用于多模态人体活动识别的蒸馏中期融合Transformer网络

Jingcheng Li*, [a]、Lina Yao*, [a, b]、Binghao Li[a]、Claude Sammut[a]

[a]澳大利亚新南威尔士州悉尼新南威尔士大学，*2052*[b]澳大利亚新南威尔士州悉尼 *CSIRO* 数据*61*，*2015*

arXiv:2305.03810v1 [cs.HC] 5 May 2023

摘要

人体活动识别是许多人机协作场景中的重要任务，同时具有广泛的实际应用。尽管单模态方法已得到广泛研究，但这类方法容易受到数据质量的影响，并且需要针对特定模态进行特征工程，因此在实际部署中不够稳健和高效。通过利用多种传感器，多模态人体活动识别能够利用互补信息，构建具有良好泛化能力的模型。虽然深度学习方法已展现出良好效果，但其提取显著多模态时空特征以及更好地融合互补信息的潜力尚未得到充分挖掘。此外，降低多模态方法的复杂度以便在边缘设备上部署，也是一个尚未解决的问题。为解决这些问题，本文提出一种基于知识蒸馏的多模态中期融合方法 DMFT，以高效完成信息特征提取与融合，从而解决多模态人体活动识别任务。DMFT首先将多模态输入数据编码为统一表示。随后，DMFT教师模型采用注意力多模态时空 Transformer 模块，提取显著的时空特征。本文还提出了时间中期融合模块，以进一步融合时间特征。然后，应用知识蒸馏方法，将教师模型学习到的表示迁移至更为简单的 DMFT学生模型中，该学生模型由多模态时空 Transformer 模块的轻量版本构成，并据此生成结果。我们在两个公开的多模态人体活动识别数据集上，将 DMFT 与多种最先进方法进行了评估。实验结果表明，该模型在有效性、可扩展性和稳健性方面均取得了具有竞争力的性能。

关键词：
活动识别，神经网络，知识蒸馏，多模态学习

1. 引言

人体活动识别（HAR）是许多人机协作场景中的一项重要任务，可带来多种有益应用，例如医疗保健、辅助生活、老年护理和野外工程。在多模态环境中，模型以多种活动数据为输入，旨在准确预测人类执行的活动。

许多现有方法通过分析各种单模态传感器数据（如 RGB、深度、骨架、惯性和 Wi-Fi 数据），对人体活动识别任务进行了广泛探索。传统的基于机器学习的方法通过手动设计特征提取方法来预测活动标签。由于这些方法高度依赖手工特征工程，因此在应用于新任务或数据质量较差时，难以实现良好的泛化。当前方法采用卷积神经网络（CNN）和长短期记忆网络（LSTM）等深度学习技术来解决这一任务，无需手工特征工程即可进行特征提取。然而，基于 CNN 的模型将时间维度和空间维度一视同仁，这限制了它们捕捉活动顺序的能力。LSTM 等序列模型侧重于时间信息，却忽略了空间通道信息。此外，当面对现实场景时，单模态方法可能不够稳健，难以实现泛化

*通讯作者电子邮箱： jingcheng.li@unsw.edu.au (Jingcheng Li)，lina.yao@unsw.edu.au (LinaYao)

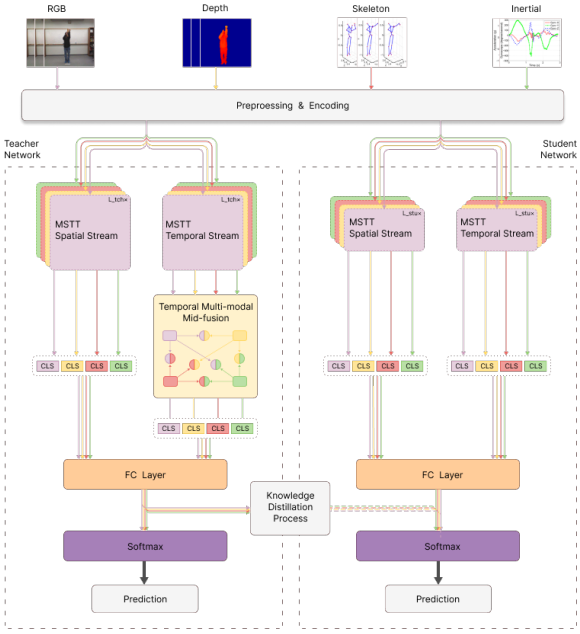


图 1：DMFT 模型的整体框架

输入数据包含噪声或质量较低。例如，在低光照环境下，基于视觉的方法可能无法实现准确识别，因为摄像机无法很好地捕捉人的动作。此外，惯性传感器也可能因电磁环境的负面影响而无法记录和传输信号。因此，有必要探索更有效的方法，充分利用来自多模态数据源的时间和空间信息，以提高实际设置下人体活动识别的准确率和鲁棒性。

因此，研究人员探索了多模态人体活动识别模型，以克服单模态方法所面临的挑战。与单模态方法不同，多模态方法能够利用来自不同模态的输入数据提取互补信息，从而实现更稳健的性能。

尽管多模态人体活动识别方法在从多个数据源中提取有信息量的时空特征并获得更好的结果方面具有诸多优势，但要实现更稳健、准确的预测，仍面临许多挑战。首先，许多现有方法会为每个模态流采用单模态特征提取方法，这不仅增加了模型复杂度，还需要额外投入来开发模态专用方法。与此同时，现有方法设计了复杂的网络架构，以提取时间和空间序列特征，导致整体复杂度上升，尤其是在多模态场景下。如果模型无法扩展，就不适合部署在真实环境中，因为边缘计算会使用许多低端设备。此外，尽管多模态学习能够利用全面且互补的信息，但如何有效地提取显著特征并进行特征融合，仍有待进一步探索。现有方法主要先分别进行特征提取，再着重融合提取出的特征，这可能无法充分共享互补信息。因此，有必要进一步研究可扩展、稳健且足够准确的人体活动识别方法，使其能够广泛部署于真实环境中。

为应对这些挑战，我们提出了用于多模态人体活动识别的蒸馏中期融合Transformer（DMFT）网络。这是一种新颖的方法，能够以有效、高效且稳健的方式解决多模态人体活动识别任务，并可推广到许多不同的多模态环境中。在特征提取和融合阶段实现具有竞争力的性能的同时，该模型还具备可扩展性，并可直接部署于现实场景中。特征编码层首先对多模态输入数据进行预处理，以各模态的原始数据作为输入，并将其编码为统一结构，以便进一步进行特征提取。与使用长短期记忆网络或卷积神经网络学习特征嵌入相比，这有助于提高效率并减少额外的数据工程工作，而且该过程可以离线并行完成。随后，我们应用多模态空间—时间Transformer（MSTT）模块，提取模态特定的空间和时间特征。通过利用注意力机制，该模块能够提取显著信息，并构建多模态输入数据的高层表示。随后，DMFT 教师网络采用时间中期融合模块，进一步提取并融合高层多模态时间信息。不同于采用后期融合方法的其他方法，时间中期融合模块执行中期融合，即在特征提取过程中完成融合。DMFT 学生网络采用轻量级 MSTT 模块，以实现高效的特征提取。然后，我们应用知识蒸馏方法，将 DMFT 教师网络学习到的特征表示迁移至更小的 DMFT 学生模型。最后，我们采用多模态集成投票方法，聚合各模态特定的输出，以生成最终预测。图 1 展示了整体

DMFT 框架。

我们在两个公开的多模态人体活动识别数据集 UTD-MHAD [1] 和 MMAct [2]，上开展了广泛实验，使用两种主体独立的评估协议和各种模态组合来评估模型的性能。结果表明，与当前最先进的方法相比，我们提出的模型取得了具有竞争力的性能。此外，我们的研究表明，通过应用知识蒸馏方法，可以提升学生网络的性能，同时显著降低计算与空间成本。

本文其余部分安排如下：第 2 节进行文献综述；第 3 节介绍所提出模型的细节；第 4 节介绍数据集、评估协议和实验设置；第 5 节比较并报告评估结果；最后，第 6 节总结全文。

2. 相关工作

2.1. 人类活动识别

人体活动识别多年来一直是面向人机交互的一项长期任务，并在各种应用中具有广阔的应用前景。解决活动识别任务的方法可分为三类：基于视觉的人体活动识别、基于传感器的人体活动识别和多模态人体活动识别。

