时空自适应图卷积与Transformer结合的动作识别网络

韩宗旺 杨 涵 吴世青* 陈 龙
(上海理工大学机械工程学院，上海，200093)

摘 要：在一个以人为中心的智能工厂中，感知和理解工人的行为是至关重要的，不同工种类别往往与工作时间和工作内容相关。该文通过结合自适应图和Transformer两种方式使模型更关注骨架的时空信息来提高模型识别的准确率。首先，采用一个自适应的图方法去关注除人体骨架之外的连接关系。进一步，采用Transformer框架去捕捉工人骨架在时间维度上的动态变化信息。为了评估模型性能，制作了智能生产线装配任务中6种典型的工人动作数据集，并进行验证，结果表明所提模型在Top-1精度上与主流动作识别模型相当。最后，在公开的NTU-RGBD和Skeleton-Kinetics数据集上，将该文模型与一些主流方法进行对比，实验结果表明，所提模型具有良好鲁棒性。

关键词：智能工厂；工人动作识别；深度学习；自适应图；Transformer
中图分类号：TP391.41; TP18 **文献标识码**：A **文章编号**：1009-5896(2024)00-0001-09
DOI: 10.11999/JEIT230551

Action Recognition Network Combining Spatio-Temporal Adaptive Graph Convolution and Transformer

HAN Zongwang YANG Han WU Shiqing CHEN Long
(*School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)

Abstract: In a human-centered smart factory, perceiving and understanding workers' behavior is crucial, as different job categories are often associated with work time and tasks. In this paper, the accuracy of the model's recognition is improved by combining two approaches, namely adaptive graphs and Transformers, to focus more on the spatiotemporal information of the skeletal structure. Firstly, an adaptive graph method is employed to capture the connectivity relationships beyond the human body skeleton. Furthermore, the Transformer framework is utilized to capture the dynamic temporal variations of the worker's skeleton. To evaluate the model's performance, six typical worker action datasets are created for intelligent production line assembly tasks and validated. The results indicate that the model proposed in this article has a Top-1 accuracy comparable to mainstream action recognition models. Finally, the proposed model is compared with several mainstream methods on the publicly available NTU-RGBD and Skeleton-Kinetics datasets, and the experimental results demonstrate the robustness of the model proposed in this paper.

Key words: Intelligent manufacturing; Recognition of worker activity; Deep learning; Adaptive graph; Transformer

1 引言

在一个以人为中心的智能工厂中，感知和理解工人的行为是至关重要的，它不仅可以给技术不熟练的工人提供在线的及时指导[1–4]，还可以起到视频检索，智能监控，人机交互等作用[5–9]。但是，以图像或者视频为输入的方式都占用了大量时间和空间，而基于骨架节点的方式，不仅可以通过18或者25多个骨架节点就可以描述个体的运动状态，也可以通过这些节点计算瞬时运动速度和关节角度[10–18]。

基于骨架动作识别方法的发展可以概括为两个阶段[19–26]：基于传统机器学习的阶段和基于深度学习的阶段。在第1阶段，动作的特征需要根据自己的先验知识进行手工设计。这些特征包括运动轨迹、视角、几何形状等，如：Seidenari等人[24]对关

收稿日期：2023-06-05；改回日期：2024-03-11
*通信作者： 吴世青 wsq07599@usst.edu.cn
基金项目：国家自然科学基金(52005338)
Foundation Item: The National Natural Science Foundation of China (52005338)

节构建笛卡尔坐标系，通过这些坐标变化状态达到对动作分类的效果。Slama等人[25]利用骨架节点间的3D位姿关系，对3D骨架空间进行重新建模和识别。通过将人体动作序列建模为已生成3维关节轨迹的线性动力学系统。Stiefmeire等人[27]提出了一种基于字符串匹配的分割和分类方法用于识别在汽车装配任务中的工人行为。但由于时空信息挖掘尺度不够以及模型泛化能力弱，很难推广到不同的动作识别任务。基于深度学习方法主要包括卷积神经网络、循环神经网络、图卷积神经网络等。其中Wang等人[28]为了提高智能制造中的人机协作运行效率和安全，引入了深度卷积神经网络的方法去识别工人的行为，以便提前预测工人的下一步动作并提供即时的帮助。Jiang等人[29]将加速度计和陀螺仪的信号序列组装到一个新颖的活动图像中以便运用深度卷积神经网络从自动组装的图像中学习用于活动识别任务的最佳功能。Tao等人[30]为了解决工人的行为感知和理解问题，将智能臂章捕获的多个IMU信号和相机获取的视频信息结合起来，然后采用卷积神经网络去学习嵌入在信号中最有区别性的特征，进一步提出了多模态的工人活动识别算法，并取得了显著的效果。Yan等人[18]提出了一种基于骨架动作识别的时空图卷积网络模型，将人体动力学的知识和时空卷积的方法结合应用于动作识别中。

如上一些方式虽然在动作识别任务中展现较好的性能，但是，在智能工厂任务场景中往往需要在克服环境背景干扰达到对动作进行精准的理解的目的，并且网络需要达到端到端的效果。而使用IMU设备会影响工人的正常工作，基于深度学习一些方法忽略了对时空维度信息加工和利用，导致在特定的应用场景中性能下降。

