图 学 学 报
JOURNAL OF GRAPHICS

基于全局时空编码网络的猴类动物行为识别

孙　峥[1,2]，　张素才[3]，马喜波[1,2]

(1. 中国科学院自动化研究所，北京 100190；2. 中国科学院大学人工智能学院，北京 100049；
3. 北京昭衍新药研究中心股份有限公司，北京，100176)

摘　　要：猴类动物行为的准确量化是临床前药物安全评价的一个基本目标。视频中猴类动物行为分析的一个重要路径是使用目标的骨架序列信息，然而现有的大部分骨架行为识别方法通常在时间和空间维度分别提取骨架序列的特征，忽略了骨架拓扑结构在时空维度的整体性。针对该问题，提出了一种基于全局时空编码网络(GSTEN)的骨架行为识别方法。该方法在时空图卷积网络(ST-GCN)的基础上，并行插入全局标志生成器(GTG)和全局时空编码器(GSTE)来提取时间和空间维度的全局特征。为了验证提出的全局时空编码网络的性能，在自建的猴类动物行为识别数据集上开展实验。实验结果表明，提出的全局时空编码网络在基本不增加模型参数量的情况下，准确率(Accuracy)指标达到 76.54%，相较于基准模型时空图卷积网络提升 6.79%。

关　键　词：行为识别；骨架序列；全局时空编码网络；猴类动物；药物安全评价

中图分类号：TP 391　　**DOI：**10.11996/JG.j.2095-302X.0000000000

文献标识码：A　　**文 章 编 号：**2095-302X(0000)00-0000-00

Monkey Action Recognition Based on Global Spatiotemporal Encode Network

SUN Zheng[1,2], ZHANG Su-cai[3], MA Xi-bo[1,2]

(1. CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China;
2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China;
3. JOINN Laboratories (Beijing) Co., Ltd., Beijing 100176, China)

Abstract: Accurate quantification of caged monkey behaviors is a primary goal in the preclinical drug safety assessment. Skeleton information is important to analyze the behaviors of monkey. However, most of the existing skeleton-based action recognition methods usually extract the features of the skeleton sequence in the spatial and temporal dimensions, ignoring the integrity of the skeleton topology. To address this problem, we propose a skeleton action recognition method based on Global Spatiotemporal Encode Network (GSTEN). Based on the Spatial Temporal Graph Convolutional Network (ST-GCN), the proposed method inserts Global Token Generator (GTG) and several Global Spatiotemporal Encoders (GSTE) in parallel to extract the global features in spatiotemporal dimension. To verify the performance of the proposed method, we conduct experiments on a self-built monkey action recognition dataset. The experimental results show that the proposed GSTEN achieves an accuracy of 76.54% without increasing the amount of model parameters, which is 6.79% improvement to the baseline model ST-CGN.

收稿日期：2022-xx-xx；定稿日期：2022-xx-xx

Received：xx xx, 2022；**Finalized**：xx xx, 2022

基金项目：国家自然科学基金资助项目(#82090051，#81871442)；中国科学院青年创新促进会 (#Y201930)

Foundation items: the Chinese National Natural Science Foundation Projects (#82890051, #81871442); the Youth Innovation Promotion Association CAS (#Y201930)

第一作者：孙峥(1996–)，男，硕士研究生。主要研究方向为姿态估计和行为识别等。E-mail：zheng.sun@nlpr.ia.ac.cn

First author：Sun Zheng (1996−), master student. His main research interests cover pose estimation and action recognition, etc. E-mail: zheng.sun@nlpr.ia.ac.cn

通信作者：马喜波(1981–)，女，研究员，博士。主要研究方向为多模态融合的医学成像方法及设备开发等。E-mail：xibo.ma@nlpr.ia.ac.cn

Corresponding author：Ma Xi-bo (1981−), researcher, Ph.D. Her main research interests cover development of multi-modal fusion medical imaging methods and equipment, etc. E-mail：xibo.ma@nlpr.ia.ac.cn

Keywords: action recognition; skeleton sequence; global spatiotemporal encode network; monkey; drug safety assessment

在临床前药物安全评价中，猴类动物在用药前后的行为变化是必不可少的观测指标[1-3]。长时间的人为观察在成本和随机性方面都有不可忽视的缺陷。因此需要研发可行的人工智能方法对猴类动物表现出来的与药物安全评价相关的行为进行实时、定量分析。目前针对人类的行为识别方法已经得到了广泛发展，然而在猴类动物上相关方法的研发却发展缓慢。因此，使用人工智能方法自动识别猴类动物的行为对临床前药物安全评价具有重要的现实意义和应用前景。

近些年一些学者使用人工智能方法进行了动物行为识别任务的研究[4,5]，这些方法在各自的动物数据集上达到了较高的性能指标，但是临床前药物安全评价场景下猴类动物行为识别任务仍然有一些特有问题亟待解决。在该类任务中，猴类动物所处的场景单一，背景扰动、光照变化以及目标外观差异较小，导致连续的视频帧和光流图中会包含冗余信息。此外，猴类动物的行为识别需要充分考虑动作在时空维度的整体性。针对这些问题，本文使用猴类动物的骨架序列信息进行行为识别，并提出基于全局时空编码网络 (Global Spatiotemporal Encode Network, GSTEN)的骨架行为识别方法。该方法本质上是使用训练好的姿态估计模型对一段视频中的每一帧进行关键点的识别，再基于上述关键点形成的骨架序列信息进行行为识别，其中骨架序列信息包括每一帧中目标关键点的二维坐标和置信概率。骨架行为识别方法关注目标的肢体动作变化，丢弃了视频背景和目标外观中的冗余信息，降低了数据对模型参数量的要求。然而，现有的大部分骨架行为识别方法[6]通常在时间和空间维度分别提取骨架序列的特征，忽略了骨架拓扑结构在时空维度的整体性。本文在时空图卷积网络 ST-GCN[6]和 Transformer[7]等相关工作的基础上提出基于全局时空编码网络的猴类动物骨架行为识别方法。该网络主要包括时空图卷积网络 ST-GCN 和全局时空编码器(Global Spatiotemporal Encoder, GSTE)。时空图卷积网络 ST-GCN 负责提取时空维度的局部特征来识别简单动作；全局时空编码器由少量的线性算子和自注意力计算模块组成，对时空维度的全局特征进行建模分析进而识别一些困难动作。全局时空编码器可以作为即插即用的轻量级模块来配合骨架行为识别模型 ST-GCN 使用，提高模型在时空维度整体性建模分析的能力。实验结果证明，全局时空编码网络在基本不增加模型参数量的情况下，可以显著提高基准模型 ST-GCN 的行为识别准确率，并且优于其他的基于视频帧和基于骨架序列的行为识别方法。综上所述，本文主要有两点贡献：

