文章编号：2095-2163(2024)01-0185-06 中图分类号：TP399 文献标志码：A

基于混合采样的城市功能区识别系统研究

吴俊杰，魏山山
（太原师范学院 计算机科学与技术学院，山西 榆次 030619）

摘 要：随着城市功能分区研究在时空尺度上不断细化，多源数据融合有利于推动城市功能分区研究的精细化发展，但多源数据存在数据不平衡现象，本文通过混合采样算法，减少数据不平衡带来的影响。通过空间面积和公众认知度对兴趣点（POI）数据进行权重赋值，利用 K 近邻和潜在狄利克雷（Latent Dirichlet Allocation，LDA）语义分析对两类数据构建数据集，分别对两类数据进行频数密度分析，对多源数据进行加权平均特征融合，将融合后的数据按频数密度差值的方法划分单一和混合功能区，并在 ArcGIS 平台渲染展示，从而实现对城市功能区可视化识别划分。利用高德地图与功能区识别结果进行精度验证，结果表明：应用该方法能快速及有效地识别城市功能区，功能区识别总体精度为 86.6%，证明该方法现实可行，可为城市未来发展规划与管理提供借鉴。

关键词：POI；数据不平衡；频数密度；城市功能区

Research on identification system of urban functional area based on mixed sampling

WU Junjie，WEI Shanshan
（College of Computer Science and Technology，Taiyuan Normal University，Yuci Shanxi 030619，China）

Abstract：With the continuous refinement of urban functional zoning research on the temporal and spatial scale，multi-source data fusion is conducive to promoting the refined development of urban functional zoning research. However，there is data imbalance in multi-source data，and this paper uses hybrid sampling algorithms to reduce the impact of data imbalance. The POI data is weighted by spatial area and public awareness，the two types of data are clustered and analyzed by K-nearest neighbor and LDA semantic analysis，the frequency density analysis of the two types of data is carried out respectively，and then the weighted average feature fusion of multi-source data is carried out，and the fused data is divided into single and mixed functional areas according to the method of frequency density difference，and rendered and displayed on the ArcGIS platform，so as to realize the visual identification and division of urban functional areas. The accuracy verification is carried out by using the Amap and the functional area recognition results. The results show that the application of this method can quickly and effectively identify urban functional areas，and the overall accuracy of functional area recognition is 86.6%，which proves that the proposed method is realistic and feasible and can provide reference for future urban development planning and management.

Key words：POI；data imbalance；frequency density；urban functional areas

0 引 言

随着中国经济的快速发展，中国城镇化的脚步也逐渐加快，城市功能分区及城市空间结构规划变得越来越重要。城市功能的分布和布局是衡量一个城市发展的重要因素，因此城市功能区的划分具有重要意义，同时也对城市功能区的识别提出了更高的要求[1-2]。以往的城市功能区识别方法主要是专家经验和统计调查，人力、物力消耗大。

近年来，基于遥感影像、兴趣点、社交媒体等数据的城市功能区识别方法得到发展，为城市功能区的规划研究与应用提供了新的思路[3]。冯慧芳等[4]利用城市出租车 GPS 轨迹数据和城市 POI 数据采用 Apriori 关联规则算法进行城市功能区识别研究；杨振山等[5]融合手机信令数据和 POI 数据，使用强度的日夜差异和内部功能混杂程度，完成区域主导功能类型判定及功能混合度评价。本研究 POI 数据量大，微博签到数据量小，少数类的数据特征容易被多数类的数据特征所覆盖，而微博签到数据往往又蕴藏着极具价值的特征信息，这种数据不

作者简介：魏山山（1996-），男，硕士研究生，主要研究方向：软件开发。
通讯作者：吴俊杰（1974-），男，硕士，副教授，硕士生导师，主要研究方向：数据挖掘与人工智能处理。Email：964966451@qq.com

收稿日期：2023-07-26

平衡现象导致功能区识别偏向 POI 数据特征。WU J 等[6]采用微博签到和 POI 两类数据融合进行城市功能区的识别,但未考虑真实权重。目前的研究大都是主观赋值,这就导致真实的权重和给定的权重有很大的误差。

本文提出数据不平衡中的混合采样算法,减少数据不平衡的现象;根据功能区空间面积和公众认知度计算权重的方法计算各类 POI 权重,提高城市功能区识别率。

1 研究区域和数据

1.1 研究区域概况

太原市位于山西省中部,是山西省政治、经济、文化和国际交流中心,也是中国重要的能源、重工业基地之一。本文的研究区域是太原市中心主城区,包括小店区、尖草坪区、迎泽区、万柏林区、杏花岭区和晋源区,东中环路、西中环路、南中环街、北中环街 4 条主干道路围成一个闭合的区域,如图 1 所示。