本文首先利用自适应图卷积使网络不仅仅关注人体骨架本身，还关注与骨架相连接的其他部分。进一步，采用Transformer框架来捕捉工人骨架在时间维度上的动态变化信息，这意味着使用了Transformer的强大能力来分析工人骨架在不同时间点上的变化情况，以更全面地理解和描述动作的动态特征。本文将时空自适应图卷积与Transformer结合，提出适用于智能工厂的动作分类模型，即STA-GCN-Transformer。为了验证在该场景的本文模型在该场景的适用性，采集和制作了基于智能工厂的工人动作数据集，并在该数据集上验证本文算法的优越性，进一步，为了测试模型的通用性，在公开数据集与其他方法进行对比。最后，为了验证本文提出模块的有效性，在消融实验中验证了该模型的工作机理。

2 工人动作数据集制作

2.1 数据采集

为了测试动作识别模型在智能工厂场景中应用的效果，选择智能工厂装配任务中常见的6种工人活动，分别是：拿工具(Grab a Tool, GT)，敲钉子(Hammer Nail, HN)，使用电钻(Use a Power-screwdriver, UP)，喝水(Drink Water, DW)，使用螺丝刀(Turn Screwdriver, TS)和使用卷尺(Use a Tape, UT)。其中，视频的采集使用海康威视公司推出的DS-2CD2T26WDA2-I相机获取操作员动作，其工人的活动信息如表1所示。

为了采集工人活动的数据集，实验场景布置如图1(左)所示，操作员站在工作台前以自然流畅的方式执行一个装配任务，位于正上方的相机记录了每个操作员的活动轨迹。图1(右)展示了工人在执行装配任务中的6种行为的样本。

2.2 数据的注释

为了使网络可以达到端到端的效果，并且网络抗环境干扰性强，本文选择了基于骨架数据作为训练数据。首先对采集的每类动作的视频进行裁剪处理，重新以15 Frame/s的帧率将原始视频剪辑为10 s小视频片段。表2记录了工人数据集所采集的样本分别。为了降低计算量，将原始视频的分辨率从640×1 080重新缩放为240×320，然后用位姿估

表 1 工人活动任务

编号	任务	活动
1	从工具箱中拿出工具	GT
2	在板上钉4颗钉子	HN
3	使用电钻拧紧10颗螺丝	UP
4	喝水15 s	DW
5	使用螺丝刀拧紧10颗螺丝	TS
6	使用卷尺	UT

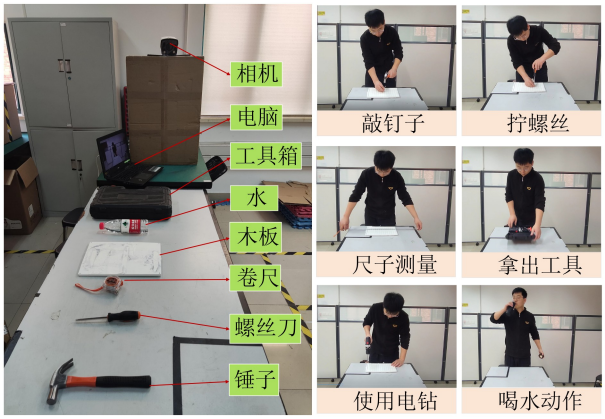

图 1 数据收集场景布置与6种工人活动示例图

计算法OpenPose[21]提取工人的骨架组成$C\times T\times V$的张量(如图2)。其中，C, T和V分别表示骨架节点2D或者3D像素坐标、视频帧数和骨架关节点的数量。为了泛化模型，将提取的人体骨架关节点进行随机移动和缩放以便增强数据集。

3　本文模型

如图3所示为本文提出模型的整体流程。模型的工作的主要流程是：首先利用时空自适应图卷积网络聚合空间上相邻关节点的信息，然后将提取到的特征展成一个序列，用Transformer网络去捕捉骨架序列的时间上的依赖性，将输出的特征利用Softmax计算类别概率。

3.1　算法模型框架

如图4所示，所提出的方法主要由5个主要组件组成，即特征嵌入模块、位置编码模块、编码器模块、动作查询解码模块和预测模块。首先，利用特征嵌入模块主要完成从输入的RGB视频转为高维的具有邻域聚合关系的骨架特征图。其中，包含RGB视频变为骨架信息和时空自适应图卷积对时空邻域特征图的提取。其次，将提取后的特征图加上位置编码，以获得逐帧逐节点的位置关系。随后，将带有位置信息的特征图输入编码模块中，将经过编码高维映射的特征图送到解码器中学习逐帧骨架节点与时序骨架节点的长时间依赖关系。最后，将解码后的特征进行分类动作所属的类别。

3.1.1　特征嵌入

考虑一个包含工人行为的骨架序列$\mathcal{X}=\{\mathcal{X}_0,\mathcal{X}_1,\cdots,\mathcal{X}_T\}\in\mathbb{R}^{N\times C}$，其中$T$是视频的帧数，$N$骨架的关节点数量，$C$是每个关节点的特征数量。进一步，观察到人体骨架不仅仅是节点之间物理连接关系，动作类别与各个骨架节点关系影响巨大[31]，为此本文采用一个自适应图的方式，去记录和学习各个骨架节点在高维中相互的依赖关系。考虑一个视频帧$\mathcal{X}_i\in\mathbb{R}^{N\times 3}$，嵌入的特征图$J_i\in\mathbb{R}^{N\times d_{model}}$可以用式(1)表示，FE用于从原始数据中提取有用特征的模块。

表 2　工人的行为数据集样本分布

自愿者编号	GT	HN	UP	DW	TS	UT
1	39	27	25	18	28	41
2	21	35	15	24	17	32
3	25	20	37	31	33	35
4	31	30	18	29	46	29
合计	116	112	95	102	124	137