(1)提出了一种基于全局时空编码网络的猴类动物行为识别方法，即插即用的轻量级全局时空编码器可以搭配骨架行为识别模型 ST-GCN 使用，在基本不增加模型参数量的同时提高模型在时空维度整体性建模分析的能力；

(2)在自建的猴类动物行为识别数据集上进行了大量的对比试验，比较了基于视频帧的行为识别方法、基于骨架序列的行为识别方法以及本文提出的全局时空编码网络。实验结果充分验证了全局时空编码网络在猴类动物行为识别任务上具有准确率高、参数少等优势。

1 相关工作

1.1 人类行为识别

人类行为识别任务通常是识别一段视频中包含的行为类别。由于视频中包含丰富的信息，不同方法利用不同角度的信息对视频中的行为进行建模分析，如外观、光流以及骨架等。SIMONYAN 等[8]提出经典双流网络，该模型模仿了人类大脑皮层理解视频信息的机制，在处理视频帧图像空间信息的基础上，对视频时序信息也做了建模理解。单独的视频帧作为表述空间信息的载体，其中包括背景和目标外观等空间信息，称为空间卷积网络；另外光流图作为时序信息的载体输入到另一个卷积神经网络中，用来理解行为的动态特征，称为时间卷积网络。针对行为识别任务中的长范围依赖问题，WANG 等[9]在经典双流网络的基础上提出了稀疏时间采样和视频级监督策略，即从整个视频段中稀疏采样了一系列片段来促使模型学习行为的全局特征。HARA 等[10]在 2D 卷积神经网络 ResNet[11]的基础上拓展了一个时间维度得到 3D 卷积网络。ResNet3D 在提取时间维度特征的同时还使用了 2D 网络中的一系列技巧，如使用残差结构来缓解梯度消失问题。TRAN 等[12]在 ResNet3D 的基础上进一步将 3D 卷积核分解为两个独立且连续的操作：2D 空间卷积和 1D 时间卷积。卷积分

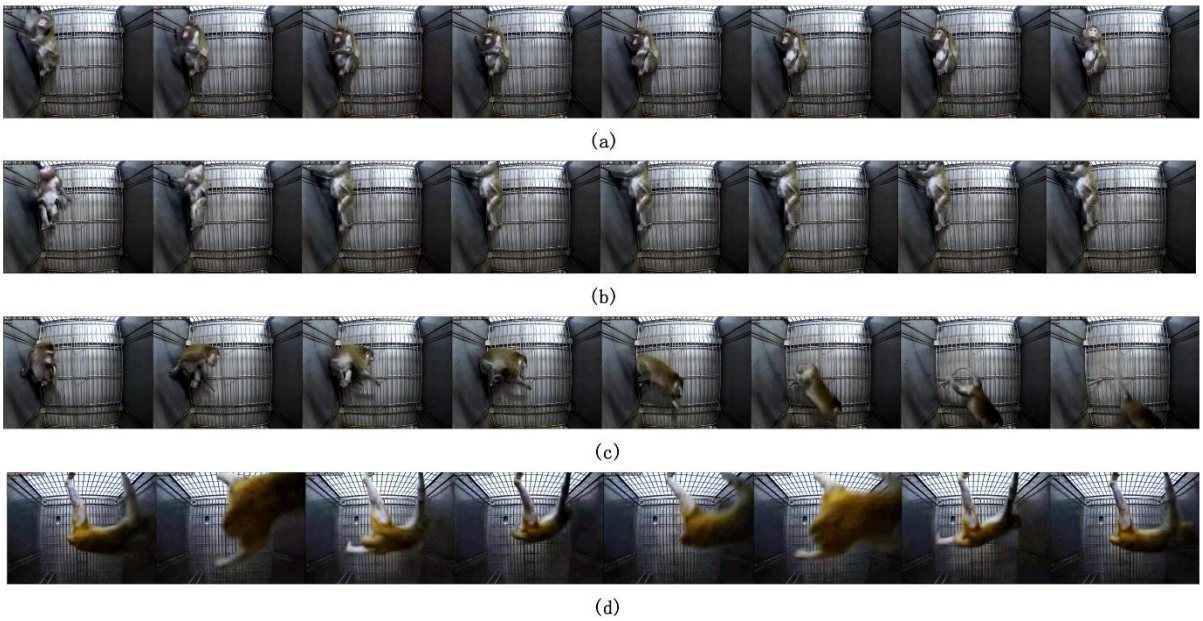

图 1 猴类动物行为识别数据集中的不同行为样本((a)蹲坐；(b)扶立；(c)向下攀爬；(d)悬挂)

Fig. 1 Different action samples in the monkey action recognition dataset((a)Squatting; (b)Standing up; (c)Climbing down; (d)Hanging)

解不仅减少了模型运算参数，同时提高了模型的拟合能力。FEICHTENHOFER 等[13]探索了视频的高低帧率对行为识别的影响，设计了低帧率的慢支路来捕获空间维度的语义信息以及高帧率的快支路捕获时间维度的运动信息。

