扩大人类活动识别：合成数据生成和增强技术的比较评估

Zikang Leng
zleng7@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Archith Iyer
aiyer351@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Thomas Plötz
thomas.ploetz@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

人类活动识别（HAR）常常受到标记数据集稀缺的限制，这主要是因为现实世界中数据收集的高成本和复杂性。为了解决这个问题，最近的工作探索了通过跨模态传输生成虚拟惯性测量单元（IMU）数据。尽管基于视频和基于语言的管道都显示出潜力，但它们在假设和计算成本上有所不同。此外，它们相对于传统传感器级数据增强的有效性尚不明确。在本文中，我们对这两种虚拟 IMU 生成方法与经典数据增强技术进行了直接比较。我们构建了一个大规模的虚拟 IMU 数据集，涵盖了从 Kinetics-400 中提取的 100 种多样化活动，并在 22 个身体位置模拟传感器信号。使用四种流行的模型，在基准 HAR 数据集（UTD-MHAD、PAMAP2、HAD-AW）上评估这三种数据生成策略。结果表明，虚拟 IMU 数据在有限数据条件下尤其显著地提升了性能，相较于仅使用真实或增强数据。在文中，我们提供了关于选择数据生成策略的实用指导，并强调了每种方法的独特优劣势。

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Virtual IMU Data; Activity Recognition; Data Augmentation

ACM Reference Format:

Zikang Leng, Archith Iyer, and Thomas Plötz. 2025. 扩大人类活动识别：合成数据生成和增强技术的比较评估. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 介绍

人类活动识别 (HAR) 中的一个持续挑战是缺乏大规模、详细注释的 IMU 数据集。收集此类数据成本高且劳动密集，需要精确的传感器放置、多样化的参与者招募和人工标注，这往往导致数据集的活动覆盖范围有限且传感器设置受限 [7] 。因此，在这些数据集上训练的模型通常在用户和环境之间的泛化能力较差，阻碍了实际应用 [30] 。

为了解决标记 IMU 数据集稀缺的问题，最近的研究探索了跨模态迁移技术，通过更丰富的来源如视频和文本生成虚拟 IMU 数据。基于视频的方法如 IMUTube [18, 19] ，将活动的 2D RGB 视频——来源于 YouTube 等平台或如 Kinetics-400 [16] 等数据集——通过姿态估计和生物机械信号建模转化为虚拟 IMU 信号。相反，基于文本的流程如 IMUGPT [20, 22] 利用大型语言模型（LLMs）和文本到动作模型如 T2M-GPT [44] ，从自然语言描述中合成人体动作，生成语义多样的虚拟 IMU 数据。虽然两种方法在扩展性上远超传统 IMU 数据集，但它们在基本假设和计算需求上存在差异。

跨模态迁移方法可以被理解为数据增强的一种替代形式，它通过外部模态（如文本或视频）合成新的训练样本，而不是修改现有的 IMU 信号。通过生成反映活动多样性变化的虚拟 IMU 数据，这些方法丰富了训练数据，并在与现实世界数据一起使用时可以提升下游的人体动作识别（HAR）性能。虽然基于文本和视频的流程已经分别展示了各自的优势，但它们尚未在一起进行评估。目前，没有研究直接比较这两种方法。此外，尽管它们与传统传感器级数据增强有共同的目标，跨模态方法尚未系统地与其进行对标分析。这种缺乏比较的分析使得从业者在选择提升 HAR 模型的策略时缺乏指导。

在本文中，我们比较了这三种用于大规模 IMU 数据生成的方法。我们的结果表明，与仅使用真实数据或传统数据增强进行训练相比，结合虚拟 IMU 数据可以持续提高模型性能，特别是在真实数据稀缺的情况下。视频生成的虚拟 IMU 数据通常提供更准确的运动信息，而文本生成的虚拟 IMU 数据则贡献了语义和上下文的多样性；结合两者能使模型达到最佳性能。此外，我们发现虚拟 IMU 数据极大地惠及表现不足的活动类别，并能帮助缓解类别不平衡问题。我们的研究结果突出了跨模态传递方法的实际优势，并提供了关于如何利用它们的互补优势来进行可扩展的 HAR（人类活动识别）的见解。

我们的主要贡献是：

(1) 我们生成了一个大规模的虚拟 IMU 数据集，涵盖了来自 Kinetics-400 的 100 种不同的人类活动，在 22 个不同的传感器位置生成了模拟信号。该数据集将在获得接受后公开提供。
(2) 我们在三个 HAR 数据集上评估了三种数据生成方法—经典的传感器级增强，使用 IMUTube 的视频虚拟 IMU 数据生成，以及使用 IMUGPT 的文本虚拟 IMU 数据生成：PAMAP2 [32] ，HAD-AW [25] ，和 UTD-MHAD [4] 。
(3) 我们为研究人员和从业者提供实用指南，以便在资源限制、所需活动复杂性和部署场景的基础上选择最合适的数据生成技术。

2 相关工作

2.1 人类活动识别

HAR 传统上依赖于活动识别链（ARC），包括记录、信号处理、分割、特征提取和分类。早期的方法采用经典的机器学习模型，这些模型是在手工提取的时间域和频域统计特征上训练的。

随着深度学习的兴起，直接处理原始传感器信号的端到端模型变得普遍。卷积神经网络（CNNs）[26] ，长短期记忆网络（LSTMs）[8] ，以及混合架构如 DeepConvLSTM [29, 35] ，带注意力机制的 BiLSTM [42] ，以及基于 transformer 的模型 [11] ，通过捕捉局部时间动态、长程依赖关系和通道间关系，达到了最先进的结果。