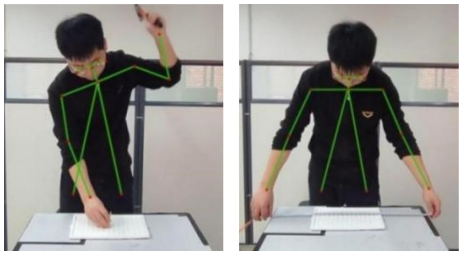

图 2 OpenPose提取的人体骨架示例

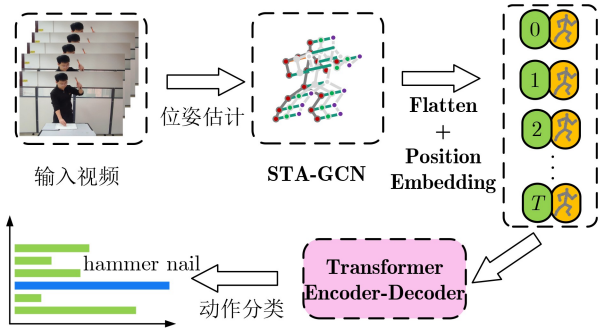

图 3 模型的整体流程

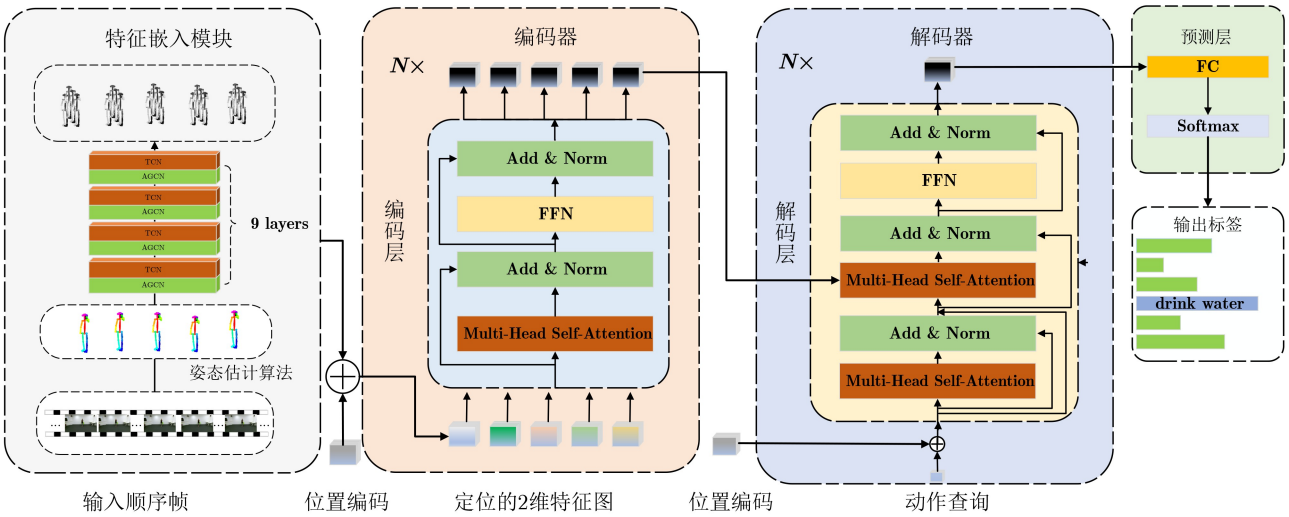

图 4 STA-GCN-Transformer动作识别网络的详细架构

$$J_i = FE(\mathcal{X}_i), \qquad i \in [1,T] \tag{1}$$

3.1.2 图卷积神经网络

人体骨架序列可以看为一系列的静态姿态在时间维度上的有序排列。在时空图上存在两种关系，一种是人体骨架的自然连接性所构成的空间上的关系，另一种是相同节点在不同帧之间所构成的时间上的关系。将一个人体骨架利用一个图去描述可记作$\varsigma=(\nu,\varepsilon)$，$\nu=\{v_1,v_2,\cdots,v_N\}$是$N$个人体关节点集合，而$\varepsilon$则是代表节点与之间的物理连接的集合。为了记录这些物理连接之间的邻域关系，使用一个邻接矩阵$\boldsymbol{A}\in\mathbb{R}^{N\times N}$去表示，如果第$i$个关节点和第$j$个关节点被连，$\boldsymbol{A}_{i,j}$=1，否则为0。进一步来说，一个节点特征在序列中的集合表示为$\mathcal{X}=\{x_{t,n}\in\mathbb{R}^c\,|t,n\in\boldsymbol{Z},1\le t\le T,1\le n\le N\}$，其中对于每个元素可以表示为特征张量$X\in\mathbb{R}^{T\times N\times C}$，其中$x_{t,n}=X_{t,n,}$表示为整个$T$帧中$t$时刻下$v_n$节点的$C$维特征向量。因此，骨架之间的邻域连接关系在结构上使用连接矩阵$\boldsymbol{A}$去描述，在节点元素的特征上利用X去描述。

在特征X和图结构$\boldsymbol{A}$所定义的骨架输入上，GCN的逐层更新规则可在t时刻的特征上进行转换，如式2所示

$$\boldsymbol{X}_t^{(l+1)}=\sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}_t^{(l)}W^{(l)}) \tag{2}$$

