arXiv:2603.05217v1 [cs.DC] 5 Mar 2026

在边缘–云计算架构上扩展城市级摄像头网络的实时交通分析 *

Akash Sharma[1],Pranjal Naman[1],Roopkatha Banerjee[1],Priyanshu Pansari[1],Sankalp Gawali[1], Mayank Arya[1], Sharath Chandra[1], Arun Josephraj[3],Rakshit Ramesh[3], Punit Rathore[2,4] Anirban Chakraborty[1], Raghu Krishnapuram[3,4],Vijay Kovvali[4] 和 Yogesh Simmhan[1,2]

[1]计算与数据科学系[2]罗伯特·博世网络物理系统中心 (RBCCPS)[3]公共利益数据中心(CDPG)[4]基础设施、可持续交通与城市规划中心印度科学理工学院，班加罗尔，印度

Email: {akashsharma, pranjalnaman, simmhan}@iisc.ac.in

摘要

城市级实时交通分析需要在严格的延迟、带宽和计算资源限制下处理数百至数千路闭路电视视频流。我们提出了一种可扩展的人工智能驱动的智能交通系统（AIITS），旨在应对边缘–云计算架构上的多维扩展需求。我们的平台通过深度神经网络推理流水线，将多摄像头实时视频流转化为动态交通图，并结合时空图神经网络实现实时临近预报和短时预测。我们在班加罗尔一个社区搭建测试平台进行验证，从树莓派接入100+ RTSP视频流，同时由 Jetson Orin 边缘加速器执行高吞吐量的检测与跟踪，并生成轻量级流量摘要，供基于云的图神经网络推理使用。容量感知调度器在异构设备间协调负载均衡，使系统在数据流数量增加时仍能维持实时性能。为确保持续适应，我们集成了 SAM3 基础模型辅助标注和持续联邦学习，以更新边缘端的深度神经网络检测器。实验表明，Jetson Orin 上的稳定接入速率可达 2000 帧每秒，同时实现低延迟聚合，并能针对多达 1000 路数据流生成准确且可扩展的时空图神经网络预测。计划中的现场演示将把完整流水线扩展至 1000 路数据流，展示其实用的跨架构可扩展性。

*已被 TCSC SCALE 挑战赛 2026 接收。将发表于 IEEE/ACM CCGRID研讨会论文集，悉尼，2026年。

1 引言

交通拥堵的社会影响。 在快速城市化地区，交通拥堵与安全问题仍是严峻挑战，影响生活质量、公共卫生和能源可持续性。在新兴特大城市中，这一问题尤为严重，这些城市拥有 >10百万人口。印度班加罗尔是全球交通拥堵程度第二高的城市，平均通勤速度为每小时 14 公里（8.7 英里），导致每位通勤者每年损失近 7 天 [21]。道路运输占全球能源消耗的 $\approx 30\%$ ，因此交通拥堵也成为一项可持续发展问题 [7]。交通密度增加还会使空气质量恶化 [11]。印度的道路交通死亡人数居世界首位，这凸显了采取系统性干预措施的必要性 [6]。

人工智能驱动的智能交通系统方法。 这些现实情况凸显了对智能交通系统（*ITS*）的需求，使其能够在异构、无序且动态变化的交通状况下有效运行。与发达地区遵循车道规范的交通不同，印度道路上车型混杂，从两轮车、三轮车到公交车和卡车，车辆行驶时并不遵循车道。雪上加霜的是，发展中城市在用于交通监控的专用传感基础设施方面也投资不足，例如车道摄像头或自动车牌识别（自动车牌识别）摄像头，而是依赖数千个并非为交通流分析而设计的通用安全摄像头。因此，传统交通计算模拟器（如 SUMO、Vissim）无法针对这些状况进行有效校准，以优化交通 [17]。这些限制促使我们采用人工智能驱动的智能交通系统（*AIITS*），从规模庞大、异构的摄像头网络中推断交通状态。

人工智能驱动的智能交通系统对速度的需求。 人工智能驱动的智能交通系统流水线在计算和通信方面都十分密集。我们的案例研究城市班加罗尔拥有 > 5000 安全摄像头，可持续生成高清视频流。如今，即使是人工监控，集中式架构也依赖高带宽光纤网络、GPU服务器以及 PB 级归档存储。向人工智能驱动的智能交通系统解决方案转型，需要对数据流进行主动实时计算，而非仅仅被动存储。扩展到数百 Gbps 和每秒 100k + 帧（FPS）的视频流规模，意味着云端GPU服务器将出现不可持续的增长。此外，要进行可操作的交通预测，需要低延迟的数据摄取和推理；随着摄像头部署增加，集中式系统难以通过广域网提供这些能力。这定义了我们的问题：在严格的延迟和资源约束下，实现来自城市数千个摄像头的实时分析，并需要协调横向扩展、纵向扩展和弹性流分发。

贡献。 我们提出了一种可扩展的边缘—云框架，用于在全市范围的摄像头网络中进行实时 AI 驱动的交通分析。我们的系统通过采用边缘优先设计，在严格的延迟、带宽和计算资源约束下处理数百路异构视频流：在分布式 Jetson 加速器上执行高吞吐量检测与跟踪，同时由树莓派承载大规模 RTSP视频流。该框架展现了多维可扩展性：(i) 跨50+ 异构边缘设备横向扩展，可支持 100 路数据流；(ii) 图神经网络流水线纵向扩展，可持续实现多达 1000 路数据流的实时推理；以及 (iii) 从边缘到