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

然而，这些有监督的深度学习方法需要大量标注数据才能有效地执行 [30] 。为减轻这种依赖性，出现了替代学习范式。自监督学习（SSL）框架通过使用辅助目标（如对比预测或时间排序）进行预训练，然后在有限的标注数据上进行微调 [5, 6, 14, 27] 。迁移学习通过将一个领域或用户群中学习的表示适应到新环境中，进一步扩展了模型的泛化能力 [3, 10, 36] 。

在这些方向的基础上，我们的工作研究了跨模态传输技术生成替代模态虚拟 IMU 数据的有效性。

2.2 跨模态转换

跨模态传输方法旨在通过利用来自其他传感模态的现有标注数据集来解决标注 IMU 数据集的稀缺性。其核心思想是将来自源模态的数据——如视频 [18, 19, 24, 33] 、文本 [20, 22] 或动作捕捉 [37, 38] ——转换为虚拟 IMU 信号。

视频方法的一个突出的例子是 IMUTube [18, 19] ，它将二维视频转换为虚拟的 IMU 数据。该方法首先从视频中提取 2D 和 3D 的人体姿势，然后估计关节旋转和全身运动，由此生成虚拟 IMU 信号。这个流程支持从公开的视频来源（例如 YouTube）或大型视频数据集（如 Kinetics-400 [16] ）中进行可扩展的数据生成。生成的虚拟 IMU 数据已被证明显著提升了下游 HAR 模型的性能，特别是在涉及大规模移动的活动中，如移动或健身练习。然而，后续的评估显示，对于涉及细微动作的活动，例如驾驶 [21] ，其益处较为有限。

与此同时，基于文本的方法，例如 IMUGPT [20, 22] ，利用大型语言模型 (LLM) 生成日常活动的多样化文本描述。使用文本到动作的模型，这些文本描述随后被转换为 3D 动作序列，并最终成为虚拟 IMU 信号。文本提示的多样性旨在反映人类在现实生活中执行活动的多种方式。将模型训练在这一多样化的虚拟 IMU 信号集上，可以提升活动识别的泛化能力和鲁棒性。

视频和文本两种方法在提高模型性能方面都显示出潜力，特别是在真实 IMU 数据稀少或不平衡的情况下。然而，现有的评估都是分别考察这些不同的方式，使得直接比较变得困难。这项工作首次在一致的实验设置下，对视频和文本的虚拟 IMU 生成方法进行了正面对比的评估，从而能够更严格地评估它们的相对有效性。

2.3 数据增强

数据增强是一种广泛使用的策略，通过合成增加训练数据的多样性来提高模型泛化能力。在计算机视觉中，常见的转换包括裁剪、旋转、翻转和彩色抖动 [34] ，而基于传感器的 HAR 则采用了设计用于干扰从而增强传感器信号的领域特定增强技术。

典型的 IMU 信号增强方法包括坐标轴置换、添加高斯噪声、时间扭曲、信号缩放和旋转变换。这些方法给输入信号引入了可控的变化，并已被证明可以提高模型对跨个体和个体内变异性的鲁棒性 [2, 9, 12, 26] 。除了这些传统方法外，一些研究还探索了使用生成对抗网络（GANs）来合成全新的传感器数据作为一种数据增强形式 [15, 23] 。然而，总体而言，这种方法的成功仅为中等或不成功。

在本研究中，我们直接比较了经典的数据增强方法和两种跨模态迁移管道——IMUTube 和 IMUGPT，以评估它们对下游 HAR 模型性能的各自优势。

3 跨模态迁移和数据增强的比较研究

本文展示了一项比较研究，旨在评估三种用于扩展 HAR 训练数据的方法：(1) 基于文本的跨模态转移（IMUGPT）；(2) 基于视频的跨模态转移（IMUTube）；以及 (3) 经典的传感器级数据增强。我们使用 (1) 和 (2) 生成了大规模的虚拟 IMU 数据集，并通过对三个现实世界数据集中的四个 HAR 模型进行训练和评估，与传统的增强技术进行比较。Figure 1 提供了我们研究的概述。

3.1 大规模虚拟 IMU 数据生成

为了构建我们的大规模标记虚拟 IMU 数据集，我们利用了 Kinetics-400 数据集 [16] 作为参考视频源。Kinetics-400 包含 400 种多样的人体动作类别，每种类别由 200 到 700 个视频片段表示，最初是为了支持基于视频的人体动作识别模型的训练而收集的。

从 Kinetics-400 数据集中，我们选择了 100 个动作类别的子集，作为我们虚拟 IMU 数据集的基础，旨在捕捉多样化的人类活动，同时保持在我们的计算限制范围内。为了确保选择的代表性和平衡性，我们首先将这 400 个类别分为更广泛的语义类别，如体育、日常活动、职业任务和社交互动。在每个类别中，我们选择那些与现有 HAR 数据集中常见活动最大重叠的动作类别。这种方法使我们能够关注那些在语义上有意义且对基于传感器的识别有相关性的类别。

为了生成这些活动的虚拟 IMU 数据，我们采用了两种方法：一种基于文本描述，另一种基于视频输入（详细信息如下所示）。然后，我们比较了使用这两种方法生成的数据来提高下游 HAR 模型性能的效用。

3.1.1 **基于文本的虚拟 *IMU* 数据生成** 为了生成基于文本的虚拟 IMU 数据，我们采用了 IMUGPT 流程 [20, 22] ，该流程可以通过自动生成人类活动的自然语言描述，自主生成虚拟 IMU 数据，然后将文本描述转换为 3D 运动，随后生成虚拟 IMU 数据。核心思想是通过首先为每个活动类别生成多样化的文本描述集合，并将这些文本描述转换为运动序列，从而捕捉活动执行方式中固有的变动性。

