中国石油大学（北京）
CHINA UNIVERSITY OF PETROLEUM , BEIJING

本科生科研训练结题报告

题　　目　图卷积神经网络聚类算法的研究

学院名称　信息科学与工程学院

专业名称　计算机科学与技术

申报人　赵新龙

指导老师　张丽英

起止时间　2022 年 1 月 10 日至 2022 年 6 月 5 日

1. 引言

聚类是机器学习/数据挖掘的一个基础性问题，作为经典的无监督学习算法在数据挖掘/机器学习的发展历史中留下了不可磨灭的印记。其中，经典的聚类算法 K-Means 也被选为数据挖掘十大经典算法。随着深度学习的兴起和计算机计算能力的提高，为聚类算法的发展注入了新活力，从深度学习方法中获得灵感的深度聚类获得了最先进的性能，并引起了广泛的关注。

深度聚类算法分为 two-step，首先学习数据的特征表示 embedding，然后基于特征表示进行数据聚类。然而这种方法所学习的数据 embedding 没有根据特定的聚类任务设计、优化图嵌入，不是任务导向的。如果能够在学习 embedding 的过程中，针对聚类任务做一些针对性的设计，学习到的 embedding 自然可以实现更好的聚类。

2. 背景

近些年，图神经网络已经成为深度学习领域最热门的方向之一，能不能利用图神经网络强大的结构捕获能力来提升深度聚类算法的效果呢？一些研究者围绕此问题开展了相关的工作。本课题的研究目的是通过研究现有图神经网络的聚类算法，使用地铁人流量数据，基于图神经网络来建模地铁站点聚类的问题。

本课题通过构建地铁站点的图结构，出站和入站客流量作为结点的属性值，基于 SDCN 模型构建图神经网络建模完成地铁站点的分类应用问题。

3. 研究环境

基于 pytorch 框架下的图卷积网路相关模型，cuda 单元进行计算。数据集来源为北京地铁 2013 年 3 月 1 日至 14 日各站人流量。

4. 研究方法与过程

以聚类为导向的深度算法 Deep Attentional Embedded Graph Clustering (DAEGC)是引入图神经网络提升深度聚类算法效果的代表成果之一。 DAEGC 一边通过图神经网络来学习节点表示，一边通过一种自训练的图聚类增强同一簇节点之间的内聚性。下图 4.1 展示了 two-step 和 DAEGC 之间的差异。

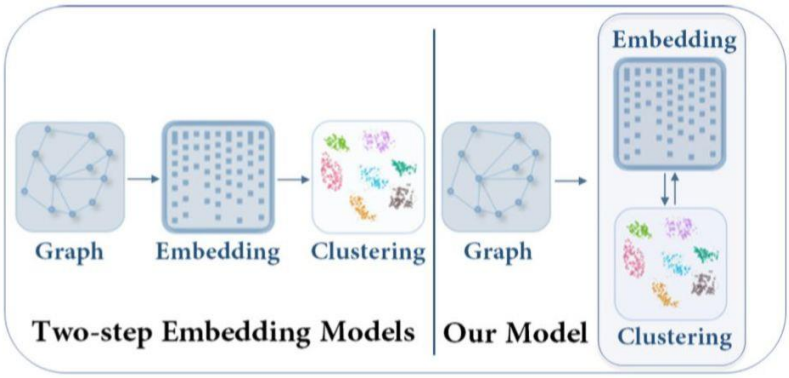

图 4.1 two-step（左） 和 DAEGC（右）

下图 4.2 是 DAEGC 的模型框架细节图。

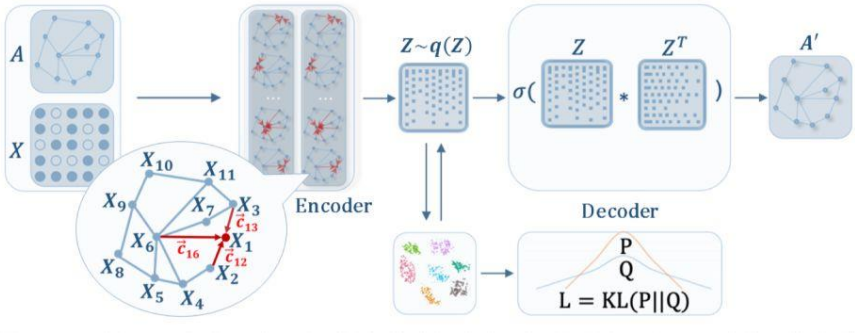

图 4.2 DAEGC 模型框架

整个 DAEGC 主要包含两大模块:带有注意力机制的图自编码器和自训练聚类。

带有注意力机制的图自编码器通过对邻居的聚合来学习节点表示,然后利用节点对的内积来重构原始网络结构,结合注意力机制来学习邻居的权重, 这样可以更好的学习节点表示。自训练聚类模块就是对GAE所学习到的embedding进行约束和整合,使其更适合于聚类任务。

Structural Deep Clustering Network (SDCN) 是 2020 年提出的基于图神经网络的聚类方法,它在保持现有深度聚类框架优点的同时,加入结构化信息。它设计了一个传递算子,将自动编码器学习到的表示转换到相应的 GCN 层,并设计了一个双自监督机制来统一这两种不同的深层神经结构,引导整个模型的更新。通过这种方式,从低阶到高阶的多种数据结构自然地与自动编码器学习到的多种表示相结合。

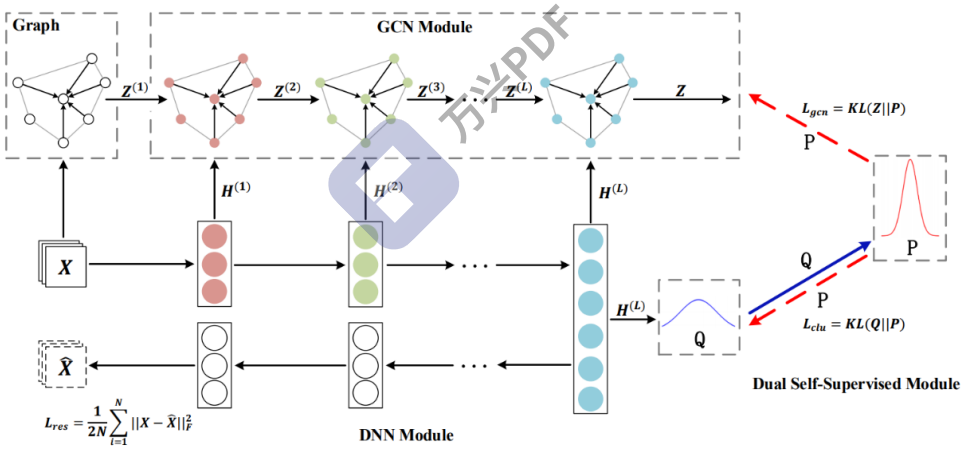

图 4.3 SDCN 模型框架

以上是 SDCN 的模型结构图。X 和 X’是输入数据和输出的重构数据，H(1)和 Z(1)代表第 1 层的 GCN 和 DNN 模型不同的颜色代表不同的的 DNN 和 GCN 层，蓝色实线代表目标分布 P 被计算基于 Q 的分布，两条红色虚线代表双向自监督机制。目标分布 P 调整 DNN 和 GCN 模型参数是同时进行的。