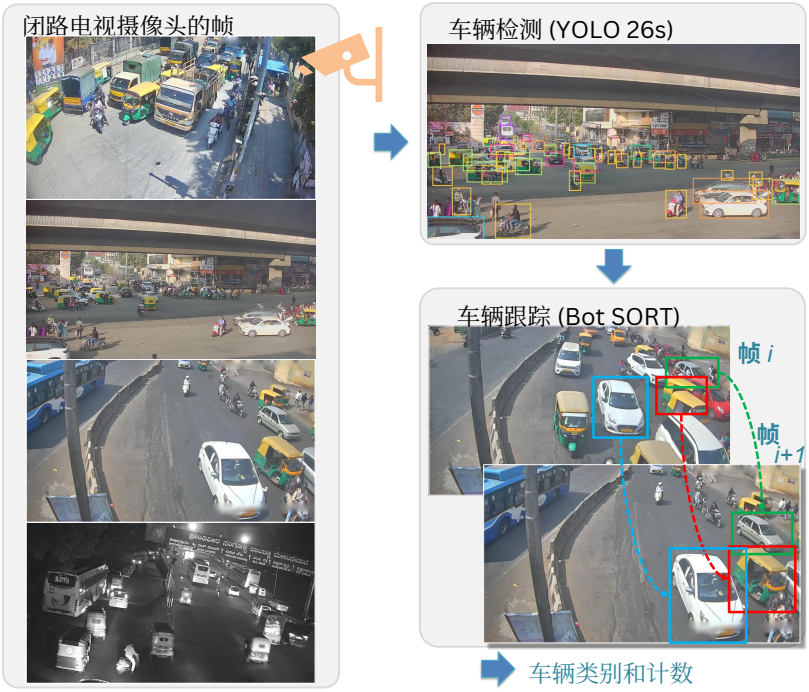


图 1：展示班加罗尔复杂交通状况的示例图像，以及用于检测车辆类别和数量的深度神经网络流水线。

到云端。

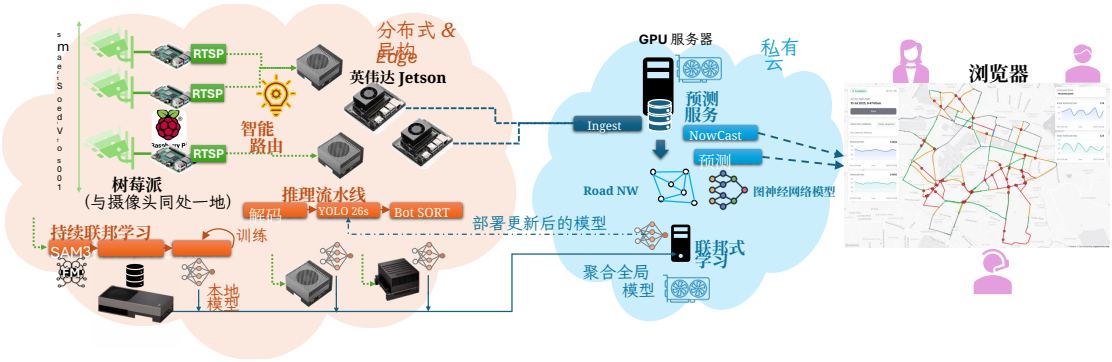
一个容量感知流调度器 根据实测的帧每秒阈值，智能地将视频流分配给边缘加速器，使数据流从数十路扩展到数百路时仍能维持实时性能。系统将结构化的单摄像头交通流摘要融合为动态的城市级时序交通图，并将其输入基于图神经网络的预测 服务，以生成短时域拥堵预测。为保持分析结果的时效性，持续联邦学习 流水线利用基础模型 检测新出现的 车辆类别，并在边缘端重新训练深度神经网络检测器，同时执行推理。

我们在班加罗尔一个街区 部署 100 个摄像头进行端到端评估，结果表明系统能够持续保持实时吞吐量、低延迟和稳健的扩展能力。这证明了 AIITS 平台具备可部署性，并能产生积极的社会影响。我们计划在现场演示中将规模扩展至 1000 路数据流。

2 班加罗尔人工智能驱动的智能交通系统

印度科学理工学院出行人工智能团队（*AIM@IISc*）是一个跨学科团队，致力于开发可扩展的数据驱动工具，以改善印度道路上的交通拥堵和安全状况。班加罗尔是理想的真实世界测试平台，团队与城市交通警察密切合作，设计并验证新颖且可扩展的方法。我们在此介绍 AIITS 的高层需求，随后于 §3中介绍拟议的设计。

可扩展 AIITS 的一项关键要求，是将 5000+安全摄像头视频流转换为能够呈现实时网络状况的时序图 。这一转换



Fig图 2：面向班加罗尔交通摄像头视频流的可扩展 AIITS 测试平台与架构 .

将高容量、占用大量带宽的多维数据精简为简洁的时序图形式，以便进行下游分析。我们通过边缘优先推理流程（图 1）与基于云的聚合相结合来实现这一目标。

多台主要部署在道路交叉口的摄像头仅覆盖了 700 km^2 区域内 2000+ kms 道路网络的一小部分。因此，需要将空间外推从 ≈ 1000 已观测到的顶点和边扩展至全市 10k+ 条未观测到的道路。鉴于为印度的异质交通校准经典交通模拟器存在局限，我们采用时空图神经网络（*ST-GNN*），以便对更广泛道路网络中的交通进行实时临近预报和短期预测。

这些预测为更高层级的分析提供支持，有助于制定干预措施。

对预测时序图进行异常检测，可以标记交通拥堵异常严重的路段，从而支持有针对性的交通警力部署或远程信号控制。此类大范围、实时的信号优化可以逐步实现自动化。ST-GNN 模型还支持假设分析，使规划人员能够评估各种政策选项，例如实施单向交通、调整车道比例或增设公交专用车道。这些措施可降低解决方案的成本，并推动基于证据的城市出行决策。

最后，由于车辆类型异构且交通状况不断变化，训练适用于印度交通的高精度视觉模型颇具挑战。我们此前开展的大规模标注工作（例如 UVH-26 [20]）将最先进的 YOLO 和 DETR 模型的准确率从 8.4提升至 31.5%，超过了在 COCO 数据集上训练的同等基线模型[16], ，但这仍然需要大量人工投入，且无法扩展至多个城市。因此，我们考虑使用多模态基础模型（如 SAM3、CLIP）作为自动教师模型，以生成初始的本地化训练数据。我们将其整合到持续联邦学习设置中，从而在不中断实时推理的情况下，利用边缘加速器开展可扩展的分布式再训练。这展示了可扩展的应用—基础设施协同设计。