对于每个活动，我们使用 GPT-4o [28] 生成了 500 个简短且多样化的描述（≤ 15 词），这些描述描绘了活动可能如何进行。文本描述生成过程是通过少样本提示策略指导的，在这个策略中，我们向 LLM 提供了示例描述和高层次的指导（例如，``the activity should involve only a single person, with minimal environmental detail''）。

生成的文本描述随后传递给运动合成模型 T2M-GPT [44] ，以产生 3D 人体运动序列。T2M-GPT 首先通过 CLIP [31] 嵌入文本，然后使用嵌入，变压器自回归地生成代码索引，这些索引随后通过一个学习的代码本解码为潜在向量。一个学习的解码器将这些潜在表示映射到包含 22 个人体关节的 3D 位置的运动序列。每个生成的文本描述被转换为一个运动序列（持续五到十秒），具体长度取决于变压器生成终止标记的时间。

我们使用逆运动学 [40] 估计了关节旋转，其中骨盆为根关节，这遵循了 3D 人体姿态估计中的惯例。利用这些旋转和根关节的平移，我们使用 IMUSim [41] 模拟虚拟 IMU 信号。IMUSim 为每个关节计算线性加速度和角速度，并引入真实的传感器噪声。通过这种方式，我们为如 Figure 1 所示的身体 22 个选定传感器位置生成了虚拟 IMU 数据。我们将由此产生的虚拟 IMU 数据集称为虚拟 IMUGPT 数据集。

为了生成基于视频的虚拟 IMU 数据，我们使用了 IMUTube 管道 [17, 18] 。对于每个选定的活动类别，我们从 Kinetics-400 数据集中获取相应的视频片段，并通过 IMUTube 进行处理。由于许多剪辑包含遮挡、跟踪错误或不稳定相机运动的片段，我们首先应用了一个自适应视频选择模块 [19] 来过滤掉低质量片段。该模块会自动删除包含无关内容、扭曲或不完整姿态、

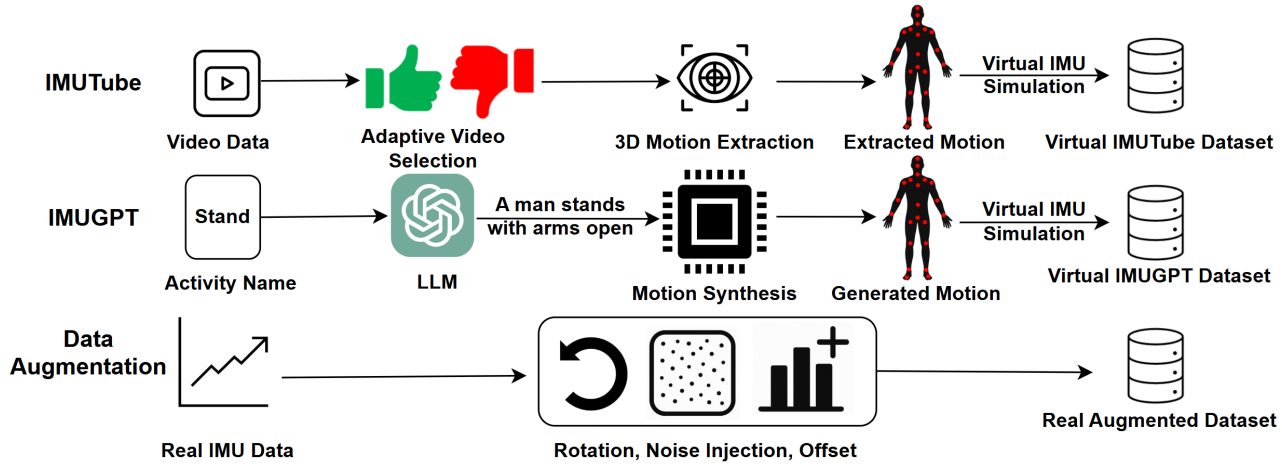

Figure 1: 对数据进行缩放的三种方法概述。

遮挡或显著前景/背景移动的序列，以确保仅保留适合姿态跟踪的片段。

经过筛选的视频片段随后被传递到主要的 IMUTube 管道。首先，使用基于转换器的二维姿势估计模型 ViTPose [39]，从每一帧中提取二维人体姿势。这些二维姿势随后使用最先进的三维姿势估计模型 MixSTE [43] 提升到三维。接下来，通过三维场景重建技术 [45] 估计相机自运动。结合重建的三维姿势和估计的相机运动，IMUTube 恢复了主体在空间中的全局运动轨迹。

与 IMUGPT 一样，我们使用 IMUSim [41] 来模拟虚拟 IMU 传感器读数。虚拟 IMU 数据是从人体的 22 个指定关节位置提取的，如 Figure 1 所示。从现在开始，我们将把这个生成的数据集称为虚拟 IMUTube 数据集。

对于给定的传感器数据段，我们应用三种形式的传感器级数据增强。每个数据段表示为一个矩阵 $\boldsymbol{X} \in \mathbb{R}^{T \times C}$，其中 T 是段中的样本数量，C 是总通道数量。每个三轴传感器贡献了 3 个通道，即 $C = 3S$，其中 S 是传感器的数量。我们应用的三种数据增强形式如下所示：

(1) 旋转增强。对于每个传感器 $s \in \{1, \ldots, S\}$，我们提取相应的三轴数据 $\boldsymbol{X}^{(s)} \in \mathbb{R}^{T \times 3}$，并应用绕 z 轴的固定旋转。旋转矩阵定义为：

$$\boldsymbol{R}_z = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \text{with } \theta = \frac{\pi}{6}.$$