由于 DAEGC 的训练机制仅根据 GAE 学习到的 embedding 进行约束，相比之下 SDCN 通过自动编码器学习到的数据转换成目标分布，进而对两个网络实现双向自监督。因此 SDCN 可以做到编码器和训练聚类过程相互独立，从而有更准确的结果，因此本研究只采用了 SDCN 模型。

本课题通过构建地铁站点的图结构，出站和入站客流量作为结点的属性值，基于 SDCN 模型构建图神经网络建模完成地铁站点的分类应用问题。

5. 实验数据与分析

5.1 数据描述

本研究使用的数据为北京市轨道交通站点 2013 年 3 月 4 日至 17 日完整两周的出行(包含完整进出站刷卡 记录)记录数，对数据进行清洗后，最终选择具有完整的进出站刷卡记录的轨道交通站点共 195 个站点作为研究对象。对轨道刷卡数据以小时为单位窗口，按进站和出站分别对站点刷卡数据进行汇总，形成进站和出站客流量二元时间序列。

5.2 实验结果

5.2.1 实验方法及评测结果

针对北京 195 个轨道交通站点，每个站点进出站的客流量形成二元时间序列，使用 SDCN 模型来应用于地铁站点得聚类问题上，GCN 模型用于提取北京地铁站之间的图结构信息，而 DNN 模型的自动编码器用来挖掘每个节点的二元时间时间序列信息，两者同时影响目标分布 P 的值，之后再通过 KL 损失值作为损失函数来训练该模型。聚类个数设置为 2 3 4 5 6 个类别，然后看他们的轮廓系数（SC）、Davies-Bouldin 指数（DB）和 Calinski Harabasz 指数（CH）。下图是各个指数具体数据。

由于此应用场景是地铁站点的聚类，所以每个站点的信息除了自身的入站出站信息外，还有北京地铁的图结构信息，根据这些站点建立邻接矩阵表示图结构信息，然后通过矩阵变换将图结构信息融入节点信息中，此举在 GNN 模型中实现，之后在评测时主要看带有图结构的原始数据。就是下表 1 中的 CH、DB、SC，其中带 raw 的三个指标是不带图结构的节点信息。可见加入图结构会使效果更好，此部分会在 5.3 实验分析中详细介绍。

表 1 聚类指标对比表

	2	3	4	5	6
CH	320.20	183.52	121.81	127.63	71.87
rawCH	153.37	95.67	62.18	54.03	36.13
	2	3	4	5	6
DB	0.74	1.50	1.22	1.54	1.25
rawDB	1.03	1.86	1.93	2.34	1.76
	2	3	4	5	6
SC	0.51	0.29	0.28	0.19	0.08
rawSC	0.38	0.22	0.19	0.10	0.14

5.2.2 评测指标介绍

①轮廓系数（Silhouette Coefficient）结合内聚度和分离度两种因素。下图是轮廓系数计算公式，其中 b(i)代表类间最小距离，a(i)代表类内平均距离，又称样本 i 的簇内不相似度。所有样本的 si 的均值称为聚类结果的轮廓系数，定义为 S，是该聚类是否合理、有效的度量。聚类结果的轮廓系数的取值在【-1,1】之间，值越大，说明同类样本相距约近，不同样本相距越远，则聚类效果越好。

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad s(i)=\begin{cases}1-\frac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \frac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases}$$

②CH 分数（Calinski-Harabaz Index）被定义为簇间离散与簇内离散的比率，是通过

评估类之间的方差和类内方差来计算得分。类别内部数据的协方差越小越好，类别之间的协方差越大越好，这样的 Calinski-Harabasz 分数会高。 总结起来一句话：CH index 的数值越大越好。

$$S(K) = \frac{T_r(B_K)}{T_r(W_K)} * \frac{N-K}{K-1}$$

B_K是组间离散矩阵
W_K是组内离散矩阵

$$W_K = \sum_{q=1}^{k} \sum_{x \in c_q} (x - C_q)(x - C_q)^T$$

$$B_K = \sum_q n_q (C_q - c)(C_q - c)^T$$

其中 k 代表聚类类别数，N 代表全部数据数目。n 是样本点数，cq 是在聚类 q 中的样本点，Cq 是在聚类 q 中的中心点，nq 是聚类 q 中的样本点数量，c 是 E 的中心（E 是所有的数据集）

③DBI(Davies-bouldin-score) 计算任意两类别的类内距离平均距离(CP)之和除以两聚类中心距离求最大值。DB 越小意味着类内距离越小同时类间距离越大。

$$DBI = \frac{1}{N} \sum_{i=1}^{N} \max_{j \neq i} \left(\frac{\overline{S_i} + \overline{S_j}}{\|w_i - w_j\|_2} \right)$$

分子：簇内所有点到该簇质心点的平均距离之和分母 d(ci,cj)：两类别质心间的距离 max()最大值部分：选取每组比例中的最大值（即选取最糟糕的一组）1/n 求和部分：将所选比例加和除以类别数

5.2.3 分类结果图

下面每个图分别是分成 2、3、4、5、6 类的数据分布图，每个点代表每天每个单位时间每个站点的进出站人流数量，上部分的是出站人数下部分是进站人数。不同的颜色代表不同类别，time 轴代表时间，数据集是从 4 点到 23 点，subway data 轴代表进出站人数。