3 可扩展的 AIITS 架构与测试平台

图 2 展示了我们的人工智能驱动的智能交通系统平台架构，该平台涵盖边缘与云资源。我们搭建了一个*AIITS* 测试平台，用于模拟所提出的城市规模部署，其规模为 ◎（100） 路视频流。每台摄像机都配有一台嵌入式计算机，用于运行实时流传输协议服务器以提供数据流；在我们的测试平台中，这些数据流通过树莓

派进行代理。核心计算架构由*Nvidia Jetson* 边缘加速器组成，包括 Orin AGX 和 Thor 开发者套件；这些设备采用紧凑型（11cm×11cm×7.2cm，800–1800 gms）低功耗（峰值 60–130W）设计，适合现场部署，并配备多核 Arm CPU、1792–2560 个 Ampere/Blackwell CUDA 核心以及 32–128GB CPU–GPU 共享 LPDDR 内存。这些设备与部分路侧单元上的摄像头共置，并通过本地光纤网络接收视频流，以执行实时推理和训练。

部署在市政私有云中的少量集中式 GPU 工作站负责协调平台、接收轻量级结构化流量更新、维护时序交通图、使用图神经网络模型执行临近预报和预测，并管理联邦学习聚合。这些工作站还承载基于浏览器的服务，向城市运营人员、居民及其他利益相关者提供实时预测、警报和系统控制功能。接下来，我们将讨论这些不同的组件。

3.1 从边缘传输实时流传输协议流

为了在不依赖实际闭路电视部署的情况下评估系统在大规模场景中的表现，我们构建了一个分布式测试平台，通过数百路并发的实时流传输协议（*RTSP*）视频流来模拟大型城市摄像头网络。每路数据流都部署在由树莓派组成的异构集群上，每台树莓派均运行$MediaMTX$这一开源 RTSP 服务器。每台服务器通过唯一端点提供一路或多路数据流，供下游处理。视频由$FFmpeg$提供；它会在特定时长内重放本地存储于每台树莓派上的预录摄像头影像。借助 FFmpeg 的流复制功能，我们避免了转码开销，使低功耗树莓派也能运行多路实时数据流。这种架构支持集群中的高并发。总计 42 台 3B 和 4B 型号的树莓派能够提供 100 路分布式数据流，构成一种可扩展、低成本的模拟方案，用于模拟城市范围内的摄像头网络邻域。增加更多树莓派以支持更多数据流时，该方案也能实现弱扩展。

3.2 在边缘加速器上扩展检测模型

为了将原始的多摄像头视频流转换为适用于基于 GNN 的预测的数据，我们在测试平台中的异构 Nvidia Jetson 设备上部署了分布式边缘推理层。该层接入实时 RTSP 流，执行实时检测与跟踪，并生成精简的车辆流量摘要，供下游分析使用。通过在本地执行推理，系统无需向云端传输高带宽视频，从而降低了端到端延迟。

3.2.1 数据流接入

终端用户通过前端界面选择摄像头流，随后由我们的调度器动态分配给 Jetson 加速器（§ 3.2.3）。在每台 Jetson 上（图 2），硬件加速的 $GStreamer$ 流水线利用板载 $nvdec$ 加速器来

将数据流直接解码到 GPU 原生内存中。DeepStream 的 $nvstreammux$ 会将来自多个摄像头流的帧批处理为统一的推理批次，从而提高 GPU 利用率，使 Jetson 能够以实时帧率处理多个并发数据流。

3.2.2 检测与跟踪

每个批处理帧均使用 YOLO26s 模型进行处理 [10], 该模型在我们的 UVH-26 数据集上训练 [20] 并使用 DeepStream 的推理引擎进行部署。虽然目标检测可提供逐帧边界框和类别，但实时交通估算需要跟踪唯一车辆在视频流中。为此，每条数据流均集成 BoT-SORT 多目标跟踪 [1], 从而在连续帧中为检测到的车辆分配持久 ID。对于每个已处理帧，系统会输出由跟踪 ID、类别标签和边界框组成的车辆级记录。这些具备身份信息的观测结果由我们的流水线在短时间间隔内聚合（可调，5–30 秒），并以低开销摘要的形式转发至 GNN 推理流水线，从而无需传输原始视频或密集特征图。

3.2.3 智能路由

由于推理层运行于异构 Jetson 设备之上，有效的流分配对于维持实时性能至关重要。我们通过采用容量驱动调度策略将数据流分配给加速器，从而实现弹性扩展，该策略基于经验得出的每设备上限。对于每种 Jetson 型号，我们通过逐步增加输入工作负载进行离线性能剖析，直至设备无法再维持原生帧率，由此定义其稳定吞吐量。在我们的部署中，Orin AGX 32GB 与 64GB 型号分别支持 ≈ 200 帧每秒和 400 帧每秒。随后，数据流分配被表述为装箱问题 [5], ，其中每条数据流的帧每秒构成物品重量，每台设备则代表一个容量已知的箱子。我们评估了两种经典启发式方法：

最佳适配 数据流被放置在仍能容纳它们的最小容量设备上。在我们的测试平台设备（§ 4.1）中，这会首先将负载紧密打包到 Orin 32GB 设备上，仅当累计需求超过 ≈ 1000 帧每秒时，才激活更大的 Orin 64GB 设备。这可在中等负载下尽量减少活跃设备数量，并降低总体功耗。

最差适配 数据流会分配给剩余容量最大的设备，从而将负载分散到各个设备上，即使是容量较大的设备也会较早参与分担。这能改善负载均衡和热量分布（例如，在炎热天气下，IP67 外壳会带来热限制时），但也会更早启用更多设备，从而增加集群的基线功耗。