多种基于视觉的架构多年来得到了广泛研究，它们通过处理图像或视频来解决人体活动识别任务。传统方法侧重于手工设计的机器学习模型 [3, 4, 5, 6, 7, 8, 9]。然而，这些方法通常需要手工设计特征工程方案，不仅耗时，而且部署到不同场景时鲁棒性较差。近年来，CNN 和 LSTM 等深度学习架构已被广泛用于更好地学习特征表示 [10, 11, 12, 13, 14, 15]。

基于传感器的方法以可穿戴传感器、环境传感器或物体传感器采集的数据作为输入，并进行人体活动识别。与基于视觉的方法相比，基于传感器的方法能够缓解计算效率和隐私方面的问题。早期也有研究探索传统机器学习方法 [16, 17, 18, 19, 20, 21]。近期方法则采用基于深度学习的架构，例如卷积神经网络（CNN） [22, 23, 24, 25, 26, 27] 和长短期记忆网络（LSTM） [28, 29, 30, 31, 32]，以更好地提取时间信息。

人们研究了多模态方法 [1, 33, 34, 35, 36, 37, 38]来解决人体活动识别任务。与单模态方法不同，后者容易受到噪声或数据丢失的影响，难以良好泛化；多模态方法则能够从不同模态的数据中学习稳健的特征表示。同时，不同模态的特征可能包含互补信息，从而提升模型的性能。作为早期尝试，Guo等人[33]为每种模态构建了一个神经网络分类器，随后通过分类器分数融合得到最终输出。Memmesheimer 等人 [35]利用骨架和惯性数据构建信号图像，并将该任务视为图像分类问题，以预测人体活动。

2.2. 注意力机制

作为提取注意信息的早期尝试，Chen 等 [39] 采用了窥视网络 [40] 来解决基于传感器的人体活动识别问题，其中每次窥视都会以高分辨率编码特定区域，同时对其余区域逐步采用较低分辨率。Long等 [41] 提出了一种无键注意力方法，从不同模态中提取时空特征，包括视觉、声学和片段级特征，然后将它们拼接起来进行视频分类。

当代基于注意力的方法 [42, 43, 36, 37, 44, 45, 38, 46], 利用自注意力方法来更好地提取显著特征。Islam 等。 [37]首先构建了单模态自注意力模块，以顺序提取单模态时空特征，然后引入混合专家模型来提取显著特征，并使用跨模态图注意力方法融合这些特征。他们的扩展工作[45]增加了一个活动组分类任务，并利用该任务来指导整体活动识别任务。Li 等。[44] 提出了一种CNN增强型Transformer方法，从Wi-Fi信号数据的信道状态信息（CSI）中提取显著的时空特征，以执行单模态人体活动识别。他们的工作展示了Transformer捕获显著时空特征并解决人体活动识别任务的能力，但他们仅使用了一种模态，而Wi-Fi信号通常可能包含噪声数据，因此鲁棒性会成为一个问题。因此，模型的复杂度和可扩展性成为问题，尤其是在多模态场景下。基于白名单自注意力的多模态方法似乎能够捕获互补信息并产生更鲁棒的结果，但它们通常采用复杂的架构，从而导致较高的空间复杂度和计算复杂度。

因此，由于硬件成本高昂，这些模型不适合在现实场景中部署。

2.3. 知识蒸馏

知识蒸馏（KD）是模型压缩方法之一，其将知识从计算开销大的教师模型迁移到更小的学生网络中。通常，学生模型能够通过学习预训练教师模型生成的更优特征分布来提升性能。目前，只有少数基于 KD 的方法[2, 47, 48, 49, 50, 51, 52]专注于多模态人体活动识别。Liu 等 [50]首先生成惯性数据的虚拟图像，然后利用这些图像构建基于 CNN 的教师网络，并使用 RGB 数据训练学生网络。Ni 等 [51] 提出了一种渐进式学习方法，首先利用骨架数据和惯性数据构建多教师模型，然后使用自适应置信度语义损失，使学生模型能够自适应地选择有用信息。然而，这些方法仅使用 1 或 2 种模态的数据作为教师模型的输入，并使用单模态数据作为学生模型的输入。因此，学生模型仍被视为单模态方法，其性能可能受到噪声或数据缺失的影响。也就是说，尽管这些方法尝试使用多模态数据进行知识蒸馏，但其模型并未利用多模态学习的优势，即利用互补信息并提升泛化能力。此外，一些工作在性能比较中没有遵循统一的或跨受试者实验设置。

因此，尽管多模态方法能够从互补信息中受益，并可能带来更稳健的性能，但仍有许多问题亟待解决。首先，现有方法通常分别从每种单模态输入数据中提取显著的时空特征，然后通过简单拼接或相加高层特征来进行后期融合。然而，互补信息并未在中间阶段实现共享与融合。其次，尽管基于深度学习的方法需要大量训练数据，但在现实生活中很难构建高质量标注数据集，尤其是在多模态场景下。目前，多模态人体活动识别数据集寥寥无几，样本数量和活动类别也相当有限。因此，复杂架构可能无法得到充分优化，在引入更复杂的新活动时也难以实现良好的泛化。第三，尽管最先进水平的多模态人体活动识别方法取得了有竞争力的性能，但它们采用了复杂架构，导致计算和空间成本高昂。在现实场景中，例如日常生活环境或现场部署，设备难以承担如此高昂的成本。知识蒸馏方法虽然能够降低模型复杂度，但学生模型仍使用单一模态的数据，而其他模态仅用于训练教师模型，以指导单模态网络。因此，目前尚无稳健有效的方法，能够以高效方式提取显著的时空多模态特征。

因此，我们提出蒸馏式中层融合 Transformer 网络，先提取并融合显著的时空多模态特征，再采用知识蒸馏方法构建相对简单的学生网络，以降低模型复杂度，同时保持具有竞争力且稳健的性能。据我们所知，这是首项将知识蒸馏方法应用于完整多模态人体活动识别任务的研究。

3. 方法

本节提出用于多模态人体活动识别的多模态中期融合 Transformer 网络，以及相应的知识蒸馏学习流程。图 2 展示了 DMFT 教师网络的整体结构。图 3 展示了 DMFT 学生网络的整体结构。图 1 展示了从教师网络到学生网络的完整知识蒸馏过程。

该框架包含四个组成部分：

(i) 一种通用的特征编码方法，用于接收原始多模态数据，并将其编码为统一结构。

(ii) 一个多模态时空 Transformer 模块和一个用作教师模块的多模态中期融合 Transformer 编码器，用于提取显著的时空特征，在特征提取过程中采用时序中间融合方法对多模态时序特征进行中期融合，并生成预测。