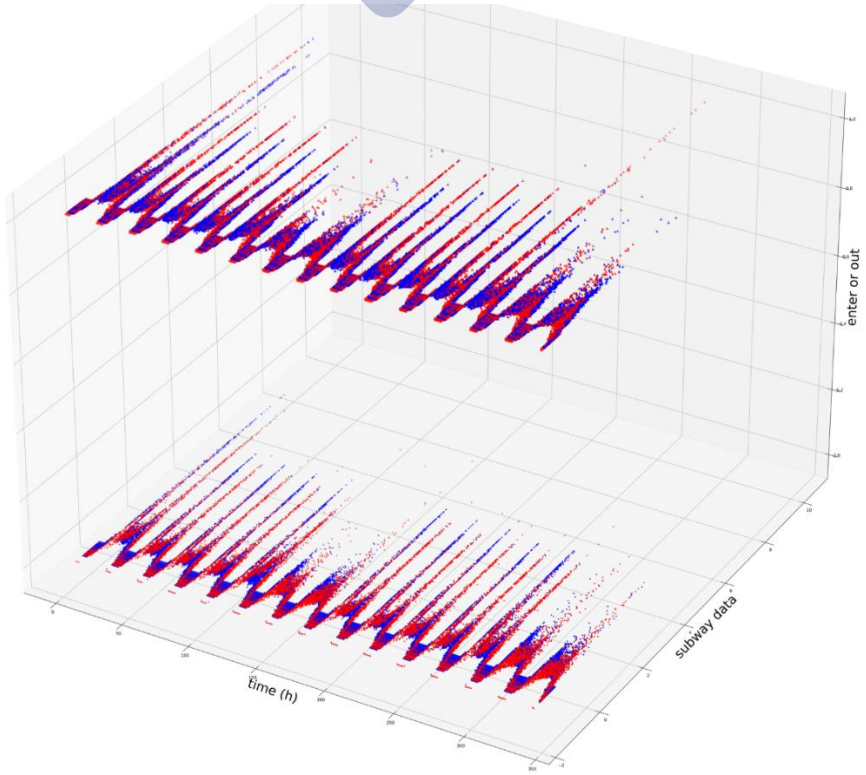

图 5.1 分两类

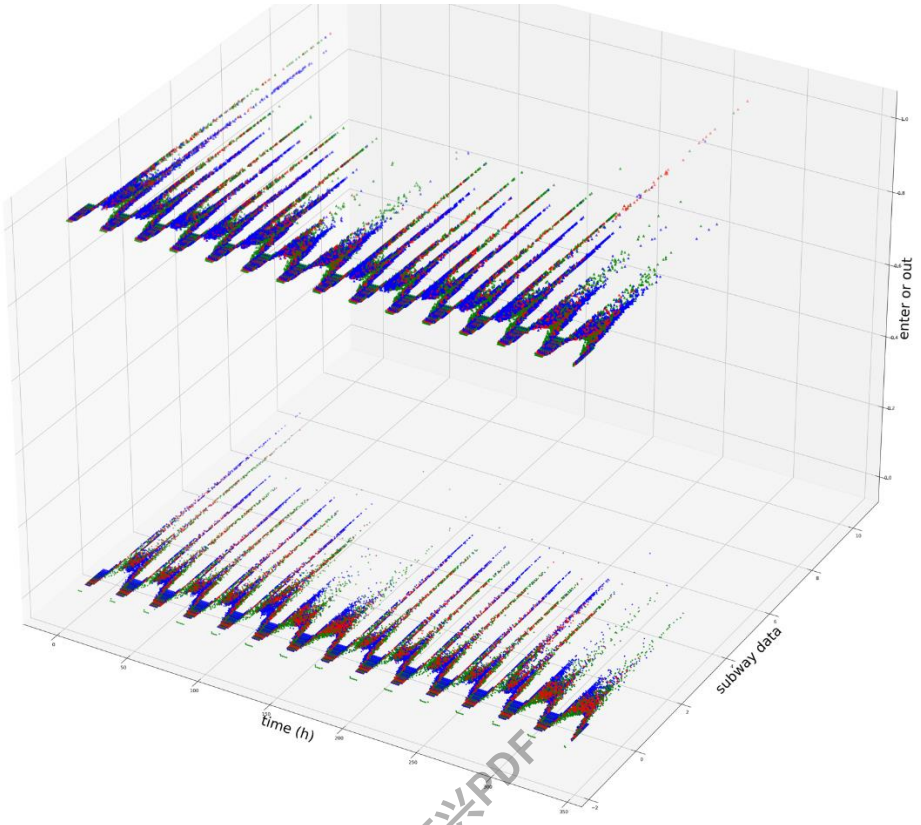

图 5.2 分三类

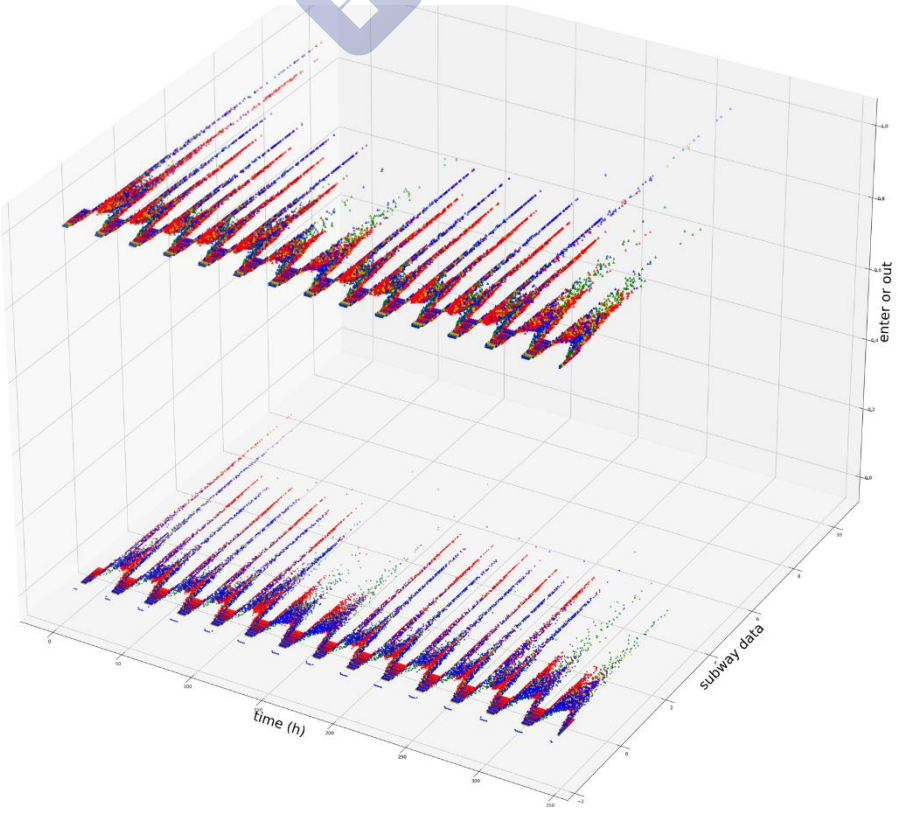

图 5.3 分四类

①其中分成两类的地铁站点名字为：

x1:['安定门', '安华桥', '安贞门', '奥林匹克公园', '奥体中心', '巴沟', '白石桥南', '北海北', '北京大学东门', '北京站', '北土城', '北新桥', '北苑路北', '菜市口', '朝阳门', '车公庄', '车公庄西', '崇文门', '磁器口', '大葆台', '大屯路东', '大望路', '大钟寺', '灯市口', '东大桥', '东单', '东四', '东四十条', '东直门', '动物园', '阜成门', '复兴门', '鼓楼大街', '光熙门', '郭公庄', '国家图书馆', '国贸', '国展', '海淀黄庄', '和平里北街', '和平门', '和平西桥', '呼家楼', '花梨坎', '花园桥', '惠新西街北口', '惠新西街南口', '积水潭', '建国门', '健德门', '金台路', '金台夕照', '劲松', '军事博物馆', '亮马桥', '灵镜胡同', '柳芳', '牡丹园', '木樨地', '南礼士路', '南锣鼓巷', '农业展览馆', '平安里', '前门', '人民大学', '荣昌东街', '荣京东街', '三元桥', '上地', '芍药居', '双井', '四惠', '四惠东', '苏州街', '太阳宫', '陶然亭', '天安门东', '天安门西', '天坛东门', '同济南路', '团结湖', '万寿路', '万源街', '王府井', '望京', '望京西', '魏公村', '五道口', '五棵松', '西单', '西四', '西土城', '西直门', '新街口', '宣武门', '雍和宫', '永安里', '张自忠路', '长椿街', '知春里', '知春路', '中关村']

x2:['安河桥北', '八宝山', '八角游乐园', '八里桥', '北宫门', '北京南站', '北苑', '草房', '常营', '传媒大学', '慈寿寺', '次渠', '次渠南', '崔各庄', '褡裢坡', '稻田', '俸伯', '高碑店', '高米店北', '高米店南', '公益西桥', '巩华城', '古城路', '管庄', '广阳城', '果园', '海淀五路居', '后沙峪', '黄村火车站', '黄村西大街', '黄渠', '回龙观', '回龙观东大街', '霍营', '角门西', '经海路', '九棵树', '旧宫', '梨园', '篱笆房', '立水桥', '立水桥南', '良乡大学城', '良乡大学城北', '良乡大学城西', '良乡南关', '林萃桥', '临河里', '刘家窑', '龙泽', '马家堡', '马泉营', '南法信', '南邵', '苹果园', '蒲黄榆', '青年路', '清源路', '森林公园南门', '沙河', '沙河高教园', '生命科学园', '生物医药基地', '十里堡', '石门', '双桥', '顺义', '宋家庄', '苏庄', '孙河', '天宫院', '天通苑', '天通苑北', '天通苑南', '通州北苑', '土桥', '西二旗', '西红门', '西小口', '西苑', '肖村', '小红门', '新宫', '义和庄', '亦庄桥', '亦庄文化园', '永泰庄', '玉泉路', '育新', '圆明园', '枣园', '长阳', '朱辛庄']

