Ancilia：面向物联网人工智能的可扩展智能视频监控

Armin Danesh Pazho*，*IEEE*学生会员，Christopher Neff*，*IEEE*学生会员，Ghazal Alinezhad Noghre，*IEEE*学生会员，Babak Rahimi Ardabili，*IEEE*学生会员，Shanle Yao，Mohammadreza Baharani，*IEEE*会员，Hamed Tabkhi，*IEEE*会员

arXiv:2301.03561v2 [cs.CV] 9 Mar 2023

摘要—随着基于视觉的人工智能不断进步、物联网联网摄像头日益普及，以及社会对快速且公平安保的需求不断增长，对精准实时智能监控的需求达到了前所未有的高度。本文介绍 Ancilia，这是一个面向物联网人工智能的端到端可扩展智能视频监控系统。Ancilia 将最先进的人工智能应用于真实世界的监控应用，同时兼顾伦理问题，并实时执行高级认知任务。Ancilia 旨在革新监控领域，为该领域带来更有效、更智能、更公平的安保，从而打造更安全、更有保障的社区，同时无需人们牺牲其隐私权。

索引词——**监控、人工智能、物联网、计算机视觉、应用、真实世界、实时、边缘、异常。**

I. 引言

我们越来越需要有效且高效的监控技术，以保护我们的城市、人和基础设施。例如，在韩国梨泰院，一场节日庆祝活动因严重拥挤造成 150 多人死亡，许多人将这场悲剧归咎于政府监管不力 [1]。在北卡罗来纳州摩尔县，针对两座变电站的蓄意袭击导致 45,000 多名居民断电数日，技术人员忙于恢复供电，当局则努力寻找袭击源头 [2]。如果能通过智能视频监控提前充分预警，这些事件本可以避免。

随着物联网人工智能（AIoT）近期兴起，一些监控解决方案提供商开始在其系统中加入基础形式的人工智能。然而，这些方法仍然十分初级，无法以真正有意义的方式提升安保水平 [3]。这是因为，尽管许多研究关注的是有助于监控系统的任务，但大多数研究都聚焦于

实验室环境中的算法改进，而不是关注真实场景中普遍存在的因素 [4], [5]。大多数研究都聚焦于单一算法，以及如何对其进行调整，以便在现成的数据集上取得尽可能好的结果；然而，这些数据集往往无法反映真实的监控环境。很少有研究探讨不同算法如何影响多算法系统中其他下游算法的性能。仍然很少有研究探索噪声（包括数据产生的噪声和系统产生的噪声）对端到端准确率的影响。此外，现实世界智能监控还需要实时性能。先进人工智能的认知能力只有在能够足够迅速地提供给安保人员，使其在为时已晚之前采取适当行动时，才真正有用。

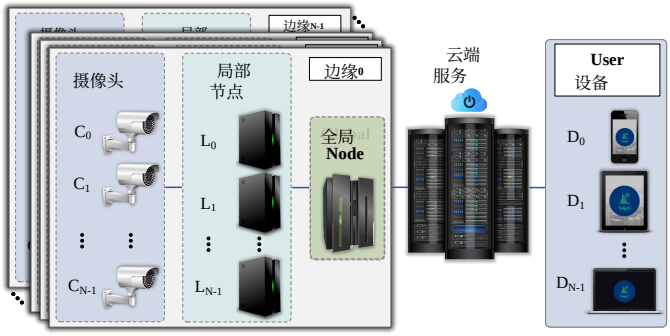

图 1. Ancilia 的概念概览。

本文介绍了 Ancilia，这是首个端到端、可扩展的智能视频监控系统，能够实时执行高级认知任务，同时达到最先进水平的成果。Ancilia 利用物联网（IoT）中摄像头的广泛普及，通过本地化服务器与现有物联网摄像头生态系统集成，从而便于在边缘进行处理。目前的物联网方法通常采用云计算，这可能导致延迟和隐私问题；或者需要配备强大处理能力的定制传感器。Ancilia 提供了一种利用现有物联网传感器的解决方案，最大限度地减少对昂贵基础设施升级的需求以及对云端处理的依赖。Ancilia 不受设备限制；只要能够访问摄像头的视频，Ancilia 就能提供智能分析。如图 1 所示，Ancilia 由三个逻辑和物理部分构成：边缘、云端和用户设备。边缘端利用大量先进的人工智能算法处理摄像头传来的数据，以实现智能安防。使用单台工作站进行边缘处理时，在中等和高人群密度的场景中，Ancilia 可实时监控多达 4 台摄像头（30 帧每秒），或多达 8 台摄像头（15 帧每秒）。Ancilia 执行高级认知任务（即动作识别、异常检测）时，与当前最先进水平（SotA）的准确度偏差为 ～ 1%。

作者均任职于美国北卡罗来纳州夏洛特市 28223 的北卡罗来纳大学夏洛特分校电气与计算机工程系。{adaneshp, cneff1, galinezh, brahimia, mbaharan, htabkhiv}@uncc.edu

* 共同通讯作者贡献相同。

Ancilia 从设计之初就致力于尊重受监控的人和社区的隐私。Ancilia 不会在任何数据库中存储个人身份信息，也不会采用人脸识别或步态检测等侵入性人工智能技术。与常见的身份信息识别不同，Ancilia 严格仅为高级任务（即动作识别、异常检测）提供姿态和位置信息。Ancilia 关注的是一个人在做什么，而不是他们是谁。这使 Ancilia 能够充当缓冲机制，帮助消除基于种族、族裔、性别、年龄和社会经济因素的偏见，从而有助于减少当局与边缘化社区之间不必要的冲突；此类冲突已日益成为严重问题。数据在边缘处理后，会发送至云端，以便与用户设备进行通信和服务管理。移动应用可让用户设备从云端接收数据，包括潜在安全问题出现时发出的警报。

总而言之，本文的贡献如下：

- 我们提出了 Ancilia，这是首个端到端、可扩展的真实场景智能视频监控系统，能够实时执行高级认知任务，同时达到最先进的准确率。
- 我们从隐私和公平性两个角度分析了智能视频监控的伦理问题，并说明 Ancilia 的设计如何专门应对这些问题。
- 我们使用两项与智能监控直接相关的高级认知任务——动作识别和异常检测——对 Ancilia 进行端到端实证评估，并研究实现实时性能所需的准确率权衡。
- 我们对 Ancilia 在不同硬件类别和不断增强的场景强度下的实时性能与可扩展性进行了全面的系统级评估，展示了 Ancilia 如何在不同情境中满足实时智能安保需求。

II. 相关工作

关于利用人工智能进行视频监控的研究层出不穷 [4], [6]–[8]。[9] 提出使用基于区域提议的光流来抑制背景噪声，并采用双向贝叶斯状态转移策略对运动不确定性进行建模，从而增强时空特征表示，以检测监控视频中的显著目标。 [10] 提出使用行人检测器、跟踪算法和掩码分类器，在监控视频流中跟踪行人。

