基于骨骼动作识别的交叉注意多尺度时空 Transformer

高逸畅[1] 林哲煌[1]
GAO Yichang LIN Zhehuang

摘 要 近年来，Transformer 在计算机视觉各类任务中成效显著。但基于 Transformer 的方法在骨骼数据多尺度特征学习上存在局限，而多尺度时空特征蕴含着关键的全局与局部信息，这对于骨骼动作识别而言至关重要。为此，文章探索了骨骼序列在空间和时间维度上的多尺度特征表示，并提出了一种用于跨尺度特征融合的高效交叉注意力机制。此外，提出了一个多尺度特征提取和融合 Transformer（multi-scale feature extraction and fusion transformer, MFEF-Former），可以分为两种类型：（1）用于空间建模的 MFEF-SFormer，利用自注意力捕获关节间和身体部位间的相关性，然后利用交叉注意力进行多尺度空间特征融合，以建模关节和身体部位之间的相关性；（2）用于时间建模 MFEF-TFormer，利用自注意力捕获多尺度时间特征，并通过交叉注意力融合多尺度特征。这两个组件在一个双流网络中结合，并在两个大型数据集 NTU RGB+D 和 NTU RGB+D 12 上进行了评估。实验表明，文章所提方法在基于骨骼的动作识别方面优于其他基于 Transformer 的方法。

关键词 动作识别；Transformer；人体骨架；多尺度特征

doi：10.3969/j.issn.1672-9528.2025.09.011

0 引言

随着计算机视觉技术的快速发展，人类动作识别已成为一项日益流行和重要的任务，其应用涵盖交通运输、医疗、娱乐、教育、安全监控和人机交互等多个领域。动作识别可以通过多种数据模态实现，包括 RGB 视频和骨骼数据。与 RGB 视频相比，骨骼数据展现出一些优势，它对身体特征的变化具有鲁棒性，并且几乎不受变化环境、复杂背景、光照条件和其他噪声源的影响。此外，深度传感器和人体姿态估

1. 广东工业大学自动化学院 广东广州 510006

计算机与数字工程 , 2023, 51(2): 440-444.

[7]TANG D W, KUPPENS P, GEURTS L, et al. End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network[J/OL].Eurasip journal on audio, speech, and music processing.2021[2025-03-26]. https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-021-00208-5.

[8]WEI S Y, ZOU S, LIAO F F, et al. A comparison on data augmentation methods based on deep learning for audio classification[J].Journal of physics: conference series, 2020, 1453: 012085.

[9]ZHAO J, ZHANG W Q.Improving automatic speech recognition performance for low-resource languages with self-supervised models[J].IEEE journal of selected topics in signal processing, 2022,16(6):1227-1241.

[10]HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J].IEEE/ACM transactions of audio, speech, and language processing,2021,29:3451-3460.

[11]YANG S L, YU Z T, WANG W J, et al. Sequence modeling[C]// Proceedings of the 23rd Chinese National Conference on Computational Linguistics. Brussels: ACL, 2024: 625-636.

[12]CHEN Z, SHAO Y F, MA Y, et al. Improving acoustic scene classification in low-resource conditions[C]//ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2024:12-30.

【作者简介】

汪慧娟（1999—），女，河南信阳人，硕士研究生，研究方向：计算机技术、移动互联应用及开发技术。

李宏滨（1968—），男，山西晋中人，硕士，副教授，研究方向：智能图像处理。

（收稿日期：2025-04-10 修回日期：2025-09-08）

计技术[1–3]的快速发展使准确捕获骨骼的运动数据更加容易。因此，基于骨骼的人类动作识别近年来受到了广泛关注。

早期深度学习算法将人体骨骼关节信息视为一组独立的特征，并将其作为特征序列或伪图像进行处理。然后将这些特征输入到卷积神经网络（CNN）[4-5]或循环神经网络（RNN）[6-7]中进行动作分类。然而，人体具有天然的拓扑结构，而基于 CNN 或 RNN 的算法忽略了运动过程中关节之间固有的时空相关性。关节之间的相关性由人体骨骼的不规则拓扑结构表示，这是基于骨骼的动作识别中不应忽略的关键信息。