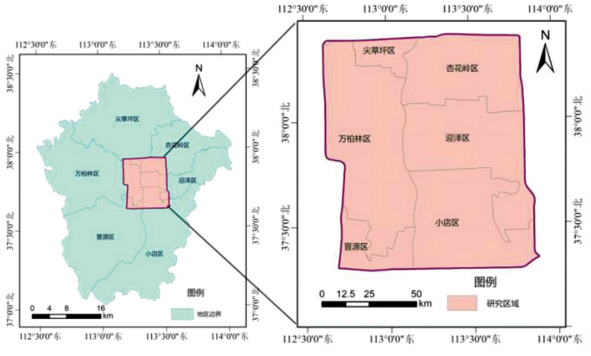

图 1 研究区域

Fig. 1 Study area

1.2 数据来源

使用的 POI 数据从高德地图提供的应用程序编程接口(API)获取,总计 32 787 条数据;通过爬虫新浪微博获取 2021 年 11 月到 12 月的 4 033 条新浪微博签到数据。POI 数据主要包括 4 方面信息:名称、类别、坐标、分类;微博签到数据包含用户名、用户 ID、签到地点、日期、时间、经纬度等数据项。

1.3 数据预处理

首先将获取的 POI 数据进行清洗,剔除和功能区不相关的 POI 数据,如公共厕所、ATM、公交站等无分类意义的 POI。本文基于 GB 501137—2011《城市用地分类与规划建设用地标准》将 POI 数据分为居住用地、公共管理与公共服务用地、商业服务业设施用地、工业用地、交通设施用地、广场绿地 6 大类,见表 1。

表 1 POI 点代表功能分类表

Table 1 POI points represent functional classification table

大类	中类	小类
居住用地	商务住宅	小区、社区等
公共管理与公共服务用地	科教文化	图书馆、博物馆、美术馆、大学、中学、小学、科研机构等
	医疗诊所	医院、药店、诊所等
	政府机关	政府机构、社会团体等
商业服务业设施用地	金融保险	银行、保险公司等
	餐饮	饭店、小吃店等
	购物	商场、便利店、市场等
	宾馆酒店	酒店、宾馆等
工业用地	公司企业	工厂、公司企业等
交通设施用地	交通设施	汽车站、火车站等
绿地广场	风景名胜	景点、纪念馆等
	公园广场	公园、广场等

微博签到数据的预处理主要是对于缺失的信息进行修改,对大量重复的数据进行合并处理,剔除一些没有意义的签到数据。

2 研究方法

2.1 混合采样算法

2.1.1 N-SMOTE 数据增量

SMOTE (Synthetic Minority Oversampling Technique)是一种经典的少数类过采样技术,通过全面布局进行数据扩增。微博签到数据各时间段数量不同,某些时间段签到数据少,如果一味的增加数据会造成数据过拟合。本文针对时间段签到数据的比例进行不同倍率扩增的 N-SMOTE 算法:

(1)将少数类数据通过 K-means 算法分为若干个簇,从每个簇中依次选取样本 x_i 作为根样本,通过欧氏距离计算该样本到本簇其他样本的距离,得到 k 个单位的 K 近邻;

(2)根据时间段不平衡的样本比例计算来设置采样倍率 N_t,以每一簇样本 x_i 为根样本,通过 K 近邻随机选择若干个少数类样本 x_n,N_t 为第 t 个时间段的采样倍率,式(1):

$$N_t = round\left(\frac{C}{Q} \times \frac{Q_{rt}}{Q_r}\right) \tag{1}$$

其中,r 为簇的研究区域;C 表示 POI 的数据量;Q 表示微博签到数据量;Q_r 表示第 r 个簇内签到数据的数据量;Q_{rt} 表示 t 时间段第 r 个簇内签到数据的数据量。

(3)根样本 x_i 与每一个随机选出的近邻 x_n，通过式(2) 进行线性插值，生成新的样本点 x_{new}：

$$x_{new} = x_i + rand(0,1) \times |x_i - x_n| \quad (2)$$

N-SMOTE 算法的基本思想是通过将少数类数据分时间段决定倍率，然后生成新的样本添加到数据集中，如图 2 所示。通过该算法得到了无任何属性的虚拟样本点，本文利用微博签到的真实数据扩充少数类样本。