上述行为识别方法中输入模型的是视频帧或者是视频帧中间的光流图。由于受背景扰动、光照变化以及目标外观差异的影响，完整输入的视频帧或者光流图中有时会有一些信息冗余。近些年一些学者开始使用视频中目标的骨架序列信息来识别目标的行为。目前骨架行为识别方法大致分为四类：基于循环神经网络、基于卷积神经网络、基于图卷积网络以及基于 Transformer 的方法。ZHU 等[14]将每个关键点的形成的时间序列输入到循环神经网络中，同时使用稀疏连接的全连接层来融合不同循环神经网络输出的特征，最后使用$softmax$函数对提取的特征进行分类。LI 等[15]使用双流卷积神经网络提取骨架伪图像在时间和空间上的局部特征，最后融合时空维度的特征来识别行为。YAN 等[6]提出时空图卷积网络 ST-GCN，首次将图卷积网络用于骨架行为识别。ST-GCN 模型对帧内骨架拓扑结构进行空间卷积，对帧间关键点序列进行时间卷积来提取时空维度的局部特征。SHI 等[16]在 ST-GCN 的基础上提出了一种双流自适应图卷积网络 2s-AGCN，其中自适应是指图的拓扑结构可以由梯度反传算法进行端到端的学习。这种数据驱动的方法增加了图构造的灵活性，使模型可以适应各种数据版本。PLIZZARI 等[17]提出 ST-TR 模型，该方法将骨架的拓扑结构和关键点形成的时间序列分别输入 Transformer 模型来提取时空维度的全局特征。ZHANG 等[18]提出 STST 模型，该方法在空间维度和时间维度上分别使用特定的 Transformer 模型来捕捉整个骨架的动态变化，同时提出自监督学习模块提高模型对于残缺的骨架结构和扰乱的视频帧序列等情况的鲁棒性。

1.2 动物行为识别

动物行为识别任务是识别一段视频中目标动物的行为类别。方法大致分为两类：单阶段和两阶段方法。单阶段方法直接使用 3D (或 2D)卷积核来提取视频中目标动物行为的局部特征，再使用全连接层输出对应的行为类别。缺点是很多与行为无关的冗余信息也会输入到模型当中，增加了模型参数量和运算量的同时也会导致模型的误识别。LI 等[4]构建了家猪的行为识别数据集 PBVD-5，该数据集包含 5 类行为：喂食、躺卧、运动、抓以及攀爬。同时，该工作提出了一种基于 SlowFast[13]的时空卷积网络来对家猪行为进行建模分析。两阶段方法首先对视频的每一帧提取目标的关键点得到骨架序列，再对目标的骨架信息进行分析得到行为类别，缺点是两阶段方法需要耗费更多的时间。BALA 等[5]提出恒河猴 3D 姿态估计数据集，先对 3D 姿态信息进行降维，再对降维后的特征进行聚类分析得

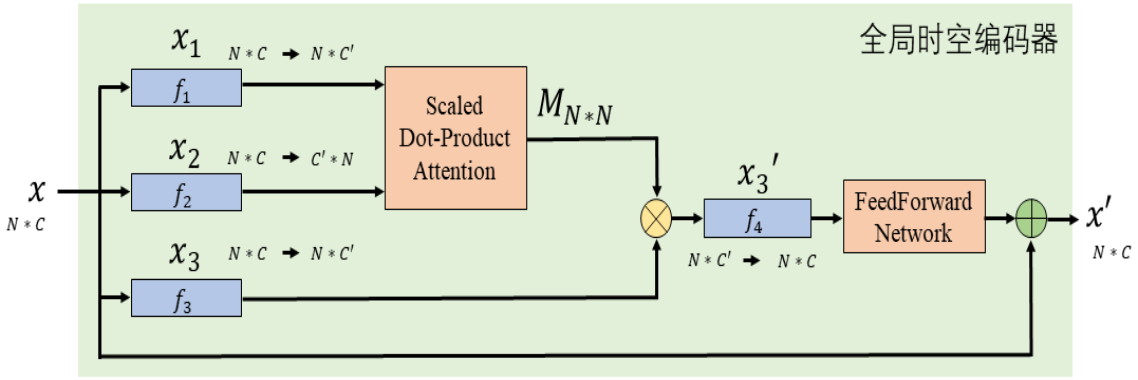

图 2 全局时空编码器

Fig. 2 Global spatiotemporal encoder

到恒河猴的 6 类行为：站立、行走、攀爬、颠倒攀爬、坐下和跳跃。

2 方法

2.1 问题分析

临床前药物安全评价场景下，猴类动物行为识别任务有如下特点：

(1)猴类动物行为识别数据集中的连续视频帧包含冗余信息。在临床前药物安全评价中，猴类动物通常处于室内的铁笼当中，导致行为识别数据集中的场景较单一，视频背景扰动和光照变化小。此外，相同品种不同猴类动物个体的外观非常相似，如恒河猴的毛发普遍呈现棕黄色，食蟹猴的毛发一般呈现灰白色。如图 1 所示，视频数据中背景扰动、光照变化以及目标外观差异较小，导致连续的视频帧中会包含冗余信息。特别地，一些运动量较小的行为 (如蹲坐、扶立)对应的连续视频帧序列冗余信息较明显。

(2) 猴类动物行为识别需要考虑时空维度的全局信息。在临床前药物安全评价场景下，一些猴类动物的动作定义需要考虑全局的空间信息和时间信息。特别地，时间维度上攀爬行为开始的动作类似于蹲坐或者扶立，训练时提取前几帧的特征可能会导致模型误识别，比如图 1(c)中向下攀爬行为的前四帧和图 1(a)中蹲坐行为类似。空间维度上如蹲坐行为的脚踝和臀部对应的关键点联系紧密，然而在 ST-GCN 中无法建模骨架上非直接连接的关键点之间的联系。此外，一些动作需要考虑骨架上不同关键点在不同时刻的联系，如四肢触地行走(行走动作可以视为四肢关键点的周期性运动)和仅下肢触地扶立的局部特征相似，但是在空间上行走时四肢距离更近，且行走动作具有时序信息。现有的一些基于 Transformer 的骨架行为识别方法[17,18]针对时空维度分别提取全局特征，无法建模骨架上不同关键点在不同时刻的联系。