在 [4], 中，研究指出，为解决实时视频监控的延迟问题，需要转向边缘计算。Nikouei 等人 [11]–[13]探讨了使用低功耗边缘设备在监控场景中执行目标检测和跟踪的可行性。他们认为，在最坏情况下，5 帧每秒的吞吐量也足以在监控应用中跟踪人员，因此可以将计算推向边缘。然而，他们的结果表明，即使是轻量级卷积神经网络也可能给低功耗设备带来问题，往往会使吞吐量降至 5 帧每秒的阈值以下。 [14] 提出了一种系统，利用低功耗嵌入式 GPU，在监控环境中执行检测、跟踪、路径预测、姿态估计和多摄像头重识别，同时着重关注实时执行以及被跟踪行人的隐私。 [15]提出了类似的系统，专注于目标检测、跟踪和多摄像头重识别，以提高吞吐量。 [16] 提出在边缘端结合使用轻量级目标检测模型，并在云端使用计算开销更大的模型，在两者之间分配计算任务，从而在建筑工地环境中实现实时视频监控。 [17] 提出采用背景检测、车辆检测和卡尔曼滤波器 [18]跟踪来进行停车场监控并确定停车场占用情况。 [19] 提出了一种系统，在监控场景中对行人进行目标检测、行人跟踪、场景分割以及联合轨迹与活动预测。

智能监控的未来正朝着能够执行高级认知任务的系统发展。一项聚焦于真实世界视频监控的近期调查 [4] 指出，尽管视频监控领域涵盖对静止物体、车辆、个人和人群的理解，但确定异常事件何时发生的能力对于智能监控系统至关重要。其他研究也支持这一观点 [6]。 [20] 利用无限隐马尔可夫模型和贝叶斯非参数因子分析，在视频流中寻找模式并检测异常事件。 [21]提出使用主动学习和模糊聚合，持续学习何种情况构成异常，并适应标准数据集中未出现的场景。[22] 提出了一个用于在商场监控环境中检测可疑行为的系统，采用分割、斑块融合和基于卡尔曼滤波器的跟踪等轻量级算法[18]。AnomalyNet [23] 是一种近期提出的循环神经网络，结合自适应迭代硬阈值和长短期记忆，直接基于像素信息运行，以消除背景噪声、捕捉运动，并学习稀疏表示和字典，从而执行视频监控中的异常检测。

III. 伦理问题

无论是传统形式还是较新的智能形式，视频监控始终与社会和伦理问题相关。在改善公共安全和安保的同时，尊重公民隐私和自主权，是这一领域最广为人知且长期存在的伦理问题[24]–[27]。只有考虑到这些问题，才能开发出成功的智能视频监控解决方案，解决公共安全问题，并在一定程度上让社区参与其中。

学者们日益关注在设计阶段纳入隐私问题，这一理念被称为“隐私保护设计” [28]。许多数据驱动型和基于人工智能的系统（例如智能视频监控技术）造成歧视和隐私侵犯，源于对个人身份信息（PII）的使用 [29], [30]。使用个人身份信息（PII），例如

在技术发展的任何阶段获取的人日常活动的真实影像，都可能增加隐私侵犯的风险。关于在不同领域使用人脸识别技术所带来的伦理挑战，以及使用这项技术如何导致隐私侵犯，长期以来一直存在争论 [31]–[34]。

在智能视频监控中执行高级认知任务（如动作识别和异常检测）的方法，可根据所用数据分为两类 [35]。第一类直接使用像素数据。一个常见示例是人脸识别 [36]，算法会查看人的面部图像以识别其身份。只要有足够的历史数据，这些算法就能表现良好，但通常被认为具有侵入性，并会增加识别个人人口统计信息的风险 [33]。第二类仅利用经过处理的信息，例如 Ancilia 使用的姿态数据，这类信息往往能够去标识化个人人口统计信息 [37]。这并不意味着个人身份信息已被彻底移除，因为一些研究仅凭步态 [38] 或轮廓 [39]，就能识别个人；但与基于像素的方法相比，这显著降低了隐私风险。

同样，不采用人脸识别技术并不能保证系统完全保护隐私。存储行人图像也是造成伦理违规的一个来源。从歧视的角度来看，使用任何形式的个人身份信息都可能加剧警务系统中的边缘化问题 [40]。因此，设计非歧视性系统的关键步骤，是确保系统不依赖个人身份信息。这要求在选择算法、所用数据类型以及数据存储方式时，采取针对性的技术设计方案。

Ancilia 通过不存储任何个人身份信息，也不在网络中传输任何个人身份信息来解决这一问题。此类数据在使用后即被销毁。对于所有高级认知任务，Ancilia 均采用基于姿态的方法，从而严格限制此类算法所使用的个人身份信息的数量和质量。这使得处理过程不会产生任何性别、族裔或基于阶级的歧视。因此，Ancilia 能够解决智能视频监控所涉及的许多隐私问题，同时也能应对歧视这一伦理问题。

IV. ANCILIA 算法框架

Ancilia 的算法核心分为两个概念系统：包含每个摄像头算法流程的本地节点，以及负责处理需要理解多个摄像机视角的所有任务的全局节点。这两个系统共同构成 Ancilia 的算法核心，也是实现所有更高层次理解的基础。图 2 展示了该算法核心的可视化表示。

A. 单摄像机视觉流程

如图 2 所示，局部算法流程始于从摄像头提取图像。首先，将图像输入目标检测器，以定位场景中的人、车辆、动物及其他重要物体。这一点很重要，不仅因为它是算法流程其余部分的基础，还因为它可用于基本的态势感知。有时，仅场景中出现某个特定物体就值得关注，例如有人出现在未经授权的区域、无人看管的包，或出现枪支。Ancilia 为此使用 YOLOv 5 [41] （不过也可以使用任何检测器）。请注意，YOLOv5 的作者提供的默认权重并未涵盖许多值得关注的物体类别。不过，已有其他研究针对枪支等类别训练了该架构 [42]–[44]，也始终可以训练自定义权重以适配目标应用。人员的位置坐标会发送给跟踪器，由跟踪器创建轨迹片段，并将每个人与其在先前图像中的检测结果进行匹配。Ancilia 使用了不包含帧相似度的 ByteTrack 版本 [45] 。在这种配置下，ByteTrack 不执行特征提取，从而显著减少计算量。如 [45], 所示，位置相似度足以实现单摄像头跟踪。跟踪使系统能够了解一个人在整个场景中的移动情况，这对许多监控应用至关重要。它还使 Ancilia 能够了解不同时刻的姿态分别属于哪些人，这对许多可提供急需态势感知的高级任务至关重要。图像中检测到的人员图像裁剪也会发送给人体姿态估计器，以生成二维姿态骨架。Ancilia 使用 HRNet[46]提取二维骨架。与直接使用原始像素数据相比，将姿态数据用于高级任务有两大优势。首先，姿态数据的维度远低于像素数据，因此计算成本低得多，使 Ancilia 能够实时运行。其次，姿态数据有助于去除像素数据中固有的基于外观的个人身份信息，使高级任务更难基于族裔、性别、年龄或其他身份相关指标形成无意的偏见。 [47], [48] 等研究尝试根据主体的姿态识别其身份，这类研究属于步态识别领域；但如第 III 节所述，与其他方法相比，基于姿态的方法具有更强的隐私保护能力。

