用的最全面的无线人体姿态估计基准。该数据集经过精心标注，涵盖二维/三维人体姿态关键点、动作分类、三维人体空间坐标以及计算得到的三维稠密姿态。MM-Fi 是首个提供五种不同非侵入式四维人体姿态估计（HPE）模态的数据集。

Kinetics-Skeleton [27]: Kinetics 是一个全面的人体动作数据集，包含 40 个类别中的 300,000 个视频片段。这些视频片段均来自 YouTube，呈现出广泛的多样性。值得注意的是，该数据集仅提供原始视频片段，不包含骨骼数据。研究在 [3] 中使用了公开可用的 OpenPose 工具箱 [4], , 估算每一帧中每个片段的 18 个关节的位置。在该数据集中，每个骨架由 18 个人体关节。每个人体关节由一个三维向量（x, y, c）表示，其中 x 和 y 表示二维关节坐标，c 表示置信度分数。此配置与 NTU RGB+D 数据集相符 [28], ，因为其输入数据表现为维度为 $3 \times 300 \times 18$的张量。该数据集划分为包含 240,000 个片段的训练子集和包含 20,000 个片段的验证子集。

B. 实验设置

本文实验使用 PyTorch 深度学习框架 [54]完成。对于 Kinetics-Skeleton 数据集，我们采用 [3], 中的输入张量尺寸，即包含 150 帧、每帧含两个对象，并镜像采用 [3] 所述的数据增强方法。训练采用随机梯度下降（SGD），Nesterov 动量设为 0.9。我们选择交叉熵损失进行识别，在搜索和训练阶段均将权重衰减设为 0.0001，批量大小设为 128，并在第 45_{th} 和 55_{th} 训练轮次将学习率 0.1 降低十倍。训练过程在第 65_{th} 训练轮次结束。

对于 MM-Fi 数据集的二维和三维人体姿态标注，我们首先采用大规模深度学习模型 HRNet-w48 提取二维关键点 [55], [56]。生成的 17 个关键点符合 COCO 协议 [57]。

随后，我们采用 [21], 中描述的方法，对这些二维关键点的两个视角进行三角测量，最终获得三维关键点。

根据识别出的三维关键点，我们收集了270个样本，涵盖10名受试者的27个动作。每个动作张量的大小为 $3\times297\times17$。为与Kinetics-Skeleton的输入格式保持一致，我们采取了两个步骤：对张量的最后三帧进行零填充，并引入一个 17_{th} 点，该点被确定为 $\mathcal{V}_{17}$ = 点与 $\left\{\frac{v_{0\,t}+v_{7\,t}}{2}1\le t\le 300\right\}$ 点之间的中点（即 0_{th} 7_{th} $_{,,}$）。经过处理后，每个动作张量的大小调整为 $3\times300\times18$。不同于NTU-RGBD等数据集，我们的270个样本具有一个显著特征：同一名受试者会多次重复执行相同动作，而不是执行不同动作。我们对这些重复动作进行分段并镜像处理，以增加样本数量。最终得到5,400个样本，每个样本的张量维度为 $3\times30\times18$。Kinetics-Skeleton的关节位置、我们估计的三维关键点及其自然连接关系如图3所示。红色节点表示新增节点。图4展示了数据处理后三种模态的可视化结果。

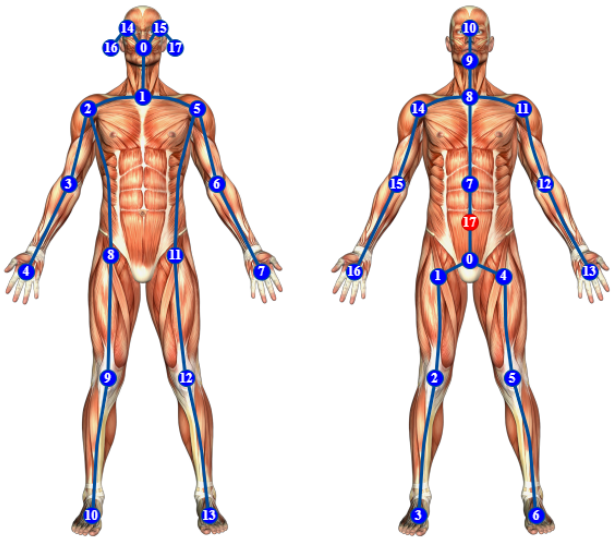


图3. 左侧示意图展示了 Kinetics-Skeleton 数据集的关节标签，右侧示意图展示了从 MM-Fi 数据集中提取的关节标签。

我们首先使用七名受试者作为训练集进行标准模型训练，其余三名受试者作为验证集。为验证 SkeFi 框架，我们使用一名受试者作为训练集，其余九名受试者作为验证集。对于 MM-Fi 数据集上的标准训练实验，我们将批量大小设为 16。学习率设为 0.05，并在第 10_{th} 和第 20_{th} 轮训练时缩小为原来的 1/10。训练过程在第 30_{th} 轮结束。对于迁移学习实验，学习率在第 5_{th}、第 10_{th} 和第 15_{th} 轮训练时缩小为原来的 1/10。训练过程在第 20_{th} 轮结束。此外，我们采用 [3] 中的数据增强方法来提高数据多样性。具体而言，我们将输入骨架序列补零至 40 帧，并通过随机旋转和平移对关节坐标进行细微调整。