然后为每个传感器计算旋转后的信号：

$$\boldsymbol{X}^{(s)}_{\text{rot}} = \boldsymbol{X}^{(s)} \cdot \boldsymbol{R}_z^{\top}.$$

所有旋转后的传感器信号被连接以形成 $\boldsymbol{X}_{\text{rot}} \in \mathbb{R}^{T \times C}$。

(2) 高斯噪声注入。我们添加从 $\mathcal{N}(0, \sigma^2)$ 中采样的元素级高斯噪声，标准差为 $\sigma = 0.05$，得到：

$$\boldsymbol{X}_{\text{noise}} = \boldsymbol{X} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.05^2).$$

(3) 传感器偏差模拟。为了模拟校准漂移或偏差，我们添加一个常数偏移 $\boldsymbol{b} \in \mathbb{R}^{1 \times C}$，其中每个分量从均匀分布 $\mathcal{U}(-0.1, 0.1)$ 中采样。偏差增强后的信号为：

$$\boldsymbol{X}_{\text{bias}} = \boldsymbol{X} + \boldsymbol{1}_T \cdot \boldsymbol{b},$$

Table 1: 训练数据集的大小（以分钟为单位）。IMUGPT 和 IMUTube 指的是虚拟数据集的子集，只包含与每个真实数据集重叠的活动。Real Augmented 表示应用数据增强后的大小。

Datasets	Real	IMUGPT	IMUTube	Real Augmented
UTD-MHAD	21	544	196	85
PAMAP2	53	272	42	212
HAD-AW	399	943	255	1,595

，其中 $\boldsymbol{1}_T \in \mathbb{R}^{T \times 1}$ 是一个全为 1 的向量，用于在所有时间步长上广播偏差。

对于给定的真实 IMU 数据集，增强应用于通过滑动窗口方法获得的所有分段传感器窗口。最终的训练数据集包括 $\boldsymbol{X}$、$\boldsymbol{X}_{\text{rot}}$、$\boldsymbol{X}_{\text{noise}}$ 和 $\boldsymbol{X}_{\text{bias}}$，它们沿批量维度连接用于训练。

3.2 数据集

我们使用了三个公开可用的 HAR 数据集进行实验：UTD-MHAD [4]、PAMAP2 [32] 和 HAD-AW [25]。UTD-MHAD 数据集包含来自八个受试者执行 27 种活动的记录，每种活动每个受试者重复四次。IMU 传感器在六项下半身活动中被放置在右腿上，剩下的 21 项上半身活动中则放置在右手腕上。PAMAP2 数据集包括来自 9 个受试者执行 18 种不同活动的数据，他们佩戴了三个分别置于胸部、惯用手腕和惯用踝部的 IMU 传感器。HAD-AW 数据集包含来自 16 名受试者的记录，每人佩戴一块 Apple Watch Series One 在右手腕上进行了 31 项活动，每项活动进行了十次。

对于每个数据集，我们仅选择与可在 Kinetics-400 数据集中获得的活动重叠的活动，以确保生成的虚拟 IMU 数据集包含这些活动的数据。最终的活动集列在 Table 2 中。我们将所有真实的 IMU 记录降采样到 20 Hz，以匹配虚拟 IMU 数据的采样率。

3.3 实验设置

3.3.1 **分类模型** 我们在 HAR 研究社区中使用四个常用模型进行研究：随机森林分类器、DeepConvLSTM [29]、带有自注意力的 DeepConvLSTM [35] 和带有注意力的 Bi-LSTM [42]。为了对 IMU 数据进行分段，我们采用滑动窗口方法，窗口大小为两秒，相邻段之间有一秒的重叠。随机森林模型在 ECDF 特征上训练 [13]，包括从窗口中提取的 15 个组件。DeepConvLSTM、带有自注意力的 DeepConvLSTM 和带有注意力的 BiLSTM 直

Table 2: 从每个数据集中选择与 Kinetics-400 中存在的活动类别重叠的活动类。

Dataset	Selected Activity Classes
HAD-AW	Riding a bike, Dancing ballet, Drawing, Driving car, Eating burger, Playing piano, Playing guitar, Jogging, Using computer, Washing hands, Writing, Playing violin, Taking a shower, Making bed
PAMAP2	Walking, Jogging, Biking, Using computer, Folding clothes
UTD-MHAD	Clapping, Throwing ball, Shooting basketball, Drawing, Boxing, Hitting baseball, Hitting tennis ball, Push, Jogging, Walking, Squat

接在原始 IMU 信号上训练，Adam 优化器用于训练最多 30 个周期，学习率通过 ReduceLROnPlateau 调度器动态调整 [1] 。

通过网格搜索选择超参数，包括学习率（范围从 10^{-6} 到 10^{-2} ）和权重衰减（从 10^{-4} 到 10^{-3} ）。对于 UTD-MHAD 和 PAMAP 数据集，我们对所有三个模型进行逐个主题的交叉验证。而对于 HAD-AW 数据集，由于发布的数据集中并不是所有受试者都执行了完整的活动集，我们采用 5 折分层交叉验证。对每个模型的交叉验证使用不同的随机种子重复三次，我们报告这些运行的平均宏 F1 得分及其标准差。