图卷积网络（graph convolutional networks, GCN）[8]已证明了其在处理像骨骼数据这样的不规则结构图方面的有效性。基于 GCN 的方法[9–14]在其发展中已经达到了一定的成熟度，并且广泛应用于基于骨骼的动作识别。基于 GCN 的方法的基本思想是将人体骨骼序列视为一系列图，其中身体关节代表图的节点，而这些关节之间的骨骼连接代表图的边，遵循身体的拓扑结构。Yan 等人[9]最初提出了时空图卷积网络（spatial-temporal GCN, ST-GCN），该网络通过 GCN 从关节中提取空间特征信息，并将其与时间卷积网络（temporal convolutional networks, TCN）结合起来以捕获时间特征信息，从而实现人体动作识别。这种方法取得了非常高的性能，超过了当时基于 RNN 和基于 CNN 的方法。因此，以 ST-GCN 为代表的基于 GCN 的方法成为基于骨骼的动作识别任务的主流，将 GCN 与时间卷积网络（TCN） 相结合，以提取关节的时空特征并将其应用于人体动作识别。然而，尽管基于 GCN 的方法擅长建模空间关节间相关性，但其在提取和融合多尺度特征方面遇到了挑战。多尺度特征在图像和视频数据中起着至关重要的作用，并且在各种计算机视觉任务中得到了广泛研究，包括图像分类、目标检测和姿态估计[15–17]。骨骼序列数据是一种具有独特格式的数据，同时涉及时间和空间维度，表现出多尺度时空特征。由于现有基于 GCN 的方法的局限性，如何有效地处理骨骼数据的多尺度时空特征以进行基于骨骼的动作识别仍然是一个开放性问题。

Transformer[18]是一种新型深度学习模型，自引入以来在自然语言处理方面取得了巨大成功。由于其在序列建模和全局信息感知方面的强大能力，Transformer 及其变体[19–23]在各种计算机视觉任务中表现出卓越的性能。然而，在骨架动作识别领域，基于 Transformer 的方法[24–26]仍相对有限，且未能充分利用多尺度时空特征，而这些特征对从骨骼数据中识别动作至关重要。因此，将 Transformer 应用于骨架动作识别，尤其是增强其处理多尺度特征的能力，仍然是一个值得深入探索的方向。

在本文中，为了解决人体骨骼数据的多尺度时空特征表示问题，提出了一个多尺度特征提取和融合 Transformer（multi-scale feature extraction and fusion transformer, MFEF-Former），由用于空间处理的 MFEF-SFormer 和用于时间处理的 MFEF-TFormer 组成。MFEF-SFormer 中，首先将短距离连接（关节）和长距离连接（身体部位）的空间特征作为两个独立分支输入，用于多尺度自注意力特征提取，从而同时捕获不同关节和身体部位之间的依赖关系。然后，利用本文提出的交叉注意力机制，在跨尺度特征融合层中将这些多尺度特征进行融合。每个分支生成一个非补丁（non-patch）token，即 CLS token，作为与其他分支交换信息的媒介，并通过注意力机制进行信息传递。MFEF-TFormer 采用相同的流程，但重点处理多尺度时间特征，特别是长期和短期序列。最后，结合 MFEF-SFormer 和 MFEF-TFormer，提出了交叉注意力多尺度时空 Transformer。

本文的贡献总结如下：

（1）提出了一个双分支 MFEF-Former 网络，该网络将原始骨骼序列数据划分为时间和空间两个分支，并进一步将其分割为时间维度和空间维度上不同尺度的特征表示。

（2）提出了专门用于处理骨骼序列多尺度时空特征的交叉注意力机制，从而实现高效的跨尺度特征融合。

1 本文方法

1.1 多尺度时空特征表示

在空间维度上，基于人体的物理结构和先验知识，将数据划分为单关节序列和身体部位序列，并考虑关节和特定身体区域之间的关系。能够捕获不同粒度级别的空间信息。

将人体所有关节划分为 p 个身体部位。例如，当 p=10 时，身体部位包括头部、左臂、右臂、左手掌、右手掌、躯干、左腿、右腿、左脚底和右脚底。每个身体部位由多个关节组成，如图 1 所示。由于每个身体部位中的关节数量不同，采用线性投影将这些部位的特征映射到统一的 C_s 维特征，从而增加潜在特征信息的表达能力。对于身体部位序列，可以被视为 $X_P=\left\{X_1^p,X_2^p,\ldots,X_t^p,\ldots,X_T^p\right\}\in\mathbf{R}^{C_s\times T\times p}$，其中 $X_t^p\in\mathbf{R}^{C_s\times p}$ 是第 t 帧的骨骼序列，具有 C_s 维特征和 p 个部位。