其中，$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$是添加了自环以保持一致性的骨架图，$\tilde{\boldsymbol{D}}$是矩阵$\tilde{\boldsymbol{A}}$的度矩阵，$\tilde{\boldsymbol{D}}_{ii}=\sum_j(\boldsymbol{A}_{ij}+\boldsymbol{I}_{ij})$，$\sigma(\cdot)$是激活函数。$\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}_t^{(l)}$可以解释为直接邻域的近似空间平均特征聚集，再经过激活的线性层进行变换。$W^{(l)}\in\mathbb{R}^{C_l\times C_{l+1}}$表示网络$l$层的可学习权值矩阵。

3.1.3 自适应图卷积神经网络

进一步从图卷积的式(2)可以观察到，一个图主要描述的是骨架节点之间的物理连接关系，这个关系也包含着相邻一边。不同类别的动作都是由身体的所有骨架节点组成，某一些的骨架节点对该动作参与度不高，但是在某一些时序中，它与运动较为明显的骨架节点有着一定的时空依赖性。如图5在“敲钉子”动作中，在整个时序中左躯干运动幅度是巨大的，它负责左手拿锤子抬手的动作。虽然左手与右手之间并没有直接的物理连接，但是，对于拿钉子来说动作是依赖右手的。所以，两者的依赖对于该类动作的区分是必不可少的。为了让图之间的连接更为广泛，参考如文献[18]中的方式利用注意力机制去关注骨架节点之间的非物理连接。矩阵初始化为1的自注意力图$\boldsymbol{M}$来关注骨架节点在时空中长时间依赖关系。所以，新的矩阵为$\boldsymbol{A}=\boldsymbol{A}\tilde{\boldsymbol{M}}$。显然，乘法操作不能改变邻接矩阵中为0的值，即不能产生除物理连接之外的连接关系。受文献[26]所启发，应用归一化的嵌入高斯函数去计算两个关节点之间的相似度，这样不仅可以反映出两个关节点之间的连接性，还可以决定它们之间的连接强度。

$$f(v_i,v_j)=\frac{\mathrm{e}^{\theta(V_i)^T\varphi(V_j)}}{\sum_{j=1}^{N}\mathrm{e}^{\theta(V_i)^T\varphi(V_j)}} \tag{3}$$

其中，v_i, v_j代表两个不同节点，N是节点个数。如果$\boldsymbol{P}$来表示嵌入高斯函数后计算出的相似度矩阵，则新的邻接矩阵表示为

$$\boldsymbol{A}=\boldsymbol{P}+\tilde{\boldsymbol{A}} \tag{4}$$

如图5所示，彩色的点代表图的顶点，这些顶点包含了工人关节点的2D或者3D坐标信息，彩色的粗线即图上两个顶点之间的边，表示人体关节点之间的骨头的连接关系。这些边反映了关节点在物理上相互连接依赖关系。图中的虚线表示物理上不连接的两个关节点之间的外部依赖关系，它反映的是针对不同工人动作产生的物理上不连接关节点之间的一种依赖关系。

3.2 位置编码

在视频帧序列中，为了让模型能够利用帧的顺序信息，将相对或绝对位置信息嵌入到视频帧的特征映射图中。与自然语言处理中Transformer模型使用的词向量位置编码不同，视频帧序列中的位置编码是一个3D张量，与视频帧特征映射图具有相同的维度，这使得位置编码能够与特征映射图进行相加操作。本文使用了不同频率的正弦和余弦函数来生成位置编码，具体而言，对于每个位置和维度，位置编码的数值由以下公式计算得出

$$\mathrm{PE}(\mathrm{pos},(\mathrm{i},\mathrm{j},2\mathrm{k}))=\sin(\mathrm{pos}/10000^{2k/d\ \mathrm{mod}\ el}) \tag{5}$$

$$\mathrm{PE}(\mathrm{pos},(i,j,2k+1))=\cos(\mathrm{pos}/10000^{2k/d\ \mathrm{mod}\ el}) \tag{6}$$

其中，pos是位置，(i,j)表示特征的空间位置，$2k$

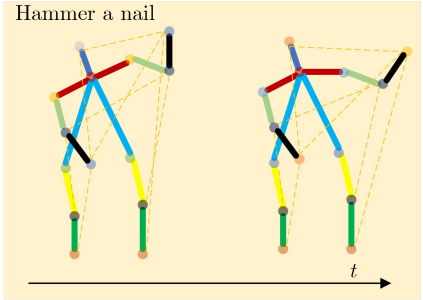

图 5 建模工人骨架作为图结构

表示特征通道维数。从上式中可以看出位置编码中的每个通道维度对应着一个正弦曲线。选择本函数是因为它能让模型很轻松的关注到相对位置的信息，因为对于任何一个固定的偏移量m，PE_{pos+m}都可以被PE_{pos}线性表示。考虑一个嵌入的特征图J_i，位置编码能够用式(7)表示，其中$\oplus$为张量的逐元素相加操作。

$$Z_i = J_i \oplus PE(i), i \in [1, n] \tag{7}$$

3.3 编码器和解码器

编码器模块是一个由N个层堆叠而成的结构，每个层包括1个多头注意力机制和1个全连接前馈网络层。这两个子层之间采用残差连接，随后接1个归一化层。为了促进残差连接，每个子层以及嵌入层产生同样的输出维度$d_{\mathrm{mod}\,el} = 256$。若加入位置编码的骨架特征序列$Z = \{Z_0, Z_1, \cdots, Z_T\} \in \mathbb{R}^{N \times d}$，编码器的输出结果$\hat{Z}$可以表示为

