DOI:10.19651/j.cnki.emt.2314259

用于骨架行为识别的时空卷积 Transformer 网络*

刘斌斌[1] 赵宏涛[2] 王 田[3] 杨 艺[2]
(1.郑州恒达智控科技股份有限公司 郑州 450000;2.河南理工大学电气工程与自动化学院 焦作 454003;
3.北京航空航天大学人工智能研究院 北京 100191)

摘 要:针对基于图卷积的骨架行为识别方法在建模关节特征时严重依赖手工设计图形拓扑,缺乏建模全局关节间依赖关系的缺点,设计了一种时空卷积 Transformer 实现对空间和时间关节特征的建模。空间关节特征建模中,提出一种动态分组解耦 Transformer,通过将输入骨架序列在通道维度进行分组并为每个组动态生成不同的注意力矩阵,允许建模关节之间的全局空间依赖关系,无需事先知道人体拓扑结构。时间关节特征建模中,通过多尺度时间卷积实现对不同时间尺度行为特征的提取。最后,提出一种时空-通道联合注意力模块,进一步对所提取到的时空特征进行修正。在 NTU-RGB+D 和 NTU-RGB+D 120 数据集的跨主体评估标准上达到了 92.5%和 89.3%的 Top1 识别准确率,实验结果表明了所提方法的有效性。

关键词:行为识别;人体骨架;自注意机制;Transformer
中图分类号:TP391.4 **文献标识码**:A **国家标准学科分类代码**:520.2040

Spatial temporal convolutional Transformer network for skeleton-based action recognition

Liu Binbin[1] Zhao Hongtao[2] Wang Tian[3] Yang Yi[2]
(1. Zhengzhou Hengda Intelligent Control Technology Company Limited, Zhengzhou 450000, China;
2. School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454003, China;
3. Research Institute for Artificial Intelligence, Beihang University, Beijing 100191, China)

Abstract: In the methon of skeleton action recognition based on graph convolution, the rely heavily on hand-designed graph topology in modelling joint features, and lack the ability to model global joint dependencies. To address this issue, we proposed a spatio-temporal convolutional Transformer network to implement the modelling of spatial and temporal joint features. In the spatial joint feature modeling, we proposed a dynamic grouping decoupling Transformer that grouped the input skeleton sequence in the channel dimension and dynamically generated different attention matrices for each group, establishing global dependencies between joints without requiring knowledge of the human topology. In the temporal joint feature modeling, multi-scale temporal convolution was used to extract features of target behaviors at different scales. Finally, we proposed a spatio-temporal channel joint attention module to further refine the extracted spatio-temporal features. The proposed method achieved Top1 recognition accuracy rates of 92.5% and 89.3% on the cross-subject evaluation criteria for the NTU-RGB + D and NTU-RGB + D 120 datasets, respectively, demonstrating its effectiveness.

Keywords: action recognition; human skeleton; self-attention mechanism; Transformer

0 引 言

人体行为识别是视频理解中一个重要的研究课题,已被广泛的应用在智能视频监控、人机交互和机器人等领域。近年来,随着深度相机和先进姿态估计[1]算法的出现,基于骨架的行为识别受到越来越多研究者的关注。骨架数据采用一系列人体关键关节坐标作为行为信息载体,这种高度抽象的表示使得骨架数据在具有极高紧凑性的同时,对于光照、背景和人体外观变化具有很好的鲁棒性。

骨架数据包含了用于行为识别的全部时空信息,有效

收稿日期:2023-07-31
*基金项目:国家自然科学基金(61972016)项目资助

建模关节的时空依赖关系对于从骨架序列中识别行为至关重要。近年来，基于深度学习的传统解决方案以循环神经网络（recurrent neural network，RNN）和卷积神经网络（convolutional neural network，CNN）为主干，将原始骨架数据转化为向量序列或伪图像以满足输入要求。然而，人体骨架具有一定的自然拓扑结构，关节和骨骼之间具有一定的相关性。通过将骨架数据转化为向量序列或伪图像，人体骨架所包含的结构信息将被破坏，这无疑对于正确识别类间差异较小的行为是不利的。

人体骨架作为一种非欧几里德结构数据，如何有效建模所包含的结构信息有利于更好的提升识别结果。得益于图卷积网络（graph convolutional network，GCN）在建模非欧几里德数据方面的优势，通过将骨架序列表示为一个时空拓扑图，有效的利用了人体骨架的拓扑信息，使基于GCN的深度学习方法获得了最先进的性能。其中，Yan等[2]首次将GCN引入骨架行为识别，将骨架序列表示为一个时空拓扑图，提出一种时空图卷积网络（spatial temporal graph convolutional network，ST-GCN）来分别实现对空间和时间特征的建模。此后，在ST-GCN的基础上一些改进方法[3]被陆续提出，这些基于GCN的方法通过预先设定的拓扑图进行关节之间的信息交互，从而实现关节特征的更新。然而，预设的拓扑图通常需要一定的人体结构先验知识，且简单的以骨架为基础的图结构通常只能够得到关节之间的局部空间依赖关系，缺乏建模全局空间关节依赖关系的能力。