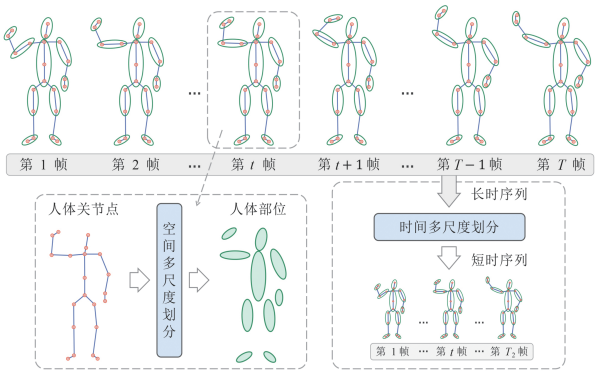

图 1 骨架序列的多尺度时空特征表示

由于骨骼序列数据在时间和空间维度上的固有差异，时间维度和空间维度所采用的多尺度特征划分方法是完全不同的。将每个关节的时间序列信息划分为长时序列和短时序列。长序列由单帧组成，指骨骼序列的原始数据$X_{\text{long}} \in \mathbf{R}^{C\times T\times N}$。另一方面，短序列包含来自多个帧的信息，并且通过从长序列中利用时间卷积网络（temporal convolutional networks, TCN）提取特征获得。对于短时序列，可以被视为具有较少帧数 T_2 的 $X_{\text{short}} \in \mathbf{R}^{C\times T_2\times N}$。

1.2 多尺度特征提取和融合 Transformer

1.2.1 MFEF-Former 概述

MFEF-SFormer 和 MFEF-TFormer 共享相同的结构，但参数不同。以 MFEF-SFormer 网络为例，图 2 为本文提出的 MFEF-SFormer 的网络架构。

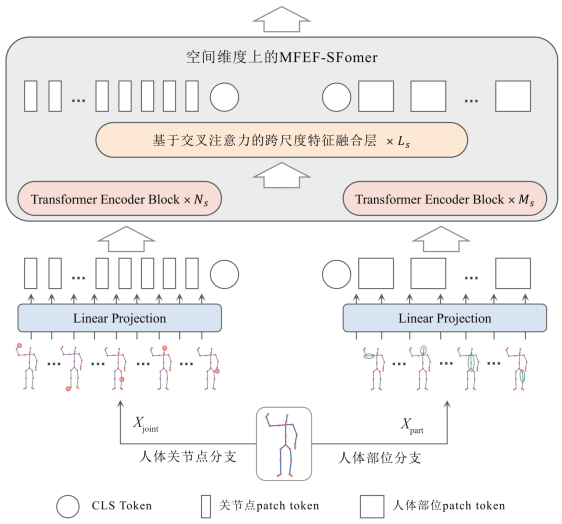

图 2 MFEF-SFormer 网络架构

MFEF-SFormer 是所提出的双流网络中空间流的关键组成部分。每个 MFEF-SFormer 都由一个双分支的常规 Transformer 编码器模块组成，该模块被堆叠 N_s 和 M_s 次用于多尺度空间特征提取然后进行跨尺度特征融合。根据不同的输入尺度，这两个分支分为关节分支和身体部位分支，如图2所示。原始的人体骨骼信息首先被处理成关节序列数据和身体部位序列数据。然后，这些序列通过 Embedding 层进行线性投影，将其转换为与编码器模块对应的 token 序列。MFEF-SFormer 执行两个核心任务：

（1）通过在 MFEF-SFormer 的两个分支中堆叠编码器模块，使用自注意力进行多尺度特征提取。

（2）使用交叉注意力进行跨尺度特征融合。从这两个分支中提取结果，并通过线性投影将其恢复到骨骼序列的原始大小。后结果被传递到下一个模块以进行进一步计算。

1.2.2 使用交叉注意力进行跨尺度特征融合

通过多个堆叠的 Transformer 编码器模块，关节分支和身体部位分支通过从骨骼序列中捕获关节（短距离）或身体部位（长距离）级别的特征来提取多尺度空间特征。因此，长序列分支和短序列分支捕获各自尺度上的特征。