$$\hat{Z} = \mathrm{Encoder}(Z) \tag{8}$$

解码器模块同样也是由N个层堆叠而成。除了上述两个子层之外，加入了第3个子层，该子层在编码器输出结果上执行多头注意力，然后接1个归一化层。解码器在每2个子层之间也应用了残差连接，考虑1个动作查询嵌入$Q \in \mathbb{R}^{1 \times d_{\mathrm{mod}\,el}}$，则解码器的输出结果可以表示为

$$\hat{Q} = \mathrm{Decoder}(\hat{Z}, Q) \tag{9}$$

3.4 多头注意力机制

注意力函数可以描述为一种将查询和一组键值对映射为输出的方式，其中查询、键、值和输出都是张量。输出结果可以通过对值进行加权求和得到，其中每个值的权重由查询和对应的键兼容函数所计算得到。在本文中，将编码器输出结果中的时间和空间维度展成一个1维序列，即$Q, K \in \mathbb{R}^{d_k \times TV}$，$V \in \mathbb{R}^{d_v \times TV}$，其中$d_k, d_v$分别是查询、键和值嵌入的维度，$T$是视频帧数，$V$是单帧骨架关节点个数。计算查询和所有键的点积，然后除以$\sqrt{d_k}$，最后运用一个Softmax函数去获得值上的权重。计算输出矩阵可以用式(10)表示

$$\mathrm{attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

为了使模型能够学习到来自不同位置的不同表示子空间信息，加入多头注意力。

$$\mathrm{multiHead}(Q, K, V) = \mathrm{C}(h_1, h_2, \cdots h_h)W^O \tag{11}$$

$$h_i = \mathrm{attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{12}$$

其中，投影矩阵$W_i^Q \in \mathbb{R}^{d_{\mathrm{mod}\,el} \times d_k}$，$W_i^k \in \mathbb{R}^{d_{\mathrm{mod}\,el} \times d_k}$，$W_i^V \in \mathbb{R}^{d_{\mathrm{mod}\,el} \times d_v}$，$W_i^Q \in \mathbb{R}^{hd_v \times d_{\mathrm{mod}\,el}}$，$h$=6，它表示平行注意力层，C为Concat操作。

4 实验

为了验证本文提出方法的有效性，在自制的工人数据集、公开的NTU-RGBD[23]和Skeleton-Kinetics[18]数据集进行实验。最后，在工人行为数据集上进行了消融实验，验证自适应图和Transformer框架对模型的有效性。

4.1 数据训练

本文模型是基于PyTorch深度学习框架搭建，采用带有Nesterov动量(0.9)的随机梯度下降法作为模型优化策略，选择交叉熵作为梯度反向传播的损失函数，权重衰减因子设置为0.000 5。对于采集的数据集中少于300帧的样本，从样本中随机选择帧直到每个样本都达到300帧。学习率被初始化为0.01，然后在第30和50个epoch时，学习率除以10，直到80个epoch后结束训练。所有实验都在一台带有4张2080Ti显卡的服务器上运行。

4.2 与先进方法的比较

在工人数据集上的评估结果。本节选择动作识别领域目前性能先进的一些网络，即Deep LSTM[23], TCN[13], ST-GCN [18], DSTANet[7]和PoseConv3D[17]网络与所提方法进行比较。在工人数据集的难点主要在于遮挡和类似行为等因素的干扰，网络需要强大的理解能力来克服类似行动的影响。本章只提供测试类别Top-1的准确性，因为总共只有6个动作，而Top-5指标对于这个数据集过于宽松。

如表3所示，在Top-1指标在中与其他方法相比，本文所提出的模型比ST-GCN模型高出8.69%，优于DSTANet模型3%之上，低于最优PoseConv3D模型0.39%，证明该模型可以很好地对动作进行分类，在理解视频中工人的行为方面更加稳健。

为了评估本文方法的泛化性能，与其他具有竞争力的基于骨架的动作识别方法在NTU-RGBD和Skeleton-Kinetics数据集上进行了比较，比较结果被分别列在表格4和表格5中。

表 3　在工人数据集和其他有竞争力的方法进行比较(%)

方法	Top-1
Deep LSTM [23]	79.89
TCN [13]	80.94
ST-GCN [18]	82.22
DSTANet [7]	87.73
PoseConv3D[17]	**91.30**
本文模型	90.91

这些方法主要划分为以下4个类别，并在表格中使用横线进行分隔：包括基于手工设计特征的方法、基于RNN的方法、基于CNN的方法和基于GCN的方法。从表4中的结果可以看出，在X-Sub和X-View子数据集指标上，本文方法要优于基于CNN所列举的方法，远高于基于手工设计特征的方法。在基于GCN的方法中，超出DPRL+GCNN模型2.45%和2.05%，但是在X-Sub指标上低于先进Shift-GCN模型和MSST-Net模型，在X-View指标上，低于先进MSST-Net模型0.95%。从表5中的结果可以看出，在Top-1和Top-5指标上，本文方法超出DSTANet模型1.05%和1.35%，低于先进Co-AGCN模型2.5%和2.75%。

4.3 消融实验

本节在工人行为数据集上对提出的各个组件进行了详细的实验，以便验证模型的各个组件的有效性。

4.3.1 自适应图策略

为了让模型不仅能够捕捉到人体物理结构上相连以及不相连的关节点之间的依赖性。采用嵌入的高斯函数去计算每个关节点与其他关节点之间的相似度，构成一个可学习的图，该图能够在模型的训练过程中进行优化和更新；通过这种方式，模型可以捕捉到不同动作之间的直观的依赖关系。为了展示这个可学习图产生的新连接。如图6所示，依次可视化了根据人体物理结构构成的邻接矩阵以及模型学习到的第4层自适应图卷积的邻接矩阵，这些邻接矩阵的可视化使用颜色来表示连接的强度。