(iii) 一个简单轻量的多模态时空 Transformer 模块，用作学生网络，以可扩展且高效的方式生成预测。

(iv) 一种知识蒸馏流程，可将知识从计算开销大的教师模型迁移到较小的学生模型。

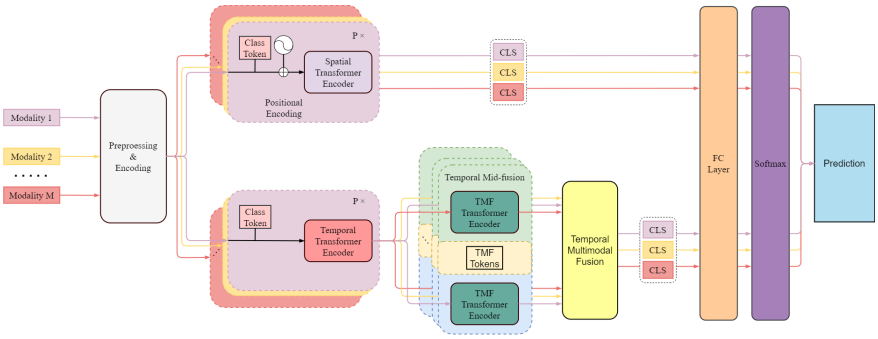


图 2：DMFT 教师网络的整体框架

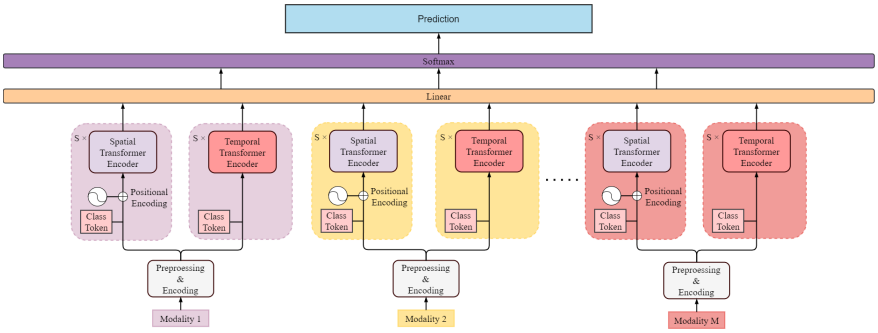


图 3：DMFT 学生网络的整体框架

首先，特征编码层将每种模态的原始数据作为输入，并将其编码为统一结构，以便进一步进行特征提取。随后，我们为空间流和时间流都添加类别标记，以捕捉显著特征的通用表示。我们还为空间流添加正弦位置编码，以保留空间顺序关系。我们采用多头自注意力机制，更好地提取显著时空特征并提高准确率。DMFT 教师网络利用时序中间融合 Transformer 模块，更好地提取并融合高层次的多模态时间信息。DMFT 学生网络利用轻量级多模态时空 Transformer 模块，以可扩展且高效的方式提取显著特征。随后，我们通过知识蒸馏流程，将 DMFT 教师模块学到的特征表示迁移到较小的 DMFT 学生模块。最后，我们采用多模态集成投票方法生成总体预测。

我们按以下顺序详细介绍该框架：通用特征编码层、多模态时空 Transformer 模块、时间中期融合 Transformer 模块，以及知识从 DMFT教师模型到 DMFT 学生模型的蒸馏过程，以及训练与优化方法。

3.1. 广义编码层

多模态数据，如 RGB 数据、深度数据、骨架数据、惯性数据和 Wi-Fi 数据，可能具有不同的表示结构、特征分布或频率。例如，惯性数据为一维数据，频率为 100HZ，而 RGB 数据具有三维结构，且频率分布不同，例如为 24 fps。我们采用先前工作 [38], 中的广义编码层，将多模态数据编码为统一表示，而无需使用复杂的特定模态特征编码器架构。

采用这种方法有三个原因。首先，该方法不需要手工特征提取或复杂的特征编码器，从而降低了模型复杂度，并且由于可以将新的模态直接集成到网络中，因此具有可扩展性。其次，该方法可以离线并行执行，并利用预训练模型；编码后的特征可供教师网络和学生网络共同使用。因此，该方法在计算和空间方面都

高效，有利于实际部署。第三，我们先前的工作 [38] 表明，这种统一编码方法无需使用复杂的模态专属特征编码器，也能取得具有竞争力的结果。因此，该方法高效、有效且具有可扩展性。

对于每种模态 $m \in M$，输入数据由 N 条记录 $R_m = [r_{m,1}, r_{m,2}, \ldots, r_{m,N}]$组成。我们首先使用固定长度的滑动窗口将原始数据转换为片段，然后对片段级数据进行平均池化。此方法可降低噪声以及计算和空间开销 [38]。此外，为了更好地支持批处理，我们对记录 R_m 进行时间对齐，使每个数据序列 $r_{m,n}$都包含相同数量的分段补丁。由于视觉数据（RGB 和深度）在时间维度上呈二维结构，我们使用预训练 ResNet 50 模型将其转换为时间维度上的一维向量，以便进一步进行特征提取。因此，对于每种模态 m，广义编码层都会生成编码后的输入特征 $X_m = [x_{m,1}, x_{m,2}, \ldots, x_{m,P_m}]$，其大小为 $(B \times P_m \times D_m)$，其中 B 为批量大小，$P_m$ 表示分段补丁的长度，D_m 表示特征维度。

3.2. 多模态时空 *Transformer* 模块

多模态时空 Transformer（MSTT）模块 [38] 采用 Transformer 编码器架构，分别提取每种模态的显著空间特征和时间特征。虽然长短期记忆网络已广泛用于解决时间序列预测任务，但其架构存在长距离依赖问题。相比之下，Transformer 可以将输入序列作为整体处理，并利用自注意力机制更好地提取显著特征。此外，不同于传统的时间 Transformer 编码器网络或序列式时空 Transformer 网络，结果 [38] 表明，MSTT 模块通过分别对空间序列特征和时间序列特征使用独立的注意力模块，能够更好地提取显著的时空特征。

对于每个模态m，编码后的特征X_m可直接用作时间流的输入特征$X_{m,T} = [x_{m,1}, x_{m,2}, \ldots, x_{m,P_m}]$。对于空间流，可通过简单的转置操作得到输入特征$X_{m,S} = [x_{m,1}, x_{m,2}, \ldots, x_{m,D_m}]$。随后，我们分别在时间序列特征和空间序列特征前添加一个可学习的类别标记x_{cls}，因为类别标记能够更好地生成输入特征的整体表示。接着，我们为空间特征流添加正弦位置编码，以保留位置信息，因为 Transformer 模型无法捕获顺序信息。

$$\begin{aligned} PE_{(pos,2i)} &= sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{1}$$

$$H_{m,S} = X_m^T = [x_{m,S,cls}, x_{m,1}, \ldots, x_{m,D_m}] + E_{pos} \tag{2}$$

$$H_{m,T} = [x_{m,T,cls}, x_{m,1}, \ldots, x_{m,P_m}] \tag{3}$$

对于空间流和时间流，我们均采用由 L_{MSTT} 层基础 Transformer 编码器堆叠而成的结构，其中每层均包含一个多头自注意力层和一个逐位置前馈层。为提取每个 Transformer 编码器层中的显著特征，我们对输入隐藏嵌入 H_m 进行线性投影，为每个头 h生成查询 Q_m、键 K_m和值V_m。

$$Q_m = H_m W_m^Q \quad K_m = H_m W_m^K \quad V_m = H_m W_m^V \tag{4}$$

然后，我们采用缩放点积来计算多头自注意力的注意力分数。其中，d 是用于平滑梯度的缩放因子。多头自注意力通过重复计算h 次，再将输出拼接起来，为下一层生成隐藏嵌入。

$$Attention_m(Q_m, K_m, V_m) = Softmax(\frac{Q_m K_m^\top}{\sqrt{d_{k,m}}})V_m \tag{5}$$

$$\begin{aligned} MultiHead_m(Q_m, K_m, V_m) &= [head_{m,1}, \ldots, head_{m,h}]W_m^O \\ where\ head_{m,i} &= Attention(Q_{m,i}, K_{m,i}, V_{m,i}) \end{aligned} \tag{6}$$

MSTT模型作为教师网络和学生网络的核心特征提取模块，为每种模态m生成显著特征表示。在下一小节中，我们将介绍时序中间融合 Transformer 模块，该模块能够基于多模态时序特征进一步进行更高层次的多模态融合。

3.3. 时序中间融合 *Transformer* 模块

虽然 MSTT 模块能够分别提取显著的多模态时空特征，但由于每个流都是独立处理的，因此多模态时空特征之间不存在交互。尽管采用特征相加或拼接（称为后期融合）是一种常见的方法来

为了融合多模态特征，这可能会丢失一些信息丰富的内容。同时，在开头拼接多模态特征（称为早期融合）虽然能让网络考虑多模态之间的关系，但计算成本会呈二次方增长，这对于多模态场景而言并不现实。此外，由于多模态特征的时间长度或特征维度可能不同，我们无法简单地拼接输入特征并进行特征提取。因此，类似于 [53], ，我们提出了时序中间融合 Transformer（TMT）模块，在特征提取过程中进一步进行多模态时序特征融合。