最近，Transformer[4]作为一种新范式被提出作为RNN的替代方案，成为自然语言处理任务中性能最先进的模型。Transformer遵循编码器-解码器结构，仅仅依赖于自注意力机制，在打破RNN并行处理局限性的同时可以更好的建模长序列。随着Transformer在自然语言处理领域中的成功，其在各种计算机视觉任务中也获得了显著的性能提升。得益于骨架数据中关节数量的稀疏性，通过将其视作一个单词，最近的一些工作[5-6]开始将其扩展到基于骨架的行为识别中。其中，Plizar等[7]提出了一种时空Transformer网络，该模型由基于骨架的行为识别任务中的Transformer编码器和GCN模块组成，使用Transformer自注意力算子代替了空间和时间上的正则图卷积。Kong等[8]通过图卷积网络块和多尺度时间嵌入模块对原始骨架进行嵌入，将多尺度时间嵌入模块设计为多个分支，提取不同时间尺度的特征，然后引入了Transformer编码器来集成嵌入及建模长期时间模式。Sun等[9]提出使用多流时空相对Transformer来克服图卷积接受域较小的缺陷，通过引入中继节点打破了空间上固有的骨架拓扑和时间维度上骨架序列的顺序。然而，这些基于Transformer的方法在进行特征更新时所有通道共享同一个注意力矩阵，相较于卷积神经网络中使用的解耦聚合机制[10]，其特征建模的灵活性受到限制。

本文针对基于图卷积的方法在建模空间关节特征时严重依赖于手工设计拓扑图及缺乏建模全局关节依赖关系的缺点，提出一种时空卷积Transformer（spatial temporal convolutional transformer network，ST-ConvTR），通过引入Transformer的自注意力机制和卷积操作分别对空间和时间关节特征进行建模。在空间关节特征建模中，受卷积神经网络解耦聚合机制启发，提出一种动态分组解耦Transformer。它将输入骨架序列在通道维度进行分组并为每个组动态生成不同的注意力矩阵，允许建模关节之间的全部空间依赖关系，而不需要事先知道人体拓扑结构。时间关节特征建模中，采用多尺度时间卷积提取不同时间尺度行为特征。最后，为进一步对所提取到的时空特征进行修正，提出一种时空-通道联合注意力模块，通过计算整个骨架序列中所有关节的权重，实现对提取到的时空特征进行加权。

1　相关工作

1.1　基于骨架的行为识别

过去，基于骨架的行为识别主要依赖手工设计特征[11]，虽然这类方法可以更好地理解行为特征，但由于骨架数据中包含大量非结构化信息，其难以捕捉所有行为特征，因此识别准确率相对较低。随着深度学习发展，基于RNN和CNN的方法被引入骨架行为识别，取得了显著优于手工设计特征方法的性能。基于RNN的方法[12]转化原始骨架数据为向量序列，旨在建模骨架序列的时间依赖关系。基于CNN的方法[13]转化原始骨架数据为伪图像，通过卷积操作提取骨架数据的时空信息。骨架数据是一种典型的非欧式数据，有效利用骨架的拓扑信息有助于更好的进行行为识别。近年来，通过将骨架序列表示为一个时空拓扑图，基于GCN的骨架行为识别方法[14]得到了广泛的应用，它可以在不丢失人体骨架拓扑信息的情况下有效的建模关节的时空依赖关系。

1.2　Transformer的自注意力机制

自注意力机制作为Transformer的重要组成部分，其计算过程为：首先，通过线性变换将输入序列映射到查询向量（query，Q）、键向量（key，K）和值向量（value，V）；然后，计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的点乘并经过一个 $Softmax(\cdot)$ 函数，得到一个包含各元素权重的注意力矩阵；最后，使用注意力矩阵实现对 $\boldsymbol{V}$ 中对应元素的加权，计算公式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Soft\max\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V} \tag{1}$$

其中，d_k 表示键向量维度，缩放因子 $\sqrt{d_k}$ 用来缓解梯度爆炸问题。

为了增强模型的表达能力和学习能力，在Transformer中提出一种多头注意力机制，允许模型同时关注输入的不同部分，从而捕捉更多的信息和关联性，计算公式如下：

$$\begin{aligned} head_i &= Attention(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i) \\ MHA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) &= Concat(head_1,\cdots,head_H)\boldsymbol{W}^O \end{aligned} \tag{2}$$

其中,H 为头数,$\boldsymbol{W}^O \in \mathbb{R}^{(H\cdot d_v)\times d_{out}}$ 为权重矩阵,d_v 和 d_{out} 分别表示值向量维度和输出维度。

2　时空卷积 Transformer 网络

2.1　动态分组解耦 Transformer 网络

在卷积神经网络的特征提取中,不同的输出通道都具有独立的卷积核,这种机制可以极大的提高空间建模能力,称之为解耦聚合。受此启发,提出一种动态分组解耦 Transformer(dynamic grouping DeCoupling transformer,DGDC-TR)。如图 1 所示,DGDC-TR 主要由位置编码(position encoding,PE)模块、空间注意力模块和空间全局注意力(spatial global attention,SGA)矩阵三部分组成。下面将分别对这 3 个部分进行阐述。