在 Transformer 中执行编码器模块计算之前，一个 CLS token 被预先添加到每个 token 序列中，这是一个与其他 token 大小相同的特殊 token。每个编码器中的 CLS token 学习全局特征信息，通常用于图像分类任务中的最终分类。同理，关节分支中的 CLS token 学习关节间的相关性，而身体部位分支中的 CLS token 学习身体部位间的相关性。两个分支中的 CLS token 代表不同尺度的全局信息，在跨尺度特征融合中起着至关重要的作用。因此，对于短距离和长距离分支中的 CLS token，以 MFEF-SFormer 的关节分支为例，提出了一种使用交叉注意力的跨尺度特征融合策略，如图 3 所示。

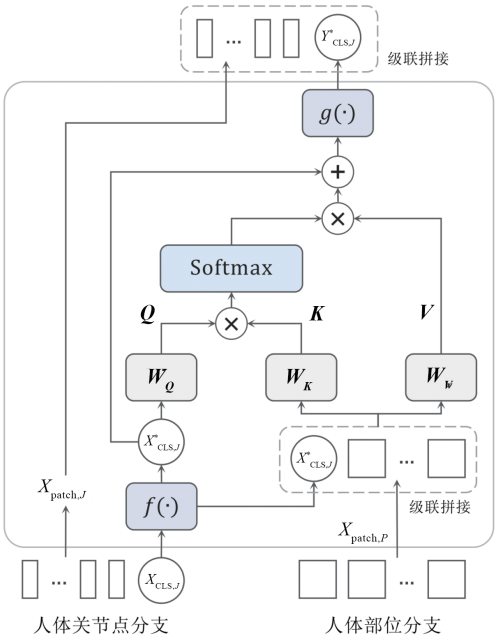

图 3 MFEF-SFormer 中交叉注意的特征融合

令$X_{\text{token},J}=\left[X_{\text{CLS},J}\,\|\,X_{\text{patch},J}\right]$为关节分支在单帧上通过常规 Transformer 编码器模块进行多尺度特征提取后的整个 token 序列，其中 $X_{\text{CLS},J}$ 是 CLS token，而$X_{\text{patch},J}=\{X_{1,J},X_{2,J},\ldots,X_{N,J}\}$是关节分支的补丁（patch）token。与之类似，令$X_{\text{token},P}=\left[X_{\text{CLS},P}\,\|\,X_{\text{patch},P}\right]$为身体部位分支的整个 token 序列，其中 $X_{\text{CLS},P}$ 是 CLS token，而$X_{\text{patch},P}=\{X_{1,P},X_{2,P},\ldots,X_{p,P}\}$是身体部位分支的补丁 token。对于关节分支，它首先从身体部位分支获取补丁 token 序列，并将其与自己的 CLS token 连接，用公式表示为：

$$X^*_{\text{token},J}=\left[f\left(X_{\text{CLS},J}\right)\|X_{\text{patch},P}\right] \tag{1}$$

式中：$f(\cdot)$ 是用于对齐维度的投影。然后，计算 $X^*_{\text{token},J}$ 和 $X^*_{\text{CLS},J}$ 之间的交叉注意力，如图 3 所示。

整个过程可以表示为：

$$\begin{aligned}\boldsymbol{Q}&=X^{*}_{\mathrm{CLS},J}\boldsymbol{W}_{\boldsymbol{Q}}\\ \boldsymbol{K}&=X^{*}_{\mathrm{token},J}\boldsymbol{W}_{\boldsymbol{K}}\\ \boldsymbol{V}&=X^{*}_{\mathrm{token},J}\boldsymbol{W}_{\boldsymbol{V}}\\ \mathrm{CAtt}\left(X^{*}_{\mathrm{token},J}\right)&=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k/h}}\right)\boldsymbol{V}\end{aligned}\tag{2}$$

式中：$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{W}_{\boldsymbol{V}}$ 是用于线性投影的参数矩阵；d_k 是 $\boldsymbol{K}$ 的维度；CAtt(·) 是交叉注意力函数；$g(\cdot)$ 是残差连接和投影。

结果 $Y^{*}_{\mathrm{CLS},J}$ 与关节分支的补丁 token 连接，用于最终融合。使用多头交叉注意力进行跨尺度特征融合的整个过程可以表示为：

$$\begin{aligned}Y_{\mathrm{CLS},J}&=f\left(X_{\mathrm{CLS},J}\right)+\mathrm{MCAtt}\left(\mathrm{LN}\left(\left[f\left(X_{\mathrm{CLS},J}\right)\|\,X_{\mathrm{patch},P}\right]\right)\right)\\ Z_{\mathrm{token},J}&=\left[g\left(Y_{\mathrm{CLS},J}\right)\|\,X_{\mathrm{patch},J}\right]\end{aligned}\tag{3}$$