②分成三类的站名：

x1:['安定门', '安贞门', '北京大学东门', '北新桥', '北苑路北', '菜市口', '崇文门', '慈寿寺', '磁器口', '大屯路东', '大望路', '高碑店', '鼓楼大街', '光熙门', '国展', '和平里北街', '和平门', '和平西桥', '花梨坎', '花园桥', '惠新西街北口', '惠新西街南口', '积水潭', '金台路', '劲松', '牡丹园', '前门', '芍药居', '双井', '四惠', '四惠东', '太阳宫', '陶然亭', '天坛东门', '万寿路', '望京西', '五棵松', '宣武门', '雍和宫', '张自忠路', '长椿街']

x2:['安和桥北', '八宝山', '八角游乐园', '八里桥', '北宫门', '北京南站', '北苑', '草房', '常营', '传媒大学', '次渠', '次渠南', '崔各庄', '褡裢坡', '稻田', '俸伯', '高米店北', '高米店南', '公益西桥', '巩华城', '古城路', '管庄', '广阳城', '果园', '海淀五路居', '后沙峪', '黄村火车站', '黄村西大街', '黄渠', '回龙观', '回龙观东大街', '霍营', '角门西', '经海路', '九棵树', '旧宫', '梨园', '篱笆房', '立水桥', '立水桥南', '良乡大学城', '良乡大学城北', '良乡大学城西', '良乡南关', '林萃桥', '临河里', '刘家窑', '龙泽', '马家堡', '马泉营', '南法信', '南邵', '苹果园', '蒲黄榆', '青年路', '清源路', '沙河', '沙河高教园', '生命科学园', '生物医药基地', '十里堡', '石门', '双桥', '顺义', '宋家庄', '苏庄', '孙河', '天宫院', '天通苑', '天通苑北', '天通苑南', '通州北苑', '土桥', '西二旗', '西红门', '西小口', '西苑', '肖村', '小红门', '新宫', '

义和庄', '亦庄桥', '亦庄文化园', '永泰庄', '玉泉路', '育新', '圆明园', '枣园', '长阳', '朱辛庄']

x3: ['安华桥', '奥林匹克公园', '奥体中心', '巴沟', '白石桥南', '北海北', '北京站', '北土城', '朝阳门', '车公庄', '车公庄西', '大葆台', '大钟寺', '灯市口', '东大桥', '东单', '东四', '东四十条', '东直门', '动物园', '阜成门', '复兴门', '郭公庄', '国家图书馆', '国贸', '海淀黄庄', '呼家楼', '建国门', '健德门', '金台夕照', '军事博物馆', '亮马桥', '灵镜胡同', '柳芳', '木樨地', '南礼士路', '南锣鼓巷', '农业展览馆', '平安里', '人民大学', '荣昌东街', '荣京东街', '三元桥', '森林公园南门', '上地', '苏州街', '天安门东', '天安门西', '同济南路', '团结湖', '万源街', '王府井', '望京', '魏公村', '五道口', '西单', '西四', '西土城', '西直门', '新街口', '永安里', '知春里', '知春路', '中关村']