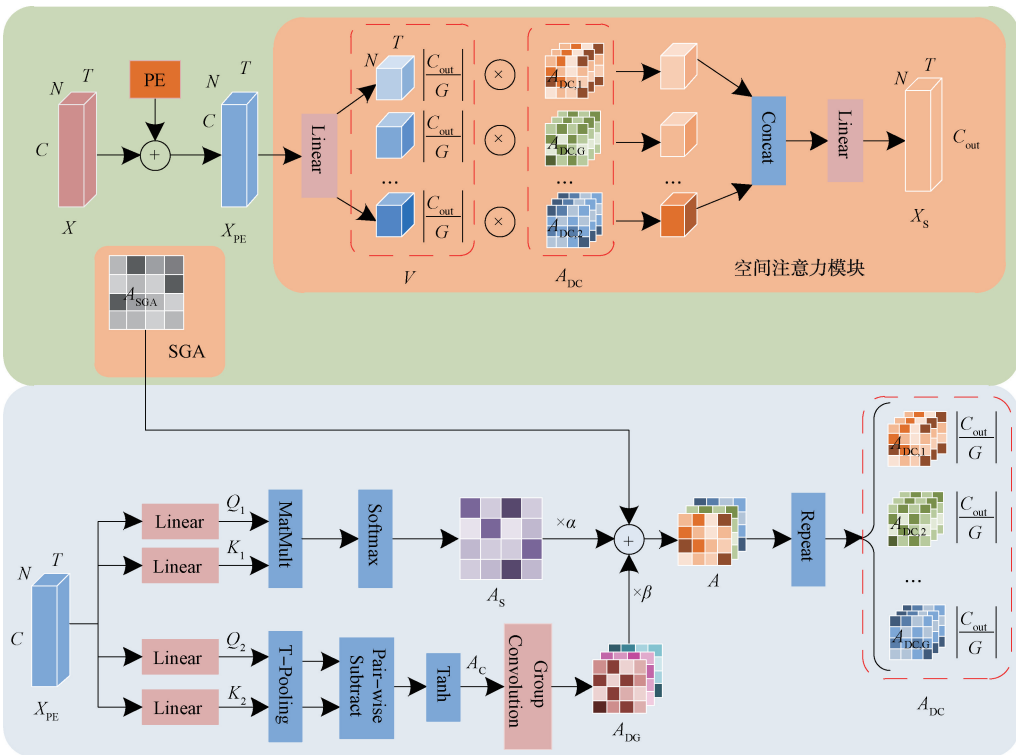

图 1　动态分组解耦 Transformer

1)位置编码

由于 Transformer 的网络结构不同于传统的神经网络,因此在处理序列数据时无法捕捉到序列顺序信息的。例如,按行打乱 $\boldsymbol{K}$、$\boldsymbol{V}$ 的顺序并不影响注意力计算的结果,但是在涉及序列信息的任务中,序列的位置信息是非常重要的,因此必须将序列的位置信息利用起来,以便在计算注意力时正确地考虑输入序列中元素的顺序和关系。由于仅对空间关节特征进行建模,这里采用所有帧共享同一组位置编码的方式,仅对空间上的关节位置进行编码。不同的是,在网络中我们将位置编码看作一个参数,采用不同频率的正弦和余弦函数[4]作为初始化,在训练过程中进行学习更新以增强位置编码对不同输入样本的适应性,计算公式如下:

$$\begin{aligned} PE(p,2i) &= \sin(p/10\,000^{2i/C_{in}}) \\ PE(p,2i+1) &= \cos(p/10\,000^{2i/C_{in}}) \end{aligned} \tag{3}$$

其中,p 表示元素的位置,i 表示位置编码向量的维度。

2)空间注意力模块

在图 1 的上半部分方框中展示了 DGDC-TR 中所使用的空间注意力模块。首先,对于原始输入骨架序列 $\boldsymbol{X} \in \mathbb{R}^{T\times N\times C}$ 经过位置编码得到输入序列 $\boldsymbol{X}_{\mathrm{PE}} \in \mathbb{R}^{T\times N\times C}$,通过一个线性变换(Linear)将 $\boldsymbol{X}_{\mathrm{PE}}$ 映射到一个值向量 $\boldsymbol{V} \in \mathbb{R}^{T\times N\times C_{out}}$,然后将值向量在通道维度分为 G 组,每组包含的通道数为 $|C_{out}/G|$,线性变化公式如下:

$$\boldsymbol{V} = \boldsymbol{X}_{\mathrm{PE}}\boldsymbol{W}_L \tag{4}$$

其中,$\boldsymbol{W}_L \in \mathbb{R}^{C\times C_{out}}$ 表示线性变换的权重矩阵,C 和 C_{out} 分别表示空间注意力模块的输入通道数和输出通道数。

经过上述操作,值向量 $\boldsymbol{V}$ 的通道被平均分为 G 组,下一步的工作在于如何根据组数生成解耦注意力矩阵 $\boldsymbol{A}_{\mathrm{DC}}$。如图 1 下半部分所示,首先,通过式(5)和(6)将 $\boldsymbol{X}_{\mathrm{PE}}$ 经过两组线性变化映射到两组不同的查询和键向量。

$$\boldsymbol{Q}_1,\boldsymbol{K}_1 = \boldsymbol{X}_{\mathrm{PE}}\boldsymbol{W}_{L1}^{Q},\boldsymbol{X}_{\mathrm{PE}}\boldsymbol{W}_{L1}^{K} \tag{5}$$

$$\boldsymbol{Q}_2,\boldsymbol{K}_2 = \boldsymbol{X}_{\mathrm{PE}}\boldsymbol{W}_{L2}^{Q},\boldsymbol{X}_{\mathrm{PE}}\boldsymbol{W}_{L2}^{K} \tag{6}$$

其中，$\boldsymbol{W}_{L1}^{Q} \in \mathbb{R}^{C\times C_m}$、$\boldsymbol{W}_{L1}^{K} \in \mathbb{R}^{C\times C_m}$、$\boldsymbol{W}_{L2}^{Q} \in \mathbb{R}^{C\times C_{out}}$ 和 $\boldsymbol{W}_{L2}^{K} \in \mathbb{R}^{C\times C_{out}}$ 分别表示所用线性变换的权重矩阵，$C_m = C_{out}/2$。

对于得到的查询向量 $\boldsymbol{Q}_1 \in \mathbb{R}^{N\times(TC_m)}$ 和键向量 $\boldsymbol{K}_1 \in \mathbb{R}^{(TC_m)\times N}$，通过式(7)可以得到一个注意力矩阵 $\boldsymbol{A}_S \in \mathbb{R}^{N\times N}$。

$$\boldsymbol{A}_S = Soft\max\left(\frac{\boldsymbol{Q}_1\boldsymbol{K}_1^T}{\sqrt{C_m}}\right) \tag{7}$$

