BigSUMO：用于大数据交通分析和并行仿真的可扩展框架

Rahul Sengupta佛罗里达大学美国盖恩斯维尔

Nooshin Yousefzadeh 佛罗里达大学 盖恩斯维尔，美国

Manav Sanghvi 佛罗里达大学 盖恩斯维尔，美国

Yash Ranjan 佛罗里达大学 盖恩斯维尔，美国

Anand Rangarajan 佛罗里达大学 盖恩斯维尔，美国

Sanjay Ranka 佛罗里达大学 盖恩斯维尔，美国

Yashaswi Karnati 英伟达公司 圣克拉拉 ，美国

Jeremy Dilmore *FDOT* 德兰 ，美国

Tushar Patel *FDOT* 德兰 ，美国

Ryan Casburn *FDOT* 德兰 ，美国

arXiv:2601.02286v1 [cs.DC] 5 Jan 2026

摘要——**随着全球城市化进程不断加快，高效管理交通基础设施对交通机构和城市规划者至关重要。我们需要能够帮助分析海量已存储交通数据并采取有效干预措施的工具。为满足这一需求，我们推出了“BigSUMO”，这是一个端到端、可扩展的开源框架，用于分析、中断检测和并行交通仿真。该系统可采集高分辨率环形检测器数据和信号状态数据，以及稀疏的探针轨迹数据。首先，系统执行描述性分析并检测潜在中断；随后，利用SUMO微观交通仿真器进行规范性分析，通过测试数百种假设情景来优化交通性能。模块化设计支持集成不同的数据处理和异常值检测算法。该流程基于开源软件和库构建，具有成本效益高、可扩展且易于部署的特点。我们希望BigSUMO能为开发智慧城市出行解决方案提供有力帮助。**

关键词—**智能交通系统、机器学习、交通运输、仿真、建模**

一、引言与相关工作

随着交通数据采集传感器日益普及，智能交通系统（ITS）正在改变我们管理和理解城市交通动态的方式。ITS 将环形线圈检测器、GPS、车联网系统、视频、激光雷达等多模态传感器等先进技术集成到交通基础设施中。

查阅已发表的研究后，我们发现有一些分析框架使用不同的数据模态，针对特定问题开展研究。 [1], [2], [3] 和 [4] 讨论了端到端软件流水线，用于处理交叉口的视频，并着重呈现安全分析。 [5] 介绍了一种在交叉口部署的基于环形检测器的中断检测系统，该系统依据历史环形检测器流量检测交通中断。

可通过 rahulseng@ufl.edu 联系 Rahul Sengupta。作者感谢佛罗里达州交通部（FDOT）提供支持。版权声明：979-8-3315-7391-1/25/$31.00 ©2025 IEEE

然而，我们发现，尚缺乏一种可扩展的端到端系统，能够在全市范围内处理 ATSPM 和轨迹数据，并使用免费开源组件提供描述性与规范性分析。在本研究中，我们提出“BigSUMO：一种用于大数据交通分析与并行仿真的可扩展框架”，以弥补这一不足。我们的贡献包括：

1）我们提出一个端到端系统，利用 ATSPM 和稀疏探针轨迹数据进行分析与仿真。2）以表格和图形形式呈现描述性交通分析，并通过交互式 Jupyter 笔记本提供，用户可根据需要扩展或自定义。3）基于运行数百次经过校准的 SUMO 仿真结果开展规范性交通分析，以帮助研究反事实场景并优化交通性能。4）我们采用可在任何运行 Python 的云平台上使用的开源软件组件。我们还使用公共交通管理机构发布的开放地理信息系统数据，例如佛罗里达州交通部（FDOT）发布的数据。5）该系统具备并行性和可扩展性。系统设计力求最大限度减少对硬件和软件的依赖，交通管理机构可借助商用云服务轻松部署。

尽管 ATSPM数据通常十分丰富，但我们承认，当前数据中网联车辆渗透率较低（占总交通量的 3% 至 7%）仍是一项局限。鉴于 V2V/V2X 基础设施部署不断增加，我们预计网联车辆渗透率将在不久的将来上升。