式中：$Z_{\mathrm{token},J}$ 是最终交叉注意力融合后的 token 序列；LN(·) 是层归一化；MCAtt(·) 是具有一个注意力头的多头交叉注意力函数。与关节分支类似，身体部位分支 $X_{\mathrm{token},P}$ 也同步进行交叉注意力融合过程，获得最终的 token 序列 $X_{\mathrm{token},P}$。

在下一层的编码器中，CLS token 再次与原始分支的补丁 token 序列融合。这使得 CLS token 能够从另一个分支的不同尺度学习特征信息，然后将其传播回自己的 token 序列，从而丰富 token 序列的信息。在 L_s 个跨尺度特征融合层中，关节分支和身体部位分支都执行此交互过程。因此，该网络能够学习在动作发生过程中关节和身体部位之间的多尺度相关性。

1.3 交叉注意力多尺度时空 Transformer 网络

为了结合 MFEF-SFormer 和 MFEF-TFormer 模块，提出了一个具有双流架构的交叉注意力多尺度时空 Transformer，名为 2s-MFEF-STFormer，它与 ST-TR-agcn [27] 的架构相似，如图 4 所示。双流网络由并行的时序流和空间流组成，每个流都专注于执行特定于其各自维度的不同任务。

（1）时序流：在时序流中，首先应用时间卷积网络（temporal convolutional network, TCN）将每个关节的时间序列信息划分为具有多个尺度的长序列和短序列两个分支，然后将输出输入一组 K_t 个时间特征学习模块，每个模块由一个图卷积网络和一个 MFEF-TFormer 网络组成。GCN 旨在从关节中提取空间信息，而 MFEF-TFormer 网络则提取和融合多尺度的时间特征信息。遵循 Transformer 结构，输入特征通过批归一化层，并通过残差连接将输入特征与 MFEF-TFormer 的输出特征相加。

（2）空间流：在空间流中，基于人体的物理结构，首先将骨骼序列划分为单关节序列和身体部位序列两个分支。这两个分支的多尺度输入被传递到 K_s 层堆叠的空间特征学习模块，其中包括 MFEF-SFormer 和一个 TCN。MFEF-SFormer 网络应用于所有关节和身体部位，以提取骨骼序列的空间信息，捕获关节间和部位间的相关性，以及关节和身体部位之间的相关性。然后，通过 TCN 进一步提取序列的时间维度信息。

2 实验

2.1 数据集

NTU RGB+D 数据集 [28] 是最广泛使用的基于骨骼的动作识别数据集，大多数模型都基于此数据集进行训练和验证。NTU RGB+D 数据集包含 56 880 个骨骼序列，这些序列来自 60 个动作类别，由 Kinect 相机采集用于 3D 动作识别。每个人体骨骼由 25 个关节的 3D 坐标表示。推荐两个基准：

（1）跨视角（X-View）：使用从相机 2 和相机 3 捕获的 37 920 个动作样本进行训练，使用从相机 1 捕获的样本进行测试。

（2）跨主体（X-Sub）：训练样本来自 20 个主体，测试样本来自其余 20 个主体。

NTU RGB+D 120 数据集 [29] 是 NTU RGB+D 的扩展，总共包含 120 个动作类别和 114 480 个动作样本，提供了更广泛的细粒度动作类别和更多的动作样本。推荐了两个基准：