对于得到的查询向量 $\boldsymbol{Q}_2 \in \mathbb{R}^{T\times N\times C_{out}}$ 和键向量 $\boldsymbol{K}_2 \in \mathbb{R}^{T\times N\times C_{out}}$，首先，经过时间平均池化(Temporal Pooling，T-Pooling)、逐对相减(Pair-wise Subtract)和 Tanh 激活函数后将得到一个通道注意力矩阵 $\boldsymbol{A}_C \in \mathbb{R}^{C_{out}\times N\times N}$；最后，将通道注意力矩阵 $\boldsymbol{A}_C$ 通过一个分组卷积将得到一个动态分组注意力矩阵 $\boldsymbol{A}_{DG} \in \mathbb{R}^{G\times N\times N}$，计算公式如下：

$$\boldsymbol{A}_{DG} = Conv2d_{(1\times1)}(\boldsymbol{A}_C, groups = G) \tag{8}$$

最后，对上述得到的注意力矩阵 $\boldsymbol{A}_S$、$\boldsymbol{A}_{DG}$ 和 $\boldsymbol{A}_{SGA}$ 经过式(9)相结合得到 $\boldsymbol{A} \in \mathbb{R}^{G\times N\times N}$，然后经过 repeat 操作将 $\boldsymbol{A}_S$ 的每个通道扩展到对应组所包含的通道数，得到一个解耦注意力矩阵 $\boldsymbol{A}_{DC} \in \mathbb{R}^{C_{out}\times N\times N}$。这里，将 $\boldsymbol{A}_{DC}$ 中的 C_{out} 个通道划分为 G 组时，每组的注意力矩阵则是相同的，和经过分组后的值矩阵 $\boldsymbol{V}$ 相对应。最后，将经过分组的值矩阵 $\boldsymbol{V}$ 和得到解耦注意力矩阵 $\boldsymbol{A}_{DC}$ 对应的组相乘即得到最后的空间建模输出 $\boldsymbol{X}_S$。

$$\boldsymbol{A} = \alpha \cdot \boldsymbol{A}_S + \beta \cdot \boldsymbol{A}_{DG} + \boldsymbol{A}_{SGR} \tag{9}$$

其中，α 和 β 表示两个可训练的标量，用来调整注意力矩阵 $\boldsymbol{A}_S$ 和 $\boldsymbol{A}_{DG}$ 的强度；$\boldsymbol{A}_{SGA}$ 为空间全局注意力矩阵。

3)空间全局注意力

原始的 Transformer 网络中使用多头注意力来获得不同层面的信息，每个头都具有特定的意义。为了迫使模型学习不同行为的更多一般注意事项，在计算分组注意力矩阵时添加了一个空间全局注意力矩阵 $\boldsymbol{A}_{SGA} \in \mathbb{R}^{N\times N}$，$\boldsymbol{A}_{SGA}$ 被添加进第 2.1 节生成注意力矩阵 $\boldsymbol{A}$ 的过程中。由于所有的数据样本共享全局注意力矩阵，它代表了人体关节间通用的联系。我们将其设置为网络的参数，直接用正态分布初始化 $\boldsymbol{A}_{SGA}$，并与模型一起对其进行优化。该模块简单且重量轻，但如消融研究所示，它是有效的。

2.2　时间关节特征建模

由于 Transformer 对输入数据的归纳偏差几乎没有假设，其泛化性能的好坏严重依赖于训练数据的数量，难以在小规模的数据集上进行训练，对于类内差异很大的骨架序列来讲进一步加大了这种限制。因此，目前将纯 Transformer 应用到骨架行为识别并未取得较好的效果，多数方法仍主要以 Transformer 和卷积的结合为主。因此，对于时间关节特征建模，采用被广泛使用的多尺度时间卷积(multi-scale temporal convolution，MST)，如图 2 所示。MST 由 5 条分支组成，1×1 卷积用于划分各分支通道，其中 d 表示卷积分支的膨胀系数，各分支通道数满足 $C = C_a + 4C_b$。MST 通过不同时间尺度大小的卷积核，有效的提取了不同时间尺度上的行为特征，在减少参数量和提高建模能力方面已被证明是有效的。

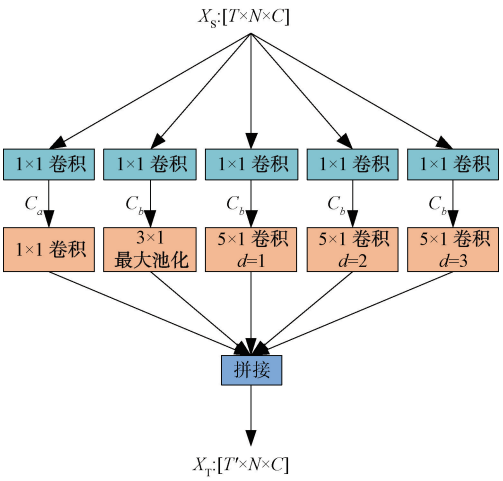

图 2　多尺度时间卷积

2.3　时空-通道联合注意力模块

在基于骨架的行为识别中，常采用一种类 SENet[15] 模块的注意力机制，该注意力不同于 Transformer 中的自注意力机制。自注意力机制计算不同关节之间的注意力分配，而类 SENet 注意力则基于关节本身进行注意力分配。研究者使用这些模块独立的在通道[16]、空间[17]或时空[18]上应用注意力。然而，直观的讲，空间信息、时间信息和通道信息之间可能彼此相关。因此，单独的考虑通道、空间和时间对于骨架序列中关节的加权可能并不是最优的。为了解决这个问题，提出一个时空-通道联合注意力(spatiotemporal-channel joint attention，STC JointAtt)模块，如图 3 所示。对于一个输入骨架序列 $\boldsymbol{X} \in \mathbb{R}^{T\times N\times C}$，对时间和空间维度进行平局池化，然后经过一个输出通道数为 1 的全连接层(fully connected layer，FC)分别生成对关节的空间注意力 $\boldsymbol{A}_{spatial} \in \mathbb{R}^{N}$ 和时间注意力 $\boldsymbol{A}_{temporal} \in \mathbb{R}^{T}$，对时空维度进行平均池化以获得通道注意力 $\boldsymbol{A}_{channel} \in \mathbb{R}^{C}$，然后对得到的三个注意力使其通过一个激活函数，最后对于这三个维度的注意力通过一个函数 $\mathcal{F}(\cdot)$ 得到一个注意力图，该注意力图用来加权骨架序列中的每个关节特征，计算公式如下：