与视频帧序列不同的是，骨架时序信息只关注目标的肢体动作，丢弃了背景以及外观中的冗余信息，降低了行为识别任务对模型参数量的要求。因此，本文使用猴类动物的骨架序列信息进行行为识别，并针对行为的时空整体性问题，进一步提出全局时空编码网络对猴类动物的行为进行整体性建模分析。

2.2 骨架序列信息提取

本文建立的猴类动物行为识别数据集只包括视频片段，缺少关键点坐标等标签，所以需要先提取目标的骨架序列信息，再进行行为识别。为了获取视频帧中每一个目标的骨架信息，本文采用人体姿态估计模型 SimpleBaseline[19]来定位和识别关键点。原数据集中视频的分辨率为 1280×960，为了适应姿态估计模型的空间尺寸，我们首先将所有视频的分辨率统一设置为 256×192，帧率设置为 30Hz；然后使用姿态估计模型来提取视频帧中每一个目标V个关键点的二维坐标(x,y)以及置信概率 p；再将提取到的二维坐标根据图像的高度 h 和宽度 w 进行归一化；最后使用一个三元组 $(x/w, y/h, p)$ 来表示骨架序列中某个关键点的特征。如果某一帧中有多个目标，则选择关键点置信概率均值最大的目标对应的关键点信息来构建三元组。假设当前帧中包含多个目标实例 $\{O_1,O_2,...,O_J\}$，目标O_j检测到V_{O_j}个关键点，第 i 个关键点的置信概率为 $p_{O_j}^i$，则最终选择的目标 O_{j^*} 可以表示为

$$O_{j^*} = \arg\max_{O_j} \frac{\sum_{i=1}^{V_{O_j}} p_{O_j}^i}{V_{O_j}}, O_j \in \{O_1,O_2,...,O_J\} \quad (1)$$

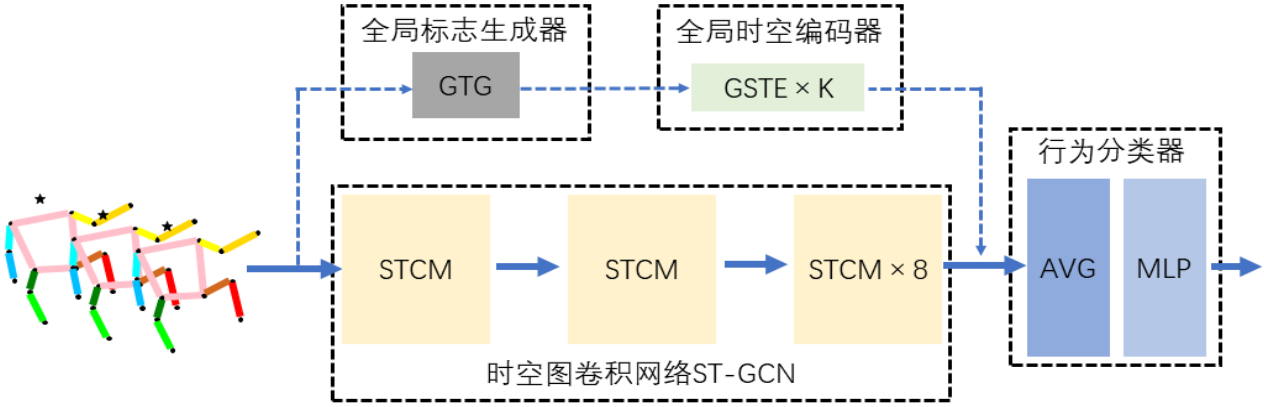

图 3 全局时空编码网络

Fig. 3 Global spatiotemporal encode network

通过逐帧的骨架信息提取，我们得到每一个视频样本的骨架序列表示。假设视频的帧数为T，关键点个数为V，则每一个视频样本的骨架序列信息可以表示为一个维度为$T*V*C$的张量，其中$C=3$表示每一个关键点的特征维数。考虑到猴类动物的动作普遍较快，我们统一设置$T=150$，不足T帧的样本视频通过从头回放的方式进行填充，超过T帧的样本视频直接截取前T帧作为输入。

2.3 全局时空编码器

本文提出全局时空编码器来提取猴类动物骨架序列信息的全局特征。与原始的 Transformer[7]不同的是，GSTE 首先舍弃了 Transformer 中的解码器，只使用串联的编码器来提取全局时空特征。其次，我们在每个单独的编码器中使用一个线性变换来连接自注意力 (Self-Attention, SA)模块和前馈神经网络(FeedForward Network, FFN)，该线性变换将 SA 特征变换为原始输入特征的维数，变换后的特征在后续的 FFN 中完成“编码-解码”的过程。因此，本文提出的 GSTE 在单个编码器内部完成“编码器编码特征-解码器解码推理”的过程。此外，我们专门设计了一个全局标志生成器 (Global Token Generator, GTG)来处理骨架序列。与现有的一些基于 Transformer 的骨架行为识别方法(如 ST-TR[17]和 STST[18]模型)分别从空间和时间维度提取全局特征不同的是，GTG 把骨架序列信息视为一个整体。GTG 不仅考虑了同一时刻的不同关键点以及同一个关键点在不同时刻之间的联系，还对骨架空间拓扑结构上处于不同时刻的不同关键点之间的联系做了建模分析。总的来说，GSTE 在单个编码器当中完成“编码-解码”的过程，同时使用 GTG 进一步增强编码器对骨架序列整体性建模分析的能力。

