Journal of Image and Signal Processing 图像与信号处理, 2023, 12(2), 188-199
Published Online April 2023 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2023.122019

Hans汉斯

基于时空注意力深度增强差分图卷积的骨架行为识别

姬仲轩[1]，杨东进[2]，周美丽[1]，白宗文[1]

[1]延安大学物理与电子信息学院，陕西 延安
[2]陕西延长石油(集团)管道运输公司，陕西 延安

收稿日期：2023年4月4日；录用日期：2023年4月24日；发布日期：2023年4月29日

摘 要

时空卷积神经网络是行为识别的主流方法之一，但传统时空图卷积神经网络在空间特征聚合存在数据冗余与时间特征提取不充分的问题，针对该问题该文提出了一种时空注意力深度增强差分图卷积网络(ST-DEdGCN)模型。首先，在空间上通过深度增强差分图卷积(DEdGC)动态地学习不同通道中节点拓扑与节点梯度信息，有效地聚合不同通道中的关节特征。其次，通过时空卷积模块在时间维度上对全局时间信息进行建模，得到高效的序列特征信息。最后在NTU RGB + D 60和NTU RGB + D 120两个数据集进行了实验，实验结果表明时空注意力深度差分图卷积网络模型在空间特征的有效聚合和时空信息的有效提取方面优于当前主流方法，为行为识别及其相关研究提供了新的技术途径。

关键词

行为识别，深度卷积，时空特征，时空注意力机

A Skeleton-Based Action Recognition with Spatiotemporal Attention Depth Enhance Differential Graph Convolution

Zhongxuan Ji[1], Dongjin Yang[2], Meili Zhou[1], Zongwen Bai[1]

[1]School of Physics and Electronic Information, Yan'an University, Yan'an Shaanxi
[2]Shaannxi Yanchang Petroleum Group Pipeline Transportation Company, Yan'an Shaanxi

Received: Apr. 4th, 2023; accepted: Apr. 24th, 2023; published: Apr. 29th, 2023

Abstract

Spatiotemporal convolution neural network is one of the mainstream methods of action recogni-

文章引用: 姬仲轩, 杨东进, 周美丽, 白宗文. 基于时空注意力深度增强差分图卷积的骨架行为识别[J]. 图像与信号处理, 2023, 12(2): 188-199. DOI: 10.12677/jisp.2023.122019

tion, but the traditional spatiotemporal graph convolution neural network while having the problems of data redundancy and insufficient temporal feature extraction. To tackle the problem, a novel Spatio Temporal attention Depth Enhance difference Graph Convolution Network (ST-DEdGCN) model is proposed in this paper. Firstly, the Depth Enhance difference Graph Convolution (DEdGC) in space is proposed to dynamically learn joint topology and joint gradient information in different channels, and the joint features in different channels are effectively aggregated. Secondly, the Spatiotemporal Attention Temporal Convolution Network is proposed to model the global temporal joint information in time, and obtain efficient temporal feature information. Finally, the proposed algorithm is verified on the public skeleton action data sets NTU RGB + D 60 and NTU RGB + D 120. The results further verify the superiority to aggregate spatial features and to extract spatial-temporal information of this model, and provide a new technical approach for action recognition.

Keywords

Action Recognition, Depth-Wise Convolution, Spatiotemporal Feature, Spatiotemporal Attention

1. 引言

人体行为识别是近年来深度学习与视觉分析领域的研究热点之一，其任务是利用神经网络模型从视频中提取人体行为动作特征，并依据提取出的特征对动作进行分类，进而达到识别的目的。行为识别技术已广泛用于视频监控、运动分析、虚拟现实和机器人技术等领域。国内外学者对此开展了大量研究并提出了一系列高效的行为识别算法，按照网络模型输入的数据模态不同，可将行为识别分为基于 RGB 视频流和基于骨架两类方法。由于骨架方法只记录人体关节点的位置坐标，其具有数据量小、语义性高、不记录背景等无关信息、模型表达鲁棒性强等优点，而且随着人体姿态评估技术的发展，可以更方便地获取到人体骨架数据。因此，基于骨架的行为识别受到越来越多专家学者的关注。

目前，骨架行为识别方法主要分为 2 类：基于手工特征的方法和基于深度学习的方法。基于手工特征的方法通过关节数据之间的关系提取动作特征。Hussein 等人[1]将骨架关节位置的协方差矩阵作为序列的判别描述符，再通过传统的分类算法进行分类。Vemulapalli 等人[2]使用旋转和平移来建模身体部位之间的几何关系，并将这种关系映射到李群代数向量空间中作为动作特征。Weng 等人[3]受到朴素贝叶斯方法的启发，通过阶段中类间距离来对动作进行分类。然而，手工特征的方法存在难以提取深层特征与过度依赖数据集的问题，因此，深度学习方法开始替代手工特征方法。

由于循环神经网络(Recurrent Neural Network, RNN)和卷积神经网络(Convolutional Neural Network, CNN)强大的特征提取能力，[4] [5] [6] [7]使用 RNN 和 CNN 的方法对骨架数据进行建模，并且取得了不错的效果。但是，这些方法在将原始骨架数据转换成伪图像作为神经网络输入时会丢失骨架的原始结构信息。为了解决此问题，Yan 等人[8]首次使用图卷积(Graph Convolution Network, GCN)将骨架数据作为图进行建模，利用骨架数据中具有图拓扑关系的邻接矩阵提取空间特征，实现了性能的提升。但是由于特征聚合共享一个原始骨架固定的图拓扑，导致图卷积无法捕获原始拓扑之外关节的联系，Lei 等人[9]和 Li 等人[10]通过构建一个可学习的图拓扑矩阵，以数据驱动的方式来寻求合适的图拓扑。由于特征图