③分成四类的站名

x1: ['安和桥北', '八宝山', '八角游乐园', '八里桥', '北宫门', '北苑', '草房', '常营', '传媒大学', '次渠', '次渠南', '崔各庄', '褡裢坡', '稻田', '俸伯', '高碑店', '高米店北', '高米店南', '公益西桥', '巩华城', '古城路', '管庄', '广阳城', '国展', '果园', '海淀五路居', '后沙峪', '花梨坎', '黄村火车', '黄村西大街', '黄渠', '回龙观', '回龙观东大街', '霍营', '角门西', '经海路', '九棵树', '旧宫', '梨园', '篱笆房', '立水桥', '立水桥南', '良乡大学城西', '良乡南关', '临河里', '刘家窑', '龙泽', '马家堡', '马泉营', '南法信', '南邵', '苹果园', '蒲黄榆', '青年路', '清源路', '沙河', '沙河高教园', '生命科学园', '生物医药基地', '十里堡', '石门', '双桥', '顺义', '宋家庄', '苏庄', '孙河', '天宫院', '天通苑', '天通苑北', '天通苑南', '通州北苑', '土桥', '西二旗', '西红门', '西小口', '西苑', '肖村', '小红门', '新宫', '义和庄', '亦庄桥', '亦庄文化园', '永泰庄', '玉泉路', '育新', '枣园', '长阳', '朱辛庄']

x2: ['安定门', '安华桥', '安贞门', '巴沟', '白石桥南', '北京站', '北土城', '北新桥', '北苑路北', '朝阳门', '车公庄', '车公庄西', '崇文门', '大葆台', '大屯路东', '大望路', '大钟寺', '灯市口', '东大桥', '东四', '东四十条', '东直门', '动物园', '阜成门', '复兴门', '鼓楼大街', '光熙门', '郭公庄', '国家图书馆', '国贸', '海淀黄庄', '和平里北街', '和平西桥', '呼家楼', '花园桥', '惠新西街北口', '惠新西街南口', '积水潭', '建国门', '健德门', '金台路', '金台夕照', '军事博物馆', '亮马桥', '灵镜胡同', '柳芳', '牡丹园', '木樨地', '南礼士路', '农业展览馆', '平安里', '人民大学', '荣昌东街', '荣京东街', '三元桥', '上地', '芍药居', '双井', '四惠', '四惠东', '苏州街', '太阳宫', '同济南路', '团结湖', '万源街', '魏公村', '五道口', '西四', '西土城', '西直门', '新街口', '雍和宫', '永安里', '张自忠路', '长椿街', '知春里', '知春路', '中关村']

x3: ['奥林匹克公园', '奥体中心', '北海北', '东单', '良乡大学城', '良乡大学城北', '林萃桥', '南锣鼓巷', '前门', '森林公园南门', '天安门东', '天安门西', '王府井', '西单']

x4: ['北京大学东门', '北京南站', '菜市口', '慈寿寺', '磁器口', '和平门', '劲松', '陶然亭', '天坛东门', '万寿路', '望京', '望京西', '五棵松', '宣武门', '圆明园']

5.3 实验结果分析

5.3.1 原始数据分析

分类结果评测显示分类数量越多效果越不好，但结合聚类的结果图来看，分成两类却是不合理的，原因在于两类仅考虑了每个站不同时间段进出站人流的相对多少，例如国贸站

早高峰出站人相对进站人多，而晚高峰则相反；天通苑站早高峰进站人多晚高峰反之。如果单纯看聚类指标的话，分类少的理应有相对更高的内聚和相对低的类间距，但实际情况要考虑的因素很多。例如每个站工作日和休息日的人流量差别，不同时间段的人流差距大小，以及类内进出站人数的方差，来反映波动情况。