对于每个融合模块（模块总数等于模态组合的数量），我们使用一组F 时序中期融合令牌$x_{TMT} = [x_{TMT,1}、x_{TMT,2}、\ldots\ldots、x_{TMT,F}]$，其作用类似于类别令牌，用于捕获共有的多模态信息，同时降低计算成本。对于L_{MSTT}层 MSTT 模块输出的任意两个时序流特征$H_{m1,T}$和$H_{m2,T}$，我们将它们与相应的 TMT令牌拼接起来，然后使用L_{TMT}层 Transformer 编码器进一步提取并融合多模态特征。由于每个时序特征$H_{m,T}$的表示维度可能不同，因此我们对 TMT令牌进行线性投影，使其在进行时序中间融合之前和之后都能匹配特征维度较小的特征D_m。这有助于融合两个维度不同的流中的特征。

$$\begin{aligned} H_{m1,T}^{TMT} &= [x_{TMF,(m1,m2)}, H_{m1,T}] \\ H_{m2,T}^{TMT} &= [LN(x_{TMF,(m1,m2)}), H_{m2,T}] \\ &where\ D_{m1} < D_{m2} \end{aligned} \tag{7}$$

然后，对于模态 \(m\)$_1$ 和 \(m\)$_2$，的每一种融合组合，多模态时序中间融合注意力流的结构如下所示。

$$H_{m,T}^{TMT'} = Transformer(H_{m,T}^{TMT}, \theta_m) \tag{8}$$

$$x_{TMF}^{'} = Average(x_{TMF,m1}^{'}, x_{TMF,m2}^{'}) \tag{9}$$

值得一提的是，当模态数量为$M > 2$时，时序中间融合将进行C_M^2 次，即从M个模态中抽样组合两个模态的次数。然后，对于每个模态m，我们对每个时序中间融合 Transformer 模块的输出取平均，以分层融合多模态信息。

$$H_{m,T}^{TMT,L} = Average(H_{m,T,1}^{TMT,L}, H_{m,T,2}^{TMT,L}, \ldots, H_{m,T,C_M^2}^{TMT,L}) \tag{10}$$

3.4. 多模态知识蒸馏

随后，我们构建多模态教师网络，该网络由 MSTT 模块和 TMT模块组成，以便更好地提取和融合多模态时空特征。教师网络的整体框架如图 2 所示。

对于每种模态m，首先将原始数据传入广义编码层，得到统一输入特征。然后，我们对$X_{m,S}$、$X_{m,T}$ 中的令牌应用含$L_{MSTT,tch}$的原始自注意力 MSTT 模块，以提取空间流和时间流的显著特征。之后，将 MSTT 模块输出的两种时间潜在特征的每种组合分别与相应的 TMT令牌拼接，并传入 TMT模块，在该模块中，令牌按照公式进行融合和更新。

然后，对于空间流和时间流，我们都会输出相应的类别令牌表示，并将其输入线性层。对于教师网络，我们对逻辑值应用 Softmax 函数，取各模态输出的平均值，从而生成总体预测 Y_t。

$$Y_m = Softmax(LN_S(h_{m,S,L,0})) + Softmax(LN_T(h_{m,T,L,0}^{TMT})) \tag{11}$$

$$\hat{Y} = \sum_{m=1}^{M} Y_m \tag{12}$$

学生网络的框架如图 3 所示。学生网络采用更简单的架构，包含一个$L_{MSTT,stu}$层 MSTT 模块。对于每种模态，输入特征都会经过 MSTT 模块，随后输出类别令牌的表示，并将其输入线性层。与教师网络类似，对逻辑值应用 Softmax 函数，以生成整体预测。

接下来，我们采用基于知识蒸馏的方法训练学生网络。训练过程如图 1 所示。我们通过 Softmax 运算将输出逻辑值转换为类别概率$P_{teacher}$，并使用温度参数$temp$对其进行软化。假设对于每种模态 m，教师网络输出的类别概率表示为

$P_{teacher,\ S}$、$P_{teacher,\ T}$。随后优化学生网络的类别概率$P_{student,\ S}$、$P_{student,\ T}$，使其与相应教师网络的 logits 分布相匹配，从而预测目标类别。

$$P = \frac{\frac{exp(h_i)}{temp})}{\sum_j exp(\frac{h_j}{temp})} \tag{13}$$

Hinton 等人使用 Kullback–Leibler（KL）散度作为损失函数，以将知识从教师网络迁移到学生网络，使学生网络的类别概率收敛到教师网络输出的类别概率。

$$KL(P_{student}, P_{teacher}) = \sum_c P_{student,c} log \frac{P_{student,c}}{P_{teacher,c}} \tag{14}$$

为了训练学生网络，我们使用加权损失，该损失由交叉熵损失 L_{CS}（硬损失）和知识蒸馏损失 L_{KD}（软损失）组成。用于优化学生网络的整体损失函数 $L_{student}$ 定义如下：

$$\begin{aligned} L_{student} &= w_{CS} L_{CS} + \\ &\sum_{m=1}^{M} w_{m,S} L_{KL}(P_{student,m,S}, P_{teacher,m,S}) + \\ &\sum_{m=1}^{M} w_{m,T} L_{KL}(P_{student,m,T}, P_{teacher,m,T}) + \\ &where\ w_{CS} + \sum_{m=1}^{M} w_{m,S} + \sum_{m=1}^{M} w_{m,T} = 1 \end{aligned} \tag{15}$$

随后，整体训练和优化过程旨在最小化损失 $L_{student}$，并针对给定的多模态输入数据生成预测 $\hat{Y}$。

4. 实验

4.1. 数据集

我们在两个公开的基准数据集 UTD-MHAD 和 MMAct 上评估 DMFT 的性能。这两个数据集是该领域目前仅有的两个主流多模态人体活动识别数据集。

UTD-MHAD [1] 数据集包含 27 项活动，每项活动由 8 名受试者（4 名男性和 4 名女性）各执行 4 次，移除损坏样本后共得到 861 个样本。该数据集包括 4 种模态：RGB、深度、骨架和惯性。使用 Kinect 摄像头捕获视觉信息，同时使用可穿戴传感器记录惯性数据，包括加速度、陀螺仪和磁数据。我们的实验设置使用了全部 4 种模态。

MMAct [2] 数据集包含 35 种活动，每种活动由 20 名受试者（10 名男性和 10 名女性）各执行 5 次。移除损坏样本后，共有 36K 个样本。该数据集包含 7 种模态：RGB、骨架、加速度、陀螺仪、方向、Wi-Fi 和压力。使用 4 个摄像头和智能眼镜记录 RGB 数据，同时使用智能手机和智能手表记录加速度、陀螺仪、方向、Wi-Fi 和压力数据。我们使用加速度、陀螺仪和方向作为惯性数据，并使用 RGB 数据进行实验。

4.2. 评估协议

对于 UTD-MHAD 数据集，我们采用两种交叉验证方法。首先，与原始论文 [1], 相同，我们采用 50-50 评估方法：使用编号为奇数的受试者（1、3、5、7）进行训练，使用编号为偶数的受试者（2、4、6、8）进行测试。同时，我们采用留一受试者交叉验证（LOSO）协议，依次选择每名受试者进行测试，并使用其余 7 名受试者进行训练。我们以 Top-1 准确率作为 UTD-MHAD 数据集的评估指标，并对结果取平均值，以便将模型性能与其他方法进行比较。

对于 MMAct 数据集，我们遵循原论文 [2]提出的两种交叉验证方案。首先，我们采用跨主体设置，其中编号靠前的 80% 受试者（1-16）用于训练，其余受试者（17-20）用于测试。此外，我们还采用跨会话设置，对每位受试者而言，前 80% 的会话用于训练，其余会话用于测试。我们使用 F1分数作为 MMAct 数据集的评估指标，并取结果的平均值，将模型性能与其他方法进行比较。

值得注意的是，50-50主体设置、留一主体交叉验证设置和跨主体设置均为主体无关设置。在现实场景中，模型用于分析新受试者所进行的活动。主体相关评估协议中，训练集和测试集都包含同一受试者的共同信息，因而忽略了参与者偏差，可能会得出不同的结论。因此，我们主要采用主体无关设置，以纳入不同受试者在现实中的差异，从而

更好地评估模型的性能。同时，尽管跨会话设置属于主体相关设置，我们仍将其纳入，以进一步展示模型性能的细节。

4.3. 实验设置

预处理。由于某些模态可能包含不同类型的数据流，我们将它们视为单一模态并进行预处理。对于每种模态，我们首先使用滑动窗口将数据流划分为多个片段，以降低采样频率。然后，对于 RGB 和深度数据，我们使用预训练的 ResNet50 网络对其进行编码。对于骨架和惯性数据，我们直接将其输入特征提取模块。对于每个空间流和时间流，我们将类别标记与输入数据拼接。