GSTE 将骨架序列信息对应的三维张量$\boldsymbol{x}\in\mathbb{R}^{T*V*C}$视为一个长度为$N=T*V$的序列，其中$C$表示通道数，$T$表示视频帧数，$V$表示每个目标的关键点个数。如图 2 所示，给定一个特征$\boldsymbol{x}\in\mathbb{R}^{N*C}$，GSTE 首先使用 3 个线性算子将特征维数变换为C'，输出 3 个尺寸为$N*C'$的特征张量，分别记为$\boldsymbol{x}_1$，$\boldsymbol{x}_2$和$\boldsymbol{x}_3$；再将$\boldsymbol{x}_1$和$\boldsymbol{x}_2$视为长度为N的序列，序列中每个元素的特征维数为C'，对$\boldsymbol{x}_1$和$\boldsymbol{x}_2$计算相关性矩阵$\boldsymbol{M}$

$$\boldsymbol{M}_{N*N}=soft\max(\frac{\boldsymbol{x}_1\boldsymbol{x}_2^{T}}{\sqrt{C'}}) \tag{2}$$

其中$soft\max$函数作用在矩阵的每一行上。使用归一化之后的相关性矩阵给$\boldsymbol{x}_3$重加权生成尺寸$N*C'$的全局特征$\boldsymbol{x}_3'$，即$\boldsymbol{x}_3'=\boldsymbol{M}\boldsymbol{x}_3$；提取到的全局特征$\boldsymbol{x}_3'$经过线性算子变换特征维数以及两个串联的全连接层组成的前馈神经网络 FFN，再加到原始的输入特征$\boldsymbol{x}$上得到 GSTE 的最终结果，即$\boldsymbol{x}'=\boldsymbol{x}\oplus\boldsymbol{x}_3'$，其中$\oplus$表示逐元素相加。特别地，GSTE 输出的特征$\boldsymbol{x}'$的尺寸仍为$N*C$,和输入特征$\boldsymbol{x}$保持一致。综上所述，GSTE 可以表示为

$$\boldsymbol{x}'=\boldsymbol{x}\oplus FFN[f_4(soft\max(\frac{f_1(\boldsymbol{x})f_2(\boldsymbol{x})^{T}}{\sqrt{C'}})*f_3(\boldsymbol{x}))] \tag{3}$$

2.4 全局时空编码网络

本文在人体骨架行为识别模型 ST-GCN[6]的基础上提出全局时空编码网络 GSTEN。如图 3 所示，全局时空编码网络由四部分组成，主网络为 ST-GCN 模型，ST-GCN 负责提取骨架序列信息时空维度的局部特征来识别简单动作；在 ST-GCN 模型上并行插入全局标志生成器和和K个串联的轻量级模块 GSTE，GSTE 针对骨架序列信息时空维度的全局特征进行建模分析进而识别一些困难动作；最后融合提取到的全局特征和局部特征，将其输入

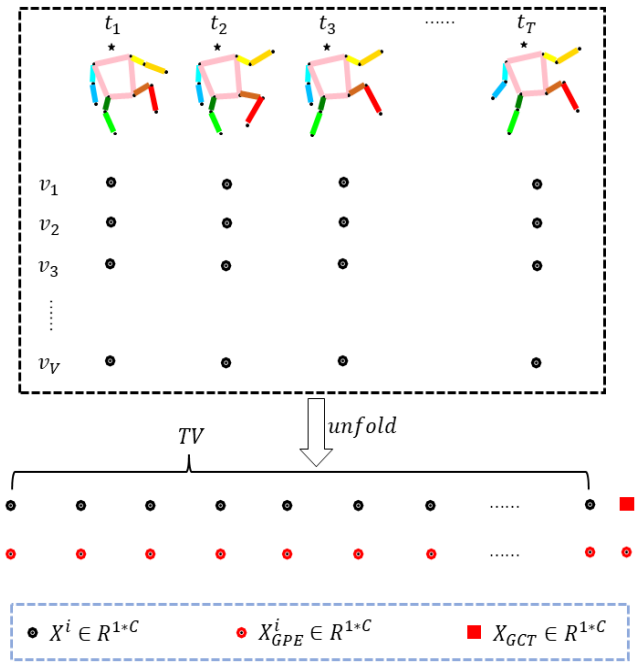

图 4 全局标志生成器

Fig. 4 Global token generator

到行为分类器中进行分类。

全局标志生成器 GTG 在骨架序列信息基础上添加可学习的全局位置嵌入 (Global Position Embedding, GPE) 和全局分类符 (Global Class Token, GCT)。全局位置嵌入学习骨架序列上所有关键点的语义信息和时序信息，全局分类符学习骨架时序信息的全局特征用于行为的分类。假设输入的骨架时序信息是一个三维张量 $\boldsymbol{X} \in \mathbb{R}^{T*V*C}$，则设置全局位置嵌入为 $\boldsymbol{X}_{\boldsymbol{GPE}} \in \mathbb{R}^{(TV+1)*C}$，设置全局分类符 $\boldsymbol{X}_{\boldsymbol{GCT}} \in \mathbb{R}^{1*C}$。如图 4 所示，在获取骨架时序信息后，将骨架序列上所有的关键点展开得到长度为 TV 的序列，在该序列上添加全局位置嵌入 $\boldsymbol{X}_{\boldsymbol{GPE}}$ 和全局分类符 $\boldsymbol{X}_{\boldsymbol{GCT}}$，最终输入全局时空编码器的特征 $\boldsymbol{X}$ 可以表示为公式(4)，其中 $Concat$ 表示序列拼接。

$$\boldsymbol{X} = Concat(\boldsymbol{X}, \boldsymbol{X}_{\boldsymbol{GCT}}) + \boldsymbol{X}_{\boldsymbol{GPE}} \quad (4)$$

猴类动物骨架序列信息通过 ST-GCN 支路可以提取时空维度的局部特征，通过 GSTE 支路提取全局特征，最后对两类特征加权融合进行行为识别。这种并行连接 GSTE 的网络结构在几乎不增加模型参数量的情况下，可以显著提高行为识别准确率。具体地，ST-GCN 模型由 10 个串联连接的局部时空卷积模块 (Local Spatial Temporal Convolution Module, STCM)组成。每一个 STCM 包括空间卷积和时间卷积。空间卷积可以表示为

$$f_{spatial}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{s}}, \boldsymbol{A} \in \mathbb{R}^{V*V}, \boldsymbol{x} \in \mathbb{R}^{V*d}, \boldsymbol{W}_{\boldsymbol{s}} \in \mathbb{R}^{d*d_s} \quad (5)$$