$$\begin{aligned}
&\boldsymbol{A}_{spatial} = \theta(FC_{(C,1)}(AvgPool_t(\boldsymbol{X})))\\
&\boldsymbol{A}_{temporal} = \theta(FC_{(C,1)}(AvgPool_s(\boldsymbol{X})))\\
&\boldsymbol{A}_{channel} = \theta(AvgPool_{st}(\boldsymbol{X}))\\
&\boldsymbol{A} = \theta(\mathcal{F}(\boldsymbol{A}_{channel} \times (\boldsymbol{A}_{spatial} + \boldsymbol{A}_{temporal}))
\end{aligned} \tag{10}$$

其中，θ 表示 $Sigmoid(\cdot)$ 激活函数。

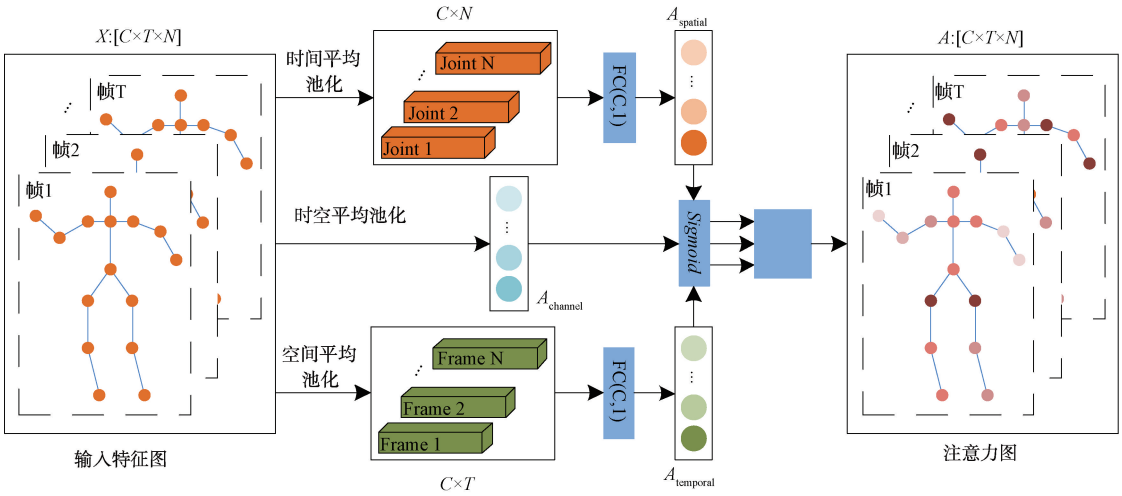

图 3 时空-通道联合注意力模块

2.4 网络总体架构

网络的总体框架如图 4 所示,在空间关节特征建模中以 DGDC-TR 为基础,采用原始 Transformer 中的多头注意力机制,整个空间建模一共包含 H 个头。在时间关节特征建模中,使用 MST 提取不同时间尺度行为特征。最后,通过 STC JointAtt 进一步的对所提取到的时空特征进行修正。此外残差连接被分别添加在空间和时间建模中,整个网络层数为 $L=10$。具体网络设置如表 1 所示,默认 $C=64$、$T=64$,$V=25$ 分别表示通道、帧和关节的数量。

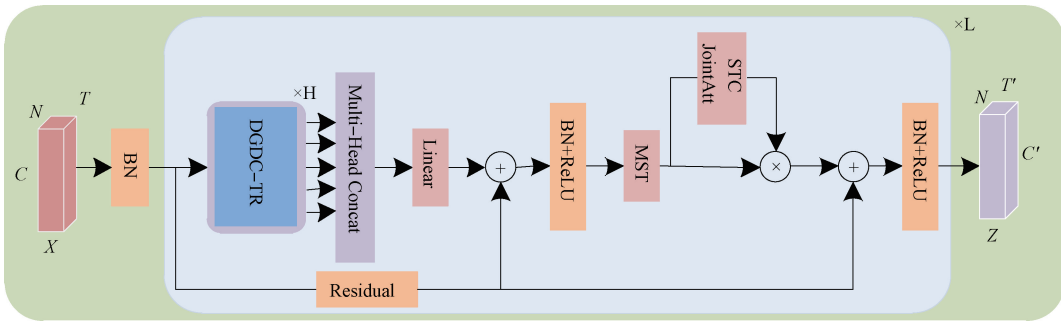

图 4 网络总体架构

表 1 网络设置

层名称	输出	头维度
L1～L4	$[C, T, V]$	$C/4$
L5～L7	$[2C, T/2, V]$	$C/2$
L8～L10	$[4C, T/4, V]$	C

3 实验结果与分析

3.1 数据集介绍及评价指标

NTU-RGB+D[19]数据集是一个大规模人类行为识别数据集,该数据集包含 60 个不同的行为类别,由 56 880 个行为片段组成。这些行为片段由 3 台不同角度的 Kinect v2 传感器从 40 名不同的志愿者身上采集得到。该数据集提供两种评估基准:1)跨主体(cross subject,CS):训练集由来自 20 名志愿者的 40 320 个行为片段组成,测试集由来自剩余 20 名志愿者的 16 560 个行为片段组成。2)跨视角(cross view,CV):训练集来自传感器 2 和 3 采集的 37 920 个行为片段组成,测试集来自传感器 1 采集的 18 960 个行为片段组成。