3.3.2 **训练配置** 我们评估以下训练配置：

(1) 仅真实数据：模型仅在真实 IMU 训练数据集上进行训练。
(2) Real + IMUGPT：训练集由真实 IMU 训练数据集与虚拟 IMU-GPT 数据集的一个子集组合而成。该子集仅包括与真实数据集中相同活动和传感器位置相对应的数据样本。
(3) Real + IMUTube: 类似于以上配置，此配置将真实 IMU 训练数据集与虚拟 IMUTube 数据集的一个子集相结合，子集是通过活动类型和传感器放置位置匹配的。
(4) Real + IMUGPT + IMUTube：该模型在一个组合数据集上进行训练，该数据集由真实 IMU 数据以及虚拟 IMUGPT 和虚拟 IMUTube 数据集的子集组成，这些子集在活动和传感器位置方面对齐。
(5) 真实 + 增强：训练数据集仅包括真实的 IMU 训练数据，并应用了数据增强技术。

对于每种训练配置，我们还进行了实验，仅使用 10 % 的真实 IMU 训练数据，以模拟获取标记 IMU 数据受限的情形。在所有情况下，评估均在未修改的真实 IMU 测试集上进行。

3.4 结果

Table 3 显示了在使用完整的真实 IMU 训练数据集时，各种训练配置的模型性能。使用完整的真实数据集时，我们发现将虚拟 IMU 数据——特别是使用虚拟 IMUTube 数据集或同时结合虚拟 IMUTube 和虚拟 IMUGPT 数据集——一般会导致下游模型性能的最大提升，特别是在 PAMAP2 和 UTD-MHAD 上。

(1) PAMAP2: 虚拟数据集 IMUTube 和 IMUGPT 的结合带来了最一致的性能提升。使用自注意力机制的 DeepConvLSTM 模型在 Real + IMUGPT + IMUTube 数据上训练，获得了最佳的宏观 F1 分数，为 90.29 % ，相较于仅在 Real、Real + Augmentation、Real + IMUGPT 和 Real + IMUTube 数据上训练，分别实现了 26.4 % 、44.83 % 、7.1 % 和 9.3 % 的相对提升。

Table 3: 使用完整真实数据的模型性能比较（宏观 F1）。绿色阴影表示每个数据集和模型组合的最佳训练配置。

	UTD-MHAD	PAMAP2	HAD-AW
Random Forest Classifier			
Real Only	65.26 ± 0.34	48.09 ± 1.68	55.68 ± 0.22
Real + IMUGPT	63.41 ± 0.49	66.22 ± 2.34	53.19 ± 0.07
Real + IMUTube	66.65 ± 0.41	86.46 ± 2.38	54.64 ± 0.10
Real + IMUGPT + IMUTube	65.06 ± 0.51	88.35 ± 2.28	52.94 ± 0.27
Real + Augmentation	67.37 ± 0.62	47.60 ± 1.78	55.34 ± 0.19
DeepConvLSTM			
Real Only	51.36 ± 1.19	48.47 ± 9.11	66.53 ± 0.20
Real + IMUGPT	52.47 ± 0.29	74.01 ± 0.54	64.54 ± 0.16
Real + IMUTube	54.23 ± 0.21	82.48 ± 3.18	65.18 ± 0.37
Real + IMUGPT + IMUTube	52.09 ± 0.26	86.59 ± 0.88	63.85 ± 0.36
Real + Augmentation	57.33 ± 1.00	55.84 ± 0.31	67.15 ± 0.33
DeepConvLSTM with Self-Attention			
Real Only	51.86 ± 0.67	71.42 ± 1.81	67.79 ± 0.11
Real + IMUGPT	49.82 ± 0.52	84.31 ± 4.74	65.63 ± 0.23
Real + IMUTube	57.37 ± 0.46	82.64 ± 2.11	65.94 ± 0.30
Real + IMUGPT + IMUTube	49.30 ± 0.54	90.29 ± 0.65	64.65 ± 0.23
Real + Augmentation	58.26 ± 1.69	62.34 ± 4.71	67.71 ± 0.36
BiLSTM with Attention			
Real Only	78.57 ± 0.18	31.07 ± 2.43	89.49 ± 0.25
Real + IMUGPT	60.09 ± 1.49	47.90 ± 8.70	82.70 ± 0.28
Real + IMUTube	84.51 ± 1.99	52.39 ± 6.53	81.19 ± 0.30
Real + IMUGPT + IMUTube	64.33 ± 7.17	69.37 ± 9.88	81.28 ± 0.27
Real + Augmentation	78.82 ± 0.78	35.58 ± 3.81	89.51 ± 0.22

(2) UTD-MHAD：对于随机森林分类器和 DeepConvLSTM 变体，在 Real + IMUTube 和 Real + Augmentation 上训练产生了可比的性能。然而，对于 BiLSTM 注意力模型，Real + IMUTube 上的训练实现了最高的宏观 F1 得分为 84.51 %，分别比 Real Only, Real + Augmentation, Real + IMUGPT, 和 Real + IMUGPT + IMUTube 提高了 7.6 %, 7.2 %, 40.6 % 和 31.4 %。
(3) HAD-AW：在所有训练配置中，性能相对稳定。添加虚拟 IMU 数据或应用数据增强技术相比单独使用未修改的真实 IMU 数据并没有带来显著的改进。

当仅使用真实 IMU 训练数据集的 10 % 时，Table 4 总结了模型在不同训练配置下的性能。总体而言，整合虚拟 IMU 数据——尤其是来自 IMUTube 数据集或同时使用两个虚拟 IMU 数据集——在所有模型和数据集上都带来了最大的改进。