其中 $\boldsymbol{x}$ 是某一帧的骨架信息，V 表示骨架上的节点个数，d 表示每个节点的特征维度。$\boldsymbol{A}$ 表示节点之间的相关性矩阵，A_{ij} 表示骨架上第 i 个节点和第 j 个节点的相关性。$\boldsymbol{W}_{\boldsymbol{s}}$ 表示空间卷积参数矩阵，d_s 表示空间卷积变换后的特征维数。时间卷积可以表示为

$$f_{temporal}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W}_{\boldsymbol{t}}, \boldsymbol{x} \in \mathbb{R}^{T*d}, \boldsymbol{W}_{\boldsymbol{t}} \in \mathbb{R}^{d*d_t} \quad (6)$$

其中 $\boldsymbol{x}$ 是某一个关键点的时序信息，$\boldsymbol{W}_{\boldsymbol{t}}$ 表示时间卷积参数矩阵，d_t 表示时间卷积变换后的特征维数。在本文的猴类动物场景中，我们使用公式(7)来计算节点相关性矩阵 $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{\Delta}^{-\frac{1}{2}}(\boldsymbol{A} + \boldsymbol{I})\boldsymbol{\Delta}^{-\frac{1}{2}} \quad (7)$$

其中邻接矩阵 $\boldsymbol{A}$ 通过人体姿态估计数据集 MS COCO[20]中关键点的连接方式来设置，$\boldsymbol{I}$ 表示维度为 $V*V$ 的单位阵，且 $V=17$，$\Delta_{ii} = \sum_j (A_{ij} + I_{ij})$。

表 1 猴类动物行为识别数据集

Table 1 Monkey action recognition dataset

行为类别	样本数	总计
卧倒	41	609
蹲坐	164	
行走	21	
向上跳跃	31	
向下跳跃	23	
向上攀爬	35	
向下攀爬	49	
悬挂	100	
扶立	111	
攀附	34	

2.5 损失函数

猴类动物行为识别任务的真值标签设计为 one-hot 编码的形式，生成的标签 $\boldsymbol{t} \in \mathbb{R}^{n*1}$，其中 n 为行为类别数。如果当前样本属于第 i 个类别，则且 $t_i = 1$，其余元素为 0。损失函数使用交叉熵损失 (Cross Entropy Loss)函数。网络最后一层的输出先经过 $soft\max$ 函数将每一个元素值映射为一个概率值，再对概率取对数，最后将输出与标签类别位置对应的值取负数相乘再求和就得到了该样本的损失。假设当前样本的标签为 $\boldsymbol{t} \in \mathbb{R}^{n*1}$，属于第 $class$ 类，即 $t_{class} = 1$，模型对样本的输出为 $\boldsymbol{y}$，则该样本的损失值计算如下

$$Loss(\boldsymbol{y}, class) = -\sum_{i=1}^{n} t_i \ln(soft\max(\boldsymbol{y})_i) \quad (8)$$

表 2　不同方法在猴类动物行为识别数据集上的实验结果

Table 2　Experimental results of different methods on the monkey action recognition dataset

方法	框架	参数量 (M)	运算量 (G)	准确率 (%)
基于视频帧的方法				
Slow[13]	ResNet-50	31.7	41.9	67.90
SlowFast[13]	ResNet-50	33.7	50.6	61.11
I3D[21]	ResNet-50	27.2	28.5	74.07
I3D[21]	ResNet-101	51.0	47.1	73.46
X3D[23]	X3D-S	3.0	2.0	71.60
MViT[24]	MViT-B	36.3	70.8	74.69
TimeSformer[25]	TimeSformer-divST	121.4	590.0	**77.16**
基于骨架序列的方法				
ST-GCN[6]	STCM*10	3.1	2.8	69.75
Js-AGCN[16]	Agcn	3.4	6.3	70.37
Bs-AGCN[16]	Agcn	3.4	6.3	63.58
2s-AGCN[16]	Agcn	6.9	12.6	70.99
MS-G3D-Joint[22]	G3d	2.7	8.7	71.60
MS-G3D-Bone[22]	G3d	2.7	8.7	72.84
MS-G3D[22]	G3d	5.4	17.4	75.31
CTR-GCN[26]	CTR-GCN	1.46	1.97	**77.78**
PoseConv3D[27]	X3D-S	0.24	0.6	75.93
GSTEN	STCM*10+GSTE*1	3.2	2.8	73.46
GSTEN	STCM*10+GSTE*2	3.3	2.8	**76.54**

3　实验

3.1　数据集和评价指标

本文建立的猴类动物行为识别数据集如表 1 所示，包括卧倒、蹲坐、行走、向上跳跃、向下跳跃、向上攀爬、向下攀爬、悬挂、扶立以及攀附 10 类行为。将采集的猴类动物行为识别数据集按照 3:1 的比例随机划分为训练集和验证集。评价指标采用准确率 $Acc = N_{cor}/N$，其中 N 为验证集的总样本数，N_{cor} 为验证集中模型预测正确的样本数。

3.2　参数设置

实验在 Geforce RTX2080Ti*8 的单节点服务器上完成，使用 Pytorch v1.8 深度学习框架进行训练。Epoch 数设置为 100，batch-size 为 16。初始学习率为 0.1，随后在 40 个 epoch 和 80 个 epoch 处衰减为原来的 0.1 倍。在全局时空编码网络 GSTEN 中如果没有特别说明，则 ST-GCN 支路特征和 GSTE 支路特征的加权和系数 w 为 0.5，即 $f_{merge} = f_{ST\text{-}GCN} + w * f_{GSTE}$，其中 $f_{ST\text{-}GCN}$ 和 f_{GSTE} 分别表示 ST-GCN 分支提取的局部时空特征和 GSTE 分支提取的全局时空特征。

