基于时频数据融合的分层图卷积手势识别方法

黄葵[1)], 刘石坚[1)*], 邹峥[2)], 荆东星[3)]

[1)] (福建省大数据挖掘与应用技术重点实验室　福州　350118)
[2)] (福建师范大学计算机与网络空间安全学院　福州　350117)
[3)] (湘西民族职业技术学院信息与智能学院　吉首　416099)
(liusj2003@fjut.edu.cn)

摘　要: 图卷积神经网络善于提取非欧几里得结构特征, 已成为基于骨架手势识别的主流方法. 针对图卷积神经网络方法训练和推理时间较长, 且准确率仍有待提升的问题, 提出一种基于时频数据融合的分层图卷积手势识别方法 HHTS-Net. 首先提出一种空间注意力块与图卷积模块相结合的框架, 降低计算负担; 然后提出针对手掌及指关节的分层图卷积模块, 提升特征提取效率; 最后提出一种融合频域特征的多流学习方案, 进一步提升模型的性能. 在公共数据集 DHG 和 SHREC'17 上与较权威的方法进行实验, 结果表明, HHTS-Net 的准确率至少提升 1.8%和 0.3%, 推理速度加快 47.6%和 52.9%以上; 通过大量消融实验, 验证了该方法的有效性. 文中算法的源代码详见 https://github.com/CoderHoooK/HTSATNet.

关键词: 手势识别; 图卷积神经网络; 多分支流特征学习; 傅里叶变换
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2024-00493

A Hierarchical Graph Convolutional Method for Gesture Recognition Based on Temporal-Frequency Data Fusion

Huang Kui[1)], Liu Shijian[1)*], Zou Zheng[2)], and Jing Dongxing[3)]

[1)] (*Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fuzhou*　350118)
[2)] (*College of Computer and Cyber Security, Fujian Normal University, Fuzhou*　350117)
[3)] (*College of Information and Intelligence, Xiangxi National Vocational and Technical College, Jishou*　416099)

Abstract: Graph Convolutional Neural Networks (GCNs) are adept at extracting non-Euclidean structural features and have become the mainstream method for skeleton-based gesture recognition. In response to the issues of long training and inference times, as well as the need for improved accuracy in GCN methods, a hierarchical graph convolutional gesture recognition method based on the fusion of temporal-frequency data, named HHTS-Net, is proposed. Firstly, a framework that combines spatial attention blocks with graph convolutional modules is proposed to reduce computational burden. Then, a hierarchical graph convolutional module tailored to the structural characteristics of the palm and finger joints is introduced to enhance feature extraction efficiency. Finally, a multi-stream learning scheme that integrates frequency domain features is proposed to further improve model performance. Experiments conducted on public datasets DHG and SHREC'17 with authoritative methods demonstrate that HHTS-Net's accuracy is improved by at least 1.8% and 0.3%, respectively, and inference speed is increased by more than 47.6% and 52.9%. Extensive ablation

收稿日期: 2024-08-22; 修回日期: 2025-01-16. **基金项目**: 国家自然科学基金面上项目(62172095); 福建省自然科学基金面上项目(2022J01932); 湖南省自然科学基金(2024JJ7549); 湖南省教育厅科学研究基金(24C1236); 湘西职院院级科研团队基金. **黄葵**(1999—), 男, 硕士研究生, 主要研究方向为深度学习、动作识别; **刘石坚**(1983—), 男, 博士, 副教授, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为计算机图形学、深度学习; **邹峥**(1984—), 女, 博士, 讲师, 硕士生导师, 主要研究方向为医学图像处理、深度学习; **荆东星**(1982—), 男, 硕士, 教授, 主要研究方向为人工智能.

experiments validate the effectiveness of this method. The source code is available at https://github.com/CoderHoooK/HHTSATNet.

Key words: gesture recognition; graph convolutional neural networks; multi-branch feature learning; Fourier transform

1 相关工作

手势识别在视觉相关领域具有重要的地位和意义，它能分析和识别手部动作，为特定意图、情感和行为的判断提供依据. 例如，在交通运输场景中，与嘈杂环境下的语言交流相比，手势沟通往往更直观高效[1]. 因此，手势识别被广泛应用于行为分析[2]、人机交互[3-4]、虚拟/混合现实[5-6]等领域.

通常，手势识别的输入数据包括彩色图像[7-8]、深度图像[9]和骨架数据[10]. 例如，图 1a 所示为一幅典型的彩色输入图像；图 1b 所示为通过 Kinect 红外相机[11]捕获的 3 帧深度图像，其中手部区域的彩色点与线组成手部骨架数据；基于上述数据和 HRNet[12]等关键点检测算法，可得到图 1b 中彩色点和线表示的骨架数据. 这 3 类数据中，彩色图像最方便获取，但也容易受光照、背景等环境因素影响；深度图像可以与彩色图像配合使用，为其提供景深信息；与上述数据相比，骨架数据更为简洁、鲁棒[13-14]，是当前普遍采用的数据模态.

a. RGB 彩色图像 b. 3 帧 Kinect 深度图像

图 1 RGB 和 Kinect 传感器捕获的手势数据示例

随着人工智能技术的发展，通过深度学习实现手势识别成为一种趋势. Devineau 等[10]使用卷积神经网络(convolutional neural network, CNN)同时对时间和空间 2 个维度进行特征提取；Lai 等[15]则选择循环神经网络(recurrent neural network, RNN)处理时间维度的特征. 由于骨架在空间维度上属于典型的非欧几里得结构，而图卷积神经网络(graph convolutional neural network, GCN)更善于处理这类数据[16]，因此本文使用 GCN 解决基于骨架的手势识别问题.

在 GCN 中常常通过堆叠图卷积层扩大感受野，使模型理解整个图的结构和节点关系，提高全局特征的抽象能力；然而计算量也会随之攀升，导致系统不堪重负. 为此，本文提出一种分层手部图卷积时空网络(hierarchical hand graph convolutional temporal-spatial network, HHTS-Net). 具体来说，为了使网络轻量化，提出一种空间注意力与分层图卷积的组合策略，替代多层图卷积，既可以避免多层图卷积造成的计算负担，又能确保模型的全局特征提取能力；在图结构上，提出一种手部分层图卷积模块，根据手部特点定义图卷积，充分提取手部的层次化特征信息；在多流策略上，提出一种时频数据多分支流特征学习方案，通过傅里叶变换拓展特征数据流，丰富特征的多样性，提升模型的准确率和鲁棒性.