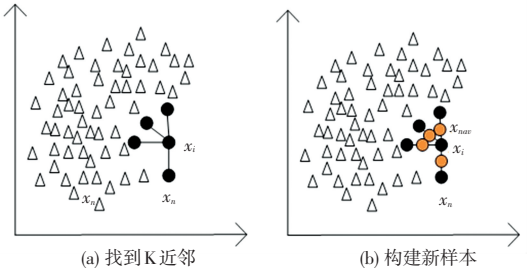

图 2　N-SMOTE 过采样原理图

Fig. 2　N-SMOTE oversampling schematic

2.1.2　ENN 欠采样

编辑最近邻(Edited Nearest Neighbor，ENN)是一种欠采样算法。首先搜索多数类样本的 3 个最近邻样本，若该样本的 3 个最近邻样本中有两个或以上和该样本类别不一样，则删除这个样本，此算法意在删除多数类样本，然而多数类样本附近往往都是多数类样本，因此 ENN 去掉的样本非常有限。本文以少数类为样本点，搜索少数类的 3 个最近邻样本，若该样本的 3 个邻样本中有两个及两个以上的多数类，则删除多数类，使数据趋于平衡。

2.2　POI 权重赋值方法

每个 POI 点都代表一个地理实体，不同的 POI 点代表的地理实体也不相同，因此需要对各类 POI 进行分类赋权，本文引入空间面积权重与公众认知度两种评价方式，经过调和得到最终的 POI 权重。

2.2.1　空间面积

POI 占地面积根据《中国现行的业态分类标准》(GB/T18106-2010) 中明确的 POI 类别建筑面积，如大型商场、综合医院等设施，通过同类对比，推算其他类别 POI 的建筑面积。本文通过计算小类面积得出中类平均面积。M 为某中类面积，式(3)：

$$M = \frac{\sum_{i=1}^{n} m_i \cdot f_i}{N} \quad (3)$$

其中，m_i 为 i 小类面积；f_i 为 i 小类数量；n 为小类的类别；N 为该中类所有数量。

本文将中类数据采用分级打分的方式进行评分，通过阅读文献确定各类 POI 点的一般面积评分见表 2。

表 2　一般面积评分

Table 2　General area score

面积/m²	用地类型	评分/分
1~499	餐饮、购物、政府机关	4
500~999	宾馆酒店、金融保险、公司企业	10
1 000~4 999	商务住宅、医疗诊所	40
5 000~9 999	科教文化、交通设施	80
≥1 0000	风景名胜、公园广场	100

2.2.2　公众认知度

不同类型的 POI 不仅具有不同的空间面积属性，还应在公众认知度和城市地标影响力方面具有不同的显著程度[7]。基于公众认知表格，见表 3，用城市地标影响力进行修正。

表 3　公众认知度

Table 3　Public awareness

POI 中类	公众认知度	POI 中类	公众认知度
商务住宅	15	交通设施	100
政府机关	35	风景名胜、公园广场	82
科教文化	67	宾馆酒店	55
医疗诊所	50	餐饮、购物	68
金融保险、公司企业	30		

2.2.3　权重调和

本文将权重比例设为 5∶5。通过权重分配原则计算得出不同大类 POI 权重：居住用地权重为 35、公共管理与公共服务用地权重为 40、商业服务业设施用地权重为 20、工业用地权重为 20、交通设施用地权重为 80、广场绿地权重为 100。

2.3　构建数据集

K 近邻算法是一种监督学习的机器学习分类算法，通过欧式距离计算样本最邻近的 k 个实例，按照已知样本类别判断新样本的类别。本文将处理后的 POI 数据利用 K 近邻算法找到每个与 POI 相邻近的数据，将得到的数据类型组成训练对。

LDA 主题模型作为一种概率模型具有潜在语义挖掘和主题提取能力。本文将微博数据利用 LDA 主题模型构建数据集，通过对微博签到数据的主题模型提取分析，判断每个区域的主要功能。困惑度是一种常用的衡量概率分布或概率模型的预测结果与文本之间拟合程度的指标，困惑度越低则表明文本的预测越准确。根据对实际数据进行困惑度指标测试，结果如图 3 所示，可见当主题个数为 10 时，模型的困惑度最低，因此本文采用 10 个主题进行分析。