NTU-RGB+D 120[20]作为 NTU-RGB+D 的扩展版本,共有 120 个行为类别,由从 106 位志愿者采集的 114 480 个行为片段组成。此外,在相机高度和水平距离上提供了 32 种设置组合,每种组合设置都具有独立的 ID。该数据集同样提供两种评估基准:1)跨主体(cross subject,CS):训练集由来自 53 名志愿者的 63 026 个行为片段组成,测试集由来自剩余 53 名志愿者的 51 454 个行为片段组成。2)跨设置(cross setting,CE):训练集由设置 ID 为偶数的 54 471 个行为片段组成,测试集本由设置 ID 为奇数 60 009 个行为片段组成。

骨架行为识别作为一个多分类问题,本文采用识别准确率(Accuracy)作为评判模型泛化性能的主要方式。对于指定测试集,识别准确率定义为正确分类的样本个数占总

样本个数的比例。

3.2　实验设置

本文中所有实验都是基于 Pytorch 深度学习框架在 Ubuntu18.04 操作系统上完成，使用 2 张 Tesla T4 GPU，显存大小为 32 G。对于消融实验，从每个行为样本中选择均匀采样一个长度为 64 的骨架序列作为输入。对于 3.4 节提供的最优识别准确率，输入骨架序列长度为 100。在两个数据集上使用一些共用设置：优化策略使用权重衰减为 0.000 4，Nesterov 动量大小为 0.9 的随机梯度下降算法；批次大小和总训练迭代轮数分别为 64 和 80，初始学习率为 0.1，在第 35 轮、55 轮和 75 轮时学习率减小到前一轮的 1/10。此外，一个热身策略[21]在前 5 轮被使用。

3.3　消融实验

为了验证所提网络的有效性，本节在 NTU-RGB+D 数据集的跨主体(CS)基准上进行了消融实验。除非特别说明，否则所有的消融实验使用头数为 4，分组数 G 等于该层网络的输出通道 C_{out}。

1)网络中模块的有效性

通过删除 ST-ConvTR 中的位置编码(PE)模块、空间全局注意力矩阵($\boldsymbol{A}_{SGA}$)、动态分组注意力矩阵($\boldsymbol{A}_{DG}$)以及时空-通道联合注意力(STC JointAtt)模块，研究了不同模块对识别性能的影响，结果如表 2 所示。从表中可以看出，完整的 ST-ConvTR 取得了最佳的识别性能，测试识别准确率为 89.9%。删除 PE 后，识别准确率仅下降 0.1%，这说明为每个关节提供位置信息对于 ST-ConvTR 仍是有效的，但骨架序列可能并不像机器翻译那样，句子中单词之间存在逻辑关系。删除 $\boldsymbol{A}_{SGA}$ 后，ST-ConvTR 的识别准确率下降幅度最大，达到了 0.6%，这说明虽然多头注意力机制可以从不同层面捕捉关键之间的依赖关系，但学习关节之间的通用依赖关系对于模型来讲是十分必要的。删除 $\boldsymbol{A}_{DG}$ 后，识别准确率和完整的 ST-ConvTR 相比下降了 0.4%，这说明为不同通道建模注意力矩阵对于捕捉不同通道上关节之间的依赖关系是有效的，极大的增强了 ST-ConvTR 的空间关节建模能力。最后，删除 STC JointAtt，识别准确率下降了 0.2%，这表明空间、时间和通道之间是存在联系，建模它们之间的联系对于骨架行为识别是有效的。

表 2　不同模块的识别准确率

模型	PE	$\boldsymbol{A}_{SGA}$	$\boldsymbol{A}_{DG}$	STC JointAtt	Acc/%
ST-ConvTR	×	√	√	√	89.8
ST-ConvTR	√	×	√	√	89.3
ST-ConvTR	√	√	×	√	89.5
ST-ConvTR	√	√	√	×	89.7
ST-ConvTR	√	√	√	√	89.9

2)不同解耦组的识别准确率

为了研究对注意力矩阵进行分组解耦的有效性，对不同的解耦组进行了消融实验，识别准确率如图 5 所示，其中 C 表示图 1 中空间注意力模块所得值矩阵($\boldsymbol{V}$)的通道数。可以看到，不进行分组时的识别准确率为 89.2%，和其他分组相比处于最低点。当将注意力矩阵分为 4 组、8 组和 16 组时，识别准确率相比于不进行分组时逐步提高，分别为 89.4%、89.5%和 89.6%。最后，将注意力矩阵分为 C 组时的识别准确率为 89.9%，在所有的分组中处于最高点。这表明了不同通道上每个关节对其他关节的关注程度是存在差异的，所提 DGDC-TR 通过将值向量的输入划分为 G 组，并为不同的组生成各自的注意力矩阵可以有效的捕捉不同通道上的运动模式。

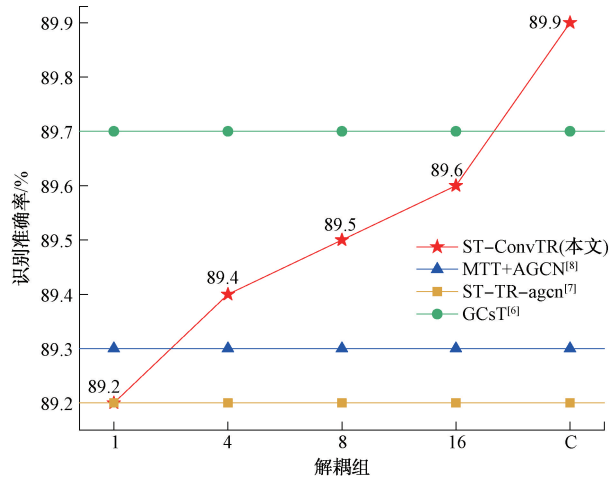

图 5　不同解耦组识别准确率

3)多头自注意力机制的有效性

为了研究多头自注意力机制对模型识别准确率的影响，在图 6 中展示了使用不同数量的头时的识别准确率。可以看到，仅使用一个头时识别准确率最低，仅为 88.7%。进一步增加使用头的数量时识别准确率持续上升，并在头数量为 4 时实现最好的性能，识别准确率为 89.9%。但是，当继续增加头数到 6 和 8 时，可以看到识别准确率开始随着头的数量增加而下降。这可能时因为 Transformer 采用的是以纯粹的点乘为基础的自注意力机制，随着头数量的增多整个网络无论是对数据量的要求还是对数据分布的要求都趋向于更加严格的限制。对于以骨架数据为基础的行为识别来讲，数据类内分布差异很大，进一步的加大了这种限制。