II. 数据处理流水线

获取的 ATSPM数据已采用结构化表格格式，只需进行少量预处理。我们根据所需的交叉口和时间范围筛选 ATSPM数据。

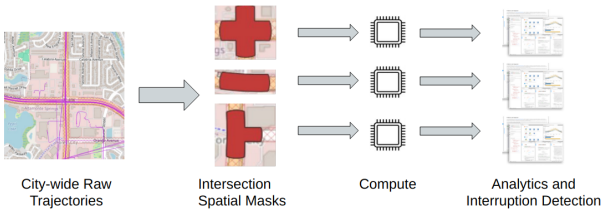

图1. 交　叉口掩膜用于分析和中断检　　测

另一方面，探针轨迹数据需要经过处理。我们的轨迹数据处理流水线可直接读取以 Parquet 或类似云存储格式存储的原始数据集。该数据集由一家商业供应商提供，每天包含 17,000 至 22,000 条唯一行程，覆盖美国佛罗里达州的 3 个县组成的广大地理区域。数据采集时间为 2024年，为期两个月。数据的时间分辨率较高，为 3 秒，空间精度约为 1 米。需要注意的是，渗透率介于 3% 至 7% 之间。这反映了供应商系统采集到的车辆占当时当地道路上所有车辆的比例。所采集的数据来自普通乘用车，而非网约车/出租车服务。

轨迹本身起始和终止于停车场、住宅区和商业区等地点。这些区域中的驾驶行为与主干道上的驾驶行为大不相同。就本研究而言，我们只关注已由佛罗里达州交通部（FDOT）编目并收录于SunStore数据库中的道路。

A. 掩膜生成过程

为了将数据处理集中于相关部分，我们需要使用道路“掩膜”来裁剪轨迹。由于手动绘制掩膜十分繁琐，下面介绍一种用于生成掩膜的半自动化流程。

首先，我们从 SunStore 下载“Basemap Route Roads”数据集，其中包含道路的地理信息系统（GIS）数据。该数据集以 M值多段线格式表示道路的几何形状，可供地理信息系统软件使用。我们使用 QGIS 加载并处理该地理信息系统数据集。

加载后，我们使用“缓冲区”功能，在表示道路两个行驶方向的两条线周围创建宽 35 米的缓冲区。随后合并缓冲区边缘，形成一个连续且不重叠的缓冲区，其边界距离两条道路中心线均为 35 米。

接下来，我们使用 SunStore 中的“交叉口”数据集，获取交叉口的中心位置。基于这些位置，我们围绕每个交叉口中心创建半径为 125 米的圆形缓冲区。我们假设交叉口宽度为 50 米，并在各侧额外延伸 100 米，以覆盖直至高级环形检测器的区域（即总直径为 125x2=250 米）。然后，我们使用“裁剪”功能将先前生成的道路缓冲区与交叉口缓冲区重叠的部分裁剪出来。

B. 原始轨迹数据预处理

原始轨迹数据按日期和时间分散存储在多个目录中的多个文件里。我们将这些文件合并到一个数据框中，以便进一步处理。随后，我们按时间戳对数据框建立索引并排序。

接着，我们筛除无关和不可靠的数据点。对于车辆点火状态显示车辆已熄火的行，我们会将其删除，因为这类点通常出现在停车场、住宅区和商业综合体中，很少出现在主干道上。我们还会删除不满足最短持续时间阈值（2 分钟）的行程。

随后，使用空间掩膜（前文已介绍）将轨迹裁剪至特定的感兴趣地理区域。这一步将分析聚焦于相关区域。最后，我们移除长度短于指定最短距离（150 米）的轨迹。

我们使用 GeoPandas 和 MovingPandas 库执行此操作。

III. 交通状态的城市尺度描述性分析

本节介绍基于裁剪后的轨迹计算的各种指标。随后，我们使用图表将这些指标可视化，以获得更深入的认识。我们重点关注交叉口级指标，因为交叉口是各种轨迹发生冲突的区域，并由交通信号灯进行调节。图 1 概述了这一过程。

分析代码使用 Pandas 构建，并在 Jupyter 笔记本中展示。该代码可用于所需时间段内任意感兴趣的交叉口。我们展示了对工作日晚上高峰期间在某交叉口采集的一小时数据的分析。