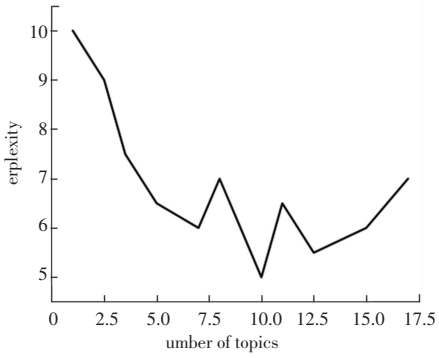

图 3 困惑度

Fig. 3 Perplexity

2.4 功能区识别

将 POI 和微博签到数据构建的数据集分别进行频数密度分析（FD），得到两类数据的特征向量，FD_i 为第 i 种类型的频数密度占该研究区域所有类型频数密度的比例，式(4)：

$$FD_i = \frac{n_i}{\sum_{i=1}^{6} n_i} \tag{4}$$

其中，i 为功能区类型，n_i 为研究区域 i 类型数量占该类型所有研究区域数量的频数密度。

不同功能区的分布情况如图 4 所示。

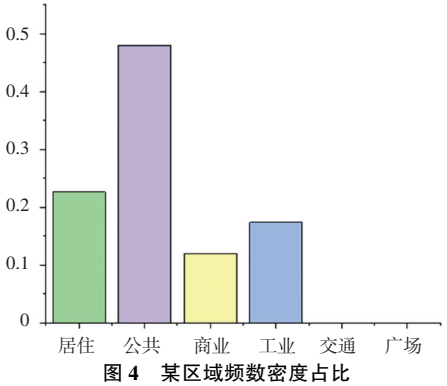

图 4 某区域频数密度占比

Fig. 4 The proportion of frequency density in a certain area

本文以 POI 特征向量和微博签到数据特征向量为数据支撑，采用加权平均的方法进行融合，即加权平均后的频数密度值，某一块区域第 i 类的特征向量值 $\boldsymbol{G}_i$，式(5)：

$$\boldsymbol{G}_i = \frac{\boldsymbol{p}_i + \boldsymbol{w}_i}{n} \tag{5}$$

式中：$\boldsymbol{p}_i$ 表示 POI 第 i 类的特征向量值，$\boldsymbol{w}_i$ 表示微博签到第 i 类的特征向量值，n 表示每种类型特征向量值个数，即为 2。

根据计算得出每一块区域频数密度比例，本文采用频数密度差值来作为功能区划分的依据，每块研究区域内频数密度数值最大的（记为 A）和第二大的（记为 B）相减，如果数值大于 20%，则判定该区域为单一功能区（A），否则判定该区域为混合功能区（$A-B$）。判断过程如图 5 所示。

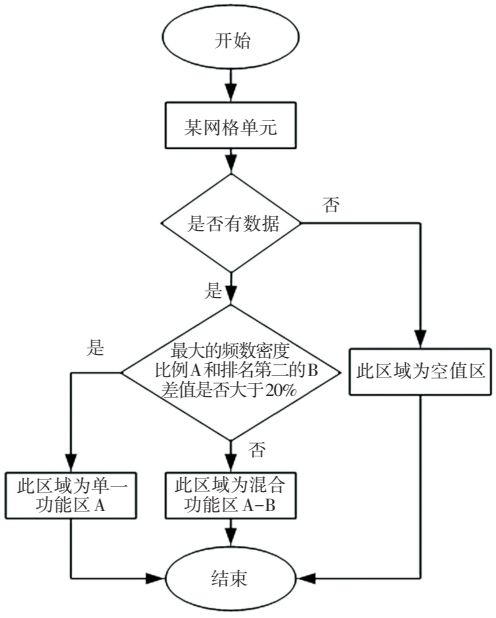

图 5 功能区判断过程

Fig. 5 Functional area judgment process

3 实验验证

3.1 验证方法

本文采用具有不同数据特点的多源数据进行实验对比，分别为未使用混合算法的多源数据，未使用混合算法但使用多因素加权法的多源数据，使用混合算法且使用多因素加权法的多源数据，见表 4。

表 4 实验方法概述

Table 4 Overview of experimental methods

方案	数据特点	方法描述
1	未使用混合采样算法的多源数据	使用 POI 和微博签到数据，分别进行机器学习和自然语言处理，最后数据融合识别城市功能区。
2	未使用混合采样算法但使用多因素加权赋值法的多源数据	使用 POI 和微博签到数据，计算权重给 POI 赋权，最后数据融合识别城市功能区。
3	使用混合采样算法且使用多因素加权赋值法的多源数据	使用 POI 和微博签到数据，使用混合采样算法来减少数据不平衡，计算权重，最后数据融合识别城市功能区。