(1) PAMAP2: 通过在 Real + IMUGPT + IMUTube 上训练的模型始终获得了最佳性能。例如，在这种配置上训练的 DeepConvLSTM 模型获得了最高的宏观 F1 分数 91.08 %，相较于仅 Real、Real + Augmentation、Real + IMUGPT 和 Real + IMUTube 分别显示出相对提高 217.1 %、132.1 %、17.9 % 和 15.4 %。
(2) UTD-MHAD：尽管由于真实 IMU 训练数据减少，总体得分较低，但使用虚拟 IMUTube 数据集训练的模型——无论是单独使用还是与虚拟 IMUGPT 数据结合使用——仍然优于其他配置。在真实数据加 IMUTube 数据上训练的随机森林分类器获得了最高的宏 F1 得分 61.48 %，分别比仅使用真实数据、真实数据加增强、真实数据加

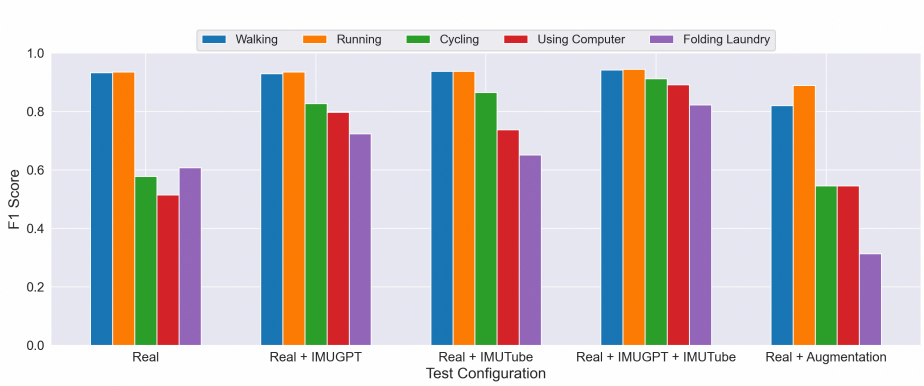

Figure 2: 不同训练配置下，带有自注意力机制的 DeepConvLSTM 模型在 PAMAP2 数据集上的类别宏 F1 得分。

Table 4: 使用真实数据的 10 % 进行模型性能（宏 F1）的比较。绿色阴影表示每个数据集的最佳训练配置。

	UTD-MHAD	PAMAP2	HAD-AW
Random Forest Classifier			
Real Only	36.13 ± 1.16	42.09 ± 4.89	33.88 ± 0.49
Real + IMUGPT	50.64 ± 0.38	77.22 ± 1.44	36.29 ± 0.37
Real + IMUTube	61.48 ± 0.69	85.30 ± 0.15	37.45 ± 0.37
Real + IMUGPT + IMUTUBE	60.65 ± 0.58	87.92 ± 2.01	39.23 ± 0.18
Real + Augmentation	53.82 ± 0.65	51.86 ± 7.35	33.80 ± 0.21
DeepConvLSTM			
Real Only	12.95 ± 1.19	28.72 ± 6.05	31.35 ± 0.05
Real + IMUGPT	30.71 ± 0.78	77.24 ± 2.27	34.58 ± 0.47
Real + IMUTube	36.40 ± 0.24	78.93 ± 2.02	37.86 ± 0.16
Real + IMUGPT + IMUTube	37.21 ± 0.44	91.08 ± 1.55	38.50 ± 0.38
Real + Augmentation	15.57 ± 4.12	39.57 ± 2.57	31.05 ± 0.04
DeepConvLSTM with Self-Attention			
Real Only	19.24 ± 0.22	37.34 ± 9.82	37.02 ± 0.35
Real + IMUGPT	29.39 ± 1.15	86.36 ± 3.16	36.88 ± 0.33
Real + IMUTube	40.92 ± 0.57	87.88 ± 0.16	40.52 ± 0.09
Real + IMUGPT + IMUTube	33.92 ± 0.41	90.27 ± 0.58	38.99 ± 0.19
Real + Augmentation	23.05 ± 1.66	54.07 ± 6.32	36.80 ± 0.48
BiLSTM with Attention			
Real Only	9.47 ± 0.90	20.90 ± 6.29	51.53 ± 0.30
Real + IMUGPT	12.61 ± 2.23	50.58 ± 3.02	45.81 ± 0.22
Real + IMUTube	37.42 ± 0.42	66.79 ± 7.16	52.88 ± 0.20
Real + IMUGPT+IMUTUBE	14.42 ± 1.22	59.64 ± 2.35	45.41 ± 0.25
Real + Augmentation	11.23 ± 0.55	28.43 ± 6.54	47.93 ± 0.28

IMUGPT 以及真实数据加 IMUGPT 和 IMUTube 提高了 70.2 %、14.2 %、21.4 % 和 1.4 %。

(3) HAD-AW: 性能提升较为温和但一致。使用 Real + IMUGPT + IMUTube 或 Real + IMUTube 的配置在大多数模型中实现了最高的宏 F1 分数，这表明当真实数据稀缺时，虚拟 IMU 数据变得尤其有利。

在完整和有限（10 %）的真实数据训练环境中，评估了三种数据扩展策略——IMUTube、IMUGPT 和经典数据增强——与仅在真实 IMU 数据上训练的模型进行比较。通过 IMUTube 和 IMUGPT 生成的虚拟 IMU 数据在各种数据集和训练条件下始终显著提升了模型性能。相比之下，经典数据增强产生了混合结果，并未能一直优于仅使用真实数据的基线，凸显了基于扰动技术在扩展数据多样性方面的局限性。