随着网络训练的深入，模型能够捕捉到更多与人体物理结构之外的连接关系，这意味着模型能够在学习过程中识别并捕捉到非相连节点的更多关联，这进一步验证了本策略的有效性。

此外，为了验证自适应策略的性能，在工人行为数据集上进行实验，将模型去掉自适应图，用原始的邻接矩阵来代替，用原始模型表示。同时可视

表 4 在NTU-RGBD数据集上和其他有竞争力的方法进行比较(%)

方法	X-Sub	X-View
Lie Group [4]	50.10	82.80
Deep LSTM [23]	60.70	67.30
ARRN-LSTM [10]	80.70	88.80
nd-RNN [11]	81.80	88.00
TCN [13]	74.30	83.10
Clips+CNN+MTLN [14]	79.60	84.80
Synthesized CNN [15]	80.00	87.20
CNN+Motion+Trans [16]	83.20	89.30
ST-GCN [18]	81.50	88.30
Shift-GCN [12]	**96.50**	90.70
MSST-Net [9]	86.60	**92.80**
DPRL+GCNN [19]	83.50	89.80
本文模型	85.95	91.85

表 5 在Skeleton-Kinetics数据上和其他有竞争力的方法进行比较(%)

方法	Top-1	Top-5
Deep LSTM [23]	16.40	35.30
TCN [13]	20.30	40.00
ST-GCN [18]	30.70	52.80
DSTANet [7]	31.10	53.20
CoAGCN [5]	**35.00**	**57.30**
本文模型	32.15	54.55

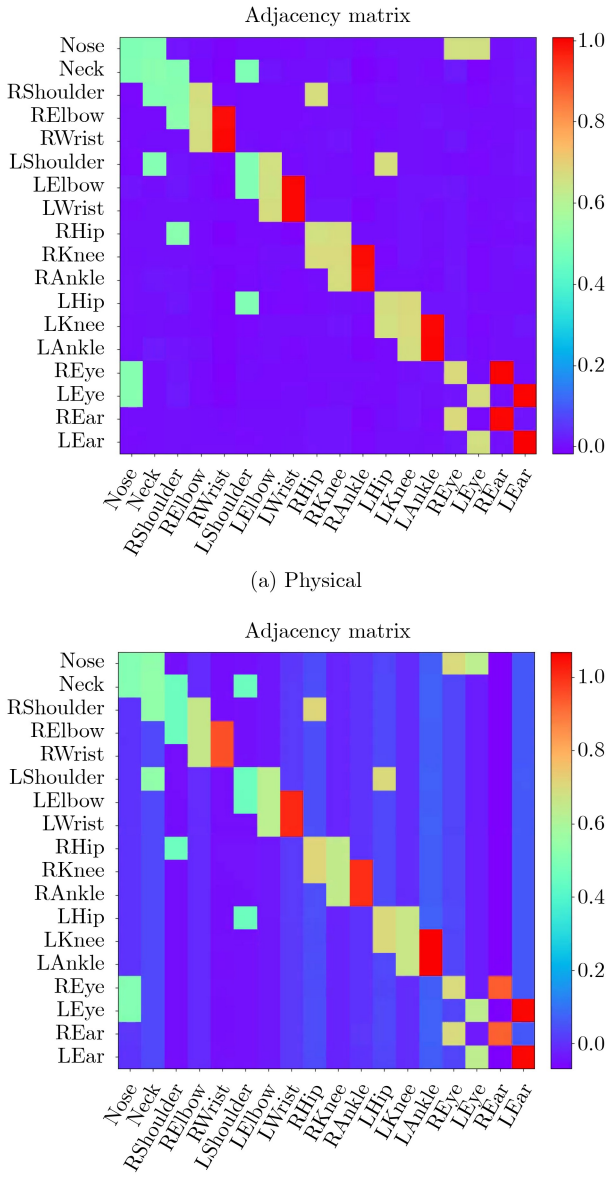

图 6 自适应图卷积的邻接矩阵示例图

化了这两种方法在工人行为数据集上训练时的Top-1和Top-5精度，如图7所示。可以明确看出自适应策略的优越性，其中在Top-1精度上，本文模型优于原始模型7%；在Top-5精度上，本文模型优于原始模型1.53%。

4.3.2 Transformer模块

为了评估Transformer模块的重要性，将Transformer中的主要组件拆开，即编码模块、解码模块、全连接的前馈网络(FFN)、多头注意力机制(MHA)和位置编码模块(PE)，以证明各个组件对动作识别的性能的影响。整个实验在工人数据集上进行。首先，将MHA替代空间图卷积Graph Convolutional Network(GCN)来聚合关节点领域的信息，如表6所示，发现其性能不及GCN，这可能是因为GCN使用手动设计的固定图作为先验知识，而MHA缺乏这种先验信息。随后，加入了FFN来扩大模型的容量，通过将MHA与FFN组合，模型的性能有所提升，但仍低于ST-GCN的准确度。之后，在前面的基础上加入了PE，如图8所示，可视化了送入Transformer之前的位置编码，其大小和特征嵌入得到的张量一致，其格式为(1,256,18,38)。表6中的MHA+FFN+PE相当于是一个编码模块，其相比于MHA+FFN，模型的Top-1和Top-5精度均提高了12%以上，这说明，模型确实学到了帧的序列信息。再者，表6中AGCN的Top-1结果优于GCN的Top-1结果4.89%，说明自适应图更有效地学习图表示。最后，将4层STA-GCN最后提取到的信息送入到Transformer，即ST-GCN+Transformer模型，其结果比AGCN+FFN+PE和ST-GCN有所改善，但是低于本文所提出的STA-GCN+Transformer模型在Top-1精度上3.3%。