3.3　对比实验

我们在猴类动物行为识别数据集上对比了基于视频帧的行为识别方法、基于骨架序列的行为识别方法以及本文提出的全局时空编码网络。如表 2 所示，一些基于视频帧的方法如 SlowFast[13]等在数据集上的性能指标明显低于骨架行为识别模型 ST-GCN。I3D[21]方法的性能优于 ST-GCN，但模型的参数量和运算量更多。特别地，TimeSformer[25]模型的准确率达到最高的 77.16%，但该模型的参数量和运算量远多于其他基于视频帧以及基于骨架序列的方法。本文提出的全局时空编码网络在 ST-GCN 模型的基础上并行添加全局时空编码器。实验结果表明，GSTEN 在基本不增加模型参数量和运算量的同时可以显著提高基准模型 ST-GCN 的准确率。此外，当全局时空编码网络搭配两个全局时空编码器时，不仅准确率比 ST-GCN 高 6.79%，并且优于大部分基于视频帧和基于骨架序列的行为识别方法。总的来说，本文构建的全局时空编码

网络在猴类动物行为识别任务上具有准确率高、参数少以及运算量小等优势。

表 3　GSTEN 消融实验

Table 3　Ablation experiments of GSTEN

实验序号	框架	参数量 (M)	准确率 (%)
Exp-1	STCM*10	3.1	69.75
Exp-2	GSTE*1	0.1	64.81
Exp-3	GSTE*2	0.3	71.60
Exp-4	STCM*10+GSTE*1	3.2	73.46
Exp-5	STCM*10+GSTE*2	3.3	**76.54**

3.4　消融实验

为了验证局部特征和全局特征在猴类动物行为识别任务上的适用性，我们对全局时空编码网络的各部分结构进行了消融实验。如表 3 所示，我们对比了只使用 ST-GCN 分支提取局部特征、只使用 GSTE 分支提取全局特征以及使用 GSTEN 模型融合局部特征和全局特征三者之间的性能差异。实验结果表明，Exp-2 中使用 1 个全局时空编码器提取全局特征时，模型的准确率低于 Exp-1 中使用 ST-GCN 提取局部特征的结果。Exp-3 中使用 2 个全局时空编码器时，模型对猴类动物行为的建模分析能力超过了 ST-GCN。当 Exp-4 和 Exp-5 中全局特征和局部特征融合时，GSTEN 模型超过单一局部特征或全局特征的结果，且性能随着全局时空编码器数量增加而提升，这表明全局特征和局部特征融合的结果比单一特征更适合猴类动物的行为识别。

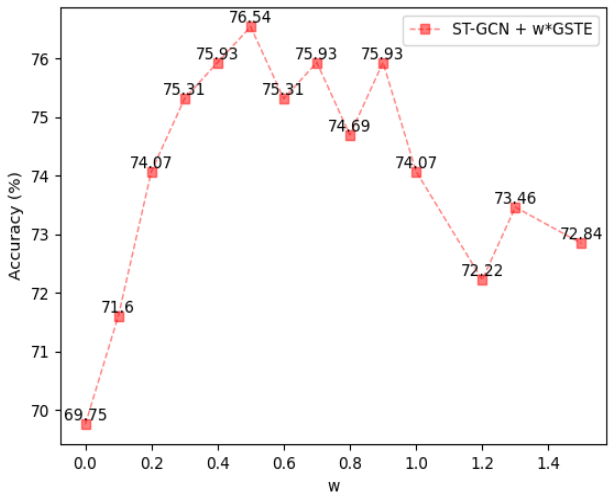

图 5 基于不同系数 w 的 GSTEN 准确率

Fig. 5　Accuracy of the GSTEN based on different coefficient w

在表 3 的 Exp-5 基础上进一步探索 ST-GCN 支路和 GSTE 支路的加权和系数 w 对构建的全局时空编码网络性能的影响。如图 5 所示，当 $w<0.5$ 时，ST-GCN 分支的特征占比较大，GSTEN 模型的准确率随着 w 增大而提高；当 $w=0.5$ 时，GSTEN 模型准确率达到最高的 76.54%；当 $0.5<w<1$ 时，GSTEN 模型准确率接近饱和；当 $w>1$ 时，GSTE 分支的特征占据主导地位，此时 GSTEN 模型准确率稍有下降。以上结果说明 GSTEN 中 ST-GCN 分支重要性较 GSTE 高。原因可能是 GSTE 分支参数较少，在建模骨架序列所有关键点之间的联系时出现欠拟合的情况。

表 4　GSTE 对不同基准模型的影响

Table 4　The impact of GSTE on different baseline models

模型	参数量 (M)	准确率 (%)
ST-GCN[6]	3.1	69.75
ST-GCN + GSTE	3.3	76.54 (+6.79)
Js-AGCN[16]	3.4	70.37
Js-AGCN + GSTE	3.7	72.84 (+2.47)
Bs-AGCN[16]	3.4	63.58
Bs-AGCN + GSTE	3.7	66.67 (+3.09)
MS-G3D-Joint[22]	2.7	71.60
MS-G3D-Joint + GSTE	3.0	75.93 (+4.33)
MS-G3D-Bone[22]	2.7	72.84
MS-G3D-Bone + GSTE	3.0	78.40 (+5.56)

此外，为了验证本文提出的全局时空编码器具有即插即用的特性，我们在不同的骨架行为识别模型上并行插入全局标志生成器和两个全局时空编码器。如表 4 所示，由于 ST-GCN 模型缺乏时空维度整体性建模分析的能力，GSTE 应用在 ST-GCN 上可以显著提升模型的行为识别准确率。当 GSTE 应用在具有全局空间建模能力的 AGCN 上时，准确率相较于 ST-GCN 提升幅度略有下降，分别为 2.47%和 3.09%。特别地，当 MS-G3D-Bone 添加 GSTE 之后模型的准确率达到 78.40%，超过本文提出的全局时空编码网络和表 2 中准确率最高的 CTR-GCN[26]方法。以上结果表明，全局时空编码器可以作为即插即用的轻量级模块配合不同的骨架行为识别模型使用，并且构建的新网络在基本不增加模型参数量的情况下，显著提高基准模型在猴类动物行为识别任务上的准确率。