B. 多摄像头行人再识别

虽然跟踪器会在单个摄像头内跟踪人，但仅凭位置信息无法准确地在多个摄像头之间对人员进行再识别。为此，发送至人体姿态估计器的相同人物裁剪图也会发送至行人重识别特征提取器，由其为每个人创建抽象特征表示。在单个批次期间，每个人仅创建一个特征表示，并且只有在能够确保该表示质量时才会创建，因为低质量的特征表示会妨碍准确的多摄像头行人再识别。Ancilia 使用特征表示过滤算法来验证人物裁剪图的两项质量指标。首先，人物裁剪图必须包含该人员的高质量视图。为此，过滤算法使用二维姿态骨架，并验证至少检测到 9 个关键点，且每个关键点的置信度至少为 60%。过滤算法会检查目标检测器生成的边界框之间的重叠（即交并比）。某个人的边界

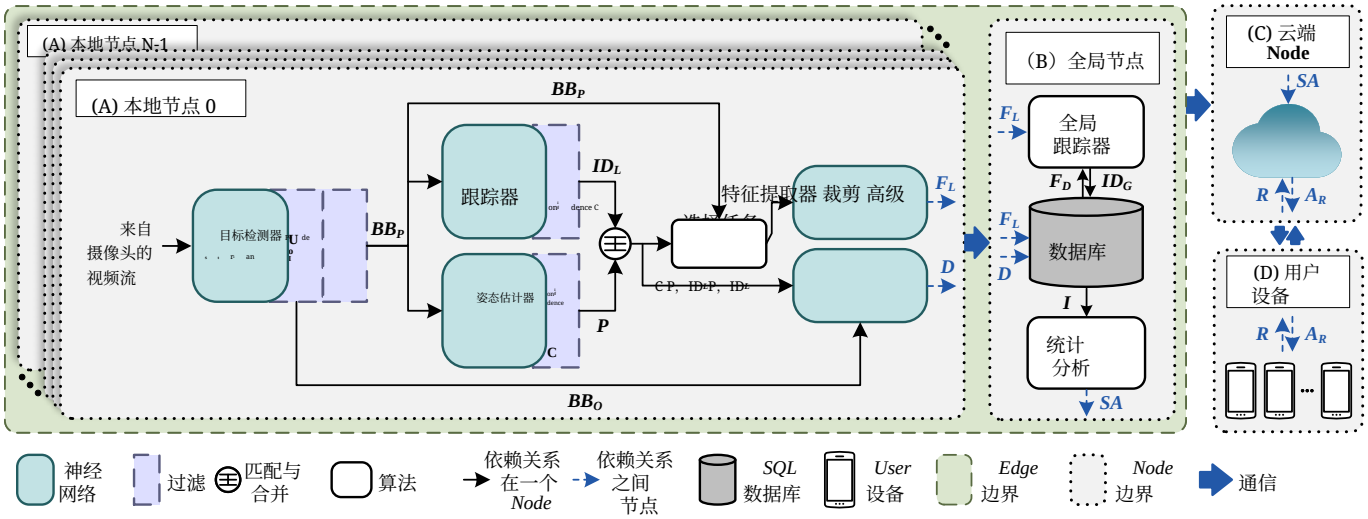

图 2. Ancilia 算法细节。 N 本地节点连接到边缘上的单个全局节点。最终分析结果会传输到云节点，为用户设备上的应用提供数据。多个边缘节点可以连接到云端，不过为清晰起见，图中仅展示了一个边缘节点。 BB_P、 BB_O、 ID_L、 P、 C、 F_L、 D、 F_D、 ID_G、 I、 SA、 R 和 A_R 分别表示行人的边界框、物体的边界框、本地身份、姿态、通过选择的人物裁剪图、本地节点的特征、下游任务的数据、数据库中的特征、全局身份、数据库中的信息、已完成的统计分析、来自用户的请求以及请求的属性。

框与任何其他人员的交并比（IoU）不得超过 0.1。如果满足这两个条件，则认为人物裁剪图的质量足够高，可以生成适当的特征表示。如果在 15 帧窗口内，同一个人的多个裁剪图都被认为合适，则选择姿态置信度最高的裁剪图。特征提取器创建的特征会发送至全局节点，用于多摄像头行人再识别。Ancilia 使用 OSNet [49]提取特征表示。

C. 更高级别的任务

为从系统层面保护隐私，所有高级任务均在本地节点上执行，从而将敏感信息保留在本地机器上。这些任务可以访问前述步骤生成的对象、跟踪和姿态数据。由于所需高级任务的选择高度依赖于具体应用，我们不将这些任务视为 Ancilia 算法核心的一部分，而是将其作为可根据预期用途定制的扩展。本文以动作识别和异常检测这两种与智能监控密切相关的常见高级任务为例。对于动作识别，我们选用 PoseConv3D [50]和 CTR-CGN [51], 这两种最先进的网络，它们能够利用 Ancilia 提供的二维人体姿态骨架。对于异常检测，我们采用 GEPC[52] 和 MPED-RNN [53], ，它们均基于二维人体姿态骨架。

许多研究采用基于像素的方法来完成这些任务，其准确率优于最先进水平的基于姿态的方法，例如 I3D [54], MVIT [55] 和 Stargazer[56]。Argus [57] 是采用像素级信息的系统的一个典型例子，随后在名为 Meva[58]的真实世界监控数据集上进行了评估。然而，鉴于第 III 节讨论的隐私优势，以及使用低维姿态数据所带来的计算优势，我们决定在 Ancilia 中继续采用基于姿态的方法。

V. 系统设计

除算法设计外，还可以从系统级设计与实现的角度分析 Ancilia。尤其是本地节点，其系统设计较为复杂，如图 3 所示。全局节点和云端则简单得多，如图 2 所示。

A. 并行性

Ancilia 的一项关键设计目标是通过平衡吞吐量与延迟来提高效率。Ancilia 采用流水线处理来利用进程并行，将任务划分为流水线系统中的六个独立阶段（$S1$, $S2$, ……, $S6$）。每个阶段均作为独立进程实现，一旦收到所需输入，便与其他进程并发执行。这些阶段通过队列相互通信，以更好地利用内存资源并实现快速的进程间通信。虽然流水线处理是一种广为人知的优化技术，但其实现会带来开销，因此需要找到平衡点。图 3 展示了本地节点系统设计的详细视图。每个流水线阶段之间都由一个队列隔开，队列大小上限为 λ 个元素，以防流水线阶段执行速度不均导致潜在的溢出。默认情况下，Ancilia 使用的值为 λ 。与常见做法一样，Ancilia 将依赖神经网络的高并行任务（即目标检测、姿态估计、特征提取以及许多高级任务）卸载到图形处理器（GPU）上执行。