本文将得到的城市功能区识别划分结果中的20 个街道与现实街区土地使用情况进行对比，验证其总体准确率。在打分评价中，满分为 3 分，即完全符合，2 分为较符合，1 分为较不符合，0 分为完全不符合。计算总体识别精确度式(6)：

$$a = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} X_i} \times 100\% \qquad (6)$$

其中，n 为街区数；X_i 为街区符合度的满分；x_i 为街区符合度的实际得分。

3.2　功能区识别结果

采用识别精确度对 3 种方法进行识别验证，3 种方案的功能区识别精确度见表 5。由表 5 可知，多源数据未使用混合采样算法功能区识别精度最低，为 80.6%；多源数据经过对 POI 的权重计算并赋值使识别精确度提高到 82.7%，证明了赋值权重对城市功能区识别的重要性；本文混合采样算法减少数据不平衡的现象，且通过计算权重实现城市功能区识别，精确度提高至 86.6%。

表 5　功能区识别精确度

Table 5　Functional area identification accuracy

方案	特点	功能区识别精确度
1	多源数据未使用混合采样算法	80.6%
2	多源数据未使用混合采样算法且使用多因素加权赋值法	82.7%
3	多源数据使用混合采样算法且使用多因素加权赋值法	86.6%

本文利用 Open Street Map(OSM)路网将研究区域划分为1 650块研究区域，根据上述方案 3 得到太原市主城区功能区识别图如图 6 所示，利用 ArcGIS 平台展示所得共 18 种功能区，包括 6 种单一功能区和 12 种混合功能区。

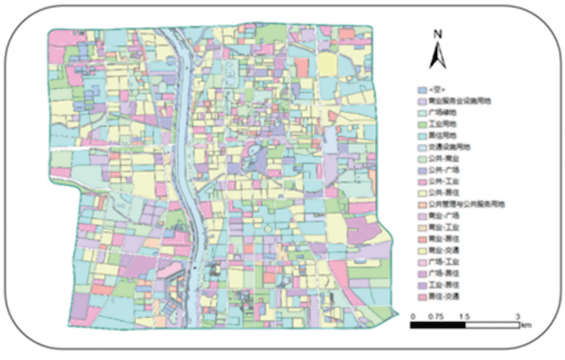

图 6　太原市主城区功能区分布图

Fig. 6　Distribution map of functional areas in the main urban area of Taiyuan City

3.3　结果分析

为了验证本文功能区识别的准确性，从识别图任意选取几处区域和高德地图对比分析，如图 7~图 9 所示。图 7 的 A 区域位于太原市杏花岭区，包含太原市第十九中学校和金刚堰十三冶小区为公共管理与公共服务用地-居住用地相关的区域，符合识别结果；图 8 的 B 区域位于太原市迎泽区，包含新月大厦、亚原大厦、中国工商银行等商业设施服务用地，因此该区域为商业设施服务用地，符合识别结果；图 9 的 C 区域位于太原市万柏林区，包含华夏银行、悦宾酒店、易捷便利店、华英图书广告等商业设施服务用地，还包括青年园，所以该区域为商业设施服务用地-广场绿地相关区域，符合识别结果。

图 7　区域 A 高德地图与功能区识别结果对比

Fig. 7　Comparison of the Gaode map of Area A and the identification results of the functional area

图 8　区域 B 高德地图与功能区识别结果对比

Fig. 8　Comparison of the Gaode map of Area B and the identification results of the functional area

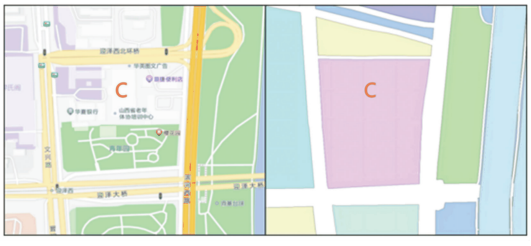

图 9　区域 C 高德地图与功能区识别结果对比

Fig. 9　Comparison of the Gaode map of Area C and the identification results of the functional area

4 结束语

本文以太原市主城区 POI 数据和微博签到数据为基础,根据 OSM 路网图进行拓扑处理构建单元网格,提出了一种基于混合采样算法的城市功能区识别方法,识别出 6 类单一功能区和 12 类混合功能区;采用混合采样算法减少多源数据融合不平衡的问题,通过 POI 多影响因素权重赋值的方法进行计算,最终对识别结果进行了精度验证,验证了该方案的可行性。