中不同通道代表不同类型的运动特征，并且不同运动特征下关节之间的相关性也不相同，上述方法对所有通道使用相同的图拓扑，这使得 GCN 在不同通道中聚合相似的运动特征，导致动作信息提取不充分，因此使用一个共享拓扑并非最佳选择。受到深度卷积的启发，Cheng 等人[11]为通道设置不同组，在每个组里对图拓扑进行学习。

基于上述，虽然国内外诸多学者在骨架行为识别上开展了大量研究并取得了优秀的研究成果，但仍存在着一些问题：1) 现有图卷积在特征聚合阶段存在部分的冗余，无法提取更精细的特征；2) 时空信息通过图卷积与时间卷积分别进行提取，但未考虑时间与其它信息之间的关联性；3) 缺少对特定时间下关节和通道的关注，无法聚焦重要的动作信息。

针对上述问题，本文提出了一个时空注意力深度增强差分图卷积网络模型。首先，受 Cheng 等人[11]和 Miao 等人[12]的启发，提出了一种深度差分图卷积网络，该网络采用深度图卷积和差分图卷积两种图卷积的形式构建空间卷积模块，能够在空间模型中更有效的提取特征；然后，通过动态通道增强模型构造时间维度自适应通道权重，来加强深度差分卷积对动作特征的建模能力；再者，通过多尺度的二维时空卷积模型与时空注意力增强模型对时空关系进行建模，以捕获更多的时空信息；最后，在 NTU RGB + D 骨架动作数据集上进行了实验。实验结果进一步验证了本文提出的行为识别模型的时空建模能力及良好的识别准确率。

2. 基于时空注意力深度增强差分图卷积的行为识别模型

骨架序列的空间特征与时间特征能够表述骨架序列中动作的完整信息，且两者之间存在着一些隐式的关联。本文提出时空注意力深度增强差分图卷积网络模型对动作信息与时空关联信息进行建模，总体框架如图 1 所示。

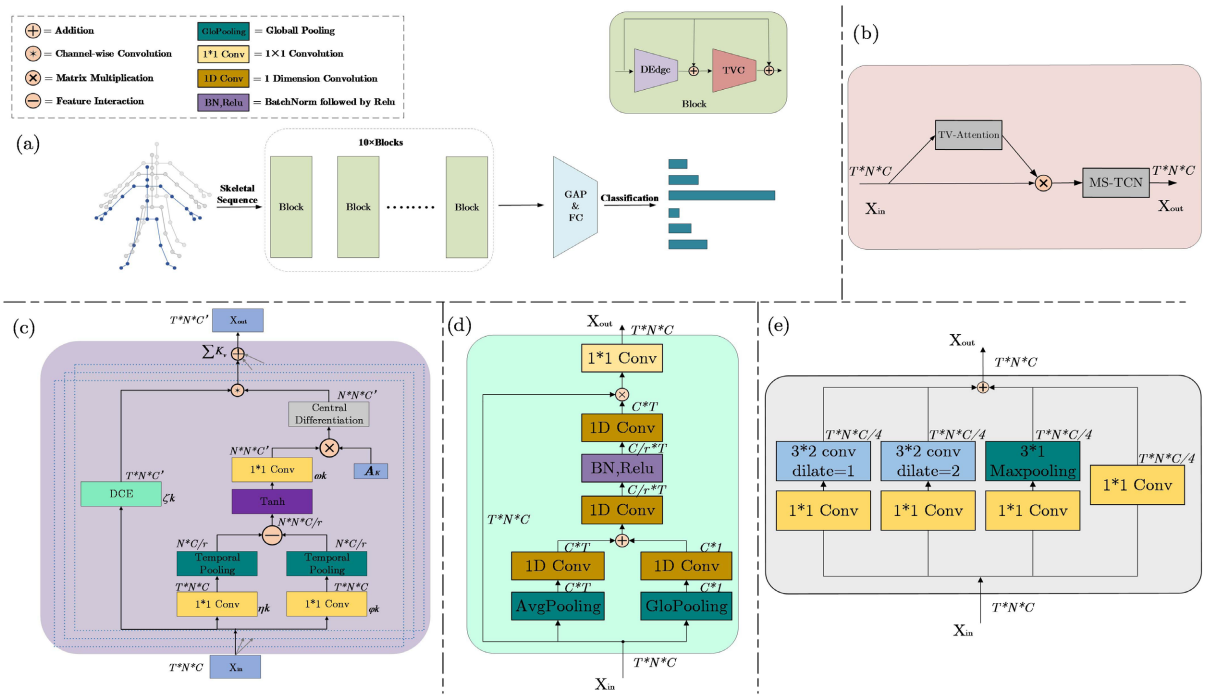


Figure 1. (a) ST-DEdGCN network structure; (b) Spatio-temporal attention enhance model; (c) Depth-enhanced differential graph convolution model; (d) Dynamic channel enhance model; (e) Multi-scalespatio-temporal convolution model

图 1. (a) 总体结构；(b) 时空注意力增强模型；(c) 深度增强差分图卷积模型；(d) 动态通道增强模型；(e) 多尺度时空卷积模型

该模型主要分为 2 部分：深度增强差分图卷积(紫色)和多尺度时空注意力增强卷积(红色)。首先骨架序列经过动态通道增强模块与深度差分图卷积得到特征图内不同通道的细粒度空间特征。然后在时空注意力增强卷积中，通过时空相关性注意力机制获取骨架序列中内在的时空注意力权重，实现针对特定时空的关注。最后经过多尺度时空卷积获取全局时空信息作为模块的输出。