实验结果表明分成四类相对而言更合理，下图是四个类别的进出站人流数量。

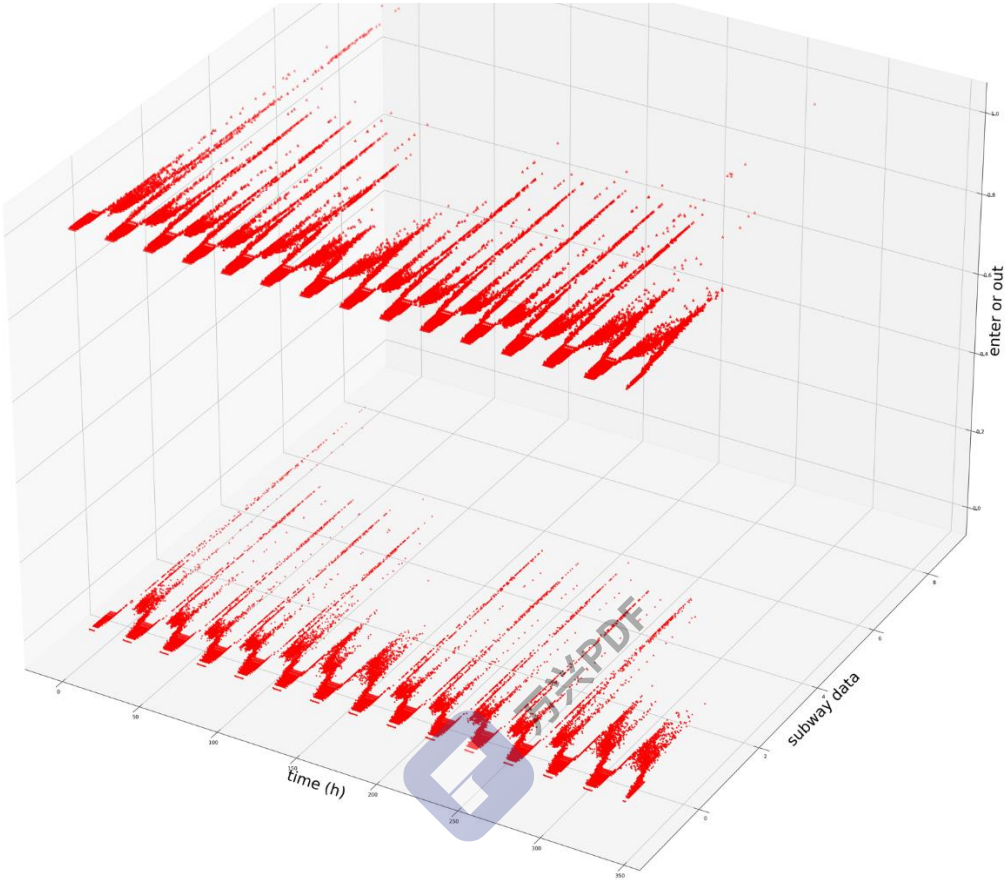

图 5.4 第一类

图 5.4 是第一类的聚类结果，可以看出此类的站（参考 5.2.3 中的③分成四类的站名）大多是一些郊区站点，在三维图中显示出早高峰进站人多，晚高峰进站人少，从图中下部分的进站人中可看出来，且波峰有明显的周期效应，每五天即工作日为一个周期，而大波峰之间的次波峰则为周末的进站人流。上半部分则与之相反，工作日波峰集中部分之前是波谷，表示工作日出站人数在每天的晚上更多，和下部分刚好对应。

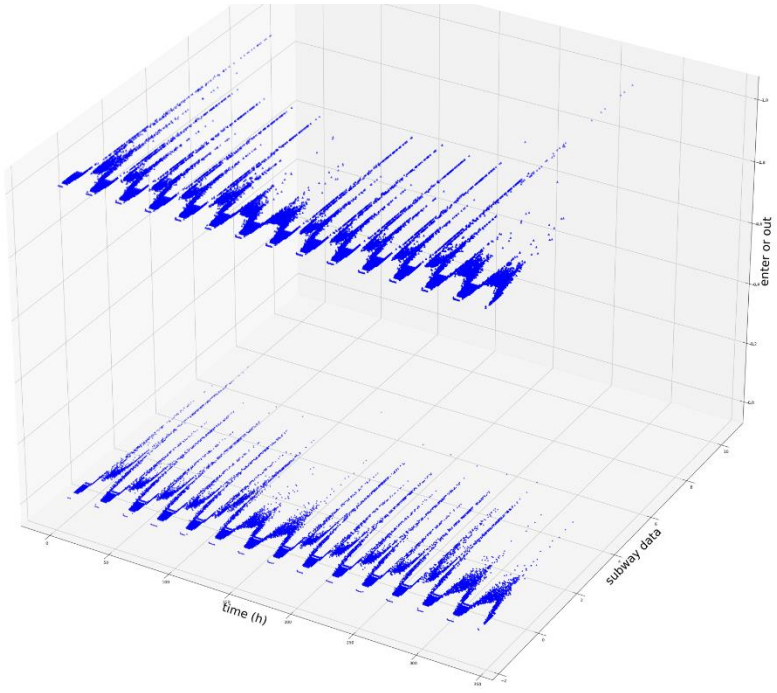

图 5.5 第二类

上图 5.5 是第二类聚类结果，该类的站大多在城区且是工作区附近，并且由图中能看出和 5.4 是完全相反的，工作日出站人数早上多晚上少，进站人数相反。周六日的趋势和上图相同

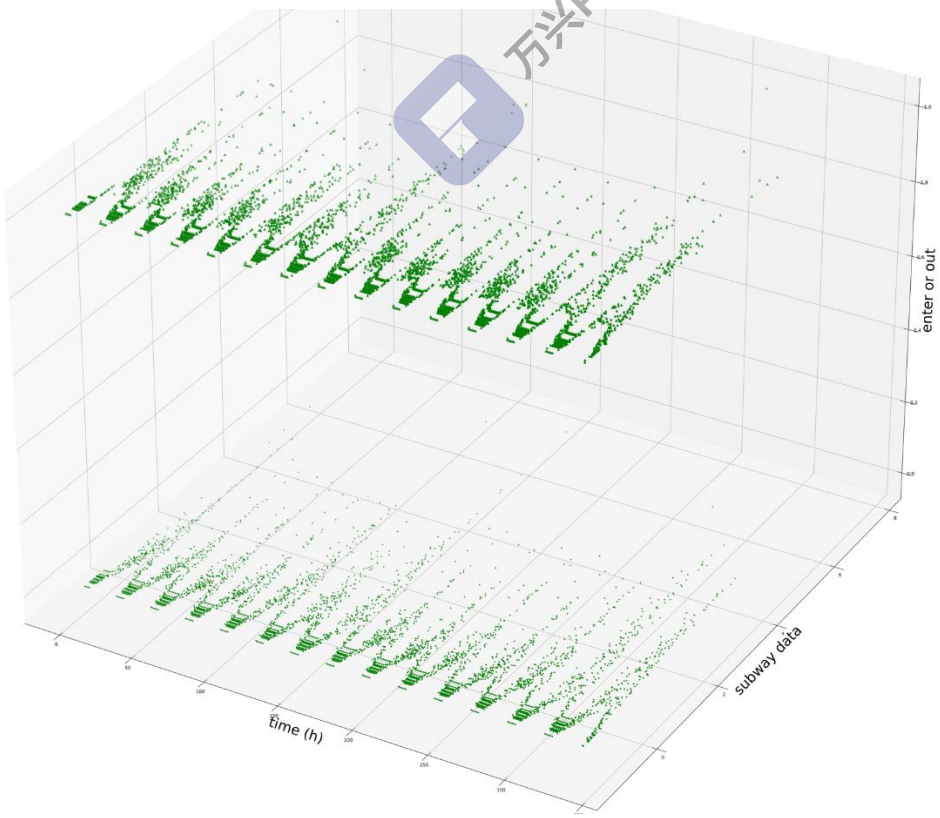

图 5.6 第三类

图 5.6 绿色是第三类聚类结果，此类站点较少，主要是旅游景点，从上图的人流分布上也能看出工作日和休息日的人流区别与前两类不同，且休息日的进出站人数还会稍多于工作