我们使用 PyTorch 框架实现该模型，并采用 Adam 优化。对于 UTD-MHAD 数据集上的实验，我们使用 NVIDIA RTX 3090 GPU；对于 MMAct 数据集上的实验，我们使用 NVIDIA A40 GPU。我们使用 Top-1 准确率作为 UTD-MHAD 数据集实验的评估指标，并使用 F1 分数作为 MMAct 数据集实验的评估指标。

5. 结果与比较

5.1. 总体比较

我们通过在两个多模态人体动作识别数据集 UTD-MHAD 和 MMAct 上进行实验，评估了 DMFT 的性能。

对于 UTD-MHAD 数据集，如上所述，我们采用 50-50被试评估协议和留一被试评估协议。UTD-MHAD数据集上的实验结果见表1和表2。我们将所提出的方法与基线方法以及较新的多模态方法进行比较。在 50-50 受试者协议下，结果表明，DMFT教师模型使用骨架和惯性数据，准确率达到 93.97%，优于其他多模态方法。DMFT学生模型的准确率为 92.12%，仅比其前身 MATN模型低 0.6%。在留一主体交叉验证设置下，DMFT教师模型使用 RGB、骨架和惯性数据，准确率达到 98.20%，优于其他多模态方法。同时，与最先进方法相比，DMFT教师模型和 DMFT学生模型均取得了有竞争力的结果。

对于 MMAct 数据集，我们采用跨主体评估协议和跨会话评估协议，并使用 F1分数作为评估指标。

表1：UTD-MHAD数据集上的50-50被试性能比较。S：骨架，D：深度，I：惯性，aug.：数据增强。

方法	模态组合	准确率（%）
MHAD [1]	I+D	79.10
MHAD [1]	I+D	81.86
Gimme Signals [35]	I+S	76.13
Gimme Signals [35]	I+S（数据增强）	86.53
MATN	I+S	92.72
DMFT（教师）	I+S	**93.97**
DMFT（知识蒸馏）	I+S	92.12

表2：UTD-MHAD数据集上的留一受试者交叉验证（LOSO）性能比较。R：RGB，S：骨架，D：深度，I：惯性。

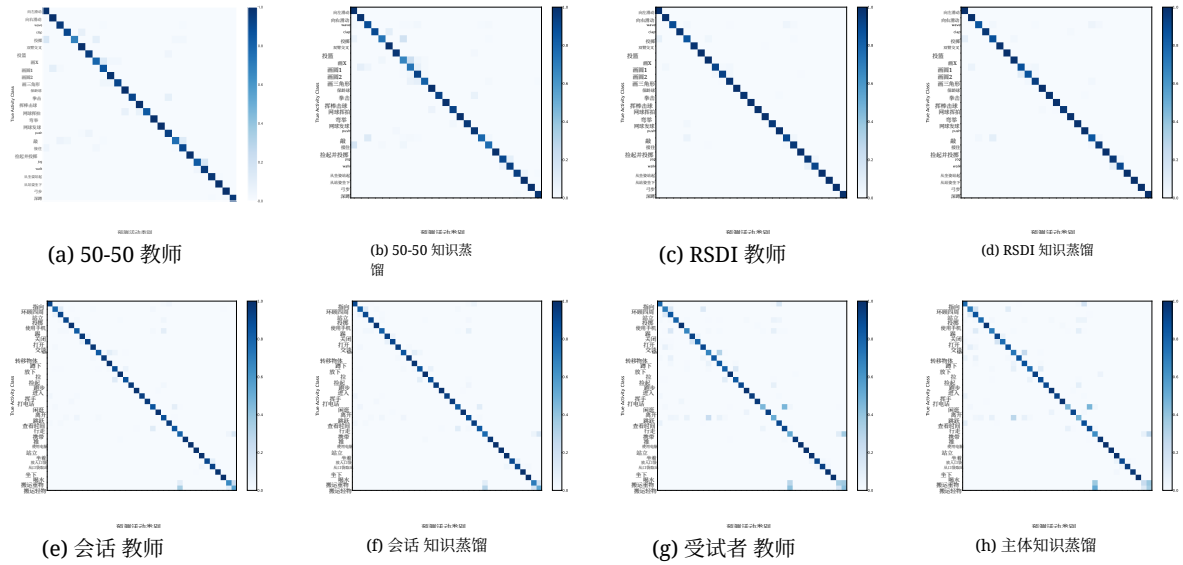
方法	准确率 (%)		
	R+S	R+S+I	R+S+D+I
Keyless [41]	90.20	92.67	83.87
HAMLET [36]	95.12	91.16	90.09
Multi-GAT [37]	**96.27**	96.75	97.56
Mumu	96.10	97.44	**97.60**
MATN	90.37	97.62	97.46
DMFT（教师）	93.06	**98.20**	97.52
DMFT（知识蒸馏）	90.26	96.52	96.53

表3：MMAct 数据集上的跨主体性能比较。R：RGB，I：惯性。

方法	模态组合	F1分数（%）
SMD [54]	I+R	63.89
学生 [2]	R	64.44
多教师 [2]	I	62.67
MMD [2]	I+R	64.33
MMAD [2]	I+R	66.45
REPDIB+MM (HAMLET) [55]	I+R	57.47
REPDIB+MM（Keyless） [55]	I+R	63.22
REPDIB+MM（REPDIB+Uni） [55]	I+R	69.39
HAMLET [36]	I+R	69.35
PSKD [51]	I+R	71.42
Keyless [41]	I+R	71.83
Multi-GAT [37]	I+R	75.24
SAKDN [50]	I+R	77.23
Mumu	I+R	76.28
MATN	I+R	**83.67**
DMFT（教师）	I+R	83.29
DMFT（知识蒸馏）	I+R	82.54

表4：MMAct 数据集上的跨会话性能比较。R：RGB，I：惯性。

方法	模态组合	F1分数 (%)
MMAD [2]	I+R +RGB	74.58
MMAD（融合） [2]	I+R +RGB	78.82
Keyless [41]	I+R +RGB	81.11
SAKDN [50]	I+R +RGB	82.77
HAMLET [36]	I+R +RGB	83.89
Multi-GAT [37]	I+R +RGB	91.48
Mumu	I+R +RGB	87.50
MATN	I+R +RGB	**91.85**
DMFT（教师）	I+R +RGB	91.62
DMFT（知识蒸馏）	I+R +RGB	91.09



(a) 50-50 教师 (b) 50-50 知识蒸馏 (c) RSDI 教师 (d) RSDI 知识蒸馏

(e) 会话 教师 (f) 会话 知识蒸馏 (g) 受试者 教师 (h) 主体知识蒸馏

图 4：在 UTD-MHAD 数据集和 MMAct 数据集上进行整体实验的混淆矩阵。子图 (a)-(d) 在 UTD-MHAD 数据集上运行，其中 (a)、(b) 是在 50-50主体设置下使用惯性数据和骨架数据得到的，(c)、(d) 是在留一主体交叉验证设置下使用 RGB、深度、骨架和惯性数据得到的。(e)-(h) 在 MMAct 数据集上使用惯性数据和 RGB 数据运行，其中 (e)、(f) 是在跨会话设置下得到的，(g)、(h) 是在跨主体设置下得到的。

实验结果如表3和表4所示。结果表明，DMFT教师模型的表现优于其他多模态方法，但不及其前身MATN模型；在跨主体协议下，其得分为 83.29%，在跨会话协议下为 91.62%。与此同时，DMFT学生模型在跨主体设置下达到 82.54%，在跨会话设置下达到 91.09%，略低于教师模型。

总体结果表明，DMFT取得了有竞争力的表现，并优于若干最先进水平的方法。总体而言，与非注意力模型相比，注意力模型表现更好，因为注意力机制有助于提取显著信息。结果表明，借助 MSTT 模块，DMFT 能够提取显著的空间特征和时间特征，取得优于非注意力方法的结果。虽然多模态方法似乎能够提升泛化能力，但该领域仍缺乏充分探索。其他基于注意力的方法采用后期融合方法来融合多模态特征，即多模态特征经过特征提取模块后再进行拼接。然而，DMFT采用TMT模块，有助于对多模态特征进行中期融合。因此，在特征提取过程中，多模态流能够共享互补信息，从而提升性能。结果表明，DMFT教师模型在具有不同模态组合的两个数据集上均表现良好。例如，在UTD-MHAD 50-50主体设置下，DMFT教师模型的准确率达到93.97%，比最先进水平的MATN模型高1.25%。当使用RGB特征时会引发隐私问题，因此这一点尤为有益。