A. 等待时间与位置

车辆通常需要在交叉口等待红灯，也会在接近队列末端时停车。

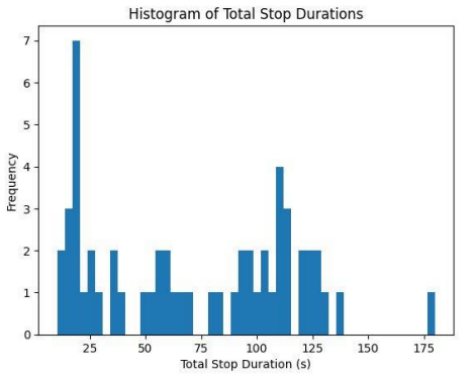

图 2. 车辆停车时间直方图。

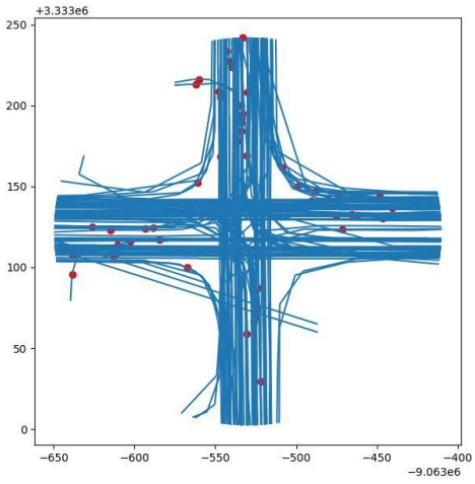

图3. 车辆停车位置图。请注意，从左侧向右下方移动的两条平行轨迹。出于隐私保护，供应商隐藏了中间轨迹点，但这些轨迹在其他方面都是有效的。我们在本分析的图中保留了这些轨迹；如有需要，也可以将其移除。

图 2 展示了车辆停车时间的直方图。这些信息有助于开展信号配时优化，因为等待时间越短越理想。可以看到，停车时间通常不超过 150 秒，与此类交叉口的周期长度大致相当。图 3 显示，被跟踪的车辆通常会在停车线前较远处停下。鉴于数据较为稀疏，可以推断，未被跟踪的车辆（位于被跟踪车辆前方）是响应交通信号灯而停车的。

B. 转向交通流分析

拥有轨迹数据的一个重要好处是能够捕捉交叉口处的转向交通流。这使我们能够根据车辆来自哪个进口道以及在交叉口采取了何种转向，对轨迹进行分组。停车线环形检测器可能无法完整捕捉这些信息，尤其是在某条车道允许多种转向（例如“直行-右转”）时。

图4展示了根据轨迹来源进口道进行颜色编码的轨迹图。矩形框表示进入车道组的位置。此外，图5以起讫点矩阵的形式对数据进行了制表。这是运行单个交叉口仿真时的一项重要输入 [6], ，因为它能够让仿真器为该交叉口生成交通流。图6展示了基于起点和终点的平均行程时间。该表为确定哪些行驶方向需要改进提供了有价值的参考。

C. 排队

根据从轨迹中计算出的停车信息，可以推断车辆从特定方向驶近交叉口时可能经历的平均排队长度。我们查找持续时间超过 10 秒的停车，以确定车辆从何处进入队列。

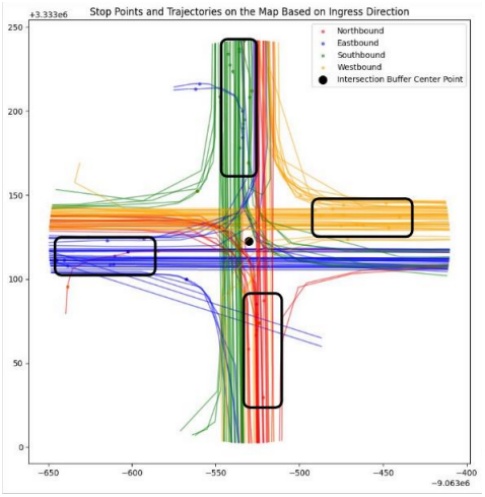

图 4. 各进口道的轨迹和停车位置。

egress_direction ingress_direction	Eastbound	Northbound	Southbound	Westbound
Eastbound	28	2	4	4
Northbound	8	10	0	12
Southbound	8	1	24	5
Westbound	3	12	5	32

图 5. 各进口道的起讫点矩阵。

egress_direction ingress_direction	Eastbound	Northbound	Southbound	Westbound
Eastbound	35.6	54.2	22.4	48.5
Northbound	51.0	22.8	NaN	59.1
Southbound	74.7	121.3	49.3	17.1
Westbound	15.5	25.7	37.1	46.0