C. 消融实验

为评估 SkeFi 中不同架构组件的作用，我们在 MM-Fi 数据集的三种模态——RGB、激光雷达和毫米波——上进行了消融研究。我们重点比较基于图卷积网络的模型，因为图卷积网络目前是基于骨架的动作识别领域中最成熟、应用最广泛的骨干网络系列，能够显式建模关节连接关系，这与 SkeFi 基于图的设计相契合。作为起点，我们纳入 ST-GCN [3] 作为代表性基线，并进一步考察 AGCN [19] 和时序相关变体 TC-AGCN，以评估更强图骨干网络的影响。为分析时序建模的作用，我们还将 MST 和 ESP-MST 模块集成到这些架构中。这种设置为分离各项设计选择的影响提供了一致的基础，比较结果汇总于表 I。

表I 使用训练集与测试集 7:3 的划分比例时，MM-Fi 三种模态上最先进方法的分类准确率比较。括号中的模块表示该模型最初采用的时序建模方法。

方法	RGB (%)	毫米波 (%)	激光雷达 (%)
STGCN（SST） [3]	85.97	63.70	45.37
AGCN（SST） [19]	92.72	72.62	71.27
AGC +MST [23]	93.64	77.11	78.25
AGC +ESP-MST	94.80	**78.37**	78.66
TC-AGCN（SST）	97.10	75.23	73.91
TC-AGC + MST	98.45	78.01	78.49
TC-AGC + ESP-MST	**98.52**	72.98	**79.21**

总体而言，尽管所有模型在 RGB 模态上均表现出色，但在毫米波和激光雷达这两种非侵入式模态上的表现仍不理想，这主要是由于关键点估计过程中的噪声和帧丢失所致。TC-AGC +ESP-MST 方法提升了准确率，在 RGB 模态和激光雷达上的准确率分别达到 98.52% 和 79.21%。尽管采用 ESP-MST 的 TC-AGCN 在毫米波模态上的表现略逊于 AGCN，但我们提出的 TC-AGCN 方法仍优于 ST-GCN 和标准 AGCN。这表明，引入时序相关模块能够有效提升模型在含噪声和帧丢失数据上的性能。引入 MST 能够有效提升模型性能。在不改变参数数量的情况下，ESP-MST 还能进一步提升性能，这表明 MST 和 ESP-MST 均能有效扩大模型的感受野，从而提升模型性能。

D. 迁移学习至 *MM-Fi* 数据集

源域/目标域选择与两阶段方案。我们选择 Kinetics-Skeleton 作为源域、MM-Fi 作为目标域，因为两者的关节数量相同，且都存在帧丢失/噪声，这使源域先验能够帮助目标域学习，同时避免手动进行特征重映射。我们先在 Kinetics-Skeleton 上进行预训练，再通过层冻结和重新初始化的分类器迁移到 MM-Fi。Kinetics-Skeleton 的评估结果如表 II 所示。