*跨模态与经典增强。*跨模态传输技术通常比经典的传感器级数据增强产生更大的性能提升。例如，在 PAMAP2 数据集中，结合来自 IMUGPT 和 IMUTube 的虚拟 IMU 数据，相对于仅使用真实数据的基线，其宏 F1 的相对提升达到 78 %，而经典增强仅有 4 % 的提升。在低数据条件下（10 % 实际数据），这种性能差距进一步扩大，其中 IMUGPT+IMUTube 实现了 163.3 % 的提升，而数据增强只有 35.5 %。这些结果表明，跨模态方法提供了比简单扰动现有信号所能实现的更丰富的数据多样性和语义。

在两个虚拟 IMU 数据源中，IMUTube 始终比 IMUGPT 表现稍好，特别是在有更复杂体育活动的数据集上。例如，在包含完整数据的 PAMAP2 上，Real+IMUTube 提供了 58.08% 的增益，而 Real+IMUGPT 为 40.65%。这可能是因为基于视频的姿态估计可以从视频中捕捉详细且真实的运动，而与之相比，基于文本的运动合成做不到。然而，将 IMUGPT 和 IMUTube 结合使用通常比单独使用任何一个取得更好的效果，这表明它们可以互为补充——IMUGPT 通过多样的生成文本描述丰富了多样性，而 IMUTube 则提供准确的运动模式。

尽管所有数据扩展方法在完整和有限的真实数据状态下都有益，但在 10 % 条件下，它们的相对有效性被放大。表现最佳的配置相对于仅使用真实数据的基线，以 10 % 的数据获得了平均 116.9 % 的提升，而在完整数据下仅有 28.8 % 的改进。这强调了在真实 IMU 数据量有限的情况下，额外训练数据的重要性。

*特定类别改进。*除了整体性能提升，合成数据显著增强了特定活动类别的表现。图 2 展示了 PAMAP2 各个训练配置的每类 F1 分数。我们观察到在骑自行车、使用电脑和折叠衣服等活动中有明显的收益，而这些活动中仅使用真实数据的模型表现较差。这些任务涉及的动作可能难以从有限的真实数据中泛化。添加虚拟 IMU 信号——尤其是来自 IMUGPT，提供了多样的活动描述——帮助提升了这些类别的表现，表明虚拟 IMU 数据可以帮助缓解类别不平衡或代表性不足的问题。

从部署的角度来看，各种方法的计算成本差异显著。经典的数据增强在计算上几乎可以忽略。在 Nvidia A6000 上，IMUGPT 生成 10 秒的虚拟 IMU 数据大约需要 10 秒。相比之下，IMUTube 生成相同输出时长大约需要 5 分钟。鉴于这些取舍，我们建议对于计算或视频资源有限的从业者，IMUGPT 是一个实用的起

点。对于那些可以使用更多资源的人，将 IMUTube 和 IMUGPT 结合使用将提供最大的益处并产生最强健的模型。

4 结论

本文对经典的数据增强方法和两种跨模态迁移方法——IMUGPT 和 IMUTube——在虚拟 IMU 数据生成中的应用进行了比较研究。我们生成了一个横跨 100 个活动的大型虚拟 IMU 数据集，并在三个数据集和四个模型上对这些方法进行了评估。结果显示，虚拟 IMU 数据在低数据情况下始终优于仅使用真实数据和增强数据的基线。IMUTube 提供了更准确的运动信息，而 IMUGPT 则有助于更广泛的数据多样性；二者结合往往会产生最佳表现。我们的研究结果为选择数据生成策略提供了实践指导，并强调了虚拟 IMU 数据在扩展 HAR 中的价值。

References

[1] 2023. REDUCELRONPLATEAU. https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html (2024, Feb 1).

[2] Luay Alawneh, Tamam Alsarhan, Mohammad Al-Zinati, Mahmoud Al-Ayyoub, Yaser Jararweh, and Hongtao Lu. 2021. Enhancing human activity recognition using deep learning and time series augmented data. *Journal of Ambient Intelligence and Humanized Computing* (2021), 1–16.

[3] Sizhe An, Ganapati Bhat, Suat Gumussoy, and Umit Ogras. 2023. Transfer learning for human activity recognition using representational analysis of neural networks. *ACM Transactions on Computing for Healthcare* 4, 1 (2023), 1–21.

[4] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*. 168–172. https://doi.org/10.1109/ICIP.2015.7350781

[5] Hui Chen, Charles Gouin-Vallerand, Kévin Bouchard, Sébastien Gaboury, Mélanie Couture, Nathalie Bier, and Sylvain Giroux. 2024. Enhancing human activity recognition in smart homes with self-supervised learning and self-attention. *Sensors* 24, 3 (2024), 884.

[6] Hui Chen, Charles Gouin-Vallerand, Kévin Bouchard, Sébastien Gaboury, Hubert Kenfack Ngankam, Maxime Lussier, Mélanie Couture, Nathalie Bier, and Sylvain Giroux. 2024. Utilizing Self-Supervised Learning for Recognizing Human Activity in Older Adults through Labeling Applications in Real-World Smart Homes. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*. 275–283.

[7] Wenqiang Chen, Shupei Lin, Elizabeth Thompson, and John Stankovic. 2021. SenseCollect: We Need Efficient Ways to Collect On-body Sensor-based Human Activity Data! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[8] Yuwen Chen, Kunhua Zhong, Ju Zhang, Qilong Sun, and Xueliang Zhao. 2016. LSTM networks for mobile human activity recognition. In *2016 International conference on artificial intelligence: technologies and applications*. Atlantis Press, 50–53.

[9] Ji Seok Choi and Jung Keun Lee. 2023. Effects of data augmentation on the nine-axis IMU-based orientation estimation accuracy of a recurrent neural network. *Sensors* 23, 17 (2023), 7458.