本文从 POI 大数据及微博签到数据的视角对城市功能区识别研究效果明显但也存在不足,该方法对城市核心区和 POI 数据多的区域划分很有效,但对于郊区或者 POI 数据稀少的区域识别不高,这也是后期进一步研究的方向。

参考文献

[1] HU Y, HAN Y. Identification of urban functional areas based on POI data: A case study of the guangzhou economic and technological development zone[J]. Sustainability, 2019, 11(5): 1385.

[2] ZHAO Haiyun, ZHANG Xinyuan, WU Bin, et al. Method of city complex function area division based on apportion of shared construction area[J]. Bulletin of Surveying and Mapping, 2017(10):89-93.

[3] 甄茂成,党安荣,许剑. 大数据在城市规划中的应用研究综述[J]. 地理信息世界,2019,26(1):6-12,24.

[4] 冯慧芳,杨文亮. 融合 GPS 轨迹和 POI 数据关联规则的城市功能区识别[J]. 测绘科学技术学报,2020,37(4):414-420.

[5] 杨振山,苏锦华,杨航,等. 基于多源数据的城市功能区精细化研究——以北京为例[J]. 地理研究,2021,40(2):477-494.

[6] WU J, ZHANG J, ZHANG H. Urban Functional area recognition based on unbalanced clustering[J]. Mathematical Problems in Engineering, 2022 (4):1-13.

[7] ZHAO Weifeng, LI Qingquan, LI Bijun. Extracting hierarchical landmarks from urban POI data[J]. Journal of Remote Sensing, 2011,15(5):973-988.

(上接第 184 页)

任务,首先对驾驶员感兴趣区域建模,用其对感知模块输入的信息进行筛选,以减少对智能车辆行为决策无用的信息,降低强化学习环境状态空间的维度,并将其融入基于 PPO 强化学习行为决策模型中;为了充分考虑周边车辆的交互关系,将注意力机制加入策略网络和价值网络中,并从车辆的安全、高效、舒适三方面设计奖励函数,以引导智能车辆的学习方向。仿真实验证明,本文提出的基于 PPO 强化学习决策框架与其他传统的强化学习决策框架相比,在训练时收敛速度更快,获得的最终收敛奖励更高,测试时通行成功率和速度更高。本文主要关注在单车道的十字路口下智能车辆无保护左转问题,在进一步的研究中,将重点关注多车道的十字路口驾驶场景。

参考文献

[1] JUNIOR. The stanford entry in the urban challenge[J]. Journal of Field Robotics, 2008, 25(9):569-597.

[2] FURDA A , VLACIC L. Enabling safe autonomous driving in real-world city traffic using multiple criteria decision making[J]. IEEE Intelligent Transportation Systems Magazine, 2011, 3(1): 4-17.

[3] 杜明博. 基于人类驾驶行为的无人驾驶车辆行为决策与运动规划方法研究[D]. 合肥: 中国科学技术大学, 2016.

[4] HOUL, DUAN J, WANG W, et al. Drivers′ braking behaviors in different motion patterns of vehicle-bicycle conflicts[J]. Journal of Advanced Transportation, 2019, 2019(PT.1):1-17.

[5] BOJARSKI M, DEL TESTA D,DWORAKOWSKI D, et al. End to end learning for self-driving cars[J]. arXiv preprint arXiv: 1604.07316, 2016.

[6] YANG Z, ZHANG Y, YU J, et al. End-to-end multi-modal multi-task vehicle control for self-driving cars with visual perceptions[C]//2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018: 2289-2294.

[7] KENDALL A, HAWKE J, JANZ D, et al. Learning to drive in a day [C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 8248-8254.

[8] HOEL C J, WOLFF K, LAINE L. Automated speed and lane change decision making using deep reinforcement learning[C]// 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 2148-2155.

[9] ISELE D, RAHIMI R, COSGUN A, et al. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 2034-2039.

[10] TRAM T, BATKOVIC I, ALI M, et al. Learning when to drive in intersections by combining reinforcement learning and model predictive control [C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019: 3263-3268.

[11] KOLEKAR S, DE WINTER J, ABBINK D. Which parts of the road guide obstacle avoidance? Quantifying the driver′s risk field [J]. Applied Ergonomics, 2020, 89: 103196.