图 6. 各进口道平均行程时间表（秒）

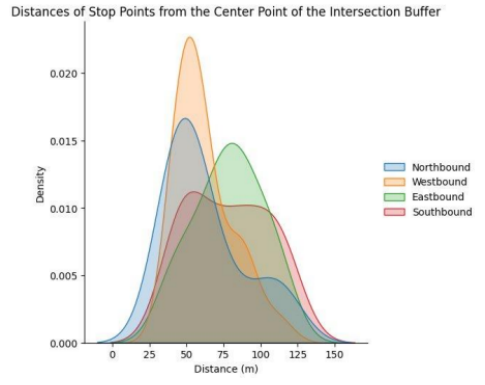

图 7. 各进口道排队长度（米）的概率分布。

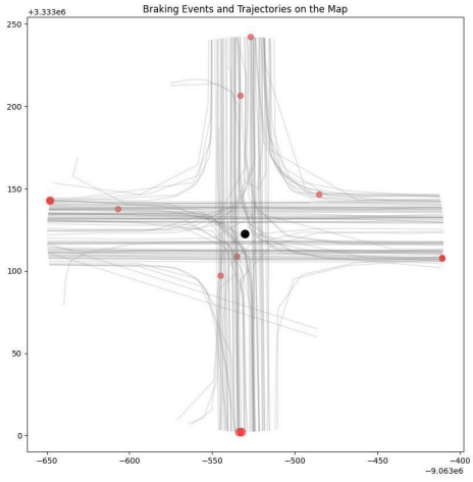

图8. 值得关注的制动事件发生的位置

。

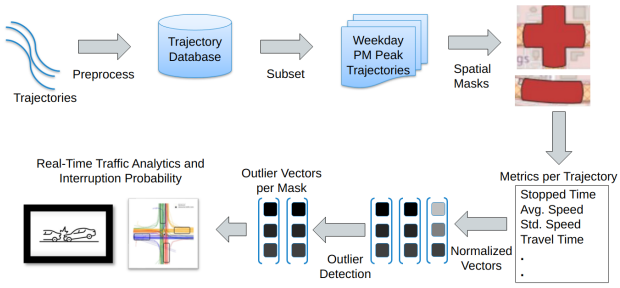

图 9. 中断检测系统概览。

利用这些数据，我们可以拟合一个正态分布，从而粗略估计车辆沿特定方向驶近交叉口时排队长度的概率分布。分布如图 7 所示。

结果显示，东向和南向的停车位置比北向和西向更远。

D. 值得关注的制动行为

由于交叉口是轨迹交汇的区域，因此可能会出现危险情况，例如险些碰撞。与交通碰撞不同，这类险些碰撞通常不会记录在交通记录中。不过，我们可以利用轨迹数据查找值得关注的制动行为。我们采用持续 2 秒、阈值为 [7] –0.47g（其中 g 为重力加速度）的制动阈值，以确定此类制动发生的位置。

图8展示了值得关注的刹车事件发生位置。由于数据较为稀疏，点数很少。不过，这些点仍能大致显示此类刹车事件发生的位置。

IV. 中断检测框架

BigSUMO整合了两个用于检测交通中断的模块，一个使用ATSPM数据，另一个使用轨迹分析。

第一个基于ATSPM数据的中断检测系统由作者在此前的一项工作中开发 [5]„ 该系统仅依赖环形检测器数据（来自 ATSPM）。这个基于环形检测器的系统通过使用一天中特定时段的细粒度分相位车辆计数，并将其与前一周及再前一周同一时段的数据进行比较来运行。这样，系统便能学习一周中这一天、这一时段通常会出现的总体模式。

现在我们介绍第二个系统，该系统使用轨迹数据。如前文所述，预处理后的轨迹存储在数据库中，并生成空间掩膜。从数据库中提取所需的轨迹数据子集，用于异常值检测。我们重点关注工作日傍晚高峰时段记录的轨迹；由于车辆流量大，这些轨迹对交通分析至关重要。我们按照前文所述，对选定的轨迹子集应用空间掩膜。针对空间掩膜内的每条轨迹，计算各项指标。重要指标包括停车时间、平均速度、速度变化（标准差）和行程时间。这些指标提供了定量测量，用于表征轨迹的行为。将计算出的指标串联起来并进行归一化，从而生成向量。因此，每条轨迹都由一个向量表示。此外，我们还针对相同的空间掩膜，在相同的时段和星期几，对一周前和两周前的数据重复这一过程。这样便可以结合上下文，推断轨迹的“正常行为”是什么样的。