B. 数据批处理

批处理是 Ancilia 用来更好地利用硬件资源的另一种技术。通常，批处理能够

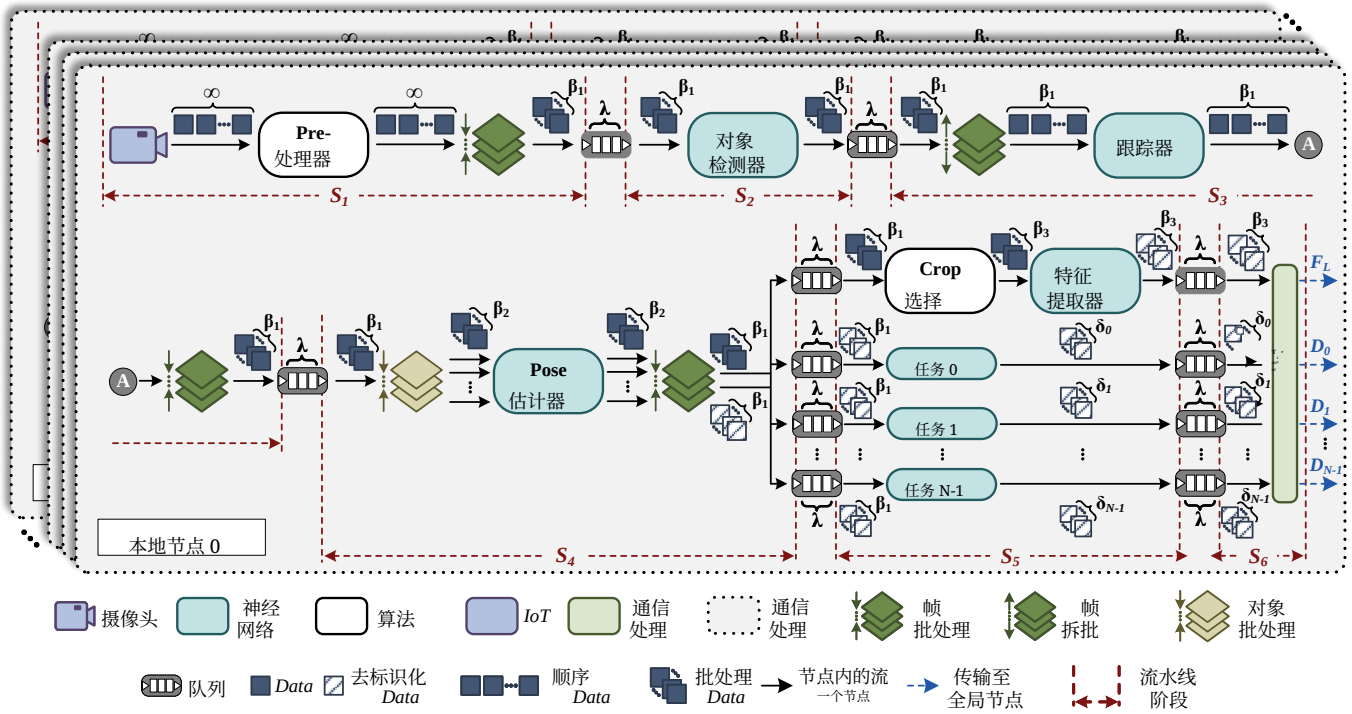

图 3. Ancilia 本地节点中系统设计的详细视图。 β 和 δ 指不同的批次大小。 λ 指队列大小。 F_L 和 D 分别表示本地特征和从下游任务接收的数据。

大幅提高系统的吞吐量，但代价是增加端到端延迟。然而，许多高级任务（例如动作识别、异常检测）需要多个视频帧的输入数据（通常称为窗口）才能开始处理。因此，只要帧批次与高级任务窗口大小相同，批处理输入帧本身会产生的延迟便已是这些高级任务固有的一部分。换句话说，如果某个高级任务需要X 帧才能开始处理，那么批次大小为 X 帧即可确保该任务同时获得所需的全部帧，不会给该任务带来额外延迟。如果高级任务的窗口大小大于批次大小，则需要处理多个批次，才能获得该高级任务的输出。此外，由于帧批处理最终会提高吞吐量，与顺序处理每一帧相比，端到端延迟会有所降低。目标检测针对整帧进行处理，而 Ancilia 中的所有其他神经网络则处理单个对象。在输入网络之前，这些对象会被批量处理，从而大幅提高硬件利用率。根据视频中检测到的相关对象数量，一个帧批次中可以包含多个对象批次。

C. 本地节点

1) S_1 - 预处理： 本地节点从摄像头接收到视频流后，预处理器负责在将帧送入算法核心之前进行所有必要的基础图像处理。这包括必要的调整尺寸、丢帧和／或颜色通道重新排序。丢帧是一种动态机制，可确保输入流水线的帧率与其吞吐量相匹配。

例如，如果帧源（即摄像头）以 60 帧每秒的速度生成帧，而 Ancilia 只能以 30 帧每秒的速度运行，那么源中的帧只有每隔一帧才会经过预处理。预处理后，帧会按顺序分组成大小为 β_1 的片段。Ancilia 设置 $\beta_1 = 15$。如第 VI-D 节所述，这样做是为了平衡吞吐量和延迟，同时也能更好地匹配高级任务的窗口大小，使这些任务只需完成两个批次即可生成输出。这也很合适，因为大多数现代安保和物联网摄像头录制视频时的帧率为 30 或 60 帧每秒。

2) S_2 - 目标检测： 批处理后的帧被发送到目标检测器，后者输出一个对象列表，其中包含类别标签和边界框坐标。行人的边界框被发送到跟踪器，而其他对象的边界框则传递至系统中，用于高级任务和统计分析。稍后阶段会将原始帧中每个行人的裁剪图像传递给姿态估计器。

3) S_3 - 跟踪： 在跟踪器处，行人的边界框会被拆分为单个样本，以适应跟踪器的顺序操作。跟踪器会对行人进行分组，并将他们与此前检测到的行人进行匹配，或为他们分配唯一的局部 ID。随后，行人会再次按帧组成批次，恢复为原始的 $\beta_1 = 15$ 帧批次大小，并发送给姿态估计器。

4) S_4 - 姿态估计： 在姿态估计器处，对人物裁剪图进行目标批处理，批次大小为 $\beta_2 = 32$。这些批次被送入姿态估计器，后者为每个人物裁剪图输出人体姿态骨架。随后，再次按帧对行人边界框、人物裁剪图、本地标识符和人体

姿态骨架进行批处理，并与目标检测器输出的目标边界框合并。选择的数据（行人边界框、人物裁剪图、本地标识符和姿态骨架）会发送至裁剪选择和特征提取；而去标识化数据（行人边界框、目标边界框、本地标识符和姿态骨架）则根据各高级任务的请求发送给相应任务。