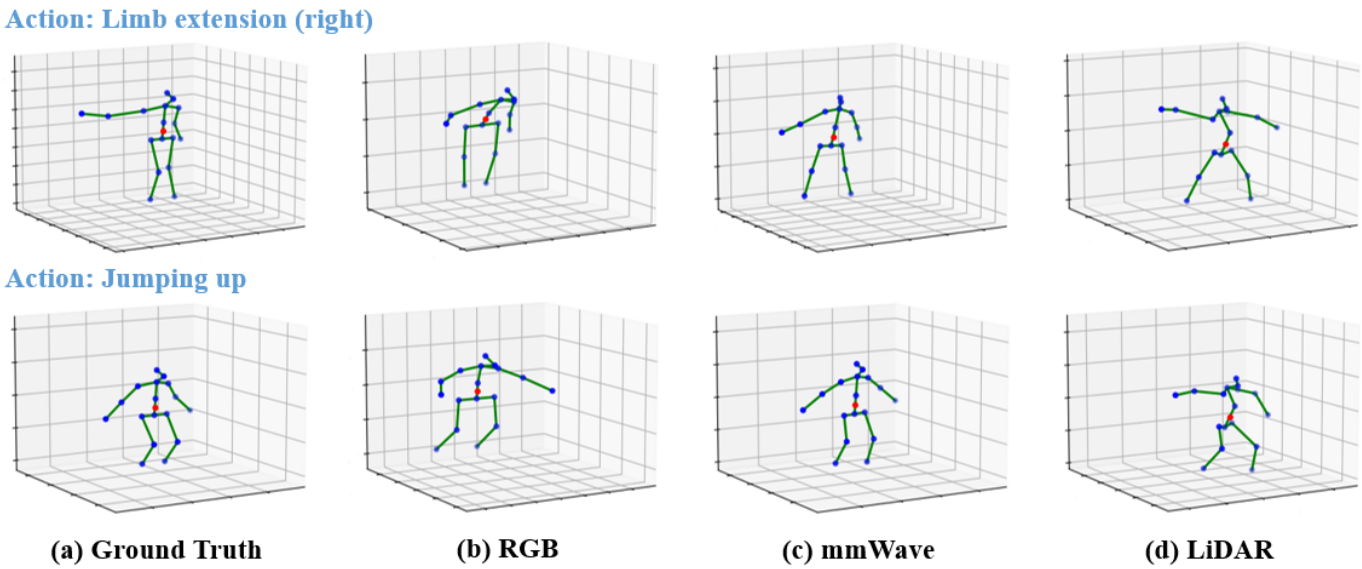


图 4. 使用三种模态进行人体姿态估计的可视化。红点表示我们添加的关键点。

表 II 在 Kinetics-Skeleton 数据集上与最先进方法的分类准确率比较。

方法	Top-1 （%）	Top-5 （%）
STGCN [3]	32.88	56.00
AGCN [19]	35.10	57.10
AGC +MST [23]	35.73	58.31
AGC +ESP-MST	35.80	58.41
TC-AGCN	35.59	58.19
TC-AGC +MST	35.58	58.35
TC-AGC + ESP-MST	35.45	**58.56**

我们发现，加入 MST 后，模型性能较基础 AGCN 提升了 0.63%。在参数数量不变的情况下，集成 ESP-MST 使性能提升了 0.7%。虽然引入 MST 和 ESP-MST 后，TC-AGCN 的 Top-1 准确率略有下降，但 Top-5 准确率进一步提高。

冻结策略。 微调的深度对迁移性能至关重要。遵循“早期层编码模态无关的骨架几何，而后期层适应目标域特性”这一普遍观点，我们冻结**第1–6层**，并微调第7–10层以及分类器 [58], [59]。为便于说明，我们将“Lk–10”定义为仅第 k–10 层（以及分类器）可训练的设置；“All”表示不冻结任何层。我们在 AGCN+ESP-MST 和 TC-AGCN+ESP-MST 上评估这一系列设置，并以 RGB 模态作为代表性示例。表III显示，**L7–10**处出现明显峰值，这支持了我们在迁移过程中冻结前六层、同时微调较高层模块和分类器的选择。

SkeFi框架的验证。为了评估本研究中所采用迁移学习的有效性，我们将训练集与验证集的比例设为 1:9。我们冻结了网络前六层的参数，以使用表 II 所示的预训练模型权重。此外，我们还进行了平行实验，在同一数据集上从头训练 AGCN模型

表III RGB模态上的微调深度与准确率。

微调层	AGCN+ESP-MST （%）	TC-AGCN+ESP-MST （%）
L10	80.78	81.52
L9–10	89.79	89.36
L8–10	90.76	91.85
L7–10	**93.97**	**94.03**
L6–10	92.16	92.45
L5–10	92.00	91.40
All	92.08	92.12

，以进行比较分析。我们将此次训练的初始学习率设为 0.05，并在第 20_{th} 和 40_{th} 训练轮次时将其降低十倍。训练在第 50_{th} 训练轮次后结束，其他训练参数在各实验中保持一致。这些实验的结果记录于表 IV，展示了迁移学习的影响与优势。

使用 SkeFi框架在 MM-Fi 数据集上进行的分类准确率比较。* 表示未使用 SkeFi框架。

方法	RGB (%)	毫米波 (%)	激光雷达 (%)
STGCN [3]	73.50	44.75	38.74
AGCN [19]	90.99	58.09	65.39
AGC + MST [23]	92.84	60.21	65.04
AGC + ESP-MST	93.97	62.22	66.54
TC-AGCN	91.64	58.67	65.08
TC-AGC + MST	93.02	62.78	65.31
TC-AGC + ESP-MST	**94.03**	**62.98**	65.02
*AGCN [19]	57.67	41.89	20.63
*TC-AGC + ESP-MST	62.61	46.23	49.49