这些策略使系统能够灵活地在异构边缘加速器上扩展检测与跟踪，同时兼顾功耗，并为大规模多摄像头部署持续提供实时吞吐量。

3.3 使用图神经网络在时序图上进行交通预测

GPU服务器承载 *Ingest*、*NowCast* 和 预报服务，用于处理从边缘推理层接收的结构化车流输出（图2）。这体现了我们架构中的纵向扩展方面：利用GPU服务器实现高吞吐量图推理。测试平台中的本地GPU工作站模拟私有云中的加速虚拟机。摄取服务 以1秒为粒度接收每个摄像头的车辆计数，每15秒批量处理一次，并维护一个存储在磁盘上的时序数据库，其中包含时间戳、摄像头ID和类别计数向量。NowCast服务通过以gRPC服务实现的低延迟流式接口，向下游应用提供聚合后的交通状态。

预测服务采用 时空图神经网络（*ST-GNN*）来预测道路网络中短时域的交通状况。ST-GNN共同建模道路图的空间结构和交通模式的时间演变。我们使用TrendGCN架构 [9], ，该架构融合顶点嵌入、时间嵌入和图卷积门控循环单元，以捕捉交叉口之间动态变化的交通相关性。模型通过将跟踪器输出汇总为交叉口级别的1分钟车辆计数时间序列进行离线训练，使用来自100个交叉口的180小时历史数据，滞后时间为5分钟，预测时域为5分钟。模型根据可配置的滞后窗口（例如数分钟）从摄取数据库中查询历史数据，并输出所选预测时域内每个交叉口的预测车辆计数。

由于摄像头覆盖不完整，在我们的验证区域中，超过 250 个交叉口里只有 ≈ 100 个可观测；我们在一张粗化图上进行预测，该图仅将配备摄像头的交叉口保留为节点。这个简化图中的边是超边，由观测到的交叉口之间的路段序列合并而成；它们既保留了短距离交通传播所需的大尺度结构，又确保每个节点都能提供可靠的输入 [15]。为了根据预测的交叉口级流量计数得到街道级（边级）流量估计，我们采用质量守恒分配，按照相邻超边的连通性比例分配车辆流量，并汇总端点贡献以计算边流量，从而在拥堵估计中保持顶点预测与边预测的一致性。

最后，边缘级流量被离散化为拥堵状态（例如自由流、中度、重度），以便集成到仪表板并用于下游分析。通过结合高频边缘摘要、时空预测以及基于图的抽象，该系统能够提供准确的短时域预测，并可在部分摄像头覆盖条件下扩展至全市范围的部署。

3.4 面向边缘联邦学习的 LLM 驱动标注

基于边缘的检测模型必须持续适应高度本地化且不断演变的车辆类别，例如机动三轮车（三轮车）或新兴商用车辆类型，这些类别在标准数据集中代表性不足。为此，我们跨 K Jetson 设备实现持续联邦学习工作流，使每个边缘设备能够整理、标注并基于自身数据进行训练，而无需将原始视频传输至中央服务器。每台 Jetson 通过时间分层采样收集训练样本。帧在可配置窗口内累积，以形成本地且多样化的数据集。

采集阶段结束后，每台设备使用 SAM3 基础模型 [3] 自动标注采样图像。标注流程接受一张图像和一组预定义的文本提示，例如

$C = \{$“一辆轿车”, “一辆运动型多用途汽车”, $\cdots\}$，涵盖常见及区域特定的车辆类别，其中一部分可能不为目标检测模型所知。这些提示被词元化为嵌入向量 E_c ，SAM3 模型 f_θ处理这些输入，为一批数据 P生成边界框坐标 B_c 和原始类别逻辑值 Z_c: $(Z_c, B_c) = f_\theta(P; E_c)$。逻辑值通过 Sigmoid 函数转换为概率，仅保留超过置信度阈值 τ（例如 $\tau = 0.30$）的预测。由此得到的每台设备的数据集为$D_k = \bigcup_{c \in \mathcal{C}} \{(c, \text{bbox}_q, p_q) \mid p_q(c) \geq \tau\}$。

随后，每台 Jetson 都会使用当前全局权重初始化基于 YOLO 的检测器，并进行本地微调。经过 E 个本地训练轮次，使用 D_k进行训练后，设备会将更新后的权重发送回服务器，由服务器使用联邦平均（FedAvg）进行聚合。随后，聚合后的全局模型会广播回边缘设备，供检测推理流程使用。

3.5 仪表板

我们实现了一个交互式 React+Mapbox 仪表板（图 2，右），作为系统的主要可视化与控制界面。输出叠加在班加罗尔的 OpenStreetMap 地图上，并以地理空间方式呈现摄像头位置和推断出的路段。更新通过 gRPC 流式传输至 FastAPI 后端，再通过 WebSocket 转发到浏览器，以保持低延迟、持续刷新。地图提供实时临近预报视图，展示各摄像头的车流摘要如何转化为城市级交通状况。彩色编码叠加层表示拥堵程度，使操作人员能够一目了然地评估路网状态。配套的时间序列面板展示实时和预测的车辆计数，并随着 Ingest 和预测服务生成新输出而更新。在当前测量值旁显示短时域图神经网络预测，有助于凸显新出现的拥堵趋势。

仪表板还支持轻量级编排。用户可以选择单个摄像头或部分数据流，触发检测与跟踪流程的定向执行。选择某个路段后，即可查看车辆流量的历史摘要，该摘要基于摄取服务维护的按秒聚合计数生成。