5) S_5 - 特征提取和高级任务：在特征提取之前，裁剪选择会根据边界框重叠和关键点置信度过滤掉低质量的人物裁剪图，如第 IV 节所述。默认情况下，交并比高于 0.1 的裁剪图，或有 9 个及以上关键点的置信度低于 60% 的裁剪图，都会被丢弃。这些阈值可以调整，以更好地适应目标应用。对于剩余的裁剪图，每个人只会选取总体关键点置信度最高的一张。剩余裁剪图会被分批处理，批次大小为动态值 β_3，根据场景中的人数确定。特征提取器接收一批 β_3 裁剪图。提取特征后，这些特征会被传输到全局节点。每个高级任务以大小为 β_1 的帧批次为粒度接收数据，并以该任务的运行粒度（δ_0，δ_1，……，δ_n）向全局节点发送数据。仅向高级任务发送去标识化数据，以符合第 III 节和第 IV 节中提到的伦理问题。每个高级任务都有自己的进程，并与其他任务以及流水线第 5 阶段的裁剪选择和特征提取并行运行。

6) S_6 - 传输：通信与流水线完全解耦，因此数据发送后，本地节点流水线会继续正常运行，无需等待全局节点的响应。重要的是，任何可识别身份的信息都不会发送至全局节点，这符合第 III 节中提到的隐私和伦理问题。

D. 全局节点

所有接收到的数据都存储在全局节点上的关系数据库中。第 IV 节所述的匹配算法会在 T 这段时间内，将接收到的特征与数据库中的现有特征进行比较，并根据结果分配全局标识符。T 的默认值设为 1 小时，但应根据应用需求进行调整。多种算法会使用关系数据库进行统计分析，详情见第 IV 节。分析结果通过云服务提供商提供的API接口传输至云节点。默认情况下，Ancilia 使用亚马逊云科技，但可根据用户／应用需求进行更改。云端（例如亚马逊云科技（AWS））接收来自全局节点的分析数据。

VI. 实验结果

A. 算法核心

Ancilia 的算法核心由多个算法组成，每个算法都基于前序算法生成的数据运行。由于这些算法依赖并不完美的神经网络，因此会产生噪声，并在整个系统中不断累积。要了解这些噪声的来源，我们必须先单独考察这些核心算法各自的准确率

表 I Ancilia 算法核心网络的独立准确率。最先进算法代表当前可实现的最高性能，此时无需考虑计算和延迟。

Task	方法	性能	数据集
Ancilia 的算法核心			
对象检测	YOLOv5 [41]	49.0（平均精度均值）	COCO [59]
跟踪	ByteTrack [45]	77.8（多目标跟踪准确度）	MOT20 [60]
Pose 估计	HRNet [46]	75.1（平均精度）	COCO [59]
人物 ReID	OSNet [49]	88.6（Top-1）	DukeMTMC [61]
最先进算法			
对象检测	InternImage [62]	65.0（平均精度均值）	COCO [59]
跟踪	SOTMOT [63]	77.9（多目标跟踪准确度）	MOT20 [60]
Pose 估计	ViTPose [64]	81.1（平均精度）	COCO [59]
人物 ReID	Centeroids-ReID [65]	95.6（Top-1）	DukeMTMC [61]

。表 I 展示了算法核心的四项主要任务——目标检测、行人跟踪、人体姿态估计和行人重识别——各自的准确率。表中还列出了每项任务中最先进水平型号的准确率。这些最先进水平的方法并不适用于智能监控应用，因为其过高的计算开销和庞大的参数量使其无法实时运行；不过，通过比较，我们可以了解当前研究所能达到的最高潜力，以及为使 Ancilia 保持实时运行而付出的准确率代价。

目标检测的准确率受影响最大，与最先进水平相比下降了 16%。这很容易理解，因为 YOLOv5 [41] 不仅是算法核心中规模最大的型号，也是唯一直接处理原始摄像头视频流的型号。因此，尽管有更大的型号可用，且能够提供更高的准确率，但即使型号规模或计算量略有增加，吞吐量也会明显下降。人体姿态估计的准确率也因类似原因而有所降低，不过幅度小得多，仅为 6%。虽然 HRNet [46] 并非处理原始摄像头视频流，但它会针对目标检测器检测到的每个人分别运行。因此，保持型号规模较小更为可取。行人重识别的准确率降幅略高于人体姿态估计，为 7%。这部分是因为采用了轻量级型号，而 OSNet [49], 这一行人重识别的最先进型号同样轻量。不过，最先进水平采用的基于质心的检索方法并不适用于笔集重识别，而大多数监控场景都属于这种情况。行人跟踪的准确率几乎没有下降，约为 0.1%。这是因为在单个摄像头下跟踪行人相对容易，像 ByteTrack [45]这样的简单轻量级算法，其性能几乎不逊于顶尖的最先进方法。

B. 高级任务

为了更好地了解算法核心产生的噪声如何影响整体性能，以及 Ancilia 在现实世界智能

监控领域中的表现如何，我们考察了在 Ancilia 上运行两项高层认知监控任务时的性能。为了使 Ancilia 能够助益智能监控任务，我们必须确保不会出现过多误报或漏掉正事件的情况。为此，我们选择了动作识别和异常检测，因为这些任务可以利用算法核心生成的人体姿态信息，从而实现更快速、偏见更少的推理。由于这两种方法都利用每个个体的人体姿态时间批次，这些实验将直接反映 Ancilia 生成的目标检测、跟踪、再识别和姿态估计数据的质量。

1）高级任务 - 动作识别：我们选择了两个最先进水平的动作识别模型 PoseConv3d [50]和 CTR-GCN [51], , 并使用 Ancilia 生成的数据对它们进行训练。对于每个模型，我们在 NTU60-XSub [66]上分别以完整（30 帧每秒）和一半（15 帧每秒）吞吐率进行训练和测试。两个模型的窗口大小均为 30，使用随机梯度下降（SGD）训练 24 轮次，动量为 0.9，并采用余弦退火调度。PoseConv3d 和 CTR-GCN 的权重衰减分别为 $3e^{-4}$ 和 $5e^{-4}$ ，初始学习率分别为 0.4 和 0.2。

表 II PoseConv3D [50] 和 CTR-GCN [51]在完整和一半吞吐率模式下于 NTU60-XSUB [66] 上的 Top-1 和 Top-5 准确率。

型号	Data	FPS	Top-1（%）	Top-5（%）
PoseConv3D [50]	[67]	15	91.96	99.47
		30	92.76	99.57
	Ours	15	88.79	98.82
		30	91.99	99.28
CTR-GCN [51]	[67]	15	86.36	98.46
		30	83.07	98.26
	Ours	15	81.58	97.52
		30	80.44	97.2

这些实验的结果见表 II。我们报告了 Top-1 和 Top-5 准确率，并比较了使用 Ancilia 生成的数据与 PYSKL 工具箱提供的原始数据所得的结果 [67]。可以看出，Ancilia 能够提供与原始数据质量相当的数据；在 Ancilia 中，作为高级任务的动作识别使用 PoseConv3D [50]以全吞吐量运行时，准确率比使用原始数据低约 1%，半吞吐量时则低约 3%。使用 CTR-GCN [51],时，与原始数据相比，Ancilia 在全吞吐量下的准确率下降 2.5%，在半吞吐量下下降 4.8%。由此可推断，PoseConv3D 比 CTR-GCN 对噪声更具鲁棒性，但两者使用 Ancilia 生成的数据时表现都相当不错，证明了 Ancilia 在智能监控应用中的有效性。