在 RGB 和毫米波模态下，与 AGCN 的空间建模相比，我们提出的 TC-AGCN 在三种时间建模策略中均表现出更高的准确率。虽然 TC-AGCN 在 Kinetics-Skeleton 数据集上的表现并不特别突出，但在上述模态中始终达到最佳准确率。从头开始训练 AGCN

在三种模态下效果都不佳。图 5 直观展示了 AGCN 在三种迁移学习模式和直接训练下的训练动态。这些发现表明，在数据集有限的场景中，若缺乏充足的数据却采用复杂的网络架构，就无法有效学习潜在的数据模式。这不仅会影响分类性能，还会导致计算资源利用效率低下。

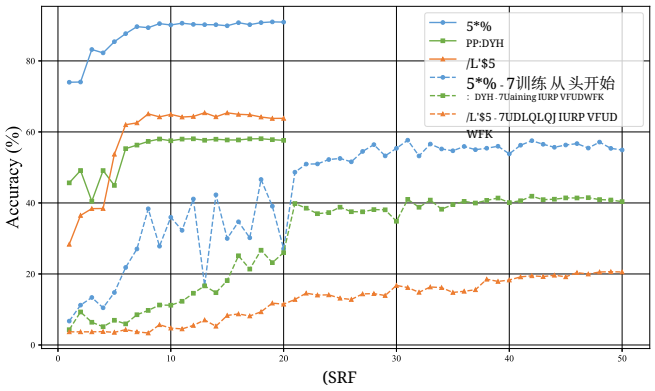


图 5. AGCN 在三种模态下的训练过程，其中虚线表示从头训练，实线表示迁移学习。

图 6 展示了采用标准训练方法和 SkeFi框架训练 AGCN 的过程。显然，在利用先验知识进行迁移学习时，尽管 SkeFi框架仅使用一名受试者作为训练集进行学习，其在 RGB模态下的性能仍与标准训练方法相当。对于毫米波和激光雷达等非侵入式数据模态，标准训练方法与 SkeFi框架之间的性能差异分别为 14% 和 6%。这些结果表明，即使训练数据有限，我们提出的 SkeFi框架也能通过迁移学习实现良好的动作分类性能。

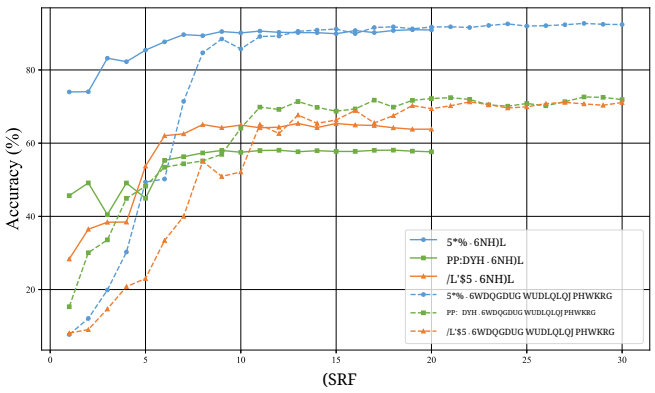


图 6. 采用标准训练方法和 SkeFi框架训练 AGCN 的过程，其中虚线表示使用标准训练方法，实线表示使用 SkeFi框架。

E. 影响迁移学习性能的其他因素

输入数据的帧数。通过考察 Kinetics-Skeleton 和 MM-Fi 数据集，我们发现每个样本中的受试者都会重复执行该动作多次。鉴于这一共同特征，我们遵循 [3]，并对 MM-Fi 采用与 Kinetics-Skeleton 相同的数据增强方法。与 Kinetics-Skeleton 从总计 300 帧中随机选取 150 帧进行训练不同，我们的研究结果表明，适当增加 MM-Fi 原始输入数据中的帧数（$3 \times 30 \times 18$），可以显著提升迁移学习的准确率。具体而言，在输入数据两端分别进行零填充，将其扩展至 40 帧（$3 \times 40 \times 18$），可获得最高准确率。不同帧数对应的结果如图 7 所示。

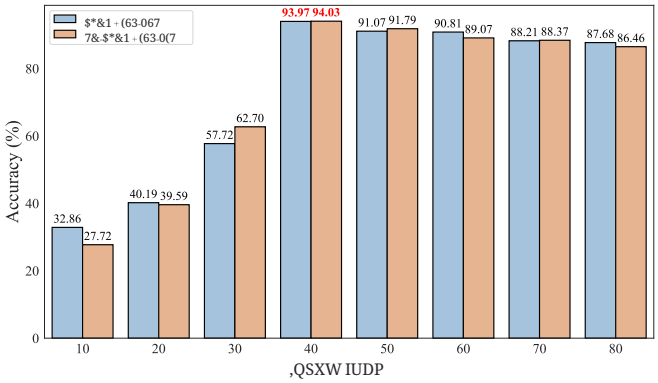


图 7. RGB模态下采用不同输入帧数的结果。

与使用原始输入数据相比，将输入数据扩展至 40 帧可使准确率提高约 30%。虽然输入数据扩展至超过 40 帧时准确率仍令人满意，但值得注意的是，帧数增加也会导致计算资源消耗相应上升。