日。波峰每天也只有一个，在中午，不会像前两类在早上和晚上是波峰波谷。

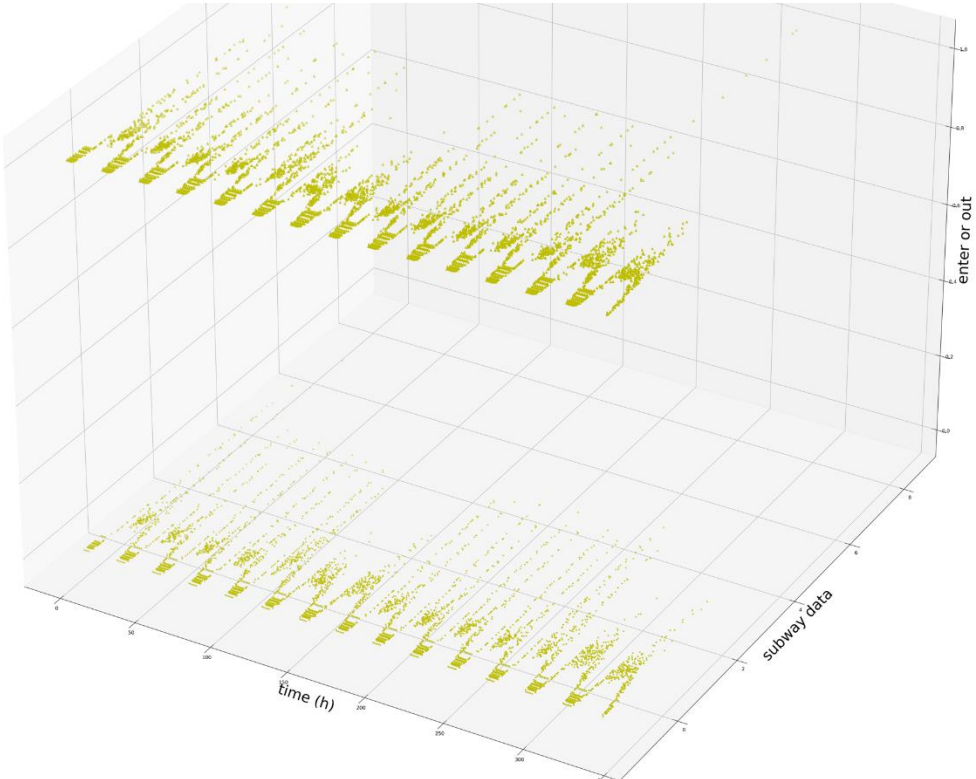

图 5.7 第四类

上图 5.7 是第四类，这类的客流量从 subway data 上能看出，比前几类普遍要少，这些站点属于不论工作日休息日人流都比较少的，说明来这里人较少且这些站附近的人乘坐地铁也较少。