另一个值得注意的发现是，CTR-GCN [51]在半吞吐率下的表现明显优于全吞吐率。这意味着，CTR-GCN 更适合利用半吞吐率所允许的更长时间窗口。在无法保证实时吞吐率达到 30 帧每秒时，选择动作识别模型就需要考虑这一点。

2）高级任务——异常检测：我们使用 ShanghaiTech 数据集 [68]训练了两种最先进的异常检测模型 GEPC [52] 和 MPED-RNN [53], , 训练数据分别采用 Ancilia 生成的数据和原作者提供的数据。训练策略与第 VI-B1 节相同，两种模型均分别以全速（20 帧每秒）和半速（10 帧每秒）模式进行训练。GEPC 使用 Adam 优化器训练 25 个轮次，窗口大小为 30 、步长为 20，学习率为 1e-4 、权重衰减为 1e-5，批次大小为 512。MPED-RNN 的输入窗口大小为 30，重建窗口为 12，预测窗口为 6。该型号使用 Adam 优化器训练 5 个轮次，学习率为 $1e^{-3}$ ，批次大小为 265。

表 III GEPC [52] 和 MPED-RNN [53] 在 ShanghaiTech 数据集的全吞吐率模式和半吞吐率模式下的 ROC 曲线下面积、PR 曲线下面积及等错误率 [68] 。

型号	Data	FPS	ROC 曲线下面积	PR 曲线下面积	EER
GEPC [52]	[52]	10	0.6906	0.5951	0.35
		20	0.7372	0.6427	0.31
	Ours	10	0.6888	0.5905	0.35
		20	0.7223	0.6023	0.32
MPED-RNN [53]	[53]	10	0.6645	0.5733	0.37
		20	0.7023	0.5869	0.36
	Ours	10	0.6685	0.5661	0.37
		20	0.6679	0.5487	0.37

本实验的结果见表 III。按照当前惯例，我们报告受试者工作特征曲线下面积（ROC 曲线下面积）、精确率-召回率曲线下面积（PR 曲线下面积）和等错误率。使用 GEPC 时可以看到，Ancilia 完全胜任这项任务：在满吞吐量下，ROC 曲线下面积仅下降 1.5%；在半吞吐量下，ROC 曲线下面积的降幅不到 0.2%。PR 曲线下面积的降幅更为明显，在满吞吐量下下降 4%，但在半吞吐量下则降至不到 0.5%。等错误率几乎完全相同，使用 Ancilia 时几乎没有变化（不到 0.01）。MPED-RNN 起初在各方面的总体准确率都较低，在满吞吐量下，其 ROC 曲线下面积降幅更为显著，下降了 3.5%。然而，在半吞吐量下，使用 Ancilia 后，ROC 曲线下面积反而有所提升，尽管仅提升了 0.5%。PR 曲线下面积的结果与 GEPC 相似：在满吞吐量下下降 3.8%，在半吞吐量下下降 0.7%。等错误率仍然几乎完全相同。能够在执行异常检测这类高级任务的同时，将准确率维持在与当前研究中的最先进水平相近的程度，这证明 Ancilia 能够生成适用于智能监控应用的高质量数据。

C. 实时系统性能

算法准确率对于确保高级认知任务提供的信息能够为监控应用带来益处至关重要。然而，Ancilia 的实时执行能力同样重要。我们开展了一系列实验，评估 Ancilia 在不同硬件、不同场景强度下，以及每个硬件设备上本地节点数量不断增加时的运行时性能。我们

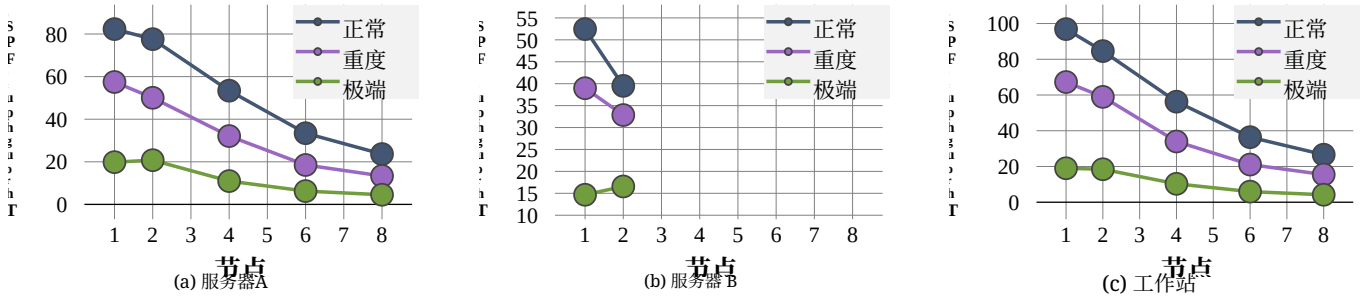

图4. 吞吐量 Ancilia在不同人群密度下随节点数量变化的吞吐量。硬件详情见表IV .

表 IV 系统配置。数据为所列型号的每个 CPU/GPU 的统计值。

Name	处理器			GPU		
	型号	核心数	时钟频率	型号	CUDA核心	VRAM
服务器A	2×EPYC 7513	32	2.6 GHz	4× V100	5120	32 GB
服务器 B	2×Xeon E5-2640 v4	10	2.4 GHz	2×Titan V	5120	12 GB
工作站	Threadripper Pro 3975WX	32	3.50 GHz	3× A6000	10752	48 GB

表 V 平均吞吐量和延迟。数据采集自本地节点数量不同的工作站。

人群密度	节点	服务器A		服务器 B		工作站	
		FPS	延迟（秒）	FPS	延迟（秒）	FPS	延迟（秒）
正常（70 次检测每秒）	1	82.31	1.17	52.45	1.52	96.88	0.87
	2	77.59	1.15	39.50	2.05	84.57	1.00
	4	53.40	1.60	-	-	56.27	1.58
	6	33.43	1.99	-	-	36.40	2.27
	8	23.64	2.05	-	-	26.60	2.84
重度（216 次检测每秒）	1	57.47	1.80	38.97	2.62	67.25	1.53
	2	50.05	2.07	32.85	3.16	58.95	1.77
	4	32.09	3.45	-	-	33.99	3.98
	6	18.51	4.17	-	-	21.08	6.89
	8	13.35	5.87	-	-	15.48	9.54
极端（每秒 744 次检测）	1	19.84	5.29	14.67	7.37	19.00	5.73
	2	20.76	5.09	16.56	6.54	18.45	5.81
	4	10.95	11.64	-	-	10.29	12.49
	6	6.25	20.70	-	-	5.93	21.87
	8	4.53	28.48	-	-	4.18	31.19

重点关注本地节点的性能，因为全局节点与算法流水线完全解耦，不会对吞吐量或延迟产生明显影响。

我们为这些实验选择了三种不同的硬件配置：高端服务器、低端服务器和高端工作站，如表IV所示。针对我们的场景，我们使用 DukeMTMC-video 数据集 [61] ，并挑选了三个具有不同人群密度的场景：正常密度、重度密度和极端密度。各场景中的检测密度分布及其对吞吐量的影响见图5。请注意，“正常密度”的定义会因应用环境而异，因此我们报告了如此广泛的密度范围。每段视频包含 32k 帧，其中 7k 帧用于预热和冷却。我们在单个系统上使用 1、2、4、6 和 8 个本地节点进行测试，以展示吞吐量和延迟在这些情况下的扩展情况。每项实验均进行三次，吞吐量和延迟取多次运行的平均值。实验结果见表V和图4。