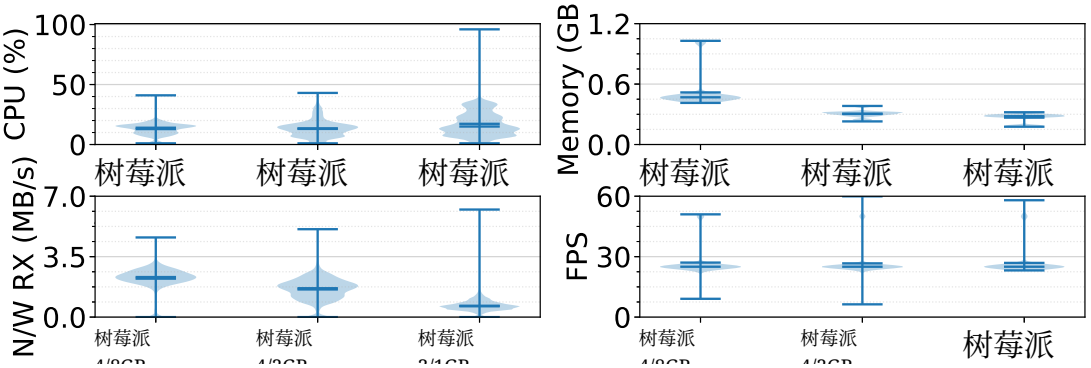


图 3：承载 100 路 RTSP 流的树莓派性能

4 结果

4.1 实验设置

我们的 RPi 集群承载着 1000 路摄像头流，其中包括 10 台 Pi 4B/8GB（每台 4 路数据流）、17 台 Pi 4B/2GB（每台 2–33 路数据流），以及 15 台 Pi 3B/1GB（每台 1 路数据流）。视频流通常为 25 帧每秒，在各台树莓派上预先同步，并由实时流传输协议服务器进行传输。测试平台还配备了加速型边缘集群 Jetson Orin AGX，其中 5 台配备 32GB 内存，4 台配备 64GB 内存，用于承载推理流水线并摄取这 100 路视频流。联邦学习设置运行在这些相同的设备上

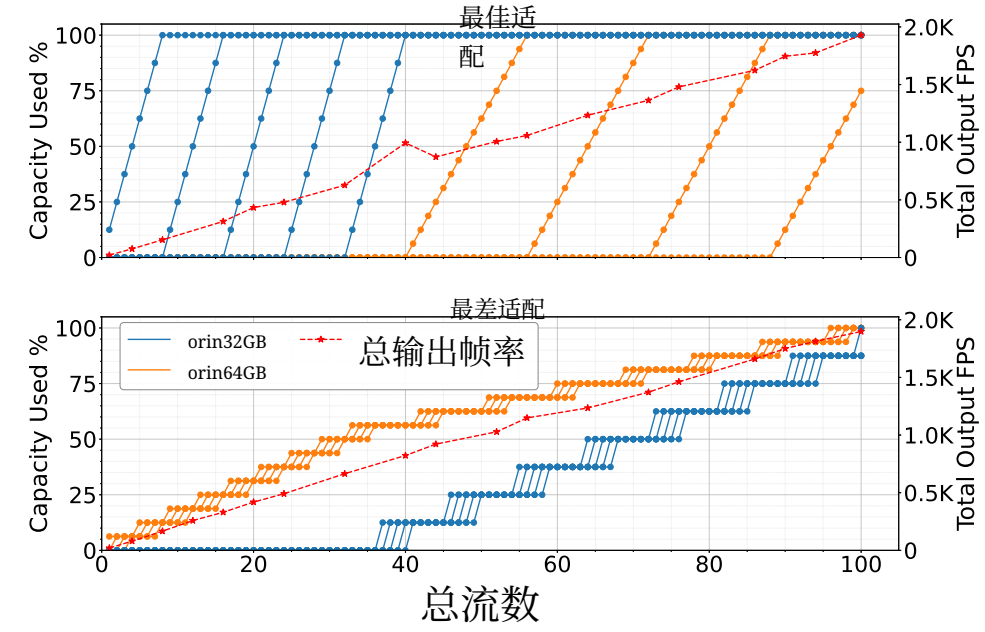
4.2 用于车辆计数的边缘分析可扩展性

4.2.1 RPiRTSP 流性能

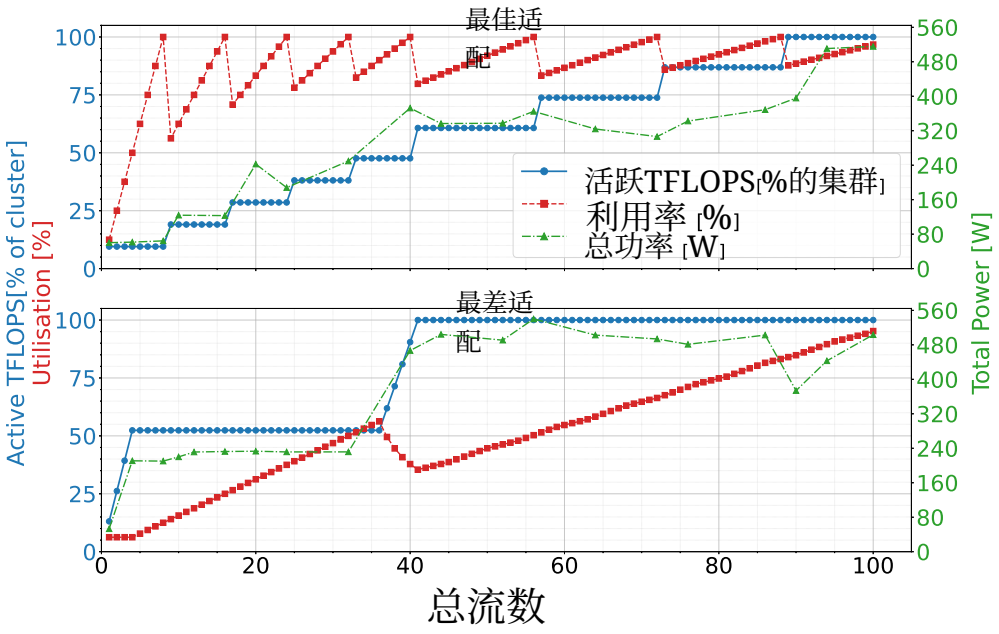
为验证测试平台的稳定性，我们评估了不同型号树莓派的资源使用情况（图 3）。结果表明，这些 RTSP 发布设备的开销极小：所有型号的 CPU 利用率中位数均低于 25%，这主要归功于流复制机制避免了高开销的编码/解码操作。内存使用量也保持在较低水平，树莓派 3/1GB 的峰值为 30%；树莓派 4/8GB 略高，因为它承载了四路数据流。网络吞吐量随数据流数量而变化：树莓派 4/8GB 发送的数据最多，树莓派 4/2GB 则发送两路或三路数据流，但所有设备的网络吞吐量均未超过带宽限制，低于 ≤ 7 MB/s；这一数值低于树莓派 3 的 100Mbps 或 12。5 MB/s 上限。帧每秒保持稳定，在 25 ± 1帧每秒的范围内，$> 90\%$ 的时间都如此。这些结果证实，即使是低功耗树莓派，也能在我们的 AIITS 测试平台中可靠地扩展至较高的流密度，同时不降低数据流质量。