5　结束语

在本文中，为提高在智能工厂中工人的动作识别的准确率，通过结合自适应图卷积和Transformer模型的方式，提高了对网络高维时空信息中的理解力。具体来说，首先，采用一个自适应的图方法去关注除人体骨架之外的连接关系；进一步，采用Transformer框架去捕捉工人骨架在时间维度上的动态变化信息。然后，采集和制作了含有6种典型常见动作的工人数据集。最后，在自制的工人数据集、公开的NTU-RGBD和Skeleton-Kinetics数据集上，将本文提出的模型与其他主流模型在该数据集上进行测试。实验结果表明，本文提出的模型有着良好的鲁棒和适用性。并且，在消融实验中也证明这种自适应图和Transformer模块对网络有效性的影响。

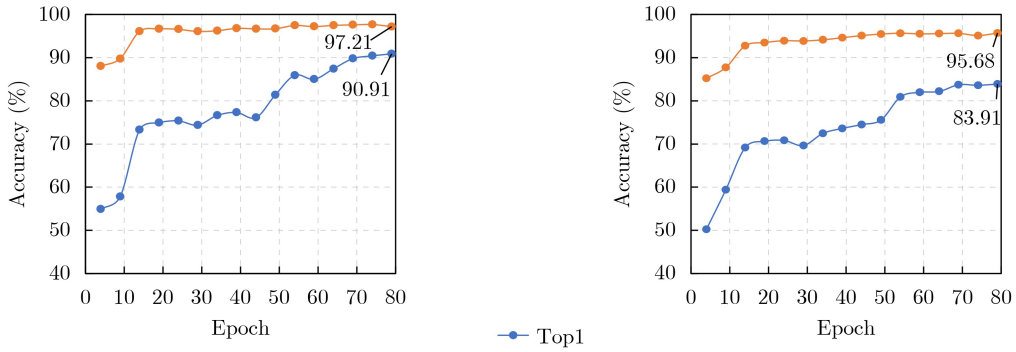

图 7 本文模型与原始模型方法在工人行为数据集上的测试精度

表 6　比较Transformer的各个组件对工人动作识别精度的影响

方法	Top-1(%)	Top-5(%)
MHA	38.28	40.95
MHA+FFN	54.67	62.61
MHA+FFN+PE	76.68	87.88
GCN+FFN+PE	70..53	81.37
AGCN+FFN+PE	75.42	86.68
ST-GCN+Transformer	87.61	92.91
STA-GCN+Transformer	**90.91**	**97.21**

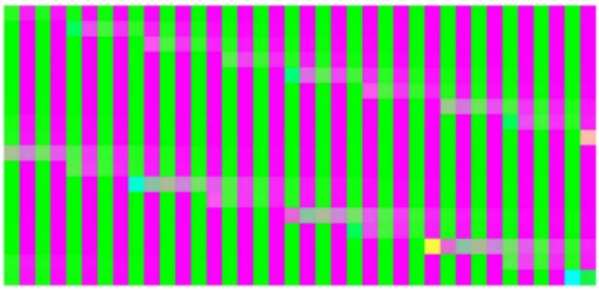

图 8 将Feature embedding最后提取到的特征加入位置编码以获取帧序列信息

参 考 文 献

[1] 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. doi: 10.11999/JEIT220270.

SHI Yuexiang and ZHU Maoqing. Collaborative convolutional transformer network based on skeleton action recognition[J]. *Journal of Electronics & Information Technology*, 2023, 45(4): 1485–1493. doi: 10.11999/JEIT220270.

[2] GEDAMU K, JI Yanli, GAO Lingling, *et al*. Relation-mining self-attention network for skeleton-based human action recognition[J]. *Pattern Recognition*, 2023, 139: 109455. doi: 10.1016/j.patcog.2023.109455.

[3] GUO Hongling, ZHANG Zhitian, YU Run, *et al*. Action recognition based on 3D skeleton and LSTM for the monitoring of construction workers' safety harness usage[J]. *Journal of Construction Engineering and Management*, 2023, 149(4): 04023015. doi: 10.1061/JCEMD4.COENG-12542.

[4] VEMULAPALLI R, ARRATE F, and CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 588–595. doi: 10.1109/CVPR.2014.82.

[5] HEDEGAARD L, HEIDARI N, and IOSIFIDIS A. Continual spatio-temporal graph convolutional networks[J]. *Pattern Recognition*, 2023, 140: 109528. doi: 10.1016/j.patcog.2023.109528.

[6] YU B X B, LIU Yan, ZHANG Xiang, *et al*. Mmnet: A model-based multimodal network for human action recognition in RGB-D videos[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3522–3538. doi: 10.1109/TPAMI.2022.3177813.

[7] SHI Lei, ZHANG Yifan, CHENG Jian, *et al*. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition[C]. Proceedings of the 15th Asian Conference on Computer Vision, Kyoto, Japan, 2021. doi: 10.1007/978-3-030-69541-5_3.

[8] 陈莹, 龚苏明. 改进通道注意力机制下的人体行为识别网络[J]. 电子与信息学报, 2021, 43(12): 3538–3545. doi: 10.11999/JEIT200431.