5.3.2 评测结果分析

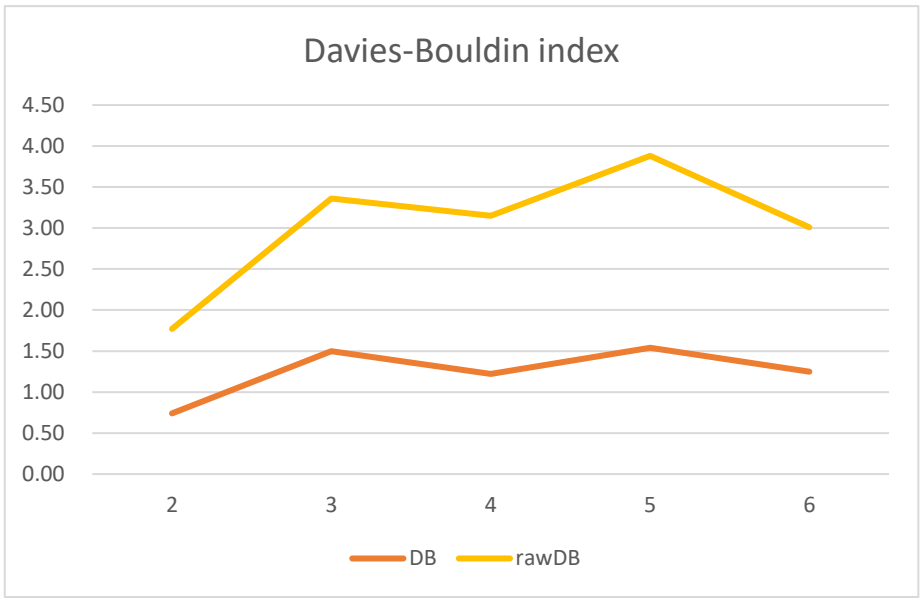

图 5.8 DB

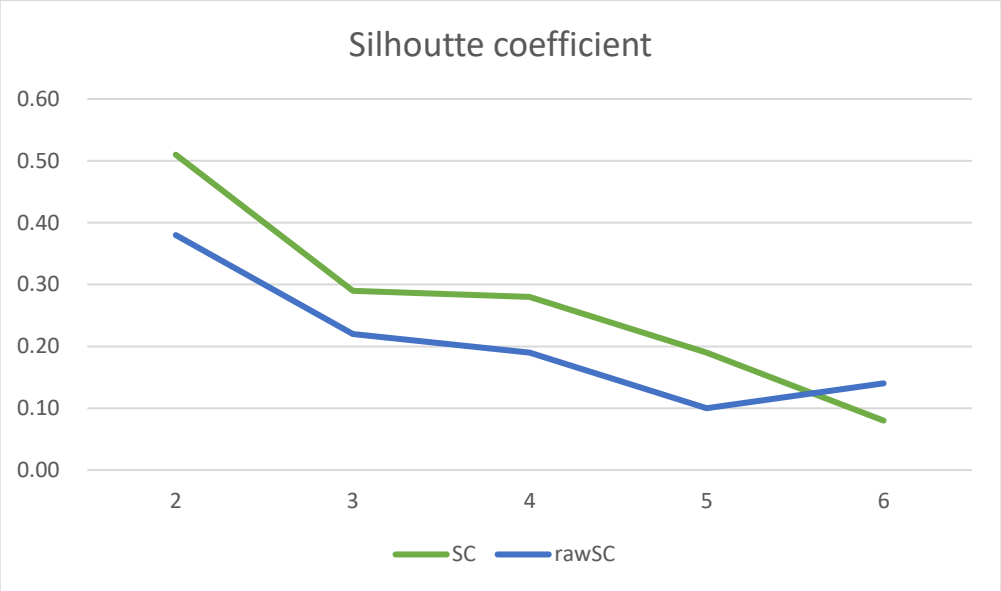

图 5.9 SC

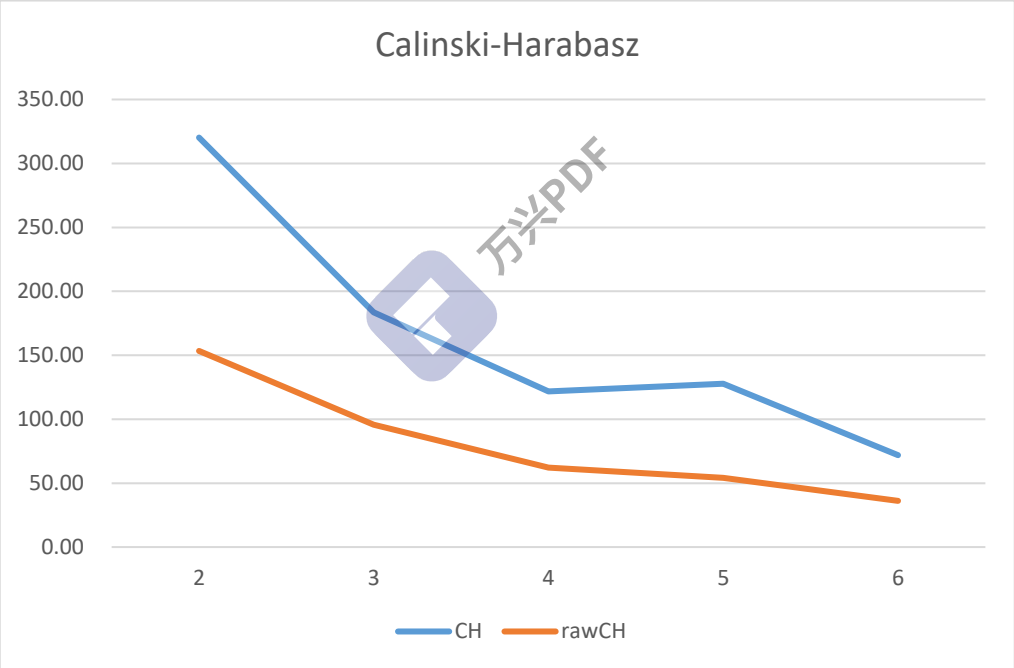

图 5.10 CH

上述三幅图 5.8-5.10 分别是 DBI、SC、CH 三个聚类指标随类数量变化而变化的结果，从中可知分成四类是合理的结论，与 5.3.1 的结果符合。虽然整体的趋势都是随着类别数量增多而聚类效果变差，但可以发现从类数为 2 变到 3 是效果下降最快的，其次变化最大的是类别从 4 变到 5。而当查看聚类数为 5 时的结果，可以发现第五类几乎没有几个站了（图 5.11）。也就是说分成五类会让类间耦合度急剧增加，通过上图三个指标能明显看出这点，DBI 的急剧增加反映了类内距离增大且类间距离减小。CH 反映出类内的离散程度，从 4 类到 5 类也是迅速扩大，而轮廓系数则反映了类内和类间的相似度。综合来看聚类数量为 4 是综合了聚类结果和实际背景的最佳选择。

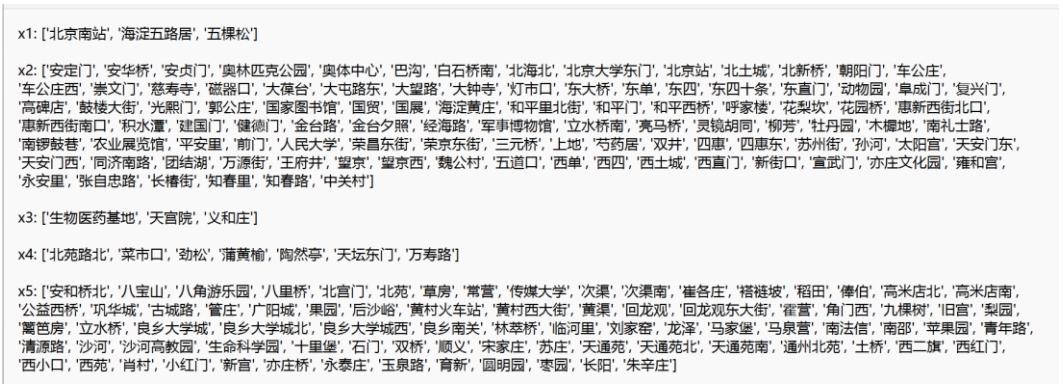

图 5.11 聚类为 5 类的站点

6 总结

本研究针对多元时间序列数据进行降维并考虑节点之间的关联性，即图结构，通过图卷积神经网络模型的算法应用于城市地铁站点分类的问题。训练用的损失函数用 KL 散度作为 loss 函数，并采用双向自监督机制训练 GCN 和 DNN。目标分布 P 是由 DNN 训练的 Q 分布得到，然后通过 target-distribution 再反向监督模型进行学习。最后的输出结果由 GCN 导出，DNN 由于只考虑节点的特征而忽略图结构信息，因此只是用来提取数据特征并作为 GCN 每次前向传播的输入。

将此模型选用北京市公交 IC 卡的轨道交通站点刷卡数据形成的进出站客流量多元时间序列进行验证，对每一类别进站和出站数据合并，作为每个站点的信息。并结合实际的地铁线路图构建图结构然后进行聚类。最终轨道交通站点分为 4 大类，分别命名为办公型、居住型、旅游型和混合型。通过聚类有效性指标轮廓系数，DB 指数以及 CH 指数来评估聚类结果，并结合不同类数量下的数据分布，选定聚类个数。其研究成果也为研究城市功能和轨道交通站点规划设计和管理服务提供了一定的参考科学依据。

参考文献：

1、 Machine Learning (stat.ML) arXiv:1906.06532 [cs.LG]
2、 Machine Learning (stat.ML): arXiv:2002.01633 [cs.LG]
3、 Machine Learning (stat.ML): arXiv:1906.01210 [cs.LG]
4、 Machine Learning (stat.ML): arXiv:2002.08643 [cs.LG]
5、L. Zhang, T. Pei, B. Meng, Y. Lian and Z. Jin, "Two-Phase Multivariate Time Series Clustering to Classify Urban Rail Transit Stations," in IEEE Access, vol. 8, pp. 167998-168007, 2020, doi: 10.1109/ACCESS.2020.3022625.