从 MMAct 数据集的结果中可以看出，跨主体设置与跨会话设置之间存在性能差距，对于所有模型而言，跨会话设置下的性能都要高得多。这主要是因为跨会话设置属于主体相关设置，因此训练集和测试集都具有同一受试者的特征。当模型部署到现实世界场景中时，所分析的将是新受试者的数据，而不仅仅是参与者的数据。在这种情况下，依据主体相关实验协议开发模型并评估其性能将导致完全不同的结果，模型的性能也会被高估。这与我们的动机一致：我们设计主体无关的实验协议来评估模型性能并检验其泛化能力，因为这样得到的结果会更加准确。为了获得更好的泛化能力，还应探索更多方法，以改进

表5：知识蒸馏方法有效性的性能比较。R：RGB，S：骨架，D：深度，I：惯性。对于 UTD-MHAD 数据集的结果，采用 Top-1 准确率；对于 MMAct 数据集的结果，采用 F1分数。

数据集	设置	模态	教师	学生	学生（知识蒸馏）
UTD-MHAD	50-50 受试者	S+I	93.97	90.93	92.12 (1.19↑)
	LOSO	R+S	93.06	89.96	90.26 (0.30↑)
		R+S+I	98.20	96.41	96.52 (0.11↑)
		R+S+D+I	97.52	96.27	96.53 (0.26↑)
MMAct	跨受试者	I+R	83.29	82.23	82.54 (0.31↑)
	跨会话	I+R	91.62	90.90	91.08 (0.18↑)

表 6：知识蒸馏方法效率的性能比较。R：RGB，S：骨架，D：深度，I：惯性 l.

数据集	设置	模态	模型	训练时间（秒）	模型大小（MB）
UTD-MHAD	LOSO	R+S	教师	6.44	1087
			学生（知识蒸馏）	2.32	364
		R+S+I	教师	11.74	1188
			学生（知识蒸馏）	3.68	375
		R+S+D+I	教师	16.08	1886
			学生（知识蒸馏）	5.88	577
MMAct	跨被试	I+R	教师	84.74	1366
			学生（知识蒸馏）	44.39	321
	跨会话		教师	81.05	2567
			学生（知识蒸馏）	43.20	588

表7：TFT令牌在UTD-MHAD数据集上的性能比较。

TFT令牌数量	2	4	8	16
准确率（%）	93.14	**93.72**	92.65	92.93

实验效果。同时，尽管在 MMAct 数据集上开展实验仍有很大的改进空间，但 UTD-MHAD 数据集的潜力似乎已经得到了充分挖掘。这是因为 UTD-MHAD 仅包含少量样本（861 个视频片段）。未来，亟需构建全面且大规模的多模态数据集。

5.2. 时间中期融合令牌的影响

在DMFT教师模型中，我们使用TMT令牌对多模态时序特征进行中期融合。本节开展消融研究，以评估TMT令牌的数量是否会对中期融合过程产生较大影响。我们在采用50-50主体设置的 UTD-MHAD数据集上进行实验。构建模型时唯一的差异在于使用了不同数量的TMT令牌。结果如表7所示。

结果表明，使用更多 TMT令牌不会对模型的性能产生显著的积极影响。这与 BMT 的结论一致：[53] 只需少量融合令牌，便足以在多模态特征之间共享通用信息。因此，我们使用 4 个 TMT令牌，这样既能降低计算成本，又能获得更好的性能。

5.3. 知识蒸馏的有效性

本节开展消融研究，以评估知识蒸馏方法对提升学生网络性能的影响。对于每种实验设置，我们训练一个未应用知识蒸馏步骤的原始学生网络。结果如表5所示。对于在UTD-MHAD数据集上进行的实验，我们使用Top-1 准确率；对于在MMAct数据集上进行的实验，我们使用F1分数。

结果表明，使用教师网络训练学生网络能够提升性能。在 UTD-MHAD数据集上采用 50-50 设置时，提升幅度最大，为 1.19%。对于 MMAct 数据集，采用跨主体设置时，性能提升了 0.31%。虽然在跨会话设置下也有小幅提升（0.18%），但该评估协议属于主体相关，因此无法反映真实世界中的情况。结果与

表 8：知识蒸馏方法效率的性能比较。R：RGB，S：骨架，D：深度，I：惯性。对于 UTD-MHAD 数据集上的结果，采用 Top-1 准确率；对于 MMAct 数据集上的结果，采用 F1分数。

数据集	设置	模态	模型	结果				
				R	D	S	I	总体
UTD-MHAD	LOSO	R+S	教师	61.02	-	91.77	-	93.06
			学生（原始）	56.46	-	89.57	-	89.96
			学生（知识蒸馏）	55.86	-	89.83	-	90.26
		R+S+I	教师	59.05	-	91.16	79.19	98.20
			学生（原始）	53.80	-	88.59	78.07	96.41
			学生（知识蒸馏）	54.95	-	89.36	78.71	96.52
		R+S+D+I	教师	57.03	39.75	90.22	78.27	97.52
			学生（原始）	53.82	31.04	88.04	77.74	96.27
			学生（知识蒸馏）	54.29	34.31	88.11	77.96	96.53
MMAct	跨受试者	R+I	教师	65.57	-	-	69.17	83.29
			学生（原始）	64.58	-	-	67.68	82.23
			学生（知识蒸馏）	65.59	-	-	67.82	82.54
	跨会话		教师	74.28	-	-	82.75	91.62
			学生（原始）	74.27	-	-	80.25	90.90
			学生（知识蒸馏）	75.42	-	-	80.94	91.08

我们的动机相符：通过采用知识蒸馏方法，可以将复杂教师模型中的知识迁移到规模更小的学生网络中，从而提升其性能。

5.4. 知识蒸馏的效率

本节将通过对比评估，展示利用知识蒸馏方法训练学生模型的效率。结果如表 6 所示，其中包括教师模型、学生模型和知识蒸馏学生模型每个训练轮次所需的训练时间及保存的模型大小。UTD-MHAD 数据集上的实验在 NVIDIA RTX3090 图形处理器上运行，MMAct 数据集上的实验在 NVIDIA A40 GPU 上运行。

结果表明，应用知识蒸馏方法后，训练时间和模型大小均显著减少，这一结果在不同设置下均有所体现。这支持了我们的动机：采用知识蒸馏方法训练学生模型，可以降低模型的时间和空间成本。鉴于现实场景中的硬件设备可能受限，我们的方法将有助于实际部署。

5.5. 多模态学习的影响

本节进一步研究 DMFT 在多模态学习中的表现。我们在 UTD-MHAD 数据集和 MMAct 数据集上，采用不同的实验方案和模态组合开展研究。对于每种实验设置，我们训练 3 个模型：教师网络、原始学生网络，以及经过知识蒸馏的学生网络。对于每种模态组合，我们比较各模态流的输出与整体输出的性能。结果见表 8。此外，我们给出图 5 和图 6，展示各项活动更详细的性能评估。对于 UTD-MHAD 数据集上的实验，我们采用 Top-1 准确率；对于 MMAct 数据集上的实验，则采用 F-1 分数。结果表明，DMFT 能够捕获各模态提供的互补信息，并充分利用显著特征，从而取得更好的结果。

多模态学习的优势之一是利用互补信息，以获得更准确、更稳健的结果。表 8 的结果表明，在所有实验设置下，整体结果均表现更佳。尽管每个模态流的性能可能有所不同，但通过聚合各模态的输出，模型能够捕获显著的模态特定特征。在这种情况下，即使某个模态输入失效，模型仍然能够利用其他模态进行特征提取并捕获互补信息，从而生成稳健预测。例如，在图 5 中，对于类别 1（向左滑动），骨骼流的表现优于其他三种模态；而对于类别 3（挥手），惯性流则优于其他模态。然而，在聚合所有模态流的信息后，整体预测优于每个模态流的表现。