CHEN Ying and GONG Suming. Human action recognition network based on improved channel attention mechanism[J]. *Journal of Electronics & Information Technology*, 2021, 43(12): 3538–3545. doi: 10.11999/JEIT200431.

[9] CHENG Qin, REN Ziliang, CHENG Jun, *et al*. Skeleton-based action recognition with multi-scale spatial-temporal convolutional neural network[C]. Proceedings of 2021 IEEE International Conference on Real-time Computing and Robotics, Xining, China, 2021: 957–962. doi: 10.1109/RCAR52367.2021.9517665.

[10] LI Lin, ZHANG Wu, ZHANG Zhaoxiang, *et al*. Skeleton-based relational modeling for action recognition[J]. arXiv preprint arXiv: 1805.02556, 2018. **(查阅网上资料, 请核对文献类型及格式)**.

[11] LI Shuai, LI Wangqiang, COOK C, *et al*. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN[C]. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5457–5466. doi: 10.1109/CVPR.2018.00572.

[12] CHENG Ke, ZHANG Yifan, HE Xiangyu, *et al*. Skeleton-based action recognition with shift graph convolutional network[C]. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 183–192. doi: 10.1109/CVPR42600.2020.00026.

[13] KIM T S and REITER A. Interpretable 3D human action analysis with temporal convolutional networks[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 1623–1631. doi: 10.1109/CVPRW.2017.207.

[14] KE Q H, BENNAMOUN M, AN S J, *et al*. A new representation of skeleton sequences for 3D action recognition[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 3288–3297. doi: 10.1109/CVPR.2017.486.

[15] LIU Mengyuan, LIU Hong, and CHEN Chen. Enhanced skeleton visualization for view invariant human action recognition[J]. *Pattern Recognition*, 2017, 68: 346–362. doi: 10.1016/j.patcog.2017.02.030.

[16] DU Yong, FU Yun, and WANG Liang. Skeleton based action recognition with convolutional neural network[C]. Proceedings of 2015 3rd IAPR Asian Conference on Pattern Recognition, Kuala Lumpur, Malaysia, 2015: 579–583. doi: 10.1109/ACPR.2015.7486569.

[17] DUAN Haodong, ZHAO Yue, CHEN Kai, *et al*. Revisiting skeleton-based action recognition[C]. Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 2969–2978. doi: 10.1109/CVPR52688.2022.00298.

[18] YAN Sijie, XIONG Yuanjun, and LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, USA, 2018: 912.

[19] TANG Yansong, TIAN Yi, LU Jiwen, *et al*. Deep progressive reinforcement learning for skeleton-based action recognition[C]. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5323–5332. doi: 10.1109/CVPR.2018.00558.

[20] BASAK H, KUNDU R, SINGH P K, *et al*. A union of deep learning and swarm-based optimization for 3D human action recognition[J]. *Scientific Reports*, 2022, 12(1): 5494. doi: 10.1038/s41598-022-09293-8.

[21] CAO Zhe, SIMON T, WEI S E, *et al*. Realtime multi-person 2D pose estimation using part affinity fields[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 7291–7299. doi: 10.1109/CVPR.2017.143.

[22] CHEN Yuxin, ZHANG Ziqi, YUAN Chunfeng, *et al*. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 13359–13368. doi: 10.1109/ICCV48922.2021.01311.

[23] SHAHROUDY A, LIU Jun, NG T T, *et al*. NTU RGB+d: A large scale dataset for 3D human activity analysis[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1010–1019. doi: 10.1109/CVPR.2016.115.

[24] SEIDENARI L, VARANO V, BERRETTI S, *et al*. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses[C]. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, USA, 2013: 479–485. doi: 10.1109/CVPRW.2013.77.

[25] SLAMA R, WANNOUS H, DAOUDI M, *et al*. Accurate 3D action recognition using learning on the Grassmann manifold[J]. *Pattern Recognition*, 2015, 48(2): 556–567. doi: 10.1016/j.patcog.2014.08.011.

[26] SHI Lei, ZHANG Yifan, CHENG Jian, *et al*. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. *IEEE Transactions on Image Processing*, 2020, 29: 9532–9545. doi: 10.1109/TIP.2020.3028207.

[27] STIEFMEIER T, ROGGEN D, OGRIS G, *et al*. Wearable activity tracking in car manufacturing[J]. *IEEE Pervasive Computing*, 2008, 7(2): 42–50. doi: 10.1109/MPRV.2008.40.

[28] WANG Limin, XIONG Yuanjun, WANG Zhe, *et al*. Temporal segment networks: Towards good practices for deep action recognition[C]. Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 20–36. doi: 10.1007/978-3-319-46484-8_2.

[29] JIANG Wenchao and YIN Zhaozheng. Human activity recognition using wearable sensors by deep convolutional neural networks[C]. Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 2015: 1307–1310. doi: 10.1145/2733373.2806333.

[30] TAO Wenjin, LEU M C, and YIN Zhaozheng. Multi-modal recognition of worker activity for human-centered intelligent manufacturing[J]. *Engineering Applications of Artificial Intelligence*, 2020, 95: 103868. doi: 10.1016/j.engappai.2020.103868.

[31] SONG Yifan, ZHANG Zhang, SHAN Caifeng, *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. doi: 10.1109/TPAMI.2022.3157033.

韩宗旺：男，博士，研究方向为机器视觉和人机协作.
杨　涵：男，硕士，研究方向为机器视觉和动作识别.
吴世青：男，副教授，硕士生导师 研究方向为机器人与数字孪生.
陈　龙：男，教授，博士生导师 研究方向为机器人与机器视觉.

责任编辑：马秀强