4.2.2 边缘分析

我们通过增加接入的 RTSP 流数量，评估边缘分析在 Nvidia Jetson 设备测试平台集群上的可扩展性。正如预期，累计输出帧每秒 处理量会随输入负载增加而提升，只要没有设备超过其经实测确定的 FPS容量（§ 3.2.3），即使每秒处理超过 *2000* 帧 (图 4a)，也能确保实时性能。为进一步量化 性能与功耗



(a) FPS容量使用率和累计处理FPS

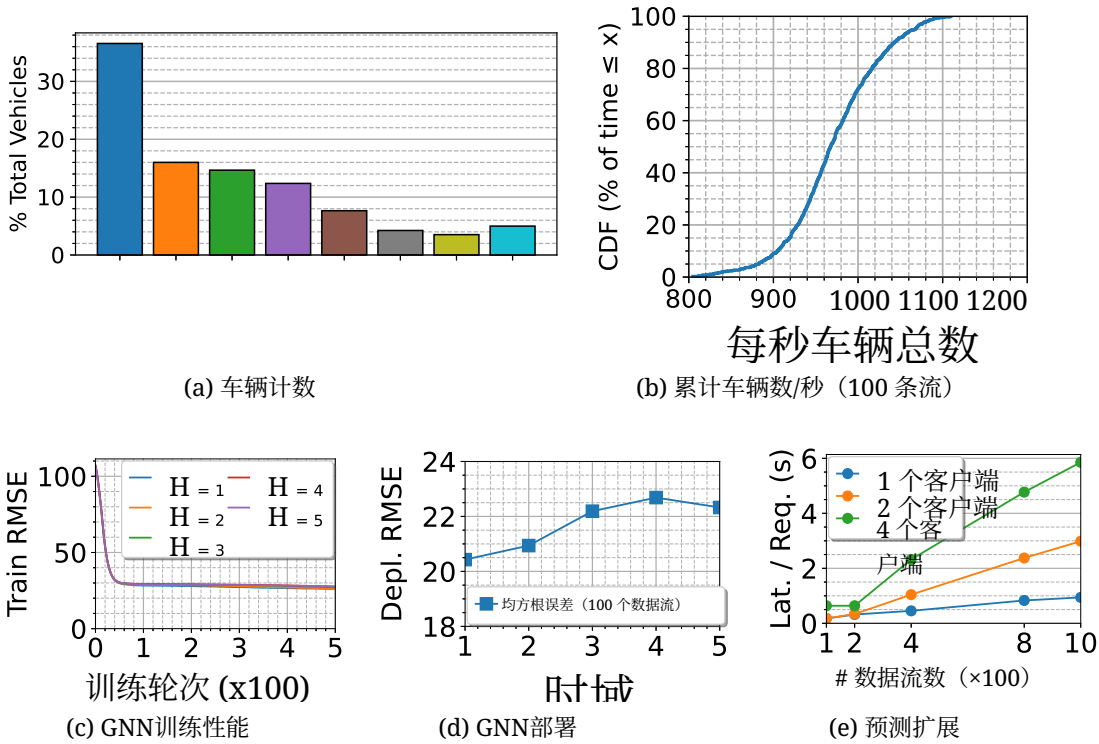


(b) 集群中的活跃与已用TFLOPS百分比，以及总功耗

图4：Jetson处理输入数据流时的效率与扩展性

效率，我们跟踪 总活跃容量 (活跃设备的 TOPS 总和)以及利用率 (相对于每台设备FPS容量的负载)，如图 4b所示，同时还跟踪总功耗。我们比较两种流分配策略。

最佳适配 （上图）会将数据流紧密地分配到最少的合适设备上，在启用 64 GB Orin 之前先使 32 GB Orin 达到饱和。这会产生阶梯式的启用模式，在中等工作负载下将活跃设备数量降至最低，并通过延后启用大容量节点来降低功耗。最差适配 （下图）



(a) 车辆计数 (b) 累计车辆数/秒（100 条流）

(c) GNN训练性能 (d) GNN部署 (e) 预测扩展

图 5：摄入速率与 GNN性能

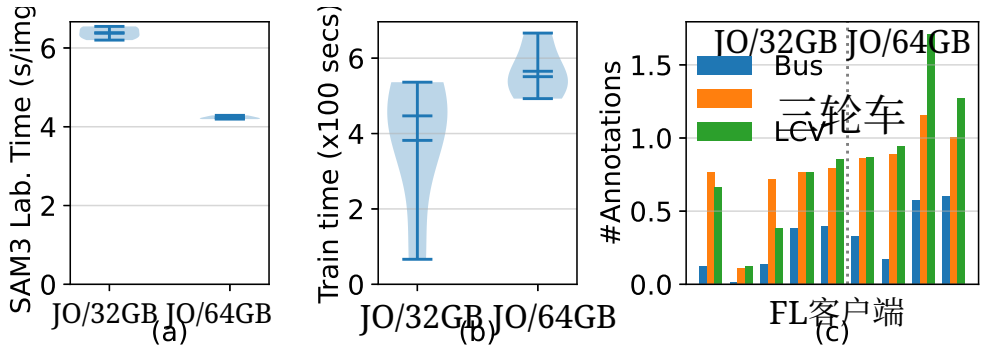


图 6：SAM3标注和联邦学习训练的性能。

会将负载分配到剩余容量最大的设备上，更早启用大容量设备，从而改善负载均衡。然而，这也会更早激活更多设备，增加基线功耗。尽管如此，在异构集群中，其影响更加复杂：最差适配可能会优先使用每流功耗效率更高的大容量设备，有时其功耗可与最佳适配相当，甚至更低。例如，在 32 个数据流下，最差适配的功耗为 231.6W，而最佳适配为 249.6W。这些结果表明，两种策略都能保持实时吞吐量：最佳适配以最少启用设备为优化目标，而最差适配则改善负载均衡，并展示了在异构边缘加速器上的有效横向扩展能力。