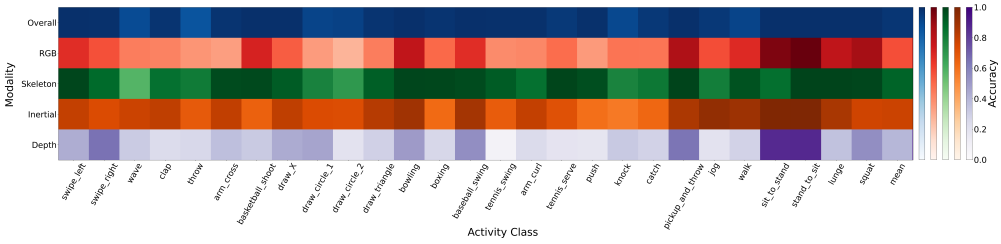


图 5：UTD-MHAD 数据集上各模态贡献的性能比较（Top-1 准确率）

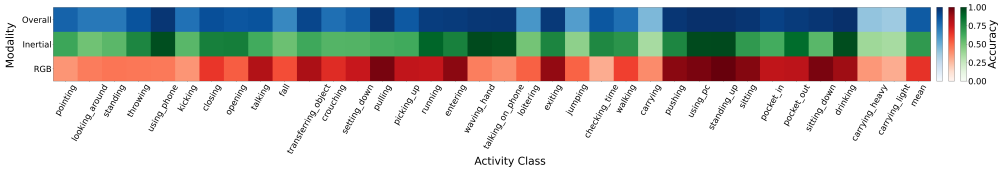


图 6：MMAct 数据集上各模态贡献的性能比较（F1分数）

此外，图 6 显示，尽管 RGB 和惯性流在类别 30（口袋）上的性能均低于 80%，但整体预测的 F1分数仍超过 90%。因此，应用多模态学习能够产生更稳健的结果，这对于实际部署十分有益，因为信号传输可能会受到影响。

6. 结论

我们的工作主要旨在开发一种有效且高效的多模态人体活动识别方法，使其能够部署于资源受限环境，并适用于主体无关设置。我们提出了 DMFT，这是一种基于知识蒸馏的注意力方法，通过对多模态特征进行中期融合来解决多模态人体活动识别任务。我们首先通过统一表征学习层对多模态数据进行编码。随后，应用多模态时序中期融合 Transformer 网络，提取各模态显著的时空特征，并进行时序中间融合，以进一步提取和融合多模态特征。我们还采用知识蒸馏方法，利用教师网络训练一个更简单的学生网络，从而在降低计算与空间成本的同时提升性能。我们在两个公开的多模态数据集 UTD-MHAD 和 MMAct 上，采用不同的实验设置进行了全面实验，以评估 DMFT 的性能。实验结果表明，我们的模型能够利用显著的多模态特征并取得有竞争力的结果，同时在有限环境下实现更优且稳健的性能，

在未来，我们计划开发有效、高效且稳健的人体活动识别模型，以更好地应对多模态人体活动识别场景中的受试者间差异挑战。

声明

本研究得到了合作研究中心项目（CRCP）资助的“DeepIoT”项目的支持。

参考文献

[1] C. Chen, R. Jafari, N. Kehtarnavaz，Utd-mhad：一个利用深度摄像头和可穿戴惯性传感器进行人体动作识别的多模态数据集，发表于：2015 年 IEEE 国际图像处理会议（ICIP），IEEE，2015 年，第 168–172 页。[2] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, T. Murakami，Mmact：一个用于跨模态人体动作理解的大规模数据集，发表于：IEEE/CVF 国际计算机视觉会议论文集，2019 年，第 8658–8667 页。[3] M. Brand, A. Hertzmann，Style machines，发表于：第 27 届年度计算机图形学与交互技术会议论文集，SIGGRAPH ’00，ACM Press/Addison-Wesley Publishing Co.，美国，2000，183 192 https:// 第 – 页。doi.org/10.1145/344779.344865 网址：。doi: 10.1145/344779. 344865 .[4] V. Pavlovic, J. M. Rehg, J. MacCormick，学习人体运动的切换线性模型，发表于：第 13 届国际神经信息处理系统会议论文集，NIPS’00，MIT Press，美国马萨诸塞州剑桥，2000 年，第 942–948 页。[5] R. Urtasun, D. J. Fleet, A. Geiger, J. Popovi´c, T. J. Dar- rell, N. D. Lawrence，拓扑约束潜变量模型，发表于：第 25 届国际机器学习会议论文集，ICML ’08，Association