2.1. 深度差分图卷积模型

现有方法通过注意力或其他机制自适应的学习人类骨架的拓扑结构，并对所有通道使用拓扑，这迫使 GCN 在不同通道中聚合具有相同拓扑的特征，既限制了特征提取的灵活性，又在部分特征聚合存在着冗余。本文提出深度差分图卷积(Depth differential Graph Convolution Network, Dd-GCN)，该方法通过构造通道级拓扑结构来探索不同通道上的运动特征，提高了模型特征提取的灵活性，并且利用更为细粒度的梯度信息对图卷积进行完善与补充，整体结构如图 1(a)所示。具体而言，该方法包含三部分：通道节点拓扑建模、中心差分化、通道整合。

通道节点拓扑建模通过构建每个通道各自的图拓扑结构来学习不同通道上的运动特征。首先通过两个核为(1 × 1)的二维卷积对输入特征 $\boldsymbol{X}_{in} \in R^{N\times C}$ 进行线性变换，然后该部分经过时间池化生成两个 $\boldsymbol{X} \in R^{N\times(C/r)}$ 作为特征交互时的输入。之后经过特征交互函数得到初步关系的图拓扑矩阵 $\overline{\boldsymbol{A}} \in R^{N\times N\times(C/r)}$。最后通过一个核为(1 × 1)的二维卷积与非线性激活函数生成最终的图拓扑矩阵 $\hat{\boldsymbol{A}} \in R^{N\times N\times C'}$，其表达式为：

$$\hat{\boldsymbol{A}} = \sigma\left(\mathrm{M}\left(\boldsymbol{X}_{in}\boldsymbol{W}_{\eta k}, \boldsymbol{X}_{in}\boldsymbol{W}_{\varphi k}\right)\boldsymbol{W}_{\omega k}\right) \tag{1}$$

其中，$\boldsymbol{W}_{\eta k}$ 和 $\boldsymbol{W}_{\varphi k} \in R^{C\times(C/r)}$ 是输入线性变换的权重矩阵，$\boldsymbol{W}_{\omega k} \in R^{C\times C'}$ 是输出权重矩阵，σ 为非线性激活函数，M 为特征交互[13]操作，目的是探索空间中各个节点之间的联系，为了更能直观的了解其原理，公式如下所示：

$$\mathrm{M} = \sigma\left(\psi\left(\boldsymbol{x}_i\right) - \xi\left(\boldsymbol{x}_j\right)\right) \tag{2}$$

$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 分别为输入线性变换后特征中 N 个节点特征之一。

由于中心差分卷积引入局部特征来增强模型识别能力，所以将中心差分化用来丰富特征信息，通过加入细粒度的梯度信息以完善原始的图卷积。该方法聚集采样点的中心梯度，只需要通过对邻接矩阵的第二维度求和来获得。因此中心差分化可以表达为：

$$\boldsymbol{Y} = \left(\boldsymbol{A}'\boldsymbol{X} - \tilde{\boldsymbol{A}} \odot \boldsymbol{X}\right)\boldsymbol{W} \tag{3}$$

其中，$\odot$ 表示元素级的哈达玛积，$\tilde{\boldsymbol{A}}$ 是通过对邻接矩阵 $\boldsymbol{A}'$ 对第二个维度求和之后($\in R^{N\times 1}$)拓展得到的($\in R^{N\times C}$)，公式如下：

$$\tilde{\boldsymbol{A}} = \left(\sum_j^N \boldsymbol{A}'_{ij}\right) * \boldsymbol{I} \tag{4}$$

$\boldsymbol{I} \in R^{1\times C}$ 向量里元素都为 1。为了提供稳健和多样化的建模能力，分配一个权重将图卷积和中心差分组合在一起，其表达式为：

$$\begin{aligned}\boldsymbol{Y} &= \gamma\cdot\left(\boldsymbol{A}\boldsymbol{X} - \tilde{\boldsymbol{A}} \odot \boldsymbol{X}\right)\boldsymbol{W} + (1-\gamma)\cdot\boldsymbol{A}\boldsymbol{X}\boldsymbol{W} \\ &= \left(\boldsymbol{A}\boldsymbol{X} - \gamma\cdot\tilde{\boldsymbol{A}} \odot \boldsymbol{X}\right)\boldsymbol{W}\end{aligned} \tag{5}$$

$\gamma \in [0,1]$ 控制分配的比重，γ 越高说明中心差分信息非常有用。

通道整合为每个通道分配了一个动态的邻接矩阵去学习每个通道中不同节点之间的关系。该动态邻

接矩阵($A_{:,:,i} \in R^{N\times N}$)是由通道拓扑建模部分所生成，每个邻接矩阵与相应特征($X_{:,:,i}^{W_\zeta} \in R^{N\times 1}$)进行特征聚合，其中下标 $i \in \{1,\cdots,C'\}$ 分别来自 i 通道。最后将各个通道上所生成的输出特征整合在一起输出最终输出 Y，其公式如下：

$$Y = A_{:,:,1}X_{:,:,1}^{W_\zeta} \,\|\, A_{:,:,2}X_{:,:,2}^{W_\zeta} \,\|\cdots\|\, A_{:,:,C'}X_{:,:,C'}^{W_\zeta} \tag{6}$$