[10] Sourish Gunesh Dhekane and Thomas Ploetz. 2024. Transfer learning in human activity recognition: A survey. *arXiv preprint arXiv:2401.10185* (2024).

[11] Iveta Dirgová Luptáková, Martin Kubovčík, and Jiří Pospíchal. 2022. Wearable Sensor-Based Human Activity Recognition with Transformer Model. *Sensors* 22, 5 (2022). https://doi.org/10.3390/s22051911

[12] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Data augmentation using synthetic data for time series classification with deep residual networks. arXiv:1808.02455 [cs.CV] https://arxiv.org/abs/1808.02455

[13] N. Y. Hammerla, R. Kirkham, P. Andras, and T. Ploetz. 2013. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proceedings of the 2013 international symposium on wearable computers*. 65–68.

[14] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2022. Assessing the State of Self-Supervised Human Activity Recognition Using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3 (2022). https://doi.org/10.1145/3550299

[15] Yifan Hu. 2023. Bsdgan: Balancing sensor data generative adversarial networks for human activity recognition. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV] https://arxiv.org/abs/1705.06950

[17] Hyeokhyen Kwon, Gregory D Abowd, and Thomas Plötz. 2019. Handling annotation uncertainty in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 109–117.

[18] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[19] Hyeokhyen Kwon, Bingyao Wang, Gregory D Abowd, and Thomas Plötz. 2021. Approaching the Real-World: Supporting Activity Recognition Training with Virtual IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–32.

[20] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. 2024. IMUGPT 2.0: Language-Based Cross Modality Transfer for Sensor-Based Human Activity Recognition. (2024). arXiv:2402.01049 [cs.CV]

[21] Zikang Leng, Yash Jain, Hyeokhyen Kwon, and Thomas Ploetz. 2023. On the Utility of Virtual On-body Acceleration Data for Fine-grained Human Activity Recognition. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers (ISWC '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3594738.3611364

[22] Zikang Leng, Hyeokhyen Kwon, and Thomas Ploetz. 2023. Generating Virtual On-Body Accelerometer Data from Virtual Textual Descriptions for Human Activity Recognition. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3594738.3611361

[23] Xi'ang Li, Jinqi Luo, and Rabih Younes. 2020. ActivityGAN: Generative adversarial networks for data augmentation in sensor-based human activity recognition. In *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*. 249–254.

[24] MinYen Lu, ChenHao Chen, Shigemi Ishida, Yugo Nakamura, and Yutaka Arakawa. 2022. A study on estimating the accurate head IMU motion from Video. *Proceedings of the Symposium on Multimedia, Distributed, Cooperative, and Mobile (DICOMO) 2022* 2022 (07 2022), 918–923. https://cir.nii.ac.jp/crid/1050011771467456512

[25] Sara Mohammed, Reda Elbasiony, and Walid Gomaa. 2018. An LSTM-based Descriptor for Human Activities Recognition using IMU Sensors. 504–511. https://doi.org/10.5220/0006902405040511

[26] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(ISWC '17)*. Association for Computing Machinery, New York, NY, USA, 158–165. https://doi.org/10.1145/3123021.3123046

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[28] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, et al. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[29] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* (2016).

[30] Thomas Plötz. 2023. If only we had more data!: Sensor-Based Human Activity Recognition in Challenging Scenarios. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. 565–570. https://doi.org/10.1109/PerComWorkshops56833.2023.10150267

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, 8748–8763.

[32] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring *(ISWC '12)*. IEEE Computer Society. https://doi.org/10.1109/ISWC.2012.13

[33] Vitor Fortes Rey, Peter Hevesi, Onorina Kovalenko, and Paul Lukowicz. 2019. Let There Be IMU Data: Generating Training Data for Wearable, Motion Sensor Based Activity Recognition from Monocular RGB Videos. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. Association for Computing Machinery, 699–708. https://doi.org/10.1145/3341162.3345590

[34] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.

[35] Satya P. Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta. 2021. Deep ConvLSTM With Self-Attention for Human Activity Decoding Using Wearable Sensors. *IEEE Sensors Journal* 21, 6 (2021), 8575–8582. https://doi.org/10.1109/JSEN.2020.3045135

[36] Elnaz Soleimani and Ehsan Nazerfard. 2021. Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. *Neurocomputing* 426 (2021), 26–34.

[37] Lena Uhlenberg and Oliver Amft. 2022. Comparison of Surface Models and Skeletal Models for Inertial Sensor Data Synthesis. In *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 1–5. https://doi.org/10.1109/BSN56160.2022.9928504

[38] Fanyi Xiao, Ling Pei, Lei Chu, Danping Zou, Wenxian Yu, Yifan Zhu, and Tao Li. 2021. A Deep Learning Method for Complex Human Activity Recognition Using Virtual Wearable Sensors. In *Spatial Data and Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-030-69873-7_19

[39] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*.

[40] K. Yamane and Y. Nakamura. 2003. Natural motion animation through constraining and deconstraining at will. *IEEE Transactions on Visualization and Computer Graphics* 9, 3 (2003), 352–360. https://doi.org/10.1109/TVCG.2003.1207443

[41] A. D. Young, M. J. Ling, and D. K. Arvind. 2011. IMUSim: A simulation environment for inertial sensing algorithm design and evaluation. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 199–210.

[42] Junjie Zhang, Yuanhao Liu, and Hua Yuan. 2023. Attention-Based Residual BiLSTM Networks for Human Activity Recognition. *IEEE Access* 11 (2023), 94173–94187. https://doi.org/10.1109/ACCESS.2023.3310269

[43] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. 2022. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13232–13242.

[44] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. 2020. Towards Better Generalization: Joint Depth-Pose Learning without PoseNet. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.