1.1 基于 GCN 的骨架特征提取

GCN 分为基于谱域的方法和基于空域的方法 2 大类. 基于谱域的方法利用基于拉普拉斯矩阵的傅里叶变换在频域上对特征进行图卷积操作. 而基于空域的方法直接聚合节点以及周围的邻居节点信息，并通过设计好的规则进行特征提取和归一化[17-18]. 由于基于空域的方法更加简单高效，因此是当前的主流方法.

基于空域方法中一般使用邻接矩阵表达图节

点之间的关系，节点之间的距离则采用跳数衡量. Yan 等[18]提出的 ST-GCN 就是典型的空域方法，它将人体骨架图的邻接矩阵应用于 GCN，并使用 GCN 提取帧内的空间信息. 虽然 ST-GCN 比 CNN/RNN 方法性能更优，但其图邻接矩阵在训练过程无法改变，难以关联距离较大的节点信息.

为了解决该问题，许多研究专注于寻找更合适的邻接矩阵构造方法. Shi 等[16]首次提出通过网络训练的方式获取邻接矩阵；在此基础上，CTR-GCN[19]实现了针对输入通道的图邻接矩阵. 考虑到关节点之间的依赖关系随时间而变化，Liu 等[20]提出的 TD-GCN 则采取逐帧计算邻接矩阵的策略. 沉重的计算负担已成为该方法的主要瓶颈.

Lee 等[21]认为，上述基于原始骨架图的 GCN 存在长程依赖问题，因此定义了多个衍生的骨架图结构，将原始骨架图中距离较远的节点直接关联起来. HHTS-Net 也借鉴该思想，但为了实现手势识别，本文重新定义了适合手部的分层骨架结构.

1.2　时空维度的信息融合

在骨架数据中节点位置随着时间不断变化. 因此，高效地提取空间和时间信息并融合对于手势识别具有重要意义.

目前主流方法是使用空间块和时间块分别提取空间信息和时间信息，并通过若干空间块和时间块堆叠融合时空信息，使用残差连接避免梯度消失[22]. 在 1.1 中提到的文献[16,18-21]使用以 GCN 为基础的空间块以及 CNN 为基础的时间块，其中 CNN 时间块可以聚合相邻多帧的信息. 鉴于 Transformer 自注意力机制[23]在自然语言处理等领域的成功应用，Song 等[24]提出在空间块与时间块之间添加一个注意力块，以此提取更加丰富的特征；Shi 等[25]使用时间注意力块与空间注意力块(spatial attention module, SA)提取时、空特征. 与 GCN 相比，SA 的优势是可以用较少的计算量提取单帧内的全局特征，但却丢失了重要的拓扑结构信息. 因此，HHTS-Net 将分层手部图卷积模块(hierarchical hand graph convolution module, HH)与 SA 相结合，提取空间信息同时发挥两者的优势.最后将 HH、SA 与 CNN 为基础的时间块堆叠融合时空信息.

1.3　多分支流策略

近年来，基于骨架序列的研究具有一个共同点，即使用多流数据融合策略. 在 ST-GCN 中，每个图节点存放对应关节点的三维或二维坐标，这种关节点的坐标被称为一阶信息. Shi 等[16]定义 2 个关节点之间的骨向量为二阶信息，并指出骨向量中的长度和方向对于动作识别具有更高的信息性和区分度. 因此，Shi 等融合关节流和骨向量流的预测向量，使性能进一步提升；Liu 等[20]在此基础之上将帧间关节点的移动信息用于手势识别. 融合关节流、骨向量流和关节移动流的预测结果后，取得了不错的性能提升；虽然数据流不断地丰富，但是这些数据流仅仅是不同关节点的坐标进行简单的相减运算得来，因此，经过复杂的模型之后，各数据流的预测结果将存在一定的同质性问题.

HHTS-Net 通过傅里叶变换将骨架序列数据从时域转到频域，并融合时、频特征进行预测，在一定程度上缓解了上述同质性问题.

2　HHTS-Net

2.1　HHTS-Net 框架

为了合理地利用手部图的拓扑结构信息，并以较少的计算量获取全局信息，本文提出一种 HHTS-Net. 如图 2 所示, HHTS-Net 由 n 个含 HH、时间卷积模块(temporal convolution module, TC)和 SA 的 HHTS 基础块组合而成，通过全连接(fully connection, FC)层输出各类预测分数. 其中, HH 和 SA 作用于空间维度, TC 则负责时间维度特征提取.

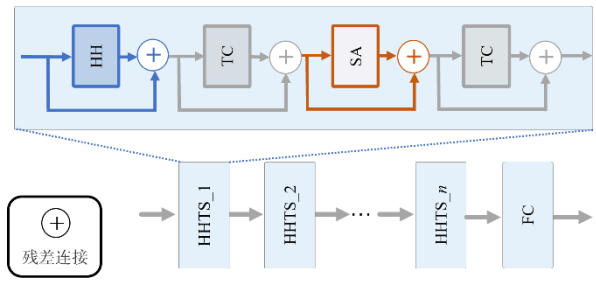


图 2　HHTS-Net 框架

2.2　HH

HH 的结构如图 3 所示，它由 1 个手部层次化分解图卷积块(hierarchically decomposed hand graph convolution, HDH-GC)和 1 个层注意力融合模块(attention-guided hierarchy aggregation, A-HA)组成. 其中, HDH-GC 通过多个图卷积高效地提取手部特征, A-HA 模块则用于确定不同图卷积的权重.

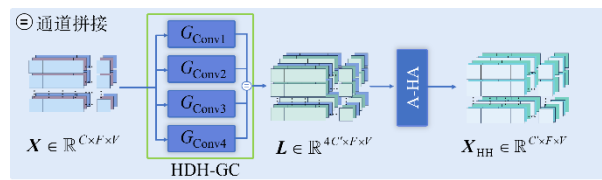


图 3　HH 结构

为了同时提取如图 4 所示的含 22 个节点的手部骨架中，生理结构上相邻及相距较远节点的特征，将 HDH-GC 拆分成 4 个图卷积块(即图 3 中的 $G_{\text{Conv1}}\sim G_{\text{Conv4}}$)，并设计对应的初始手部图 $L_1\sim L_4$，如图 5 所示. 可以看出，与图 4 所示的原始图结构不同，图 5 中的空心节点与其余节点是不连通的(如图中虚线边所示)，而斜线条纹节点与实心节点之间是全连接关系(如图中实线边所示). $L_1\sim L_4$ 的设计思路是：离掌心越远的节点，执行手势时变化的幅度一般越大. 因此，可以根据距离掌心节点的远近进行分层，且确保不同层之间具有一定的区分度，每层中既保留原来相邻的节点，又拉近了长程节点的距离. 通过学习，网络能够自行调整 $L_1\sim L_4$ 的结构.