4)多尺度时间卷积的有效性

为了证明 MST 的有效性，使用 MST 和一个在时间维度上核大小为 $K_t=9$ 的卷积分别代替时间建模(temporal modeling，TM)，在 NTU-RGB+D 数据集的 CS 评估标准上使用关节数据进行了对比试验。实验结果如表 3 所示，使用 MST 的识别准确率较单独使用卷积高，这是因为 MST 可以从多个时间尺度上提取丰富且具有差异的时间信息。同时由于每条分支上通道数较小，整体模型在参数

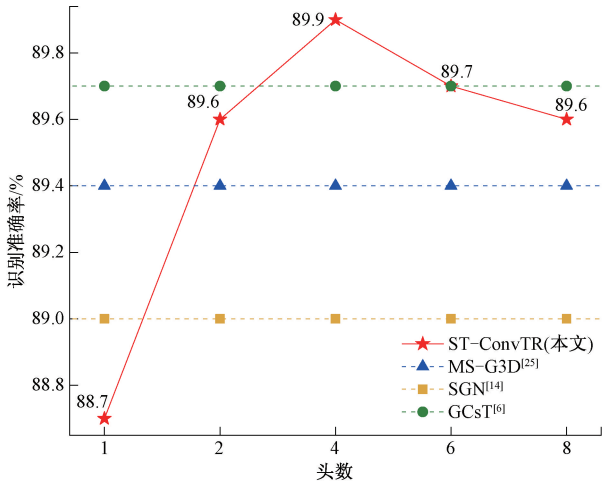

图 6 不同头的识别准确率

和计算量方面也较使用卷积出现了降低。综上所述,证明了 MST 对提高识别准确率的有效性。

表 3 MST 对模型识别准确率的影响

模型	TM	参数/M	FLOPs/G	Acc/%
ST-ConvTR	MST	1.73	1.90	89.9
ST-ConvTR	$K_t=9$	3.47	3.97	89.5

5)注意力矩阵可视化

如图 7 所示,为了对 DGDC-TR 具有更直观的认识,可视化了“喝水”行为在网络第 3 层、第 6 层和第 9 层中的空间全局注意力矩阵(SGA)和动态分组注意力矩阵(ADG)中的 3 个。其中,橙色越深表示受到的注意越大,蓝色越深表示受到的注意越小。结果表明:1)每层中的空间全局注意力矩阵分布趋于密集,它代表了人体关节间通用的联系,能够捕捉空间全局关节依赖关系,有助于更好的识别不同行为。2)对于动态分组注意力矩阵,在不同通道上关节之间的依赖关系存在差异化,表明了 DGDC-TR 可以更好的建模不同通道上的行为特征。

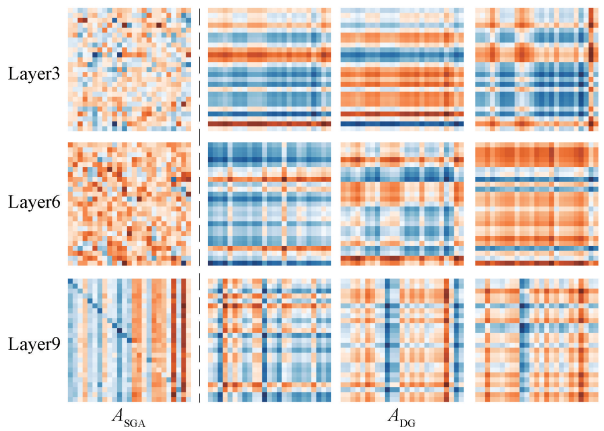

图 7 注意力矩阵可视化

3.4 与其他方法比较

本节将 ST-ConvTR 与 NTU-RGB+D 和 NTU-RGB+D 120 数据集上的最先进方法的识别准确率进行了比较,结果如表 4 和 5 所示,包括基于 GCN 的方法和基于 Transformer 的方法。提供的最优识别准确率采用广泛使用的多流融合框架[17],其中 2 s 表示使用关节和骨骼两种数据的融合得分,4 s 表示使用关节、骨骼、关节运动和骨骼运动四种数据的融合得分。可以看到,在 NTU-RGB+D 和 NTU-RGB+D 120 上所提方法使用 4 种数据融合的识别准确率均优于其它方法,表明了该模型提取时空关节特征的有效性。

表 4 NTU-RGB+D 数据集的识别准确率比较

模型	CS/%	CV/%
ST-GCN[2]	81.5	88.3
NAS-GCN[22]	89.4	95.7
DC-GCN+ADG[10]	90.8	96.6
DualHead-Net[23]	92.0	96.6
STime-TR[24]	87.1	91.8
MSST-RT[9]	88.4	93.2
ST-TR-agcn[7]	90.3	96.3
MTT+Shift-GCN[8]	90.8	96.7
ST-ConvTR (Joint)	90.0	95.3
ST-ConvTR (Bone)	90.5	94.9
ST-ConvTR (2s)	92.1	96.5
ST-ConvTR (4s)	92.5	96.9