$\boldsymbol{A} \in R^{N\times N\times C'}$ 为拓扑建模差分化后的邻接矩阵，$\|$ 表示拼接函数。

2.2. 动态通道增强模型

上一小节介绍了深度差分卷积模型，该模型首先学习到通道级的邻接拓扑矩阵，然后通过该矩阵聚合相应的动作特征。但是骨架序列是一个具有时间维度的数据，所以不同时间帧上的动作特征表达内容也不同，然而深度差分卷积模型在特征聚合时结合全局时间帧的所有通道统一进行特征聚合，没有考虑到不同时间帧间通道所聚合的运动特征的重要性。因此为了解决该问题，本小节提出了一个动态通道增强模型(Dynamic Channel Enhance Model, DCE)来增强深度差分卷积模型中的通道级运动特征聚合能力。该模型根据上下文动态校准每个时间帧的通道权重，通过该权重模型可以区分空间下不同通道的重要程度，并且该时间帧自适应权重也探索了一定的时间关系有助于时间维高效特征聚合。

因为动态通道增强模型(DCE)生成的权重随时间帧而进行改变，所以不仅要考虑当前帧，更重要的是要考虑该时间帧的上下文内容，因此通过学习上下文局部信息和全局信息生成当前帧的自适应权重，其模型结构如图 1(d)所示。此模块主要构成部分为全局特征提取、局部特征提取、特征加权。该模型通过对输入 $\boldsymbol{X} \in R^{T\times N\times C}$ 进行全局和局部平均池化来获得更高效率的时间帧描述向量，其公式如下：

$$\boldsymbol{V}_g = GloPooling\left(\boldsymbol{X}\right) \tag{7}$$

$$\boldsymbol{V}_a = AvgPooling\left(\boldsymbol{X}\right) \tag{8}$$

其中 $\boldsymbol{V}_g \in R^C$，$\boldsymbol{V}_a \in R^{C\times T}$。之后将全局和局部特征通过线性 1D 卷积后加权在一起，加权后的特征通过降维比为 r 的两层堆叠 1D 卷积来捕获不同时间的关联性，最终生成动态的时间通道权重。该权重能够表示不同时间帧下相应的通道权重，大大增强了深度差分卷积模型的动作特征聚合能力，总体公式如下：

$$\boldsymbol{Y} = Conv_2\left(Conv_1\left(\boldsymbol{V}_g \oplus \boldsymbol{V}_a\right) + \boldsymbol{d}_1\right) + \boldsymbol{d}_2 \tag{9}$$

其中输出 $\boldsymbol{Y} \in R^{T\times C}$，$\boldsymbol{d}$ 为偏重，$\oplus$ 为元素传播相加。

2.3. 时空注意力增强模型

在骨架行为识别领域中，许多现有方法只关注空间节点之间的联系，而未考虑时间上的重要性。一些方法提出时间注意力机制[14]突出重要的时间节点，但是忽视了不同时间节点之间存在的联系。由于3DCNN 体现了建模时空信息的优越性，本文提出时空注意力增强卷积对时空信息进行建模，丰富不同骨架帧之间的多样化表示，增强更具有辨识度的动作时空特征获取能力。该方法分为两部分：时空相关性注意力机制和多尺度时空卷积模型。

在骨架行为识别中，大部分动作信息可由少数关节表示，故聚焦特定关节能一定程度提升骨架行为的识别准确率。时空注意力增强模型(图 1(b))用来探索不同时间关键节点间的联系，图 2 给出了 TV-attention 方法的详细细节。在图 2 中，输入特征 $\boldsymbol{X}_{in} \in R^{N\times C\times T\times V}$，其中 N 为批次大小，C 为通道个数，T 为时间维度，V 为节点个数。针对输入特征通过简单的矩阵线性变换操作生成注意力热图 $\boldsymbol{X}_{out} \in R^{N\times 1\times T\times V}$。其公式如下：

$$\boldsymbol{X}_{out} = \text{sigmod}\left(Conv\left(\text{relu}\left(W\boldsymbol{X}_{in} + b\right)\right)\right) \tag{10}$$

其中 sigmod 和 relu 为非线性激活函数，*Conv* 是两层具有降维比的 1 × 1 卷积组成。该热图为不同时间节点分配权重，增强了不同动作特征之间的差异性，提高了模型的特征提取能力。

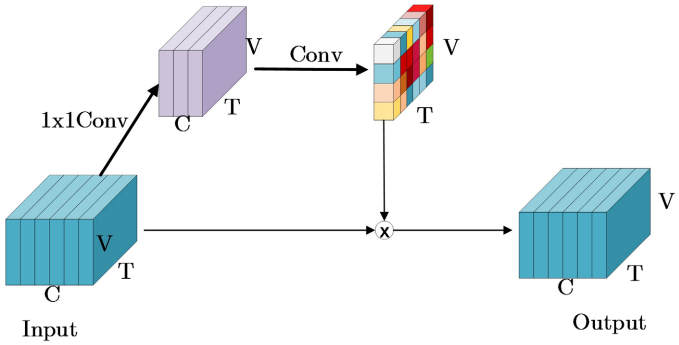


Figure 2. Schematic diagram of temporal attention mechanism
图 2. 时空相关性注意力机制示意图

在对骨架序列的时间信息建模中，许多现有的工作[8]使用固定内核大小 $K_t \times 1$ 的时间卷积来建模时间信息。作为多尺度空间聚合的自然拓展，许多工作[15] [16]通过多尺度的方法增强学习时间信息，但是这些方法仅仅在局部空间上进行时间建模，而没有考虑到全局信息的重要性。多尺度时空卷积模块通过二维的时间卷积，将内核大小固定为 3 × 2，并使用不同的扩展速率而不是更大的内核达到对更大空间视野的建模。采取瓶颈结构设计，降低额外分支带来的计算成本，并且加入剩余连接来丰富特征信息。其框图如图 1(e)所示。