在人群密度正常的情况下，服务器A和工作站在最多运行四个本地节点时，均可达到超过 50 帧每秒，端到端延迟分别为 1.60 秒和 1.58 秒。这一帧每秒数远高于动作识别和异常检测算法在满负载下所需的帧每秒数，且延迟足够低，适用于大多数监控应用；在这些应用中，主要关注的是及时通知当局，以便采取适当应对措施。在正常场景下，服务器 A 和工作站均可处理 6 个本地节点，同时保持在 30 帧每秒以上。工作站在运行全部 8 个本地节点时，仍能保持在 26 帧每秒以上；而服务器A在运行 8 个本地节点时则降至略低于 24 帧每秒。服务器 B 在运行单个节点时可达到超过 50 帧每秒，但同时处理两个节点时则略低于 40 帧每秒。由于只有两块显存有限的图形处理器，服务器 B 无法同时运行 4 个或更多节点。

高人群密度带来了更大的挑战，服务器A和工作站都只能在最多使用4个节点时达到30帧每秒以上。端到端延迟也比正常人群密度下更长，在大多数情况下，所有系统的延迟都增加了50%至100%，在最极端的情况下甚至增加了230%。服务器A和工作站分别能够在6个和8个节点下保持15帧每秒以上，而服务器A在8个节点下则降至略高于13帧每秒。服务器 B 的表现与其在

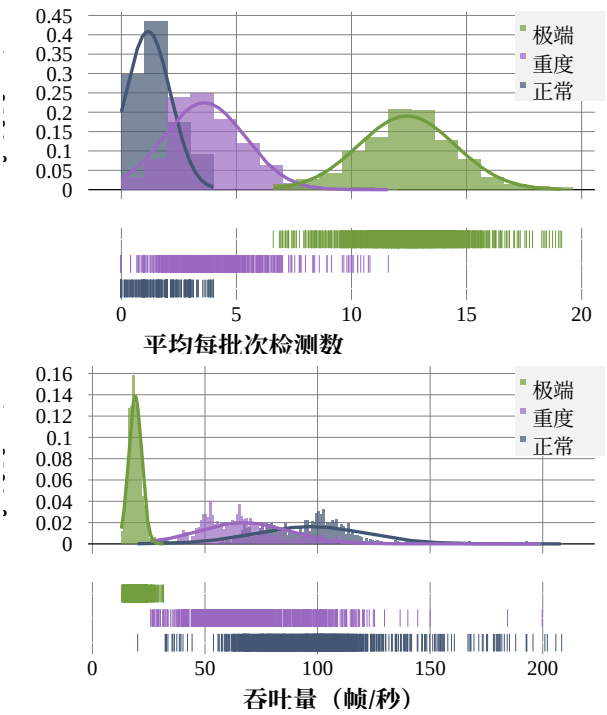

图 5. 不同人群密度下的检测分布及其对吞吐量的影响。数据由使用单个本地节点的工作站采集。

正常人群密度下类似，最多使用2个节点时仍能保持30帧每秒以上，但吞吐量略低。假设高级任务只需要一半的吞吐量，服务器 B 仍适合运行最多两个节点。

在人群密度极端的场景下，Ancilia 开始难以应对。即使只使用一个摄像头，所有系统的帧率都无法达到 30 帧每秒以上，因此无法实现满吞吐量的动作识别。使用 2 个节点时，服务器A的帧率能够达到 20 帧每秒以上（但值得注意的是，使用 1 个节点时无法达到）；而工作站即使使用 1 个节点也达不到这一水平。服务器A和工作站在使用 4 个节点时都能保持 10 帧每秒以上，但使用 6 个和 8 个节点时，帧率分别降至约 6 帧每秒和 4 帧每秒。 [69] 认为，5 帧每秒适用于行人跟踪；虽然确实如此，但依赖详细人体运动信息的高级任务（例如动作识别和异常检测）在帧率低于 10 帧每秒时，往往难以保证准确性。另一个问题是延迟增加。使用 6+ 个节点运行时，服务器 A 和工作站的延迟均超过 20 秒；这对许多监控应用而言尚可接受，但对于需要更快响应的应用来说可能过高。吞吐量偏低，再加上延迟较高，因此在极端人群密度下使用 Ancilia 时，难以建议在单个系统上运行超过 4 个节点；低吞吐量和高延迟并非主要顾虑的应用除外。服务器 B 无法达到 30 帧每秒，但在使用 1 个和 2 个节点时都能维持在约 15 帧每秒，因此适合以半吞吐量运行动作识别和异常检测。

有趣的是，在极端人群密度下，我们开始观察到异常情况：两台服务器在使用单个节点时的性能都不如使用两个节点时。这很可能是因为它们采用双CPU配置，可用的CPU资源十分充足，而单个节点无法充分利用这些资源。

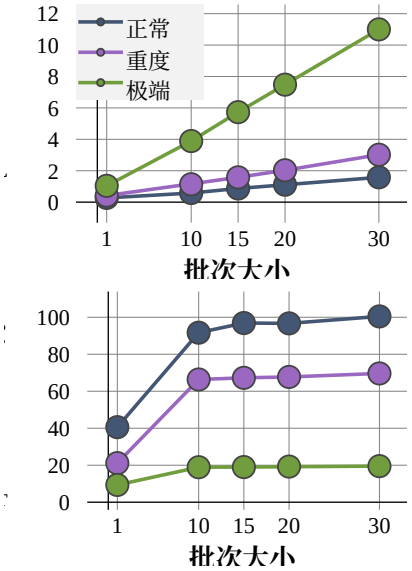

图 6. 不同人群密度下，吞吐量和延迟随批次大小的变化趋势。数据采集自配备单个本地节点的工作站。

表 VI 不同批次大小对吞吐量和延迟的影响。

人群密度	批次大小	FPS	延迟（秒）
正常	1	40.58	0.27
	10	91.66	0.58
	15	96.88	0.87
	20	96.69	1.11
	30	100.46	1.58
重度	1	21.13	0.42
	10	66.35	1.14
	15	67.25	1.58
	20	67.73	2.03
	30	69.62	3.02
极端	1	9.34	1.04
	10	18.92	3.89
	15	19.00	5.73
	20	19.28	7.48
	30	19.52	11.00

因此，在极端人群密度场景下，只有同时运行多个节点时，两台服务器的表现才开始符合预期，并与其他系统相似。对此不必过于担心，因为只运行一个本地节点时，购买这种高端服务器级机器并不划算；相比之下，更注重延迟的工作站价格更低，效果也更好。

总体而言，Ancilia 能够满足高级认知任务的需求，同时实现适用于实时智能监控应用的性能。具体性能取决于所用硬件和场景强度，但这些结果表明，即使在最极端的场景下，Ancilia 也可用于为监控应用提供智能辅助。