V. 结论

本文提出了 SkeFi，这是一种用于动作识别的非侵入式、数据驱动的跨模态学习框架，并配以时序自适应图卷积网络（TC-AGCN），该网络融合了增强的多尺度时序建模方法。该方法有效应对缺失数据帧问题，并提升了时序建模的特征学习效率。在 MM-Fi 数据集上进行测试后发现，SkeFi 的表现优于从头开始训练的基线模型，展现出显著的计算效率。通过将 SkeFi框架与标准训练方法进行比较，可以发现，即使训练数据有限，SkeFi 仍能保持较高的分类性能。此外，我们的研究结果表明，采用最佳微调和数据处理策略可显著提升 SkeFi 的性能。

致谢

本研究由新加坡教育部一级资助RG83/25、RS36/24以及南洋理工大学的启动资助共同支持。

参考文献

[1] Jun Liu、Amir Shahroudy、Dong Xu和Gang Wang，《用于三维人体动作识别的带信任门控时空LSTM》，载于《计算机视觉——*ECCV 2016*：第*14*届欧洲会议论文集，第三部分》，*2016*年*10*月*11*日至*14*日，荷兰阿姆斯特丹。Springer，2016年，第816–833页。[2] Hong Liu、Juanhui Tu和Mengyuan Liu，《用于基于骨架的动作识别的双流三维卷积神经网络》，*arXiv*预印本 *arXiv:1705.08106*，2017年。[3] Sijie Yan、Yuanjun Xiong和Dahua Lin，《用于基于骨架的动作识别的时空图卷积网络》，载于《*AAAI*人工智能会议论文集》，2018年，第32卷。[4] Zhe Cao、Tomas Simon、Shih-En Wei和Yaser Sheikh，《利用部件亲和场进行实时多人二维姿态估计》，载于《*IEEE*计算机视觉与模式识别会议论文集》，2017年，第7291–7299页。[5] 《Nuitrack全身骨架跟踪软件——适用于Android、Windows、Linux、iOS、Intel RealSense和Orbbec的Kinect替代方案》，https://nuitrack.com/。[6] George Papandreou、Tyler Zhu、Nori Kanazawa、Alexander Toshev、Jonathan Tompson、Chris Bregler和Kevin Murphy，《迈向野外环境下的精准多人姿态估计》，载于《*IEEE*计算机视觉与模式识别会议论文集》，2017年，第4903–4911页。[7] Guilherme Maeda、Marco Ewerton、Gerhard Neumann、Rudolf Lioutikov和Jan Peters，《人机协作中的快速动作识别与轨迹生成的相位估计》，《国际机器人研究期刊》，第36卷，第13–14期，第1579–1594页，2017年。[8] Tamas Bates、Karinne Ramirez-Amaro、Tetsunari Inamura和Gordon Cheng，《基于虚拟现实演示在线同步学习与识别日常活动》，载于《*2017*年*IEEE/RSJ*国际智能机器人与系统会议（*IROS*）》。IEEE，2017年，第3510–3515页。[9] Amr Elkholy、Mohamed E Hussein、Walid Gomaa、Dima Damen和Emmanuel Saba，《人体动作表现中基于骨架的高效稳健质量评估与异常检测》，《*IEEE*生物医学与健康信息学期刊》，第24卷，第1期，第280–291页，2019年。[10] Jianfei Yang、Xinyan Chen、Dazhuo Wang、Han Zou、Chris Xiaoxuan Lu、Sumei Sun和Lihua Xie，《SenseFi：深度学习赋能的WiFi人体感知库与基准》，《*Patterns*》，第4卷，第3期，2023年。[11] Jialian Li、Jingyi Zhang、Zhiyong Wang、Siqi Shen、Chenglu Wen、Yuexin Ma、Lan Xu、Jingyi Yu和Cheng Wang，《LidarCap：利用激光雷达点云进行远距离、无标记的三维人体动作捕捉》，载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》，2022年，第20502–20512页。[12] Sizhe An和Umit Y Ogras，《MARS：面向智慧医疗的毫米波辅助康复系统》，《*ACM*嵌入式计算系统汇刊（*TECS*）》，第20卷，第5s期，第1–22页，2021年。[13] Jianfei Yang、Yunjiao Zhou、He Huang、Han Zou和Lihua Xie，《MetaFi：通过商用WiFi进行无设备姿态估计，以实现元宇宙虚拟形象模拟》，载于《*2022*年*IEEE*第*8*届世界物联网论坛（*WF-IoT*）》。IEEE，2022年，第1–6页。[14] Isura Nirmal、Abdelwahed Khamis、Mahbub Hassan、Wen Hu和Xiaoqing Zhu，《基于无线电的人体感知深度学习：最新进展与未来方向》，《*IEEE*通信调查与教程》，第23卷，第2期，第995–1019页，2021年。[15] Rıza Alp G¨uler、Natalia Neverova和Iasonas Kokkinos，《DensePose：野外环境下的稠密人体姿态估计》，载于《*IEEE*计算机视觉与模式识别会议论文集》，2018年，第7297–7306页。[16] Kaiming He、Xiangyu Zhang、Shaoqing Ren和Jian Sun，《用于图像识别的深度残差学习》，载于《*IEEE*计算机视觉与模式识别会议论文集》，2016年，第770–778页。[17] Fuzhen Zhuang、Zhiyuan Qi、Keyu Duan、Dongbo Xi、Yongchun Zhu、Hengshu Zhu、Hui Xiong和Qing He，《迁移学习综合综述》，《*IEEE*论文集》，第109卷，第1期，第43–76页，2020年。[18] Sinno Jialin Pan和Qiang Yang，《迁移学习综述》，《*IEEE*知识与数据工程汇刊》，第22卷，第10期，第1345–1359页，2009年。