3. 实验与分析

3.1. 实验数据集

NTU RGB + D. NTU RGB + D [17]是目前使用最广泛的动作识别数据集，包含了 56,880 个 3D 骨架数据。该数据集共有 60 类和 40 个受试者，每个样本包含一个动作，最多有两个受试者，并且是由三个 Microsoft Kinect v2 深度摄像头从不同视图同时捕获。该数据集推荐了两个基准：1) cross-subject (X-sub)：数据集分为训练集和测试集。训练集包含 40,320 个视频片段，验证集包含 16,560 个视频片段，两个子集中的受试者不同。训练数据来自 20 名受试者，测试数据来自其他受试者。2) cross-view (X-view)：该基准测试中的训练集包含 37,920 个视频片段，由摄像头在 0˚，45˚时拍摄，该验证集包含 18,960 个视频片段，这些视频片段由摄像头在−45˚时拍摄。

NTU RGB + D 120. NTU RGB + D 120 [18]是 NTU RGB + D 的扩充版本，涉及更多的主题和动作类别，更具有挑战性。该数据集包含 120 个动作类中的 114,480 个动作样本，由 106 名不同的受试者执行。该数据集推荐了两个基准：1) cross-subject (X-sub)：训练数据来自 53 名受试者，测试数据来自其他 53 名受试者。2) cross-setup (X-setup)：训练数据来自设置 ID 为偶数的样本，测试数据来自设置 ID 为奇数的样本。

3.2. 实验设置

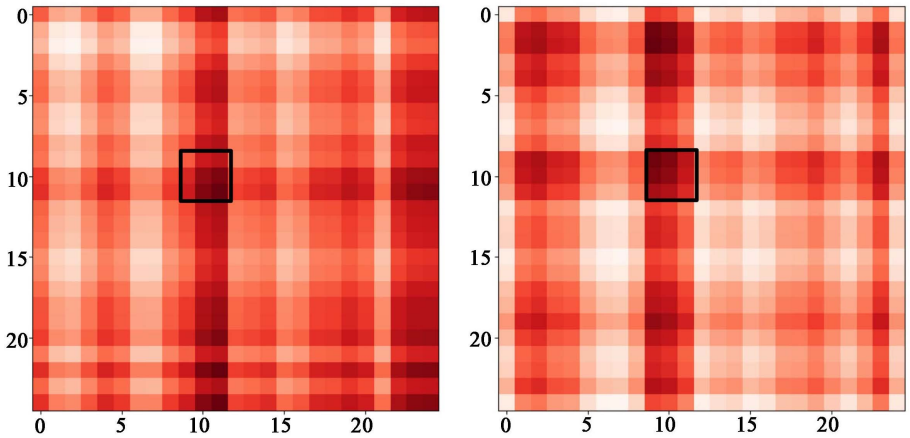
本文实验在 Ubuntu16.04 操作系统下进行，并采取 PyTorch 深度学习框架实现。所有实验均是在两张 Tesla T4 GPU 上进行。对于整个网络，使用 10 层时空注意力差分深度图卷积网络模块搭建模型，10 层的输出通道数分别为(64, 64, 64, 64, 128, 128, 128, 256, 256, 256)。使用动量(0.9)权重衰减(0.0004)的随机梯

度下降法(Stochastic Gradient Descent, SGD)对模型总共训练 70 代。利用交叉熵(cross entropy)损失，学习率设置为 0.1 并在第 30 和 50 代后以 0.1 比例衰减。对于 NTU 60/120 数据集，批量大小为 64。原始骨架序列被缩小到 64 帧的固定大小。数据预处理采用的策略与[19]中介绍的相同。

3.3. 实验结果与分析

本文的关键点主要在于：1) 使用深度差分图卷积提取高效的运动特征；2) 通过动态通道增强模块加强深度差分卷积动态时间特征聚合能力；3) 通过多尺度时空卷积获取骨架视频时空特征，并加入时空注意力机制增强对特定时空关节的关注。下面分别评估这两个模块对识别性能的影响，并将最终架构中的各个组件及其配置与现有的流行方法进行具体的对比分析。本文实验对比以 NTU RGB + D 60 数据集的 cross-subject 节点流(joints)为基准。

3.3.1. 深度差分图卷积验证



(a) 第 10 通道学习到的邻接拓扑矩阵　　(b) 第 50 通道学习到的邻接拓扑矩阵

Figure 3. Neighborhood topology matrix for different depth channel learning
图 3. 不同深度通道学习的邻接拓扑矩阵

带有深度差分的图卷积(Dd-GCN)相比于传统图卷积可以捕获到更高效的运动特征。因此为了验证 Dd-GCN 模块捕捉复杂空间特征的有效性，本小节首先对 Dd-GCN 模块进行可视化，其次与现阶段流行方法进行对比。如图 3 所示，给出了不同通道所学习到对应的自适应图拓扑矩阵的示例(第 10 和 50 通道)。该示例为一个“喝水”的动作，图中像素值越接近 0 (白色)表示关节之间的关系越弱。可以明显看出，每个通道所学习到的邻接矩阵是不同的，这表明本文的方法可以根据不同通道的特定运动特征学习相应的拓扑邻接矩阵。并且在所有通道中，与动作相关的一些联系很强的节点信息也都保存着，如黑框标出的部分，在动作“喝水”行为中，不同通道对左臂的注意十分密切。

表 1 给出了本文算法中深度差分图卷积网络(Dd-GCN)与适应图卷积(AGCN)，差分图卷积(CD-GCN)，深度图卷积(DGCN)，信道拓扑细化图卷积(CTR-GCN)的行为识别效果。本节使用各个组件逐步建立了验证模型，使用 AGCN [9]中的单流作为控制实验的基线，通过替换该方法空间特征提取层来进行更深层次的对比，其中所有模型的识别结果由本文自己运行得出。由表 1 可知：1) 相较于未使用通道级拓扑的图卷积模型，结合通道级拓扑增强结构的图卷积在实验准确率上具有优势；2) 对比其他通道级拓扑的图卷积模型，通过结合中心差分的方法，本文模型在(X-sub)的 Top-1 评价指标下识别准确率得到显著的提升，达到最高 90.2%的识别准确率，其也充分证明了深度差分图卷积对于获取高效动作特征的有效性。因此，