（1）跨场景（X-Setup）：按照场景 ID 分割采集的样本，偶数场景 ID 的样本用于训练，其余奇数场景 ID 的样本用于测试。

（2）跨主体（X-Sub）：训练样本来自 53 个主体，测试样本来自其余 53 个主体。

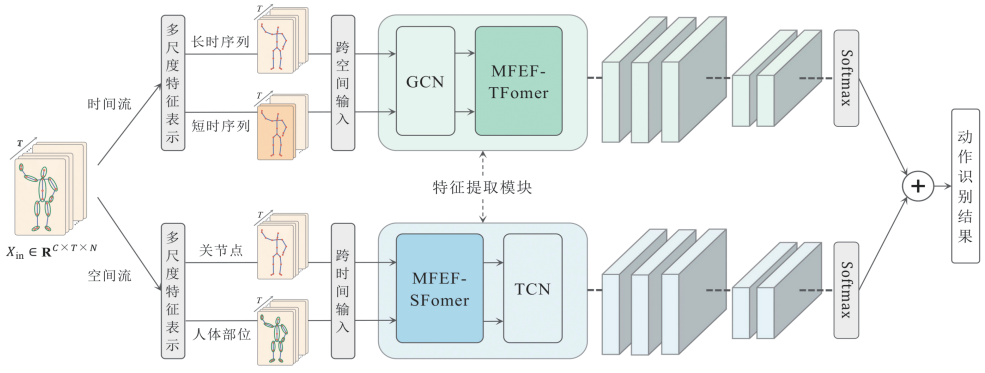

图 4 2s-MFEF-STFormer 网络架构

2.2 消融实验

在本节中，将进行消融研究，以评估所提出网络的各个组件的有效性。并与各种流行的建模方法进行比较，以分析每个提出的设计及其配置的有效性。所有的消融研究都在 NTU RGB+D 数据集的 X-View 基准上进行。

2.2.1 不同尺度的影响

本文提出了一种将时空特征划分为多尺度的方法，该方法将骨架序列在时域和空域上划分为不同尺度的特征。这些划分作为 MFEF-SFormer 的身体部位分支和 MFEF-TFormer 的短序列分支的输入数据。为了验证所提出的不同尺度的有效性，在表 1 中测试了不同的 p（$p \leq N$）和 T_2（$T_2 \leq T$）。

表 1 不同尺度大小对识别准确率的影响

不同的 p	不同的 T_2	识别准确率 /%
N	T	95.1
6	T/4	95.9
6	T/2	96.1
6	T	95.7
10	T/4	96.3
10	T/2	**96.4**
10	T	96.0

与 p=10 相比，p=6 表示左手和右手掌以及脚掌被合并到左臂和右臂区域以及左腿和右腿区域中。短序列的大小 T_2=T/2 和 T_2=T/4 是通过具有不同步幅的 TCN 提取的。需要注意的是，当 p=N 和 T_2=T 时，表示原始骨架序列在时域和空域上都没有被划分。通过观察可以发现，p=10 的准确率比 p=6 提高了 0.3，这是因为更精细的划分（包括手掌和脚掌）能够提取更详细的特征，使其更适合动作识别。某些动作，如“写字”和“阅读”，需要区分手掌和脚掌的特征。采用 T_2=T/2 的时间划分方法相比 T_2=T/4 表现出略高的准确率（0.1），但这一改进不如空域划分显著。然而，并不是尺度越大，网络性能就越好。当设置 p=N 和 T_2=T 时，准确率显著下降 0.8，网络退化为单尺度特征，两个分支接收相同的输入数据，网络只能学习关节点间和帧间的相关性，而无法捕捉时空多尺度特征的关系。这一结果也验证了所提出的多尺度特征划分方法的有效性，最佳性能是通过 p=10 和 T_2=T/2 实现的。

2.2.2 网络结构的影响

通过实验评估了单流网络的性能，并比较了不同架构的网络在动作识别中的准确率。具体来说，比较了时间流网络（MFEF-TFormer+GCN）和空间流网络（MFEF-SFormer＋TCN）的性能。此外，还进行了单流和双流网络的性能比较分析。时空单流网络由 MFEF-SFormer 和 MFEF-TFormer 串联组成。另一方面，时空双流网络则通过并行方式结合了时间流和空间流。比较结果展示在表 2 中。在骨架动作识别中，单时间流网络的准确率为 94.4%，比单空间流网络高出 0.3%。此外，时空单流网络（由 MFEF-SFormer 和 MFEF-TFormer 串联组成的单流网络）相比单时间流或单空间流网络，准确率有显著提升，增加了 1.4，达到了 95.8%。然而，时空单流网络的准确率仍然低于时空双流网络，后者通过并行方式结合了时间流和空间流，准确率达到了 96.4%。这一观察结果表明，双流网络架构使得面向不同维度信息的单流网络能够互为补充，实现相互促进。文中模型 2s-MFEF-STFormer 即时空双流网络的并行结构，如表 2 所示。

表 2 不同网络架构结果对比

	方法	网络结构的组成	识别准确率 /%
单流	空间流	MFEF-SFormer+ TCN	94.1
	时间流	MFEF-TFormer+ GCN	94.4
	时空单流	MFEF-SFormer+ MFEF-TFormer	95.8
双流	时空双流	Spatial Stream + Temporal Stream	**96.4**