[19] Lei Shi、Yifan Zhang、Jian Cheng和Hanqing Lu，《用于基于骨架的动作识别的双流自适应图卷积网络》，载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》，2019年，第12026–12035页。[20] Sachin Mehta、Mohammad Rastegari、Anat Caspi、Linda Shapiro 和 Hannaneh Hajishirzi，“Espnet：用于语义分割的空洞卷积高效空间金字塔”，载于《欧洲计算机视觉会议（ECCV）论文集》，2018年，第552–568页。[21] Jianfei Yang、He Huang、Yunjiao Zhou、Xinyan Chen、Yuecong Xu、Shenghai Yuan、Han Zou、Chris Xiaoxuan Lu 和 Lihua Xie，“Mm-fi：用于多功能无线感知的多模态非侵入式4D人体数据集”，arXiv预印本 arXiv:2305.10345，2023年。[22] Yong Du、Wei Wang 和 Liang Wang，“用于基于骨架的动作识别的分层循环神经网络”，载于《IEEE计算机视觉与模式识别会议论文集》，2015年，第1110–1118页。[23] Ziyu Liu、Hongwen Zhang、Zhenghao Chen、Zhiyong Wang 和 Wanli Ouyang，“解耦并统一用于基于骨架的动作识别的图卷积”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第143–152页。[24] Yuxin Chen、Ziqi Zhang、Chunfeng Yuan、Bing Li、Ying Deng 和 Weiming Hu，“用于基于骨架的动作识别的通道级拓扑细化图卷积”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第13359–13368页。[25] Haodong Duan、Jiaqi Wang、Kai Chen 和 Dahua Lin，“DG-STGCN：用于基于骨架的动作识别的动态时空建模”，arXiv预印本 arXiv:2210.05895，2022年。[26] Haodong Duan、Yue Zhao、Kai Chen、Dahua Lin 和 Bo Dai，“重新审视基于骨架的动作识别”，arXiv预印本 arXiv:2104.13586，2021年。[27] Will Kay、Joao Carreira、Karen Simonyan、Brian Zhang、Chloe Hillier、Sudheendra Vijayanarasimhan、Fabio Viola、Tim Green、Trevor Back、Paul Natsev 等，“Kinetics人体动作视频数据集”，arXiv预印本 arXiv:1705.06950，2017年。[28] Amir Shahroudy、Jun Liu、Tian-Tsong Ng 和 Gang Wang，“NTU RGB+D：用于3D人体活动分析的大规模数据集”，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第1010–1019页。[29] Dian Shao、Yue Zhao、Bo Dai 和 Dahua Lin，“FineGym：用于细粒度动作理解的分层视频数据集”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第2616–2625页。[30] Khurram Soomro、Amir Roshan Zamir 和 Mubarak Shah，“UCF101：包含101类野外视频人体动作的数据集”，arXiv预印本 arXiv:1212.0402，2012年。[31] Kaixuan Chen、Dalin Zhang、Lina Yao、Bin Guo、Zhiwen Yu 和 Yunhao Liu，“基于传感器的人体活动识别的深度学习：概述、挑战与机遇”，《ACM计算调查（CSUR）》，第54卷，第4期，第1–40页，2021年。[32] Chenning Li、Zhichao Cao 和 Yunhao Liu，“深度人工智能赋能的泛在无线感知：综述”，《ACM计算调查（CSUR）》，第54卷，第2期，第1–35页，2021年。[33] Athira Nambiar、Alexandre Bernardino 和 Jacinto C Nascimento，“基于步态的行人重识别：综述”，《ACM计算调查（CSUR）》，第52卷，第2期，第1–34页，2019年。[34] Sumit Majumder、Tapas Mondal 和 M Jamal Deen，“用于远程健康监测的可穿戴传感器”，《Sensors》，第17卷，第1期，第130页，2017年。[35] Ebrahim Nemati、M Jamal Deen 和 Tapas Mondal，“用于长期应用的无线可穿戴心电传感器”，《IEEE通信杂志》，第50卷，第1期，第36–43页，2012年。[36] Cheng Zhang、Fan Yang、Gang Li、Qiang Zhai、Yi Jia ng 和 Dong Xuan，“MV-Sports：基于运动与视觉传感器集成的体育分析系统”，载于《IEEE INFOCOM 2018—IEEE计算机通信会议》，IEEE，2018年，第1070–1078页。[37] Fei Wang、Yizhe Lv、Mengdie Zhu、Han Ding 和 Jinsong Han，“XRF55：用于室内人体动作分析的射频数据集”，《ACM交互式、移动、可穿戴和泛在技术会议录》，第8卷，第1期，第1–34页，2024年。[38] Qifan Pu、Sidhant Gupta、Shyamnath Gollakota 和 Shwetak Patel，“使用无线信号进行全屋手势识别”，载于《第19届移动计算与网络年度国际会议论文集》，2013年，第27–38页。[39] Hamada Rizk 和 Ahmed Elmogy，“多用户智能环境中基于WiFi的自监督身份识别”，《Sensors》，第25卷，第10期，第3108页，2025年。[40] Zhicheng Yang、Parth H Pathak、Yunze Zeng、Xixi Liran 和 Prasant Mohapatra，“使用毫米波监测生命体征”，载于《第17届ACM移动自组织网络与计算国际研讨会论文集》，2016年，第211–220页。