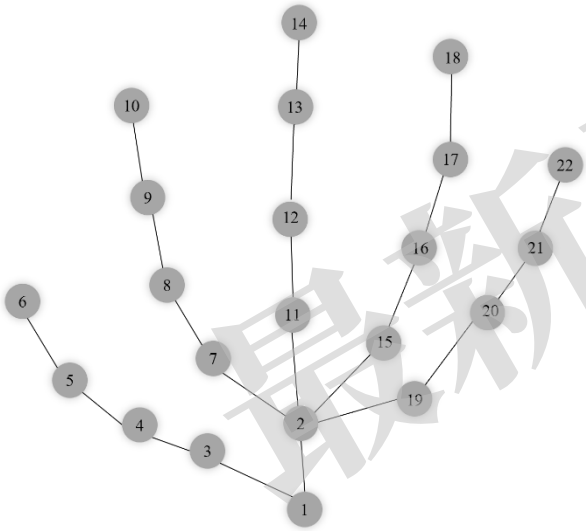


图 4 含 22 个节点的手部骨架节点编号

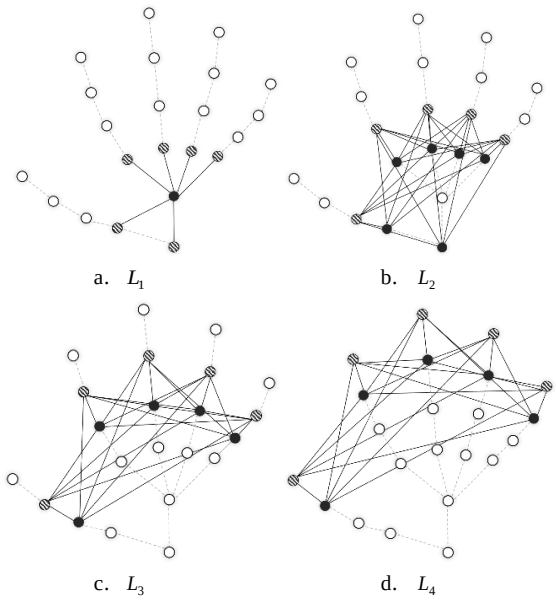


图 5 可视化手部分层图结构

图卷积块 $G_{\text{Convk}}\ (k=1,2,3,4)$ 的结构如图 6 所示. 首先，输入特征通过 4 个1×1卷积改变通道数后形成 4 流输出；然后对 $\boldsymbol{X}_1^k\sim\boldsymbol{X}_3^k$ 进行图卷积操作，公式为

$$\boldsymbol{X}_s^k=\theta_s^k\left(\boldsymbol{X}\right)$$

$$\boldsymbol{O}_s^k=\phi_s^k\left(\boldsymbol{X}_s^k\boldsymbol{A}_s^k\right).$$

其中，$\boldsymbol{X}\in\mathbb{R}^{C\times F\times V}$，$C$ 表示输入通道，F 表示帧数，V 表示关节点个数，θ_s^k 和 ϕ_s^k 均表示线性函数，$\boldsymbol{A}_s^k$ 表示图的邻接矩阵，s 分别取值 1, 2 和 3. 对第 4 个分支，执行图边卷积(edge convolution, Edge-Conv)操作. 为了基于每个卷积操作通道学习到不一样的特征，使用 3 个图卷积，而 Edge-Conv 操作可以找到语义相近的节点进行特征提取. 最后，按照

$$\boldsymbol{O}^k=\boldsymbol{O}_1^k\,\|\,\boldsymbol{O}_2^k\,\|\,\boldsymbol{O}_3^k\,\|\,\boldsymbol{E}^k,$$

将 $\boldsymbol{O}_1^k$，$\boldsymbol{O}_2^k$ 和 $\boldsymbol{O}_3^k$ 与 Edge-Conv 作得到的 $\boldsymbol{E}^k$ 进行拼接，得到最终的输出 $\boldsymbol{O}^k$. 其中，$\|$ 表示通道拼接操作.

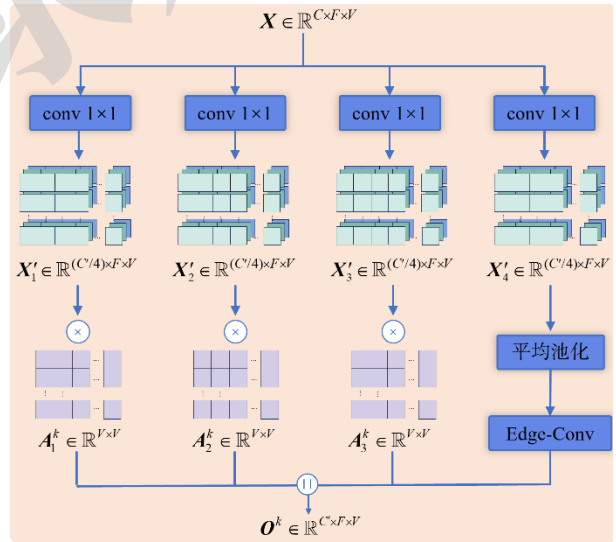


图6 第 k 层图卷积块 G_{Convk} 的结构

将 $\boldsymbol{O}^k$ 按照

$$\boldsymbol{O}=\boldsymbol{O}^1\,\|\,\boldsymbol{O}^2\,\|\,\boldsymbol{O}^3\,\|\,\boldsymbol{O}^4,$$

进行拼接，然后按照

$$\boldsymbol{X}_{\text{HH}}=Z(\boldsymbol{O}),$$

进行计算，即可得到 HH 块的最终输出 $\boldsymbol{X}_{\text{HH}}\in\mathbb{R}^{C'\times F\times V}$. 其中，$Z$ 表示 A-HA 模块[21]，其结构如图 7 所示. 首先将 $\boldsymbol{O}$ 沿着时间维度进行最大池化，然后通过 Edge-Conv 之后与 $\boldsymbol{O}$ 相乘，得到 $\boldsymbol{X}_{\text{HH}}$. 通过学习，A-HA 模块可以自行确定 $L_1\sim L_4$ 对于最终结果的贡献度.