4.3 Ingest 与 ST-GNN 预测的可扩展性

我们分析了摄取服务的车流数据，以及 ST-GNN 预测管线。在 15 分钟内的 100 条数据流中，系统捕获了多种车辆类型，其中两轮车（37%）、轿车（15%）和三轮车（14%）占多数（图 5a）。在近 30% 的时间窗口内，汇总检测量超过每秒 1000 辆独立车辆，峰值达到每秒 1110 辆，直接体现了 Ingest 的吞吐量，并展现出系统在持续负载下的可扩展性（图 5b）。TrendGCN 训练在所有时域上都快速收敛，均方根误差也很早便趋于稳定（图 5c）。部署时，均方根误差从 1 分钟时域下的 20 小幅上升至 4 分钟及更长时域下的约 23，表明多步预测性能稳定（图 5d）。为评估可扩展性，我们将数据流数量从 100 增至 1000（通过合成数据流扩充）。每 5 秒生成一次预测时，随着并发客户端数量从 1 增至 4，延迟仍处于可控范围（图 5e），证实 ST-GNN 管线能够支持更多摄像头以及更高的客户端需求。

4.4 FM与持续联邦学习的可扩展性

为评估联邦学习框架的实用性与可扩展性，我们将系统部署在由9台Nvidia Jetson Orin AGX组成的集群上，包括 5×32GB（JO/32GB）和 4×64GB（JO/64GB）。这些Jetson设备共同Ingest 300路并发视频流，其中每台JO/32GB承载28路数据流，每台JO/64GB承载40路数据流。如§3.4所述，我们采用时间分层采样，在150分钟内每隔20秒窗口随机抽取一帧，每路视频流累计45帧，每台JO/32GB共计1260帧，每台JO/64GB共计1800帧。这些帧由SAM3利用Orin的加速器进行标注，平均每张图像的标注延迟分别为6.3秒（JO/32GB）和4.0秒（JO/64GB），如图6（左）所示。对于9个FL客户端中的每一个，我们在图6（右）中展示了YOLO模型未知的三个类别的分布，并在图6（中）报告了每种Jetson类型对应的训练时间分布。由于JO/64GB接收的数据流数量更多，其采集的数据也更多，约为JO/32GB所采集数据量的1.2–5×，每个训练轮次处理的数据更多，训练时间略高。标注分布显示出分布的非独立同分布性，凸显了在此场景中采用联邦学习的必要性，以确保所有Jetson都能识别全部类别。

4.5 讨论与扩展演示

我们的 AIITS 架构通过在功能上将边缘推理与基于云的时空预测隔离开来，实现了可扩展性。这样无需重新配置整个系统，即可横向扩展加速器。我们采用新颖的感知容量弹性调度器和 SAM3 驱动的标注方法，以缓解非结构化城市环境中预期会出现的“数据漂移”。通过将繁重的视频分析部署在网络边缘，系统避免了集中式带宽瓶颈，从而实现可持续的城市级交通感知。

SCALE 挑战赛现场演示将通过将支持范围扩展至 ≈ 100 树莓派和 ≈ 25 Jetson 加速器上的 1000 路视频流，展示系统的全面可扩展性，超越论文中的实验结果。这将凸显我们的设计实时处理城市级异构多模态工作负载的能力。

5 相关工作

智能交通系统（ITS）通常依赖专用道路传感器（例如感应线圈），这些传感器可直接提供可用的时间序列数据，但在新兴国家进行城市范围的扩展部署成本高昂 [24]。尽管闭路电视网络已得到广泛部署，但如何将海量视频流转换为交通流信号仍缺乏充分研究，并且需要分布式多流检测、跟踪、同步和聚合 [2, 23]。现有系统通常只处理其中的单个组件，而没有像我们这样提供集成解决方案的端到端流水线。Li 等人 [13] 提出了一种多阶段监控框架，但扩展能力有限；Seo 等人 [19] 优化了DNN分区，却未处理并发多流工作负载；而 Rong 等人 [18] 则研究了同构假设下的边缘—云调度。

时空图神经网络 [8, 9, 12] 具有出色的预测性能，但其假设输入是同步传感器输入，而不是大规模生成的视频衍生信号。模型适配方面还存在进一步的不足。预训练检测器在异构且不断发展的城市交通场景中通常表现不佳，而构建大规模标注数据集的成本很高。开放世界发现方法可减少人工标注，但通常假设采用集中式数据访问 [22]。联邦学习支持在边缘进行本地训练，同时避免原始数据共享 [4], ，但设备异构性仍然具有挑战性 [14]。针对这些不足，我们将异构多流边缘推理、容量感知调度、图级聚合、图神经网络预测和SAM辅助标注，与轻量级联邦学习更新相结合，构建统一的端到端分析流水线，以实现城市规模预测。

6 结论

我们提出了一个 AIITS 平台，将轻量级深度神经网络、ST-GNN 预测、基础模型辅助标注和持续联邦学习相结合，从而在边缘–云计算架构中实现高效扩展，同时最大限度地减少数据传输和集中式计算。结果证实，该平台能够扩展到大规模摄像头部署和异构边缘硬件，并支持多模态扩展：流级横向扩展、设备级纵向扩展以及跨架构协同，展现了其在城市规模城市出行分析中的实际影响。未来，我们将评估更拥堵条件下的性能，将处理流程扩展至 5000+feeds，并研究推理与联邦学习之间的性能交互。我们计划进一步利用网络拓扑和能耗信号优化数据流到 Jetson 的部署位置，并启用动态模型选择，以在数据流变化时维持吞吐量。