[41] Haipeng Liu、Yuheng Wang、Anfu Zhou、Hanyue He、Wei Wang、Kunpeng Wang、Peilin Pan、Yixuan Lu、Liang Liu 和 Huadong Ma，“通过毫米波感知在智能家居场景中实时识别手臂手势”，*ACM*交互式、移动、可穿戴和泛在技术会议录，第4卷，第4期，第1–28页，2020年。[42] Zhen Meng、Song Fu、Jie Yan、Hongyuan Liang、Anfu Zhou、Shilin Zhu、Huadong Ma、Jianhua Liu 和 Ning Yang，“利用毫米波感知识别共存的多个人员的步态”，载于*AAAI*人工智能会议论文集，2020年，第34卷，第849–856页。[43] Shuai Wang、Dongjiang Cao、Ruofeng Liu、Wenchao Jiang、Tianshun Yao 和 Chris Xiaoxuan Lu，“通过联合学习解析动态毫米波雷达点云中的人体”，*ACM*交互式、移动、可穿戴和泛在技术会议录，第7卷，第1期，第1–22页，2023年。[44] Han Zou、Jianfei Yang、Hari Prasanna Das、Huihan Liu、Yuxun Zhou 和 Costas J Spanos，“面向准确、稳健的无设备人体活动识别的WiFi与视觉多模态学习”，载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，2019年，第0–0页。[45] Han Zou、Jianfei Yang、Yuxun Zhou、Lihua Xie 和 Costas J Spanos，“通过无监督对抗域适应实现稳健的WiFi无设备手势识别”，载于*2018*年第*27*届国际计算机通信与网络会议（*ICCCN*）。IEEE，2018年，第1–8页。[46] Dan Wu、Ruiyang Gao、Youwei Zeng、Jinyi Liu、Leye Wang、Tao Gu 和 Daqing Zhang，“Fingerdraw：利用WiFi信号进行亚波长级手指运动追踪”，*ACM*交互式、移动、可穿戴和泛在技术会议录，第4卷，27，2020 1，1第。–页。[47] Wenjun Jiang、Hongfei Xue、Chenglin Miao、Shiyang Wang、Sen Lin、Chong Tian、Srinivasan Murali、Haochen Hu、Zhi Sun 和 Lu Su，“基于WiFi的三维人体姿态构建”，载于第*26*届移动计算与网络年度国际会议论文集，2020年，第1–14页。[48] Yili Ren、Zi Wang、Sheng Tan、Yingying Chen 和 Jie Yang，“Winect：利用商用WiFi对自由形式活动进行三维人体姿态追踪”，*ACM*交互式、移动、可穿戴和泛在技术会议录，第5卷，第4期，第1–29页，2021年。[49] Yili Ren、Zi Wang、Yichao Wang、Sheng Tan、Yingying Chen 和 Jie Yang，“GoPose：利用WiFi进行三维人体姿态估计”，*ACM*交互式、移动、可穿戴和泛在技术会议录，第6卷，第2期，第1–25页，2022年。[50] Abhi Kamboj、Anh Duy Nguyen 和 Minh Do，“C3T：面向人体动作识别的跨模态时间迁移”，*arXiv*预印本 *arXiv:2407.16803*，2024年。[51] Geonho Bang、Kwangjin Choi、Jisong Kim、Dongsuk Kum 和 Jun Won Choi，“RadarDistill：通过从激光雷达特征进行知识蒸馏，提升基于雷达的目标检测性能”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第15491–15500页。[52] Lingjun Zhao、Jingyu Song 和 Katherine A Skinner，“CRKD：通过跨模态知识蒸馏增强相机-雷达目标检测”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第15470–15480页。[53] Maosen Li、Siheng Chen、Xu Chen、Ya Zhang、Yanfeng Wang 和 Qi Tian，“用于基于骨架的动作识别的动作-结构图卷积网络”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第3595–3603页。[54] Adam Paszke、Sam Gross、Soumith Chintala、Gregory Chanan、Edward Yang、Zachary DeVito、Zeming Lin、Alban Desmaison、Luca Antiga 和 Adam Lerer，“PyTorch中的自动微分”，载于*NIPS 2017*自动微分研讨会，2017年。[55] Ke Sun、Bin Xiao、Dong Liu 和 Jingdong Wang，“面向人体姿态估计的深度高分辨率表示学习”，载于*CVPR*，2019年。[56] Jingdong Wang、Ke Sun、Tianheng Cheng、Borui Jiang、Chaorui Deng、Yang Zhao、Dong Liu、Yadong Mu、Mingkui Tan、Xinggang Wang、Wenyu Liu 和 Bin Xiao，“面向视觉识别的深度高分辨率表示学习”，*TPAMI*，2019年。[57] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick，“Microsoft COCO：常见物体及其环境”，载于计算机视觉—*ECCV 2014*：第*13*届欧洲计算机视觉会议，苏黎世，瑞士，*2014*年*9*月*6*日至*12*日，论文集，第*V*部分 *13*。Springer，2014年，第740–755页。[58] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，“BERT：面向语言理解的深度双向变换器预训练”，载于*2019*年北方会议论文集