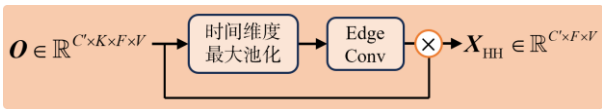


图7　A-HA模块结构

2.3　SA 和 TC

SA 和 TC 的作用分别为提取单帧内全局特征和聚合前后多帧特征. 其中, SA 参考 DSTA-Net[25] 中使用注意力机制提取全局特征的思想，它们的区别在于: DSTA-Net 在时、空 2 个维度上均使用注意力机制；而帧间手势特征关联的有效性大多局限于数帧范围内. 因此，HHTS-Net 仅在空间维度使用注意力，而把时间维度上的特征提取任务交给后续的 TC.

SA 的结构如图 8a 所示，首先将骨架特征 $\boldsymbol{X}'_{\mathrm{TC}}$ 输入到并行的 U 头注意力块，$U=3$ 为经验值，再通过一个线性层改变通道数并进行残差连接，之后依次通过激活函数 Leaky ReLU、线性层和批标准化(batch normalization, BN)，最后通过激活函数 Leaky ReLU 激活后进行残差连接后输出. SA 的公式化表示为

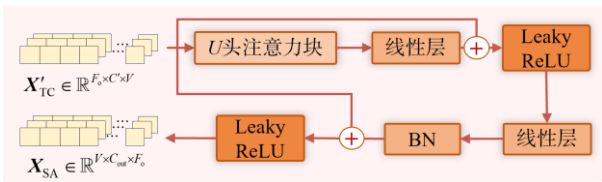
$$\boldsymbol{X}'_{\mathrm{SA}}=\mathrm{softmax}\left(\sum_{f}^{F}\left(\sigma\left(\boldsymbol{X}'_{f}\right)^{\mathrm{T}}\mu\left(\boldsymbol{X}'_{f}\right)\right)\right)\boldsymbol{X}'_{\mathrm{TC}},$$

$$\boldsymbol{X}''_{\mathrm{SA}}=\lambda_1\left(\boldsymbol{X}'_{\mathrm{SA}}\right)+\boldsymbol{X}'_{\mathrm{TC}},$$

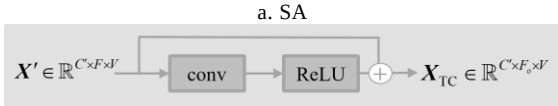
$$\boldsymbol{X}'''_{\mathrm{SA}}=\mathrm{BN}\left(\lambda_2\left(\alpha\left(\boldsymbol{X}''_{\mathrm{SA}}\right)\right)\right),$$

$$\boldsymbol{X}_{\mathrm{SA}}=\alpha\left(\boldsymbol{X}'''_{\mathrm{SA}}+\boldsymbol{X}'_{\mathrm{TC}}\right).$$

其中 μ 、σ 、λ_1 和 λ_2 均表示线性函数，α 表示 Leaky ReLU 激活函数, BN 表示 BN 操作.



a. SA



b. TC

图 8　SA 和 TC 的结构示意图

TC 的结构如图 8b 所示，首先将输入特征 $\boldsymbol{X}'$ 通过大小为 5×1 的卷积核，然后使用激活函数进行激活，最后进行残差连接. TC 的公式化表达为

$$\boldsymbol{X}_{\mathrm{TC}}=\mathrm{ReLU}\left(\mathrm{conv}\left(\boldsymbol{X}'\right)\right).$$

2.4　基于快速傅里叶变换的关节流及多流学习

如果将手势动作通过手部关键点的振动进行表征，振动的点产生波，那么就可以通过傅里叶变换在频域中对波的组成进行分析. 由于频域特征与原始输入特征具有较大的差异，因此可以解决多流学习中的特征同质性问题.

令 $\boldsymbol{j}$ 表示关节流数据，对其进行快速傅里叶变换(fast Fourier transform, FFT)，公式为

$$\mathrm{FFT}\left(\boldsymbol{j}\right)=\boldsymbol{X}_{\mathrm{Re}}+\mathrm{i}\boldsymbol{X}_{\mathrm{Im}}\tag{1}$$

其中，$\boldsymbol{X}_{\mathrm{Re}}$ 表示计算结果的实部，$\boldsymbol{X}_{\mathrm{Im}}$ 表示虚部. 本文提出一种基于 $\boldsymbol{j}$，$\boldsymbol{j}_{\mathrm{m}}$ 和 $\boldsymbol{j}_{\mathrm{r}}$ 的多流学习策略，其中，$\boldsymbol{j}_{\mathrm{m}}$ 表示关节运动流数据(其计算方式参考文献[20])，$\boldsymbol{j}_{\mathrm{r}}$ 即公式(1)中的 $\boldsymbol{X}_{\mathrm{Re}}$.

最终结果采用投票的方式实现，通过

$$\boldsymbol{Q}=\sum_{\boldsymbol{d}\in D}\alpha_m\boldsymbol{d},$$

构建一个预测向量 $\boldsymbol{Q}\in\mathbb{R}^n$. 其中，$\alpha_m$ 的取值范围是[0,1]，$D=\{\boldsymbol{j},\boldsymbol{j}_{\mathrm{m}},\boldsymbol{j}_{\mathrm{r}}\}$. $\boldsymbol{Q}$ 数值最大的分量即为对应的预测类别.

2.5　HHTS-Net 流程

图 9 所示为 HHTS-Net 的总体流程. 首先对原始骨架数据进行预处理，包括训练阶段的随机旋转、采样、归一化以及测试阶段的归一化等；然后基于 $\boldsymbol{j}$ 流数据计算得到 $\boldsymbol{j}_{\mathrm{m}}$ 流和 $\boldsymbol{j}_{\mathrm{r}}$ 流数据，分别使用它们训练不同的 HHTS-Net 模型(训练阶段)或者应用训练好的模型进行预测(测试阶段)；最终的预测结果由 3 个模型投票确定.