可以看出本文提出的方法不仅能够捕获到更高效的动作特征，且在空间特征提取部分识别准确率优于其他算法，表现出更好的性能。

Table 1. Comparison of spatial feature extraction ablation on NTU RGB + D 60
表 1. 在 NTU RGB + D 60 上的空间特征提取消融对比

Methods	data	Acc(X-sub)
AGCN	joint	86.3%
CD-GCN	joint	87.1%
DGCN	joint	87.2%
CTR-GCN	joint	89.3%
Dd-GCN	joint	**90.2%**

3.3.2. 动态通道增强模型验证

虽然深度差分模块的图卷积可以捕获到更高效的运动特征，但是该模型通过固定时间帧通道间的关系捕获运动特征，忽视了不同时间下运动特征代表不同的行为。因此为了验证动态通道增强模块(DCE)对深度差分卷积的加强性，列出了该模块与现阶段流行方法进行对比。

Table 2. Comparison of DCE module ablation on NTU RGB + D 60
表 2. 在 NTU RGB + D 60 上的 DCE 模块消融对比

Methods	data	Acc(X-sub)
DGCN	joint	87.2%
DGCN + DCE	joint	**88.3%**
CTR-GCN	joint	89.3%
CTR-GCN + DCE	joint	**90.0%**

由于 DCE 模块是对动态通道的探索，所以表中都是在深度图卷积方法的基础上进行对比。从表 2 中可知：相较于未使用通道增强结构的深度图卷积模型，结合了动态通道增强的深度图卷积模型在 joint 流指标下识别率均为最优。其中相较于 DGCN 方法与 CTRGCN 方法，基于动态通道增强模型在 X-sub 下分别提高了 1.1%与 0.7%。实验结果进一步验证了动态通道增强模块对深度图卷积方法的增强性。

3.3.3. 时空注意力增强模型验证

时空注意力增强卷积模块通过多尺度时空卷积获取空间与时间之间的关联信息，并加入时空注意力机制增强对特定时空关节的关注。因此为了验证 ST-ATTCN 模块对时空信息建模的有效性，本小节首先对时空注意力增强模块进行可视化，其次对现阶段时间特征获取方法上进行了消融实验。如图 4 所示，该图为“喝水”动作的时空注意力部分层可视化结果图。图中横坐标表示的是节点，纵坐标表示的是时间，像素值越接近 0 (深蓝色)表示关系越弱。并且可以清楚的看到在 9~11 帧里红框标出部分，左右胳膊节点和身体的部分节点之间存在更紧密的联系(0~5 帧经分析属于基础行为部分，没有出现特定强相关动作特征信息)。

多尺度时空卷积模型来对时间域进行特征提取和特征聚合，仍然把 AGCN 作为本次控制实验的基线。首先，使用传统的多尺度时间卷积网络(Multiple Temporal Convolutional Network, MS-TCN)替代传统的

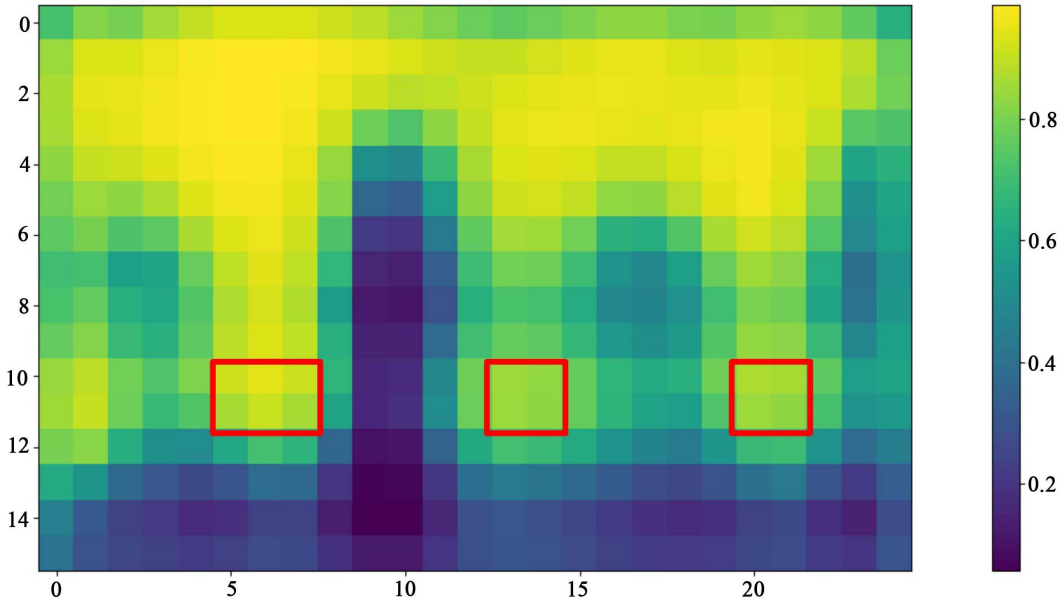


Figure 4. Visualization of temporal attention module
图 4. 时空注意力模块可视化

时间卷积网络(Temporal Convolutional Networks, TCN)，该模型受到 GoogleNet 思路的影响，把传统的时间特征提取增广成了多尺度的形式，从而丰富了特征的表达。如表 3 所示，把 MS-TCN 代替基线中的 TCN，相比于基线准确性提升了 0.5%，之后加入本文提出的时空注意力增强卷积(Spatiotemporal Attention Temporal Convolution Network, ST-ATTCN)，在基线上准确性提高了 0.8%，比 MS-TCN 准确性高出了 0.3%。可以看出，在时间特征提取部分本文提出的 ST-ATTCN 捕获了不同时间不同节点之间的联系，提高了相似动作特征之间的差异性，与其他方法相比表现出更好的性能。