异常值检测算法会评估归一化向量，以识别明显偏离正常行为的轨迹。我们使用基于角度的异常值检测（ABOD）算法 [8]。该算法通过考察数据对象的差向量之间夹角的方差来进行检测。该算法对于高维数据而言快速且有效。我们指定算法在计算角度相似性时考虑 10 个最近邻。

我们使用 PyOD 库进行异常值检测。代码足够灵活，既可以使用不同特征生成向量，也可以使用不同的异常值检测算法。

从使用空间掩膜裁剪轨迹、预处理、生成分析 Jupyter 笔记本，到向量化并执行异常值检测，整个端到端流程在每个交叉口耗时~2-3 分钟（实际运行时间）。该流程可以通过诸如“multiprocessing”的库在多个交叉口之间并行化。这样便能使用配备多线程的通用 CPU，为多个交叉口生成分析报告并进行中断检测。

这些离群点可能表示交通中断或异常事件。聚类后的离群点数据可用于后续任务，例如生成实时交通中断概率热力图。该热力图直观呈现不同区域发生中断的可能性，帮助交通管理部门及时发现并应对。图 9 展示了该框架的概览。

因此，所提出的系统具有模块化和可扩展性。它可并行用于多个交叉口，也可采用不同的异常值检测算法。目前，基于 ATSPM 的

中断检测系统比基于轨迹的系统可靠得多。未来，随着网联车辆的普及率不断提高，我们也可以期待基于轨迹的系统取得更好的性能。

V. 使用 SUMO微观仿真的规范性分析

本节介绍规范性分析模块的用法。该模块可在一系列不同的交通参数下并行运行经过校准的 SUMO 仿真，并帮助我们确定改善交通性能的最佳行动方案。在现实世界中实施重大交通干预措施（例如重新设置交通走廊的信号配时、修改交叉口几何结构等）时，我们希望先在仿真中测试这些措施可能产生的影响。理想情况下，我们还希望在各种可能的参数范围内（例如交通流模式、信号配时方案等）测试这些干预措施的稳健性。为此，我们开发了一种并行仿真模型，可利用多个 CPU核心在不同参数下运行 SUMO 仿真。该系统可在商用云基础设施上轻松运行数百个并行仿真，且不受任何许可证限制。下面简要介绍其工作流程。有关更详细的说明，请参阅 [6], 第 6 章。

我们采用城市交通仿真（SUMO） [9] 微观仿真框架，因为它是一款开源（采用 GPL 许可）、高度便携的微观交通仿真器。该软件专为处理大规模、交通繁忙的道路网络而设计，涵盖车辆和行人，并已广泛用于研究。图 10 展示了系统概览。

A. 底图校准

为仿真部分道路网络，我们需要构建感兴趣区域的底图。我们使用 SUMO NETEDIT 工具，这是一款用于手动编辑道路网络的图形化网络编辑器。也可以从 OpenStreetMap 导入底图。

创建底图后，必须对其进行校准，以确保它能反映现实情况。如果仿真未能恰当地复现实地状况，所得指标将无法使用。仿真模型的多个方面都需要校准：

1）流量校准：为了校准模拟交通流，必须确保在与实际观测到的交通流动相对应的地点，有正确数量的车辆驶入和驶出。环形检测器数据能够准确统计交叉口的车辆数量，但无法揭示车辆实际行驶路径的细节。通过使用轨迹数据，我们可以推断出各交叉口近似的起讫点（O-D）矩阵，并以转向概率的形式表示。我们使用“route-Sampler”工具输入车辆数量（来自 ATSPM数据）和起讫点概率矩阵。该工具的输出是一个文件，其中包含单辆车辆及其预设的驶入系统时间。车辆的行驶路线也经过预先设定。这确保生成的车辆总体上遵循与现实世界相同的趋势，包括车辆数量和转向流量（TMC）。