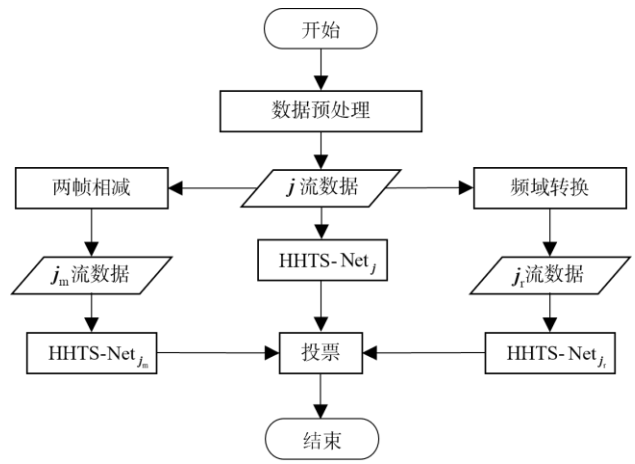


图 9　HHTS-Net 流程图

3　实验及结果分析

3.1　实验环境及评价指标

本文选择公共数据集 DHG[26]和 SHREC'17[27] 进行实验. DHG 是使用 Intel RealSense 相机以 30 帧/s 的速度采集的 2 800 个序列数据，共分为 14 个手势类别；每个序列的帧数从 20~50；序列按其手势、使用的手指数量和表演者标记. 20 个手势执

行者将每个单指手势执行 5 次，得到含 14 个类的子数据集 DHG-14；在此基础上，手势执行者再用整只手执行手势，得到含 28 个类别的子数据集 DHG-28. SHREC’17 数据集的采集设备和规格与 DHG 一致，同样分为 14 类和 28 类 2 个子集，不同之处在于 SHREC’17 数据集采集的手势执行者为 28 人.

在数据集划分方面，DHG 采用的策略是将前 19 个手势执行者被采集的数据作为训练集，最后 1 个手势执行者被采集的数据作为测试集；SHREC’17 并未给出划分方案，因此本文采用与文献[27]相同的策略. 最终，得到的训练集规模和测试集规模如表 1 所示. 在样本帧数方面，由于 DHG 和 SHREC’17 数据集的样本帧数不统一，为了实现批处理深度学习，通过采样算法分别将其规范至 50 和 180.

表 1　2 个数据集中的训练集和测试集大小

数据集	训练集	测试集
DHG	2 660	140
SHREC’17	1 960	840

本文实验使用一台配备 RTX 3090 24 GB 显卡的台式机运行. 网络模型训练采用随机梯度下降法(stochastic gradient descent, SGD)；批处理大小为 48；学习率开始采用热身策略[24]，在前 5 个训练轮次从 0.02 增长到 0.1，再使用余弦退火策略从 0.1 下降到 1×10^{-5}；权重衰退值为 5×10^{-4}.

从准确率和速度 2 方面对 HHTS-Net 的性能进行实验. 由于本文的目标是在精简 GCN 结构、降低系统复杂度的情况下提升手势识别的准确率，因此选择 3 个评价指标进行对比实验：准确率 P，指正确分类的样本数与样本总数的百分比，是衡量方法优劣的重要标准，在相关工作中被广泛采用；速度分为在训练集上每个 Epoch 的平均时间开销 T_{T} 和完成所有测试集样本测试的平均时间开销 T_{I}，它们是衡量算法时效性的重要依据.

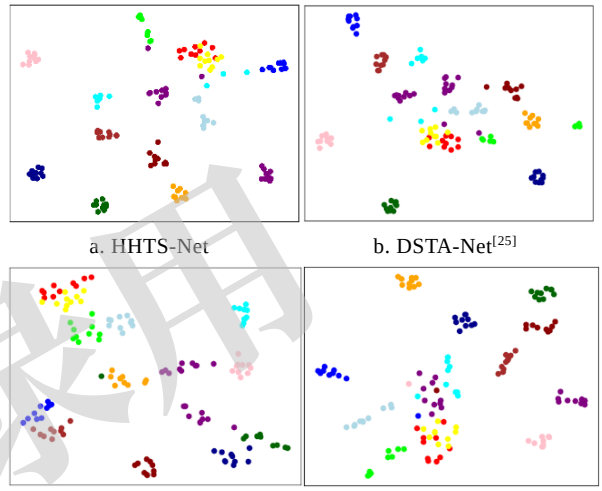
3.2　对比实验

本文采用 DSTA-Net[25]作为基线模型，将 HHTS-Net 与权威方法进行实验，对比方法的详细信息如表 2 所示.

表 2　对比方法详细信息

方法	出版平台	出版年	网络流个数
ST-GCN[18]	AAAI	2018	1
DSTA-Net[25]	ACCV	2020	4
NormalizedEdgeCN[28]	PR	2021	2
MS-ISTGCN[24]	TCSVT	2022	3
TD-GCN[20]	TMM	2023	3

(1) 在 DHG-14 数据集上，将 HHTS-Net, DSTA-Net[25], TD-GCN[20]和 ST-GCN[18]的分类效果进行实验，各模型在单流下对测试集所有样本使用 t-SNE 方法[29]降维到二维的可视化结果如图 10 所示. 可以看出，与 DSTA-Net[25]相比，HHTS-Net 各类分布更为均匀，有较大的类间差异. 与 TD-GCN[20]和 ST-GCN[18]相比，HHTS-Net 可以将相同类别的样本聚集得更紧密.



a. HHTS-Net　b. DSTA-Net[25]　c. TD-GCN[20]　d. ST-GCN[18]

图 10　DHG-14 上单流下不同模型输出特征的 t-SNE 图

(2) 在 DHG 数据集上，将 HHTS-Net 与各对比方法进行定量对比实验，结果如表 3 和表 4 所示. 其中，HHTS-Net 分为单一 $\boldsymbol{j}$ 流(简称为本文($\boldsymbol{j}$ Only))，$\boldsymbol{j}+\boldsymbol{j}_{\mathrm{r}}$ 双流(简称为本文($\boldsymbol{j}+\boldsymbol{j}_{\mathrm{r}}$))和最终的三流分支方案(简称为本文($\boldsymbol{j}+\boldsymbol{j}_{\mathrm{r}}+\boldsymbol{j}_{\mathrm{m}}$)). 可以看出，在时间开销上，无论是 T_{T} 还是 T_{I}，HHTS-Net 均为最优；在准确率方面同样如此，在 DHG-14 上，与 MS-ISTGCN[24], TD-GCN[20]和 DSTA-Net[25]相比，本文($\boldsymbol{j}+\boldsymbol{j}_{\mathrm{r}}+\boldsymbol{j}_{\mathrm{m}}$)的 P 别提升 2.1%, 1.9%和 2.0%.