计算机协会，纽约州纽约市，美国，2008，页 1080 1087 https://doi.org/10.1145/1390156.－。1390292 10.1145/1390156.1390292网址：。doi：。[6] J. M. Wang、D. J. Fleet、A. Hertzmann，人体运动的高斯过程动力学模型，《IEEE模式分析与机器智能汇刊》30（2008）283–298。doi：10.1109/TPAMI.2007.1167。[7] I. Akhter、T. Simon、S. Khan、I. Matthews、Y. Sheikh，双线性时空基模型，《ACM图形学汇刊》31 2012 https://doi.org/10.1145/2159516.（）。2159523 10.1145/2159516.2159523网址：。doi：。[8] I. Sutskever、G. Hinton、G. Taylor，循环时间受限玻尔兹曼机，载于：第21届神经信息处理系统国际会议论文集，NIPS'08，Curran Associates Inc.，雷德胡克，纽约州，美国，2008，页1601–1608。[9] G. W. Taylor、L. Sigal、D. J. Fleet、G. E. Hinton，用于三维人体姿态跟踪的动态二元潜变量模型，载于：2010年IEEE计算机学会计算机视觉与模式识别会议，2010， 631 638 10.1109/CVPR.页–。2010.5540157doi：。[10] K. Fragkiadaki、S. Levine、P. Felsen、J. Malik，用于人体动力学的循环网络模型，载于：2015年IEEE国际计算机视觉会议（ICCV），2015，页4346–4354。doi：10.1109/ICCV.2015.494。[11] X. Wang、L. Gao、J. Song、H. Shen，超越帧级卷积神经网络：用于视频动作识别的显著性感知三维卷积神经网络与长短期记忆网络，《IEEE信号处理快报》24（2016）510–514。[12] J. Martinez、M. J. Black、J. Romero，使用循环神经网络进行人体运动预测，载于：IEEE计算机视觉与模式识别会议论文集，2017，页2891–2900。[13] H. Lee、M. Jung、J. Tani，通过多时空尺度循环神经网络识别视觉感知的组合式人体动作，《IEEE认知与发展系统汇刊》10（2017）1058–1069。[14] L.-Y. Gui、Y.-X. Wang、X. Liang、J. M. Moura，对抗式几何感知人体运动预测，载于：欧洲计算机视觉会议（ECCV）论文集，2018，页786–803。[15] M. Guo、E. Chou、D.-A. Huang、S. Song、S. Yeung、L. Fei-Fei，用于少样本三维动作识别的神经图匹配网络，载于：欧洲计算机视觉会议（ECCV）论文集，2018，页653–669。[16] L. Fan、Z. Wang、H. Wang，基于决策树的人体活动识别模型，载于：2013年先进云计算与大数据国际会议，IEEE，2013，页64–68。[17] P. Paul、T. George，一种在智能手机上进行人体活动识别的有效方法，载于：2015年IEEE工程与技术国际会议（ICETECH），IEEE，2015，页1–3。[18] K. M. Chathuramali、R. Rodrigo，使用支持向量机加速人体活动识别，载于：新兴地区信息与通信技术进展国际会议（ICTer2012），IEEE，2012，页197–203。[19] L. Liu、Y. Peng、M. Liu、Z. Huang，一种采用时间序列形状片段多层模型的基于传感器的人体活动识别系统，《知识型系统》90（2015）138–152。[20] S. Fallmann、J. Kropf，使用隐马尔可夫模型对连续数据进行人体活动识别及纳入离散数据的问题，载于：2016年IEEE国际会议：普适智能与计算、先进可信计算、可扩展计算与通信、云计算与大数据计算、人联网及智慧世界大会（UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld），IEEE，2016，页121–126。[21] M. H. Kabir、M. R. Hoque、K. Thapa、S.-H. Yang，面向家庭环境中人体活动识别的双层隐马尔可夫模型，《国际分布式传感器网络杂志》12（2016）4560365。[22] W. Jiang、Z. Yin，利用深度卷积神经网络和可穿戴传感器进行人体活动识别，载于：第23届ACM多媒体国际会议论文集，2015年，第1307–1310页。[23] J. B. Yang、M. N. Nguyen、P. P. San、X. L. Li、S. Krishnaswamy，基于多通道时间序列的深度卷积神经网络用于人体活动识别，载于：第二十四届人工智能国际会议论文集，IJCAI'15，AAAI出版社，2015年，第3995–4001页。[24] L. Peng、L. Chen、Z. Ye、Y. Zhang，Aroma：一种基于深度多任务学习、利用可穿戴传感器识别简单与复杂人体活动的方法，《ACM交互、移动、可穿戴与泛在技术论文集》2（2018）1–16。[25] K. Chen、L. Yao、D. Zhang、B. Guo、Z. Yu，多智能体注意力活动识别，载于：第二十八届人工智能国际联合会议论文集，IJCAI-19，国际人工智能联合会议组织，2019年，第1344 1350 https://doi.org/10.24963/ijcai.–页。2019/186 10.24963/ijcai.2019/186网址：。doi：。[26] H. Xue、W. Jiang、C. Miao、F. Ma、S. Wang、Y. Yuan、S. Yao、A. Zhang、L. Su，DeepMV：用于无设备人体活动识别的多视图深度学习，《ACM交互、移动、可穿戴与泛在技术论文集》4（2020）1–26。[27] L. Bai、L. Yao、X. Wang、S. S. Kanhere、B. Guo、Z. Yu，用于活动识别的对抗式多视图网络，《ACM交互、移动、可穿戴与泛在技术论文集》4（2020）1–22。[28] Y. Guan、T. Pl¨otz，利用可穿戴设备进行活动识别的深度LSTM学习器集成，《ACM交互、移动、可穿戴与泛在技术论文集》1（2017）1–28。[29] V. S. Murahari、T. Pl¨otz，人体活动识别中的注意力模型，载于：2018年ACM可穿戴计算机国际研讨会论文集，2018年，第100–103页。[30] M. Zeng、H. Gao、T. Yu、O. J. Mengshoel、H. Langseth、I. Lane、X. Liu，通过连续注意力理解并改进用于人体活动识别的循环网络，载于：2018年ACM可穿戴计算机国际研讨会论文集，2018年，第56–63页。[31] M. Bock、A. H¨olzemann、M. Moeller、K. Van Laerhoven，利用浅层LSTM改进HAR深度学习，载于：2021年可穿戴计算机国际研讨会，2021年，第7–12页。[32] H. Wu、Z. Zhang、X. Li、K. Shang、Y. Han、Z. Geng、T. Pan，一种基于微分时空LSTM的新型踏板肌肉骨骼响应，用于人体活动识别，《知识型系统》261（2023 110187 https://www.sciencedirect.）。com/science/article/pii/S0950705122012837网址：。doi：https://doi.org/10.1016/j.knosys.2022.110187。[33] H. Guo、L. Chen、L. Peng、G. Chen，利用分类器集成的多样性进行基于可穿戴传感器的多模态人体活动识别，载于：2016年ACM普适与泛在计算国际联合会议论文集，2016年，第1112–1123页。[34] L. Yao、Q. Z. Sheng、B. Benatallah、S. Dustdar、X. Wang、A. Shemshadi、S. S. Kanhere，Wits：一种支持物联网的计算框架，用于个性化智能家居中的活动识别，《Computing》100（2018）369–385。

[35] R. Memmesheimer、N. Theisen、D. Paulus，Gimme Signals：用于多模态活动识别的判别式信号编码，载于：2020年IEEE/RSJ国际智能机器人与系统会议（IROS），IEEE，2020年，第10394–10401页。[36] M. M. Islam、T. Iqbal，HAMLET：一种基于层次化多模态注意力的人体活动识别算法，载于：2020年IEEE/RSJ国际智能机器人与系统会议（IROS），IEEE，2020年，第10285–10292页。[37] M. M. Islam、T. Iqbal，Multi-GAT：一种基于图注意力的层次化多模态表示学习方法，用于人体活动识别，《IEEE机器人与自动化快报》6（2021）1729–1736。[38] J. Li、L. Yao、B. Li、X. Wang、C. Sammut，用于多模态人体活动识别的多智能体Transformer网络，载于：第31届ACM信息与知识管理国际会议论文集，2022年，第1135–1145页。[39] K. Chen、L. Yao、D. Zhang、X. Wang、X. Chang、F. Nie，一种用于人体活动识别的半监督循环卷积注意力模型，《IEEE神经网络与学习系统汇刊》31（2019）1747–1756。[40] V. Mnih、N. Heess、A. Graves等，视觉注意力循环模型，《神经信息处理系统进展》27（2014）。[41] X. Long、C. Gan、G. De Melo、X. Liu、Y. Li、F. Li、S. Wen，用于视频分类的多模态无关键注意力融合，载于：第三十二届AAAI人工智能会议，2018年。[42] L. Chen、Y. Zhang、L. Peng，Metier：一种基于深度多任务学习、利用可穿戴传感器进行活动与用户识别的模型，《ACM交互、移动、可穿戴与泛在技术论文集》4（2020）1–18。[43] S. Liu、S. Yao、J. Li、D. Liu、T. Wang、H. Shao、T. Abdelzaher，Giobalfusion：一种用于多传感器信息融合的全局注意力深度学习框架，《ACM交互、移动、可穿戴与泛在技术论文集》4（2020）1–27。[44] B. Li、W. Cui、W. Wang、L. Zhang、Z. Chen、M. Wu，用于人体活动识别的双流卷积增强Transformer，载于：AAAI人工智能会议论文集，第35卷，2021年，第286–293页。[45] M. M. Islam、T. Iqbal，Mumu：一种基于协同多任务学习的引导式多模态融合，AAAI，2022年。[46] S. Suh、V. F. Rey、P. Lukowicz，Tasked：基于Transformer的对抗学习，通过自知识蒸馏利用可穿戴传感器进行人体活动识别，《知识型系统》260（2023）110143。[47] Z. Chen、L. Zhang、Z. Cao、J. Guo，从手工设计特征中蒸馏知识以进行人体活动识别，《IEEE工业信息学汇刊》14（2018）4334–4342。[48] F. M. Thoker、J. Gall，用于动作识别的跨模态知识蒸馏，载于：2019年IEEE国际图像处理会议（ICIP），IEEE，2019年，第6–10页。[49] R. Gao、T.-H. Oh、K. Grauman、L. Torresani，Listen to look：通过预览音频进行动作识别，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10457–10467页。[50] Y. Liu、K. Wang、G. Li、L. Lin，用于传感器到视觉动作识别的语义感知自适应知识蒸馏，《IEEE图像处理汇刊》30（2021）5573–5588。[51] J. Ni、A. H. Ngu、Y. Yan，用于人体动作识别的渐进式跨模态知识蒸馏，载于：第30届ACM多媒体国际会议论文集，2022年，第5903–5912页。[52] J. Ni、R. Sarbajna、Y. Liu、A. H. Ngu、Y. Yan，用于视觉到传感器动作识别的跨模态知识蒸馏

识别，载于：ICASSP 2022-2022 IEEE国际声学、语音与信号处理会议（ICASSP），IEEE，2022，第4448–4452页。[53] A. Nagrani、S. Yang、A. Arnab、A. Jansen、C. Schmid、C. Sun，《用于多模态融合的注意力瓶颈》，《神经信息处理系统进展》第34卷（2021），第14200–14213页。[54] G. Hinton、O. Vinyals、J. Dean，《蒸馏神经网络中的知识》，NIPS深度学习与表示学习研讨会（2014）。[55] R. Islam、H. Zang、M. Tomar、A. Didolkar、M. M. Islam、S. Y. Arnob、T. Iqbal、X. Li、A. Goyal、N. Heess等，《通过离散信息瓶颈进行深度强化学习中的表示学习》，arXiv预印本 arXiv:2212.13835（2022）。