计算语言学协会美国分会：人类语言技术，第*1*卷（长篇与短篇论文），2019年，第4171–4186页。

[59] Jason Yosinski、Jeff Clune、Yoshua Bengio 和 Hod Lipson，“深度神经网络中的特征可迁移性有多强？”，神经信息处理系统进展，第27卷，2014年。



Shunyu Huang 于2022年获得华中科技大学人工智能与自动化学院工学学士学位，并于2024年获得南洋理工大学电气与电子工程硕士学位。目前，他是台积电的系统工程师，负责开发智能半导体制造系统。



Yunjiao Zhou 于2021年获得西南交通大学信息科学与技术学院工学学士学位，并于2022年获得南洋理工大学电气与电子工程硕士学位。目前，她正在南洋理工大学物联网实验室攻读博士学位。她的研究兴趣包括面向多传感器应用的多模态学习，以及智能家居中的人类感知。



Jianfei Yang（M’18–SM’25）是新加坡南洋理工大学机械与航空航天工程学院助理教授，同时受聘于电气与电子工程学院，并在该校领导多模态具身人工智能与机器人系统（MARS）实验室。他于2016年获得中山大学工学学士学位，2020年获得南洋理工大学博士学位，并荣获最佳论文奖。博士毕业后，他担任新加坡伯克利教育研究联盟（BEARS）高级研究工程师，该联盟隶属于加州大学伯克利分校；随后于2021年至2023年在南洋理工大学担任校长博士后研究员。2024年，他曾赴东京大学和哈佛大学担任访问学者。

他的研究聚焦于物理人工智能，探索人工智能如何赋能机器人、物联网设备和工业平台等物理系统，使其能够感知、理解并与真实世界互动。他曾入选《福布斯》亚洲30位30岁以下精英榜（2024年），并自2023年起入选斯坦福大学全球前2%顶尖科学家名单。他还在涵盖计算机视觉、具身人工智能及跨学科研究领域的十余项国际人工智能挑战赛中获奖。