表 3　在 DHG-14 上 6 种方法的结果对比

方法	T_{T} /s	T_{I} /s	P /%
ST-GCN[18]			85.6
DSTA-Net[25]	22.6	1.6	93.8
NormalizedEdgeCN[28]			92.9
MS-ISTGCN[24]			93.7
TD-GCN[20]	49.4	2.1	93.9
本文($\boldsymbol{j}$ Only)	4.8	**0.4**	94.3
本文($\boldsymbol{j}+\boldsymbol{j}_{\mathrm{r}}$)	9.6	0.7	95.0
本文($\boldsymbol{j}+\boldsymbol{j}_{\mathrm{r}}+\boldsymbol{j}_{\mathrm{m}}$)	**14.4**	1.1	**95.7**

注. 粗体表示最优值.

表 4　在 DHG-28 上 6 种方法的结果对比

方法	T_{T} /s	T_{I} /s	P /%
ST-GCN[18]			81.2

DSTA-Net[25]	22.6	1.6	90.9
NormalizedEdgeCN[28]			91.1
MS-ISTGCN[24]			91.2
TD-GCN[20]	49.4	2.1	91.4
本文($\boldsymbol{j}$ Only)	4.8	**0.4**	90.0
本文($\boldsymbol{j}+\boldsymbol{j}_r$)	9.6	0.7	90.0
本文($\boldsymbol{j}+\boldsymbol{j}_r+\boldsymbol{j}_m$)	**14.4**	1.1	**92.9**

注. 粗体表示最优值.

(3) SHREC'17 数据集上，将 HHTS-Net 与各对比方法进行定量对比实验，结果如表 5 和表 6 所示. 可以看出，HHTS-Net 的性能依然占优；尽管在 28 类手势分类任务上，HHTS-Net 的 P 不及 TD-GCN[20]和 MS-ISTGCN[24]，但在速度方面，本文($\boldsymbol{j}+\boldsymbol{j}_r+\boldsymbol{j}_m$)的训练与推理时间分别比 TD-GCN[20]缩短 57.8%和 58.9%；若将 HHTS-Net 中，提取帧间信息部分的一维卷积替换成与 TD-GCN[20]相同的多尺度时间卷积块，并将其命名为本文($\boldsymbol{j}+\boldsymbol{j}_r+\boldsymbol{j}_m$)$^+$，则其准确率可提升至与 TD-GCN[20]相同的水平，且时间开销仍然占优，如表 6 所示. 实验结果表明，从速度和准确率均衡性上看，HHTS-Net 的综合性能最优.

表 5　在 SHREC'17-14 上 6 种方法的结果对比

方法	T_T /s	T_I /s	P /%
ST-GCN[18]	—	—	92.7
DSTA-Net[25]	216.6	7.0	97.0
NormalizedEdgeCN[28]	—	—	94.8
MS-ISTGCN[24]	—	—	96.7
TD-GCN[20]	316.5	12.6	97.0
本文($\boldsymbol{j}$ Only)	44.4	**1.8**	96.9
本文($\boldsymbol{j}+\boldsymbol{j}_r$)	89.0	3.6	97.5
本文($\boldsymbol{j}+\boldsymbol{j}_r+\boldsymbol{j}_m$)	**133.4**	5.3	**97.6**

注. 粗体表示最优值.

表 6　在 SHREC'17-28 上 6 种方法的结果对比

方法	T_T /s	T_I /s	P /%
ST-GCN[18]			87.7
DSTA-Net[25]	217.7	8.0	93.9
NormalizedEdgeCN[28]			92.9
MS-ISTGCN[24]			94.9
TD-GCN[20]	316.1	12.9	**95.0**
本文($\boldsymbol{j}$ Only)	44.3	**1.7**	93.0
本文($\boldsymbol{j}+\boldsymbol{j}_r$)	88.6	3.5	93.5
本文($\boldsymbol{j}+\boldsymbol{j}_r+\boldsymbol{j}_m$)	**133.4**	5.3	93.9
本文($\boldsymbol{j}+\boldsymbol{j}_r+\boldsymbol{j}_m$)$^+$	256.4	6.7	**95.0**

注. 粗体表示最优值.

3.3　消融实验

HHTS 基础块由处理空间维度的 HH, SA 以及处理时间维度的 TC 组成，为了验证不同模块带来的性能增益，将基线模型 DSTA-Net[25]与上述模块组合得到的模型进行消融实验，对比方法名称及其模块组合方式如表 7 所示. 为了确保对比的公平性，仅使用单一关节流进行实验.

表 7　消融实验对比方法名称及其模块组成

方法名称	空间模块	时间模块
SAT	SA	TC
HHT	HH	TC
HHTS-Net	HH 及 SA	TC

在 4 个数据集上，4 种消融实验方法的准确率对比如图 11 所示. 可以看出，HHTS-Net 及删减相关模块后的准确率均高于基线模型 DSTA-Net[25]；由于数据存在差异，SAT 与 HHT 在不同数据集上的准确率表现不一；与 SAT 和 HHT 相比，结合 3 种模块的 HHTS-Net 的 P 均高 1.3 个百分点，体现了该方法的优越性.