Table 3. Comparison of temporal feature extraction ablation on NTU RGB + D 60
表 3. 在 NTU RGB + D 60 上的时间特征提取消融对比

Methods	data	Acc(X-sub)
Baseline AGCN	joint	86.3%
Baseline + MS-TCN	joint	86.8%
Baseline + ST-ATTCN	joint	**87.1%**

3.3.4. 与主流网络对比结果

为了验证本文提出模型的性能，本文将自己提出的模型在 NTU RGB + D 60 和 NTU RGB + D 120 数据集上与其他前沿方法进行对比，对比结果如表 4 和表 5 中所示。本文模型在 NTU RGB + D 60 数据集的 X-Sub 和 X-View 两种评判标准下准确率分别达到了 92.6%和 96.7%；在 NTU RGB + D 120 数据集的 X-Sub 和 X-View 两种评判标准下准确率分别达到了 89.1%和 90.4%。在两个数据集上，本文方法明显优于基于 GCN 的基准方法[8]，与其他优秀的前沿方法相比同样有比较强的竞争力。

综上实验结果表明：基于时空注意力深度增强差分图卷积的行为识别模型相较于现阶段图卷积的行为识别方法，即实现了骨架序列中动作信息的高效提取与对时空关联信息和时空特定关节的注意力增强，又具有不错的识别准确率与泛化能力。

Table 4. Comparison of recognition accuracy of different action recognition algorithms on NTU RGB + D 60 dataset
表 4. 不同行为识别算法在 NTU RGB + D 60 数据集上识别准确率对比

Methods	Year	NTU RGB + D 60	
		X-Sub (%)	X-View (%)
ARRN-LSTM [20]	2016	80.7	88.8
Ind-RNN [21]	2018	81.8	88.0
HCN [22]	2018	86.5	91.1
ST-GCN [8]	2018	81.5	88.3
AS-GCN [10]	2019	86.8	94.2
2S-AGCN [9]	2019	88.5	95.1
AGC-LSTM [23]	2019	89.2	95.0
Shift-GCN [24]	2020	90.7	96.5
DC-GCN + ADG [11]	2020	90.8	96.6
MS-G3D [15]	2020	91.5	96.2
CD-GCN [12]	2021	90.9	96.5
CTR-GCN [13]	2021	92.2	96.6
Ta-CNN [27]	2022	90.7	95.1
ST-DEdGCN (ours)	-	**92.6**	**96.7**

Table 5. Comparison of recognition accuracy of different action recognition algorithms on NTU RGB + D 120 dataset
表 5. 不同行为识别算法在 NTU RGB + D 120 数据集上识别准确率对比

Methods	Year	NTU RGB + D 120	
		X-Sub (%)	X-View (%)
ST-LSTM [5]	2016	55.7	57.9
GCN-LSTM [25]	2017	61.2	63.3
RotClips + MTCNN [26]	2017	62.2	61.8
ST-GCN [8]	2018	72.4	71.3
2S-AGCN [9]	2019	82.9	84.9
Shift-GCN [24]	2020	85.9	87.6
DC-GCN+ADG [11]	2020	86.5	88.1
MS-G3D [15]	2020	86.9	88.4
CTR-GCN [13]	2021	88.9	90.4
CD-GCN [12]	2021	86.3	87.8
Ta-CNN [27]	2022	85.7	87.3
ST-DEdGCN (ours)	-	**89.3**	**90.6**

4. 结论

本文提出了一种新的基于骨架的时空注意力深度增强差分图卷积神经网络。通过分析现有模型的不

足，提出了深度增强差分模块和时空注意力模块来加强相关性建模能力。为了验证所提出方法的有效性，分别从可视化结果、网络精度提升等方面进行实验验证。通过在骨架动作数据集上与其他主流方法进行比较，实验结果再次证明了改进后的模块有效性。本文方法在处理骨架序列中相似轨迹的数据时，无法提取表征能力较强的特征向量，这也是今后需要深入研究的地方。

参考文献

[1] Hussein, M.E., Torki, M., Gowayyed, M.A., *et al*. (2013) Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3d Joint Locations. *Proceedings of the* 23*rd International Joint Conference on Artificial Intelligence*, Beijing, 3-9 August 2013, 2466-2472.

[2] Vemulapalli, R., Arrate, F. and Chellappa, R. (2014) Human Action Recognition by Representing 3d Skeletons as Points in a Lie Group. 2014 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Columbus, 23-28 June 2014, 588-595. https://doi.org/10.1109/CVPR.2014.82

[3] Weng, J.W., Weng, C.Q. and Yuan, J.S. (2017) Spatio-Temporal Naive-Bayes Nearest-Neighbor (st-nbnn) for Skeleton-Based Action Recognition. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 4171-4180. https://doi.org/10.1109/CVPR.2017.55

[4] Du, Y., Wang, W. and Wang, L. (2015) Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1110-1118. https://doi.org/10.1109/CVPR.2015.7298714

[5] Liu, J., Shahroudy, A., Xu, D., *et al*. (2016) Spatio-Temporal LSTM with Trust Gates for 3d Human Action Recognition. 2016 14*th European Conference on Computer Vision* (*ECCV*), Amsterdam, 11-14 October 2016, 816-833. https://doi.org/10.1007/978-3-319-46487-9_50