2.3 与其他先进算法对比

将所提出的方法——跨注意力多尺度时空 Transformer（2s-MFEF-STFormer）与其他先进的方法在 NTU RGB+D 和 NTU RGB+D 120 数据集上进行了比较。先进的方法涵盖了多种基于 RNN、CNN、GCN 和其他 Transformer 的方法，如表 3 所示。

可以发现当前骨架动作识别领域的主流方法是基于 GCN

表 3 NTU RGB+D 和 NTU RGB+D 120 数据集对比结果

方法		NTU RGB+D		NTU RGB+D 120	
		X-Sub/%	X-View/%	X-Sub/%	X-Setup/%
CNN	RotClips+MTCNN [30]	81.1	87.4	62.2	61.8
	CNN+Motion+Trans [31]	83.2	88.8	—	—
RNN	Dense-IndRNN-aug [7]	86.7	93.9	—	—
GCN	Sem-GCN [12]	86.2	94.2	—	—
	MST-GCN [13]	91.5	96.6	87.5	88.8
	2s-AGCN [11]	88.5	95.1	82.9	84.9
	SGN [32]	89	94.5	79.2	81.5
	Dynamic-GCN [14]	91.5	96	87.3	88.6
	EfficientGCN-B4 [33]	92.1	96.1	88.7	88.9
	EfficientGCN-B4 w/CEA [34]	92.3	96.2	89	89.2
Transformer	ST-TR [25]	89.9	96.1	82.7	84.7
	ST-TR-agcn [27]	90.3	96.3	85.1	87.1
	DSTA-Net [24]	91.5	96.4	89.6	89
	2s-MFEF-STFormer (Ours)	91.9	96.4	89.3	89.5
	2s-MFEF-STFormer+Js-AGCN	**92.1**	**96.7**	**89.4**	**89.7**

的方法。GCN 方法在发展上已经达到了相当成熟的水平。然而，基于 Transformer 的方法显示出了巨大的潜力，并且已经取得了与 GCN 相媲美的性能。提出的 2s-MFEF-STFormer 基于 Transformer，结合骨架序列中的多尺度时空特征表示，并通过跨尺度特征融合进行跨注意力操作。该网络在 NTU RGB+D 数据集的 X-Sub 和 X-View 设置下分别达到了 91.9% 和 96.4% 的准确率。

此外，在 NTU RGB+D 120 数据集的 X-Sub 和 X-Setup 设置下，准确率分别为 89.3% 和 89.5%。结果表明，2s-MFEF-STFormer 在性能上相比于传统 Transformer 方法有了显著提升。

此外，在双流网络的时间流中，将常规 GCN 替换为更先进的 GCN 方法，如 Js-AGCN，可以使动作识别准确率提高 0.1。这也为涉及不同类型网络有机结合的改进提供了新的思路。

3 结论

在本文中，将骨骼序列划分为时间维度和空间维度上的多尺度特征表示。此外，提出了一个多尺度特征提取和融合 Transformer，进一步分为用于空间建模的 MFEF-SFormer 和用于时间建模的 MFEF-TFormer。MFEF-Former 利用自注意力进行多尺度时空特征提取，并利用交叉注意力进行跨尺度特征融合。在 NTU RGB+D 和 NTU RGB+D 120 数据集上的大量实验证明了本文方法的有效性。

参考文献：

[1] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway:IEEE,2017: 7291-7299.

[2] XU X X, GAO Y G, YAN K, et al. Location-free human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE, 2022: 13137-13146.

[3] PENG Q C, ZHENG C, CHEN C. A dual-augmentor framework for domain generalization in 3D human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE, 2024: 2240-2249.

[4] HU G Y, CUI B, YU S. Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention[C]//2019 IEEE International conference on multimedia and expo (ICME). Piscataway:IEEE,2019: 1216-1221.

[5] DUAN H D, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2022: 2969-2978.

[6] SI C Y, JING Y, WANG W, et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning[C]//Proceedings of the European conference on computer vision (ECCV). Berlin:Springer, 2018: 103-118.

[7] LI S, LI W Q, COOK C, et al. Deep independently recurrent neural network (indRNN)[EB/OL].(2020-12-09)[2024-05-29].https://doi.org/10.48550/arXiv.1910.06251.

[8] DUAN H D, WANG J Q, CHEN K, et al. PYSKL: towards good practices for skeleton action recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia.NewYork:ACM, 2022: 7351-7354.