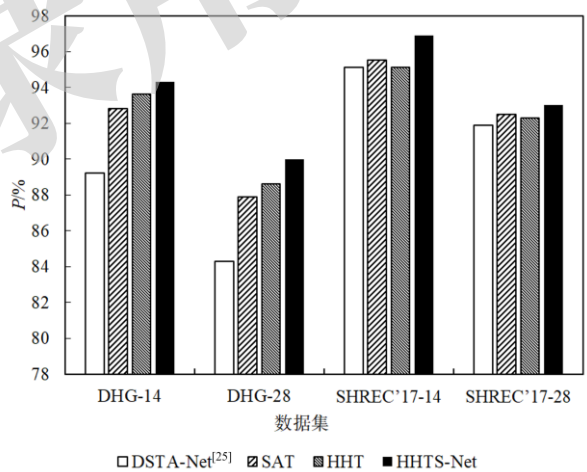


图 11　4 种消融实验方法的 P 对比

在 4 个数据集上，4 种消融实验方法的时间开销对比如图 12 和 13 所示. 可以看出，HHTS-Net 及其变体同样优于基线方法 DSTA-Net[25]；SA 的时间复杂度低于 HH，而结合 SA 与 HH 的 HHTS-Net 以牺牲少量时间成本为代价，换取了可观的准确率提升.

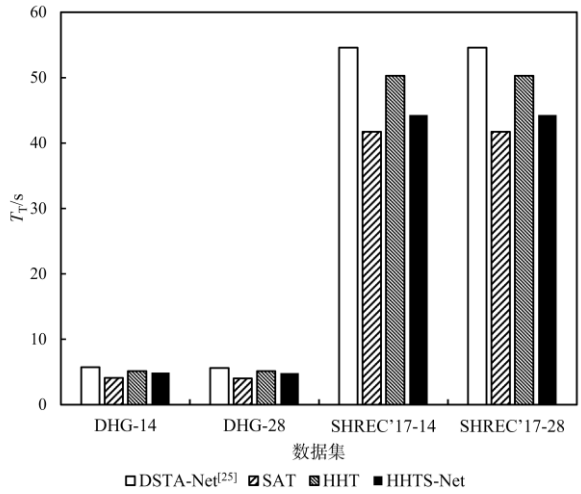


图 12　4 种消融实验方法的 T_T 对比

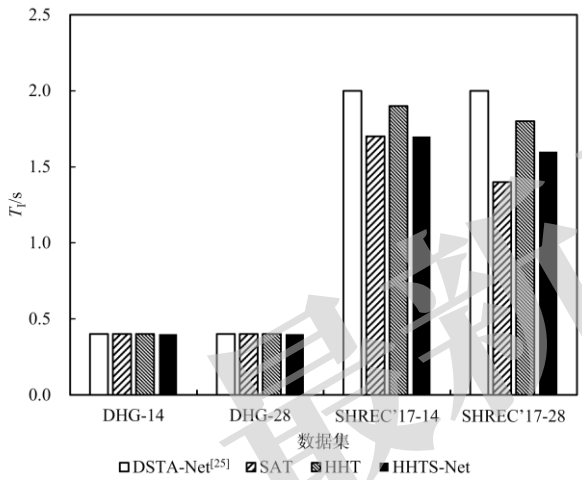


图 13　4 种消融实验方法的 T_I 对比

在 DHG-28 数据集上，4 种消融实验方法的训练进程对比如图 14 所示. 可以看出，HHTS-Net 很快就能进入收敛状态，且大部分时间的 P 都比其他方法更高.

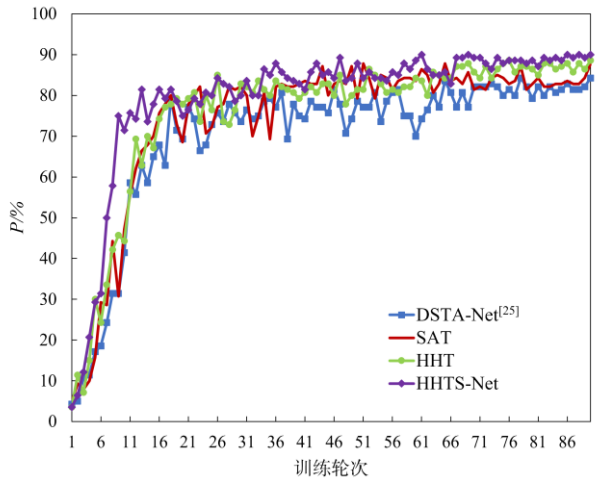


图 14　在 DHG-28 上 4 种方法的训练进程对比

在 DHG 和 SHREC'17 数据集上，HHTS 块的个数为 n，HHTS-Net 的准确率对比如图 15 所示. 可以看出，$n=4$ 时，2 个数据集 P 均取得了最优；在 DHG-14, SHREC-14 和 SHREC-28 上，当 n 取不同的值时性能相差不大. 因此, HHTS-Net 中使用 4 块 HHTS.

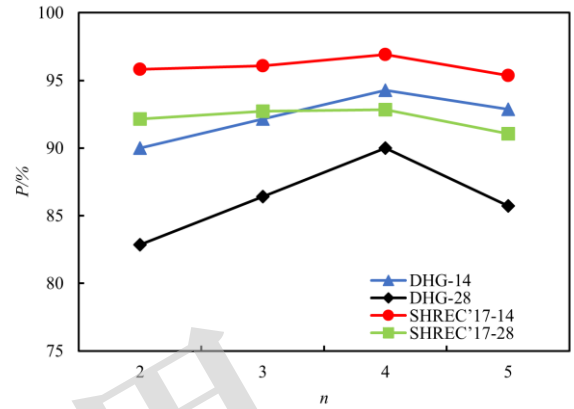


图 15　n 不同时 HHTS 块的 P 对比

在 SHREC'17 数据集上，分别使用 $\boldsymbol{j}$ 流、$\boldsymbol{j}_m$ 流和 $\boldsymbol{j}_r$ 流以及它们的组合进行训练和测试，实验结果如表 8 所示. 可以看出，在单流效果上，$\boldsymbol{j}_r$ 流在手势识别任务中与 $\boldsymbol{j}$ 流和 $\boldsymbol{j}_m$ 的效果相当；在融合多流的效果上，3 流融合的效果比单流或任意 2 流的效果都要好.

表 8　在 SHREC'17 上各分支流及其组合的 P 对比　%

方法	14 类	28 类
$\boldsymbol{j}$	96.9	92.9
$\boldsymbol{j}_m$	93.9	92.1
$\boldsymbol{j}_r$	96.9	90.7
$\boldsymbol{j}+\boldsymbol{j}_r$	97.5	93.5
$\boldsymbol{j}+\boldsymbol{j}_r+\boldsymbol{j}_m$	**97.6**	**93.9**

注. 粗体表示最优值.