[6] Li, C.K., Wang, P.C., Wang, S., *et al*. (2017) Skeleton-Based Action Recognition Using LSTM and CNN. 2017 *International Conference on Multimedia & Expo Workshops* (*ICMEW*), Hong Kong, 10-14 July 2017, 585-590. https://doi.org/10.1109/ICMEW.2017.8026287

[7] Li, C., Zhong, Q.Y., Xie, D., *et al*. (2017) Skeleton-Based Action Recognition with Convolutional Neural Networks. 2017 *IEEE International Conference on Multimedia & Expo Workshops* (*ICMEW*), Hong Kong, 10-14 July 2017, 597-600. https://doi.org/10.1109/ICMEW.2017.8026285

[8] Yan, S.J., Xiong, Y.J. and Lin, D.H. (2018) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. 2018 30*th Innovative Applications of Artificial Intelligence* (*IAAI*-18), New Orleans, 2-7 February 2018, 2-7.

[9] Shi, L., Zhang, Y.F., Cheng, J., *et al*. (2019) Two-Stream Adaptive Graph Convolutional Networks for Skeleton Based Action Recognition. 2019 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 12026-12035. https://doi.org/10.1109/CVPR.2019.01230

[10] Li, M.S., Chen, S.H., Chen, X., *et al*. (2019) Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. 2019 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 16-20. https://doi.org/10.1109/CVPR.2019.00371

[11] Cheng, K., Zhang, Y.F., Cao, C.Q., *et al*. (2020) Decoupling GCN with Drop Graph Module for Skeleton-Based Action Recognition. 2020 *European Conference on Computer Vision* (*ECCV*), Glasgow, 23-28 August 2020, 536-553. https://doi.org/10.1007/978-3-030-58586-0_32

[12] Miao, S.Y., Hou, Y.H., Gao, Z.M., *et al*. (2021) A Central Difference Graph Convolutional Operator for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, **32**, 4893-4899. https://arxiv.org/abs/2111.06995

[13] Chen, Y.X., Zhang, Z.Q., Yuan, C.F., *et al*. (2021) Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. 2021 *International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 10-17. https://doi.org/10.1109/ICCV48922.2021.01311

[14] Bai, R.W., Li, M., Meng, B., *et al*. (2021) Hierarchical Graph Convolutional Skeleton Transformer for Action Recognition. 2022 *IEEE International Conference on Multimedia and Expo* (*ICME*), Taipei, 18-22 July 2022, 1-6. https://doi.org/10.1109/ICME52920.2022.9859781

[15] Liu, Z.Y., Zhang, H.W., Chen, Z., *et al*. (2020) Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 143-152. https://doi.org/10.1109/CVPR42600.2020.00022

[16] Gao, X., Hu, W., Tang, J.Y., *et al*. (2019) Optimized Skeleton-Based Action Recognition via Sparsified Graph Regression. 27*th ACM International Conference on Multimedia*, Nice, 21-25 October 2019, 601-610.

https://doi.org/10.1145/3343031.3351170

[17] Shahroudy, A., Liu, J., Ng, T.T., *et al*. (2016) NTU RGB+D: A Large Scale Dataset for 3d Human Activity Analysis. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 27-30. https://doi.org/10.1109/CVPR.2016.115

[18] Liu, J., Shahroudy, A., Perez, M.L., *et al*. (2019) NTU RGB+D 120: A Large-Scale Benchmark for 3d Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2684-2701. https://doi.org/10.1109/TPAMI.2019.2916873

[19] Zhang, P.F., Lan, C.L., Zeng, W.J., *et al*. (2020) Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 1112-1121. https://doi.org/10.1109/CVPR42600.2020.00119

[20] Li, L., Zheng, W., Zhang, Z.X., *et al*. (2018) Skeleton-Based Relational Modeling for Action Recognition. 2019 *IEEE International Conference on Multimedia and Expo* (*ICME*), Shanghai, 8-12 July 2019. http://arxiv.org/abs/1805.02556

[21] Li, S., Li, W.Q., *et al*. (2018) Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN. 2018 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Salt Lake City, 18-23 June 2018, 18-22. https://doi.org/10.1109/CVPR.2018.00572

[22] Li, C., Zhong, Q.Y., Xie, D., *et al*. (2018) Cooccurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. 2018 27*th International Joint Conference on Artificial Intelligence* (*IJCAI*), Stockholm, 13-19 July 2018, 786-792. https://doi.org/10.24963/ijcai.2018/109

[23] Si, C.Y., Chen, W.T., Wang, W., *et al*. (2019) An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. 2019 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 1227-1236. https://doi.org/10.1109/CVPR.2019.00132

[24] Cheng, K., Zhang, Y.F., He, X., *et al*. (2020) Skeleton-Based Action Recognition with Shift Graph Convolutional Network. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 183-192. https://doi.org/10.1109/CVPR42600.2020.00026

[25] Liu, J., Wang, G., Duan, L.Y., *et al*. (2017) Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks. *IEEE Transactions on Image Processing*, **27**, 1586-1599. https://doi.org/10.1109/TIP.2017.2785279

[26] Ke, Q.H., Bennamoun, M., An, S.J., *et al*. (2018) Learning Clip Representations for Skeleton-Based 3d Action Recognition. *IEEE Transactions on Image Processing*, **27**, 2842-2855. https://doi.org/10.1109/TIP.2018.2812099

[27] Xu, K.L., Ye, F.F., Zhong, Q.Y., *et al*. (2022) Topology-Aware Convolutional Neural Network for Efficient Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 2866-2874. https://doi.org/10.1609/aaai.v36i3.20191