7 致谢

我们感谢班加罗尔交通警察局（BTP）和班加罗尔警察局为我们提供安全城市摄像头数据的访问权限，本研究所使用的视频数据即源自这些数据。

参考文献

[1] Nir Aharon 等。BoT-SORT：用于多行人跟踪的鲁棒关联。技术报告 2206.14651，arXiv，2022年。[2] Johan Barthélemy、Nicolas Verstaevel、Hugh Forehead 和 Pascal Perez。面向智慧城市实时交通监测的边缘计算视频分析。*Sensors*，19(9):2048，2019年。[3] Nicolas Carion、Laura Gustafson、Yuan-Ting Hu、Shoubhik Debnath、Ronghang Hu、Didac Suris、Chaitanya Ryali、Kalyan Vasudev Alwala、Haitham Khedr、An- drew Huang 等。Sam 3：基于概念的任意分割。*arXiv* 预印本 *arXiv:2511.16719*，2025年。[4] H. Brendan cMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas。基于去中心化数据进行深度网络通信高效学习。载于*AISTATS*，2017年。[5] Edward G. Coffman、Michael R. Garey 和 David S. Johnson。装箱问题的近似算法——最新综述。载于 Giorgio Ausiello、Mario Lucertini 和 Paolo Serafini 编，计算机系统设计的算法设计，第 284 卷，国际机械科学中心。施普林格，维也纳，1984年。[6] 全球道路安全基金（GRSF）。GRSF 2024年年度报告。技术报告，全球道路安全基金，世界银行，2024年。访问日期：2026年2月19日。[7] 国际能源署（IEA）。2025年能源效率：交通运输。 https://www.iea.org/reports/energy-efficiency-2025/transport，2025年。访问日期：2026年2月19日。[8] Jiawei Jiang、Chengkai Han、Wayne Xin Zhao 和 Jingyuan Wang。Pdformer：用于交通流量预测的传播延迟感知动态长程变换器。载于*AAAI*人工智能会议论文集，第 37 卷，第 4365–4373 页，2023年。[9] Juyong Jiang、Binqing Wu、Ling Chen、Kai Zhang 和 Sunghun Kim。通过对抗学习和联合时空嵌入提升交通预测的鲁棒性。载于第 *32* 届 *ACM*信息与知识管理国际会议论文集，CIKM ’23，第 987–996 页，美国纽约，2023年。计算机协会。

[10] Glenn Jocher 和 Jing Qiu。Ultralytics YOLO。
https://github.com/ultralytics/ultralytics，2023年。[11] Kanishka Kacker、Ridhima Gupta 和 Saif Ali。交通拥堵是否会造成健康危害？来自一座高度拥堵和污染城市的证据。讨论论文系列 DP 23-10，环境促进发展倡议，2023年7月。访问日期：2026年2月19日。[12] Shiyong Lan、Yitong Ma、Weikang Huang、Wenwu Wang、Hongyu Yang 和 Pyang Li。DSTAGNN：用于交通流预测的动态时空感知图神经网络。载于国际机器学习会议，第 11906–11917 页。机器学习研究论文集，2022年。[13] Ran Li。面向智能交通系统中实时车辆检测、跟踪和速度测量的多阶段深度学习方法。《科学报告》，15(1):22531，2025年。[14] Tian Li、Anit Kumar Sahu、Ameet Talwalkar 和 Virginia Smith。联邦学习：挑战、方法与未来方向。《IEEE信号处理杂志》，2020年。[15] Yaguang Li、Rose Yu、Cyrus Shahabi 和 Yan Liu。扩散卷积循环神经网络：数据驱动的交通预测。载于国际学习表征会议，2018年。[16] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。微软COCO：上下文中的常见物体。载于欧洲计算机视觉会议，第 740–755 页。施普林格，2014年。[17] Htet Naing、Wentong Cai、Jinqiang Yu、Jinghui Zhong 和 Liang Yu。通过结合动态数据驱动进化优化的微观交通仿真进行细粒度轨迹重建。《IEEE智能交通系统汇刊》，26(2):1930–1950，2024年。[18] Chenghao Rong、Jessie Hui Wang、Juncai Liu、Jilong Wang、Fenghua Li 和 Xiaolei Huang。通过调度海量摄像头流来优化大规模实时视频分析。《IEEE/ACM网络汇刊》，30(2):867–880，2021年。[19] Woobean Seo、Saehwa Kim 和 Seongsoo Hong。DNNPipe：基于动态规划的最优DNN分区，用于物联网网络上的流水线推理。《系统架构杂志》，166:103462，2025年。[20] Akash Sharma、Chinmay Mhatre、Sankalp Gawali、Ruthvik Bokkasam、Brij Kishore、Vishwajeet Pattanaik、Tarun Rambha、Abdul R. Pinjari、Vijay Kovvali、Anirban Chakraborty、Punit Rathore、Raghu Krishnapuram 和 Yogesh Simmhan。城市视觉黑客马拉松数据集与模型：迈向图像标注和准确

面向印度交通的视觉模型。技术报告 arXiv:2511.02563，印度科学理工学院，2025年。[21]TomTom。TomTom 2025年交通指数排名。https://www.tomtom.com/traffic-index/ranking，2025年。访问日期：2026-02-19。[22] S. Vaze、V. Balasubramanian、M. Gygli、Y. Nestarev 和 A. Vedaldi。广义类别发现。发表于计算机视觉与模式识别会议，2022年。[23] Hao Frank Yang、Jiarui Cai、Chenxi Liu、Ruimin Ke 和 Yinhai Wang。基于边缘人工智能和表征学习的协同多摄像头车辆跟踪与交通监控。《交通研究*C*辑：新兴技术》，148:103982，2023年。[24] Xuan Zhou、Ruimin Ke、Hao Yang 和 Chenxi Liu。智能交通系统感知与边缘计算的交汇：愿景与挑战。《应用科学》，11(20):9680，2021年。