[9] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[EB/OL].(2018-01-25)[2025-05-12].https://doi.org/10.48550/arXiv.1801.07455.

[10] PENG W, HONG X P, CHEN H Y, et al. Learning graph convolutional network for skeleton-based human action recognition by neural searching[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto:AAAI Press,2020, 34(3): 2669-2676.

[11] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.Piscataway:IEEE,2019: 12026-12035.

[12] DING X L, YANG K, CHEN W. A semantics-guided graph convolutional network for skeleton-based action recognition[C]//Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence. NewYork:ACM,2020: 130-136.

[13] CHEN Z, LI S C, YANG B, et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence.Palo Alto:AAAI Press,2021: 1113-1122.

[14] YE F F, PU S L, ZHONG Q Y, et al. Dynamic GCN: context-enriched topology learning for skeleton-based action recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia. NewYork: ACM, 2020: 55-63.

[15] CHEN C F R, FAN Q F, PANDA R. CrossVIT: cross-attention multi-scale vision transformer for image classification[C]//Proceedings of the IEEE/CVF international confer-

ence on computer vision. Piscataway:IEEE,2021: 357-366.

[16] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[17] NTINOU I, SANCHEZ E, TZIMIROPOULOS G. Multi-scale vision transformers meet bipartite matching for efficient single-stage action localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2024: 18827-18836.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. NewYork:ACM,2017:6000-6010.

[19] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL].(2021-06-03)[2025-01-11].https://doi.org/10.48550/arXiv.2010.11929.

[20] CHEN Y P, DAI X Y, CHEN D D, et al. Mobile-Former: bridging mobileNet and transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE,2022: 5270-5279.

[21] DING M Y, XIAO B, CODELLA N, et al. DaVIT: Dual attention vision transformers[EB/OL].(2022-04-07)[2025-02-12].https://doi.org/10.48550/arXiv.2204.03645.

[22] ZHAO S X, XING Y X, XU H Y. WTransU-Net: wiener deconvolution meets multi-scale transformer-based U-net for image deblurring[J]. Signal, image and video processing, 2023, 17:4265-4273.

[23] FAN Q H, HUANG H B, CHEN M R, et al. RMT: retentive networks meet vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2024: 5641-5651.

[24] SHI L, ZHANG Y F, CHENG J, et al. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition[EB/OL].(2020-07-07)[2025-02-22].https://doi.org/10.48550/arXiv.2007.03263.

[25] PLIZZARI C, CANNICI M, MATTEUCCI M. Spatial temporal transformer network for skeleton-based action recognition[EB/OL].(2020-12-11)[2025-01-26].https://doi.org/10.48550/arXiv.2012.06399.

[26] AHN D, KIM S, HONG H, et al. STAR-Transformer: a spatio-temporal cross attention transformer for human action recognition[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway:IEEE,2023: 3330-3339.

[27] PLIZZARI C, CANNICI M, MATTEUCCI M. Skeleton-based action recognition via spatial and temporal transformer networks[J]. Computer vision and image understanding, 2021, 208/209: 103219.

[28] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+ D: a large scale dataset for 3D human activity analysis[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway:IEEE,2016: 1010-1019.

[29] LIU J, SHAHROUDY A, PEREZ M, et al. NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(10): 2684-2701.

[30] KE Q H, BENNAMOUN M, AN S J, et al. Learning clip representations for skeleton-based 3D action recognition[J]. IEEE transactions on image processing, 2018, 27(6): 2842-2855.

[31] LI C, ZHONG Q Y, XIE D, et al. Skeleton-based action recognition with convolutional neural networks[C]//2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway:IEEE,2017: 597-600.

[32] ZHANG P F, LAN C L, ZENG W J, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2020: 1112-1121.

[33] SONG Y F, ZHANG Z, SHAN C F, et al. Constructing stronger and faster baselines for skeleton-based action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(2): 1474-1488.

[34] TANG J, GONG S H, WANG Y J, et al. Beyond coordinate attention: spatial-temporal recalibration and channel scaling for skeleton-based action recognition[J]. Signal, image and video processing, 2024,18: 199-206.

【作者简介】

高逸畅（2000—），男，广东揭阳人，硕士研究生，研究方向：机器学习、深度学习动作识别。

林哲煌（1996—），男，广东汕头人，硕士研究生，研究方向：机器学习、深度学习动作识别。

（收稿日期：2025-03-02 修回日期：2025-08-29）