2) 超速校准： 校准还必须考虑现实中的超速行为，因为 SUMO 默认会让车辆保持在限速以内。轨迹数据可以揭示交叉口处观测到的最高速度分布，尤其是在信号灯变绿之后。利用这一观测到的速度分布设置 SUMO 的“speedFactor”，使车辆能够根据由观测分布推导出的参数，按正态分布抽样速度。这样可以确保仿真中的超速情况与实际交通数据中观测到的行为相符。

3) 交叉口信号配置：我们的底图由多个交叉口组成，每个交叉口有四个进口道，采用标准 NEMA（美国国家电气制造商协会）信号相位配置。每个交叉口包含四个直行/右转交通流和四个左转交通流。在任何时刻（通常情况下），八个互不冲突的交通流中有两个显示绿灯，从而允许车辆安全通行。这些相位绿灯的顺序按照一种称为“环与屏障（Ring-and-Barrier）”的序列进行排列。这些数值可以从信号配时表中获取。如果无法获取，也可以通过研究 ATSPM数据，找出最大和最小绿灯时间及其顺序来估算这些数值。随后，可将这些信息输入 SUMO，以设置交通信号灯行为。大多数交叉口还在每个进口道设置了专用左转缓冲区，以容纳左转交通流。在缓冲区填满之前，这可以防止左转交通流阻塞直行/右转交通流。利用谷歌地图等在线地图服务，可以轻松查看这些缓冲区的长度。

B. 并行仿真

对于包含 10 个交叉口的交通走廊，每次仿真（模拟 1 小时）通常需要 3–6 分钟，具体取决于输入交通流量。如果要模拟多种参数的组合，例如流量、信号灯方案等，生成大型数据集可能需要数天。为加快这一过程，我们使用 Python 的多进程库同时运行数十个仿真，并将结果存储为 XML格式。

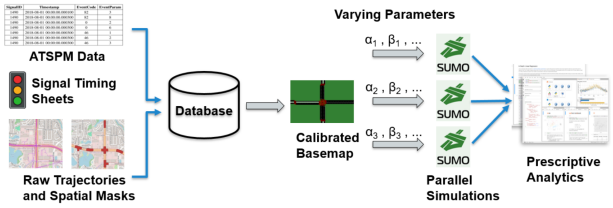

图 10. 基于SUMO的并行仿真模块概览。

C. 应用案例

我们介绍了 BigSUMO 在研究中的应用案例。未来，我们计划逐步整理、编写文档并发布更多源代码（github.com/NSH2022/BigSUMO）。

1）优化信号配时方案：并行仿真框架的一个重要应用，是能够通过网格搜索为交通走廊寻找最优信号配时方案。例如，我们可以为信号配时的各个方面选取一组可行参数，例如常见的周期长度、各相位的最短和最长绿灯时间等。我们可以对这些参数进行测试，找出能产生最优结果的参数，通常依据走廊行程时间等性能指标进行评估。BigSUMO 的这一功能可与 ReTime 相比，后者通过 VISSIM 仿真进行并行仿真。不过，ReTime 是专有软件，需要许可证。

2) 深度神经模型的数据生成：BigSUMO 支持大规模数据生成，满足训练和评估深度学习模型的需求。特别是，图神经网络（GNN）已成为一种强大的工具，可利用道路网络固有的图结构，以及其中相互依赖的智能体和属性，学习时空交通动态。

BigSUMO 可生成大规模数据，包括精细的按车道交通数据、交通信号状态，以及不同拥堵场景下的车辆行为。此外，BigSUMO 能够生成多样且高分辨率的传感器模态，使研究人员能够构建多层交通图 [10]。例如， [11] 采用图注意力网络（GAT），在单层或多层结构的图数据上完成训练后，估计出口和入口流量环形检测器所在位置的按车道交通波形时间序列。此外，模拟交通行为的“数字孪生”模型，例如 MTDT [12] 和 TGDT [13] [14],，可对城市交叉口和交通走廊的交通性能进行全面评估，并开展拥堵预测。这些研究使用 BigSUMO 生成的数据进行训练和评估。