4　结　语

本文提出一种基于时频数据融合的分层图卷积手势识别方法 HHTS-Net. 首先提出一种空间注意力与分层图卷积的组合机制，通过减少 GCN 层降低计算量，结合 SA 提升模型的全局特征提取能力；然后提出一种 HH，使得图卷积可以提取到手部特有的层次化骨架特征；最后提出一种时频数据多流特征学习方案，通过傅里叶变换将时域手部骨架序列数据转换到频域，使用融合的时频数据流进行预测，提高模型的准确率. 通过大量消融实验和对比实验，验证了 HHTS-Net 的有效性.

由于本文侧重于空间维度上的图特征提取，在时间维度上依然采用传统的卷积方式实现，因

此基于各帧之间相关性的手势识别将是未来的研究方向.

参考文献(References):

[1] Zhang Cheng, Hou Yibin, He Jian. Traffic police gestures recognition based on graph convolution with height layering partitioning strategy[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(7): 1037-1046(in Chinese)
(张丞，侯义斌，何坚. 高度分层分区的图卷积交警手势识别技术[J]. 计算机辅助设计与图形学学报，2022，34(7): 1037-1046)

[2] Han Chong, Han Lei, Sun Lijuan, *et al*. Millimeter wave radar gesture recognition algorithm based on spatio-temporal compression feature representation learning[J]. Journal of Electronics & Information Technology, 2022, 44(4): 1274-1283(in Chinese)
(韩崇，韩磊，孙力娟，等. 基于时空压缩特征表示学习的毫米波雷达手势识别算法[J]. 电子与信息学报，2022，44(4): 1274-1283)

[3] Wang Yong, Wang Shasha, Tian Zengshan, *et al*. Two-stream fusion neural network approach for hand gesture recognition based on FMCW radar[J]. Acta Electronica Sinica, 2019, 47(7): 1408-1415(in Chinese)
(王勇，王沙沙，田增山，等. 基于 FMCW 雷达的双流融合神经网络手势识别方法[J]. 电子学报，2019，47(7): 1408-1415)

[4] Fan Jingjing, Xue Haowei, Wu Xinhong, *et al*. Gesture recognition algorithm introducing ghost feature mapping and channel attention mechanism[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(3): 403-414(in Chinese)
(范晶晶，薛皓玮，吴欣鸿，等. 引入重影特征映射和通道注意力机制的手势识别算法[J]. 计算机辅助设计与图形学学报, 2022, 34(3): 403-414)

[5] Zhang Y F, Cao C Q, Cheng J, *et al*. EgoGesture: a new dataset and benchmark for egocentric hand gesture recognition[J]. IEEE Transactions on Multimedia, 2018, 20(5): 1038-1050

[6] Wan Huagen, Xiao Haiying, Zou Song. Hand gesture interaction for next-generation public games[J]. Journal of Computer-Aided Design & Computer Graphics, 2011, 23(7): 1159-1165(in Chinese)
(万华根，肖海英，邹松. 面向新一代大众游戏的手势交互技术[J]. 计算机辅助设计与图形学学报，2011，23(7): 1159-1165)

[7] Köpüklü O, Gunduz A, Kose N, *et al*. Real-time hand gesture detection and classification using convolutional neural networks[C] //Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1-8

[8] Strezoski G, Stojanovski D, Dimitrovski I, *et al*. Hand gesture recognition using deep convolutional neural networks[C] //Proceedings of the ICT Innovations 2016: Cognitive Functions and Next Generation ICT Systems. Heidelberg: Springer, 2018: 49-58

[9] Chen X H, Guo H K, Wang G J, *et al*. Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2017: 2881-2885

[10] Devineau G, Moutarde F, Xi W, *et al*. Deep learning for hand gesture recognition on skeletal data[C] //Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 106-113

[11] Tang Shoufeng, Zhou Nan, Zhao Renci, *et al*. Depth image hole repair algorithm based on Kinect camera[J]. Transducer and Microsystem Technologies, 2023, 42(3): 128-131(in Chinese)
(唐守锋，周楠，赵仁慈，等. 基于 Kinect 相机的深度图像空洞修复算法[J]. 传感器与微系统, 2023, 42(3): 128-131)

[12] Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 5686-5696

[13] Wang J D, Sun K, Cheng T H, *et al*. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349-3364

[14] Zou Z, Nie M X, Liu, X S, *et al*. Improved LDTW algorithm based on the alternating matrix and the evolutionary chain tree[J]. Sensors, 2022, 22(14): Article No.5305

[15] Lai K, Yanushkevich S N. CNN+RNN depth and skeleton based dynamic hand gesture recognition[C] //Proceedings of the 24th International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3451-3456

[16] 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. Los Alamitos: IEEE Computer Society Press, 2019: 12018-12027

[17] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs[C] //Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York: JMLR.org, 2016: 2014-2023

[18] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 7444-7452

[19] Chen Y X, Zhang Z Q, Yuan C F, *et al*. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 13339-13348

[20] Liu J F, Wang X S, Wang C, *et al*. Temporal decoupling graph convolutional network for skeleton-based gesture recognition[J]. IEEE Transactions on Multimedia, 2024, 26: 811-823

[21] Lee J, Lee M, Lee D, *et al*. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition[C] //Proceedings of the IEEE/CVF International Confer-ence on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 10410-10419

[22] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[23] 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: Curran Associates Inc., 2017: 6000-6010

[24] Song J H, Kong K, Kang S J. Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 6227-6239

[25] Shi L, Zhang Y F, Cheng J, *et al*. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition[C] //Proceedings of the 15th Asian Conference on Com-puter Vision. Heidelberg: Springer, 2020: 38-53

[26] Smedt Q D, Wannous H, Vandeborre J P. Skeleton-based dynamic hand gesture recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2016: 1206-1214

[27] Smedt Q D, Wannous H, Vandeborre J P, *et al*. 3D hand gesture recognition using a depth and skeletal dataset: SHREC'17 track[C] //Proceedings of the 10th Eurographics Workshop on 3D Object Retrieval. Aire-la-Ville: Eurographics Association Press, 2017: 33-38

[28] Guo F T, He Z X, Zhang S Y, *et al*. Normalized edge convolutional networks for skeleton-based hand gesture recognition[J]. Pattern Recognition, 2021, 118: Article No.108044

[29] Laurens V D M, Hinton G. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9(2605): 2579-2605