表 5 NTU-RGB+D120 数据集的识别准确率比较

模型	CS/%	CE/%
ST-GCN[2]	70.7	73.2
2s MS-G3D[25]	86.9	88.4
Dynamic-GCN[26]	87.3	88.6
EfficientGCN-B4[18]	88.7	88.9
MSST-RT[9]	79.3	82.3
ST-TR-agcn[7]	85.1	87.1
MTT+AGCN[8]	86.1	87.6
ST-ConvTR (Joint)	85.0	87.8
ST-ConvTR (Bone)	87.0	87.0
ST-ConvTR (2s)	89.0	90.4
ST-ConvTR (4s)	89.3	90.6

4 结 论

在基于人体骨架的行为识别中,提高模型识别准确率的关键在于如何更好的提取骨架序列的时空特征和捕捉关节间的全局依赖关系,本文提出一种时空卷积 Transformer 网络。空间上,使用动态分组解耦

Transformer 实现对空间上全局关节依赖关系的建模，与基于 GCN 的方法相比并不需要人体结构先验知识。时间上，采用多尺度时间卷积，在有效降低参数量和计算量的基础上增强了时间关节特征的建模能力。最后，通过时空-通道联合注意力模块对多提取到的时空关节特征进一步修正。实验结果表明，与现有先进的方法相比，所提方法具有更高的识别准确率。此外，详细的消融实验得以证实所提方法的每个组件在骨架行为识别中都具有重要的作用。

参考文献

[1] 张小娜，吴庆涛. 基于深度学习的自顶向下人体姿态估计算法[J]. 电子测量技术，2021，44(9)：105-109.

[2] YAN S J，XIONG Y J，LIN D H，et al. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. 32th AAAI Conference on Artificial Intelligence. Reston，USA：AAAI，2018：7445-7452.

[3] 李志晗，刘银华，谢锐康，等. 基于关节点运动估计的人体行为识别[J]. 电子测量技术，2022，45(24)：153-160.

[4] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook，NY，USA：Curran Associates Inc，2017：6000-6010.

[5] 石跃祥，朱茂清. 基于骨架动作识别的协作卷积 Transformer 网络[J]. 电子与信息学报，2023，45(4)：1485-1493.

[6] BAI R W，LI M，MENG B，et al. Hierarchical graph convolutional skeleton transformer for action recognition[C]. 2022 IEEE International Conference on Multimedia and Expo(ICME)，2022：01-06.

[7] PLIZZARI C，CANNICI M，MATTEUCCI M. Skeleton-based action recognition via spatial and temporal transformer networks[J]. Computer Vision and Image Understanding，2021，208：103219.

[8] KONG J，BIAN Y，JANG M. MTT：Multi-scale temporal transformer for skeleton-based action recognition[J]. IEEE Signal Processing Letters，2022，29：528-532.

[9] SUN Y，SHEN Y X，MA L Y. MSST-RT：Multi-stream spatial-temporal relative transformer for skeleton-based action recognition[J]. Sensors，2021，21(16)：5339-5341.

[10] CHENG K，ZHANG Y F，CAO C Q，et al. Decoupling GCN with dropgraph module for skeleton-based action recognition[C]. European Conference on Computer Vision. Cham，Switzerland：Springer，2020：536-553.

[11] EVANGELIDIS G，SINGH G，HORAUD R. Skeletal quads：Human action recognition using joint quadruples[C]. 2014 International Conference on Pattern Recognition. Stockholm，Sweden：IEEE，2014：4513-4518.

[12] DU Y，WANG W，WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2015：1110-1118.

[13] KE Q H，BENNAMOUN M，AN S J，et al. A new representation of skeleton sequences for 3D action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2017：4570-4579.

[14] ZHANG P F，LAN C L，ZENG W J，et al. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE，2020：1109-1118.

[15] HU J，SHEN L，ALBANIE S，et al. Squeeze-and-excitation networks[J]. IEEE Trans Pattern Anal Mach Intell，2020，42(8)：2011-2023.

[16] SI C Y，CHEN W T，WANG W，et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2019：1227-1236.

[17] SHI L，ZHANG Y F，CHENG J，et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing，2020，29：9532-9545.

[18] 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，2023，45(2)：1474-1488.

[19] SHAHROUDY A，LIU J，NG T T，et al. NTU RGB+D：A large scale dataset for 3d human activity analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2016：1010-1019.

[20] 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, 2020, 42(10): 2684-2701.

[21] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 770-778.

[22] PENG W, HONG X, CHEN H, et al. Learning graph convolutional network for skeleton-based human action recognition by neural searching[C]. 34th AAAI Conference on Artificial Intelligence. Reston, USA: AAAI, 2020: 2669-2676.

[23] CHEN T L, ZHOU D S, WANG J, et al. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition[C]. 29th ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery, 2021: 4334-4342.

[24] ZHANG Q, WANG T, ZHANG M, et al. Spatial-temporal transformer for skeleton-based action recognition[C]. proceedings of the 2021 China Automation Congress(CAC), 2021, 22-24.

[25] LIU Z Y, ZHANG H W, CHEN Z H, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 140-149.

[26] YE F F, PU S L, ZHONG Q Y, et al. Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition[C]. 28th ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery, 2020: 55-63.

作者简介

刘斌斌,中级工程师,主要研究方向为综采智能化。

E-mail:liubinbin202203@163.com

赵宏涛(通信作者),硕士研究生,主要研究方向为行为识别。

E-mail:894069829@qq.com

王田,博士,副教授,主要研究方向为机器视觉与人工智能。

E-mail:wangtian@buaa.edu.cn

杨艺,博士,副教授,主要研究方向为深度学习、强化学习和智能控制。

E-mail:yangyi@hpu.edu.cn