表 5　GSTE 消融实验

Table 5　Ablation experiments of GSTE

实验序号	框架	参数量 (M)	准确率 (%)
Exp-1	GSTE*1	0.13	64.81
Exp-2	GSTE*2	0.27	**71.60**
Exp-3	GSTE*3	0.40	69.75
Exp-4	GSTE*4	0.53	70.37
Exp-5	GSTE*5	0.66	66.67
Exp-6	GSTE*6	0.79	67.90

我们进一步对比了不同数量的全局时空编码器对模型的影响。如表 5 所示，Exp-1、Exp-5 和 Exp-6 中全局时空编码器的数量过少或者过多，都会对模型性能产生影响。Exp-2 中使用 2 个全局时空编码器来提取全局特征时，模型准确率指标达到最高的 71.60%。

4　结束语

本文的工作从实际临床前药物安全评价场景出发，使用深度学习方法对猴类动物行为识别任务进行了研究，对人工智能方法在药物安全评价中的应用进行了积极的探索。本文首先分析了临床前药物安全评价场景下，现有人类行为识别领域基于视频帧和基于骨架序列的方法应用到猴类动物的缺陷。然后基于这些缺陷，本文提出了一种基于全局时空编码网络的猴类动物行为识别方法，即插即用的轻量级全局时空编码器可以搭配骨架行为识别模型 ST-GCN 使用，在基本不增加模型参数量的同时提高模型在时空维度整体性建模分析的能力。最后在生成的猴类动物行为识别数据集上进行了完善的实验，对比了基于视频帧的行为识别方法、基于骨架序列的行为识别方法以及本文提出的全局时空编码网络，结果充分验证了全局时空编码网络在临床前药物安全评价场景猴类动物行为识别任务上具有准确率高、参数少等优势。

对于骨架行为识别方法而言，获取带有骨架序列标签的训练数据是前提条件。借助于训练好的姿态估计模型，我们可以大规模获取行为视频中每个猴类动物个体的骨架信息。然而，使用姿态估计模型提取骨架信息再进行行为识别的两阶段过程比较耗时，并且对于不同的临床前药物安全评价场景，我们需要训练鲁棒性和泛化性更强的姿态估计模型。因此，未来工作的一个重点方向是探索更有效的骨架信息获取方法，如使用一些穿戴式设备[28]直接获取猴类动物的骨架信息。

参考文献（References）

[1] PLAGENHOEF M R, CALLAHAN P M, BECK W D, et al. Aged rhesus monkeys: Cognitive performance categorizations and preclinical drug testing[J]. Neuropharmacology, 2021, 187: 108489.

[2] BANKS M L, HUTSELL B A, BLOUGH B E, et al. Preclinical assessment of lisdexamfetamine as an agonist medication candidate for cocaine addiction: effects in rhesus monkeys trained to discriminate cocaine or to self-administer cocaine in a cocaine versus food choice procedure[J]. International Journal of Neuropsychopharmacology, 2015, 18(8): pyv009.

[3] EBELING M, KÜNG E, SEE A, et al. Genome-based analysis of the nonhuman primate Macaca fascicularis as a model for drug safety assessment[J]. Genome Research, 2011, 21(10): 1746-1756.

[4] LI D, ZHANG K, LI Z, et al. A spatiotemporal convolutional network for multi-behavior recognition of pigs[J]. Sensors, 2020, 20(8): 2381.

[5] BALA P C, EISENREICH B R, YOO S B M, et al. Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio[J]. Nature Communications, 2020, 11(1): 1-12.

[6] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Thirty-second AAAI Conference on Artificial Intelligence. 2018.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[8] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27.

[9] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European Conference on Computer Vision. Springer, Cham, 2016: 20-36.

[10] HARA K, KATAOKA H, SATOH Y. Learning spatio-temporal features with 3d residual networks for action recognition[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 3154-3160.

[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[12] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.

[13] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6202-6211.

[14] ZHU W, LAN C, XING J, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30(1).

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

[16] SHI L, ZHANG Y, 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. 2019: 12026-12035.

[17] PLIZZARI C, CANNICI M, MATTEUCCI M. Spatial temporal transformer network for skeleton-based action

recognition[C]//International Conference on Pattern Recognition. Springer, Cham, 2021: 694-701.

[18] ZHANG Y, WU B, LI W, et al. STST: Spatial-Temporal Specialized Transformer for Skeleton-based Action Recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 3229-3237.

[19] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 466-481.

[20] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//European Conference on Computer Vision. Springer, Cham, 2014: 740-755.

[21] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[22] LIU Z, ZHANG H, CHEN Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 143-152.

[23] FEICHTENHOFER C. X3d: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 203-213.

[24] FAN H, XIONG B, MANGALAM K, et al. Multiscale vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6824-6835.

[25] GEDAS BERTASIUS, HENG WANG, LORENZO TORRESANI. Is Space-Time Attention All You Need for Video Understanding? [C]// International Conference on Machine Learning. 2021: 813-824.

[26] CHEN Y, ZHANG Z, YUAN C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13359-13368.

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

[28] 邓颖，吴华瑞，孙想．基于机器视觉和穿戴式设备感知的村镇老年人跌倒监测方法[J]．西南大学学报（自然科学版），2021, 43(11): 186-194.
DENG Y, WU H R, SUN X. Design of a Real-Time Human Falling Monitoring Method for Elderly People in Villages and Towns Based on Multi-dimensional Data Analysis[J]. Journal of Southwest University, 2021, 43(11): 186-194 (in Chinese).