VI. 结论

在本项工作中，我们利用 AT-SPM 和稀疏轨迹数据，开发了一个用于描述性分析和规范性分析的端到端系统。我们开发了一个数据管道，用于摄取和处理 ATSPM 数据及轨迹数据。随后，我们以等待时间和位置地图、转向移动比例、制动严重程度地图、排队分布等形式，展示了详细的分析结果。我们还基于这两类数据源，提出了一个中断检测模块。接下来，我们介绍了该框架如何使我们能够在不同条件下研究反事实场景，以分析交通干预措施的潜在影响。为此，我们采用了 SUMO 的并行化实现，同时运行多个仿真。该系统还可用于生成训练深度学习模型的数据。BigSUMO 的潜在应用包括自适应信号控制、拥堵收费以及车路基础设施协同。

参考文献

[1] T. Banerjee, K. Chen, A. Almaraz, R. Sengupta, Y. Karnati, B. Grame, E. Posadas, S. Poddar, R. Schenck, J. Dilmore, S. Srinivasan, A. Rangarajan, 和 S. Ranka, 《面向行人和车辆安全的现代交叉口数据分析系统》，载于*2022 IEEE*第*25*届智能交通系统国际会议 *(ITSC)*，2022年，第3117–3124页。

[2] M. V. Peppa, T. Komar, W. Xiao, P. James, C. Robson, J. Xing, 和 S. Barr, 《面向基于闭路电视的城市交通流量检测与预测的端到端框架》，传感器，第21卷，第2期，2021年。

[3] [ ] 1424 8220 21 2 629 在线。可访问：https://www.mdpi.com/ - / / / B. Tian, Q. Yao, Y. Gu, K. Wang, 和 Y. Li, 《交通流监测的视频处理技术：综述》，载于*2011*年第*14*届*IEEE*智能交通系统国际会议 *(ITSC)*，2011年，第1103–1108页。[4] R. Sengupta, T. Banerjee, K. Chen, Y. Karnati, S. Ranka, 和 A. Ran- garajan, 《利用鱼眼视频实现有效的交通信号安全与优化》。载于车辆技术与智能运输系统会议，2023年，第54–63页。[5] Y. Karnati, D. Mahajan, A. Rangarajan, 和 S. Ranka, 《用于交通中断检测的机器学习算法》，载于*2020*年第五届雾计算与移动边缘计算国际会议 *(FMEC)*，2020年，第231–236页。[6] Y. Karnati, D. Mahajan, T. Banerjee, R. Sengupta, P. Clay, R. Casburn, N. Agarwal, J. Dilmore, A. Rangarajan, 和 S. Ranka，综合走廊管理的数据分析与机器学习。CRC出版社，2024年。[7] R. Sengupta, T. Banerjee, Y. Karnati, S. Ranka, 和 A. Rangarajan, 《利用专用短程通信路侧单元数据推导制动行为》。车辆技术与智能运输系统会议，2023年1月，第420–427页。[8] H.-P. Kriegel, M. Schubert, 和 A. Zimek, 《高维数据中的基于角度的离群点检测》，载于第*14*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，KDD ’08系列。美国纽约：计算机协会，2008年，第444–452页。 [在线]。可访问：https://doi.org/ 10 1145 1401890 1401946 。/。

[9] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Fl¨otter¨od, R. Hilbrich, L. L¨ucken, J. Rummel, P. Wagner, 和 E. Wießner, 《使用SUMO进行微观交通仿真》，载于*IEEE*智能交通系统会议 *(ITSC)*。IEEE，2018年。[10] N. Yousefzadeh, M. T. Thai, 和 S. Ranka, 《多层图嵌入综合综述》，*Authorea*预印本，2023年。[11] N. Yousefzadeh, R. Sengupta, Y. Karnati, A. Rangarajan, 和 S. Ranka, 《用于按车道、拓扑不变的交叉口交通仿真的图注意力网络》，*IEEE*智能交通系统汇刊，第26卷，第4期，第5082–5093页，2025年。[12] ——, 《MTDT：多任务深度学习数字孪生》，载于*2024*年第*27*届*IEEE*智能交通系统国际会议 *(ITSC)*，2024年，第2058–2065页。[13] N. Yousefzadeh, R. Sengupta, J. Dilmore, 和 S. Ranka, 《TGDT：面向城市交通走廊的基于时间图的数字孪生》，*arXiv*预印本 *arXiv:2504.18008*，2025年。[14] N. Yousefzadeh, R. Sengupta, 和 S. Ranka, 《用于城市主干道行程时间分布预测的动态图注意力网络》，*arXiv*预印本 *arXiv:2412.11095*，2024年。