D. 批次大小对实时性能的影响

为了解批次大小对端到端延迟和吞吐量的影响，我们在工作站上使用单个节点进行测试，同时改变批次大小。结果如图 6 所示。正如预期，随着

所有密度下批次大小的增加，吞吐量和延迟都会增加。批次大小从 1 增加到 10 所带来的吞吐量跃升最为显著；使用更大的批次大小时，收益逐渐减少，而延迟的增加则趋于更加成比例。然而，由于高级任务需要 30 帧，端到端延迟并不能直接代表执行高级任务的延迟。

总而言之，需要在吞吐量、端到端延迟和批次大小之间取得平衡。端到端延迟过高会影响检测到的感兴趣目标触发警报的速度，而吞吐量较低则会影响高级任务的准确性，如第 VI-B 节所示。同样，批次大小过小意味着在高级任务开始运行之前需要处理更多批次。批次大小为 15 时，能够很好地实现这种平衡：在正常人群密度下，端到端延迟仅为 0.87 秒（不到一秒），吞吐量为 96.88 帧每秒；高级任务只需处理两个批次即可运行。对于重度和极端人群密度，结果也类似，不过正如预期的那样，吞吐量更高、延迟更低。

VII. 结论

本文介绍了 Ancilia，这是一种面向物联网人工智能的端到端可扩展智能视频监控系统。通过实证评估，Ancilia 展示了其将最先进的人工智能应用于真实世界监控应用的能力。Ancilia 能够实时执行高级认知任务（即动作识别和异常检测），同时兼顾监控应用中普遍存在的伦理和隐私问题。

致谢

本研究由美国国家科学基金会（NSF）资助，资助编号为 1831795；另获美国国家科学基金会研究生研究奖学金，奖学金编号为 1848727。

参考文献

[1] C. Alexandre，“梨泰院惨案对公共安全的影响”，2022年12月。 [在线]。获取地址：https://thediplomat.com/2022/12/the-public-safety-implications-of-the-itaewon-tragedy/ [2] N. Salahieh、J. Miller 和 H. Yan，“北卡罗来纳州居民恢复供电之际，调查人员调查针对美国各地变电站的恐怖主义行为与威胁。一位专家解释了需要采取哪些措施”，2022年12月。 [在线]。获取地址：https://www.cnn.com/2022/12/08/us/power-outage-moore-county-investigation-thursday/index.html [3] S. Feldstein 和卡内基国际和平基金会，《人工智能监控的全球扩张》。卡内基国际和平基金会，2019年。 [在线]。获取地址：https://books.google.com/books?id=W9JQzQEACAAJ [4] M. R. Patrikar、Devashree R. Parate，“视频监控系统中基于边缘计算的异常检测：综述”，《国际多媒体信息检索期刊》，第85–110页，2022年。 [在线]。获取地址：https://doi.org/10.1007/s13735-022-00227-8 [5] A. Danesh Pazho、G. Alinezhad Noghre、A. A. Purkayastha、J. Vempati、O. Martin 和 H. Tabkhi，“复杂分布式系统异常检测的基于图的深度学习方法综述”，*arXiv*预印本 *arXiv:2206.04149*，2022年。[6] X. Li 和 Z.-m. Cai，“监控视频中的异常检测技术”，载于*2016*年第九届图像与信号处理、生物医学工程与信息学国际大会（*CISP-BMEI*），2016年，第54–59页。[7] T. Li、H. Chang、M. Wang、B. Ni、R. Hong 和 S. Yan，“拥挤场景分析：综述”，*IEEE*视频技术电路与系统汇刊，第25卷，第3期，第367–386页，2015年3月。 [在线]。获取地址：https://doi.org/10.1109/TCSVT.2014.2358029

[8] B. S. Shobha 和 R. Deepu，“基于视频的车辆检测、识别与跟踪综述”，载于《2018年第三届可持续解决方案计算系统与信息技术国际会议》（CSITSS），2018年，第183–186页。[9] J. Zhang、C. Xu、Z. Gao、J. J. P. C. Rodrigues 和 V. H. C. de Al- buquerque，“基于工业普适边缘计算的智能物联网监控显著性检测”，《IEEE工业信息学汇刊》，第17卷，第7期，第5012–5020页，2021年。[10] G. T. Draughon、P. Sun 和 J. P. Lynch，“用于跟踪和可视化城市环境中口罩佩戴情况的计算机视觉框架的实现”，载于《2020年IEEE国际智慧城市会议》（ISC2），2020年，第1–8页。[11] R. Xu、S. Y. Nikouei、Y. Chen、A. Polunchenko、S. Song、C. Deng 和 T. R. Faughnan，“面向边缘智能监控的实时人体目标跟踪”，载于《2018年IEEE国际通信会议》（ICC），2018年，第1–6页。[12] S. Y. Nikouei、Y. Chen、S. Song、R. Xu、B.-Y. Choi 和 T. Faughnan，“作为边缘网络服务的智能监控：从Haar级联、支持向量机到轻量级卷积神经网络”，载于《2018年IEEE第四届协作与互联网计算国际会议》（CIC），2018年，第256–265页。[13] S. Y. Nikouei、Y. Chen、S. Song、B.-Y. Choi 和 T. R. Faughnan，“利用轻量级检测与跟踪算法，推动作为边缘网络服务的智能监控（ISENSE）”，《IEEE服务计算汇刊》，第14卷，第6期，第1624–1637页，2021年。[14] C. Neff、M. Mendieta、S. Mohan、M. Baharani、S. Rogers 和 H. Tabkhi，“Revamp2t：面向隐私保护行人跟踪的实时边缘多摄像头视频分析”，《IEEE物联网期刊》，第7卷，第4期，第2591–2602页，2020年。[15] B. Gaikwad 和 A. Karmakar，“用于边缘实时多人多摄像头跟踪的智能监控系统”，载于《实时图像处理期刊》，第18卷，2021年。[16] Y. Zhao、Y. Yin 和 G. Gui，“基于轻量级深度学习的智能边缘监控技术”，《IEEE认知通信与网络汇刊》，第6卷，第4期，第1146–1154页，2020年。[17] R. Ke、Y. Zhuang、Z. Pu 和 Y. Wang，“基于物联网设备边缘人工智能的智能、高效且可靠的停车监控系统”，《IEEE智能交通系统汇刊》，第22卷，第8期，第4962–4974页，2021年。[18] R. E. Kalman，“线性滤波与预测问题的一种新方法”，《ASME基础工程学报汇刊》，第82卷，D系列，第35–45页，1960年。[19] J. Liang、L. Jiang、J. C. Niebles、A. G. Hauptmann 和 L. Fei-Fei，“窥探未来：预测视频中人物未来的活动与位置”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019年6月。[20] V. Nguyen、D. Phung、D.-S. Pham 和 S. Venkatesh，“视频监控中异常检测的贝叶斯非参数方法”，《数据科学年鉴》，第2卷，第21–41页，2015年。 [在线]。可访问：https://doi.org/10.1007/s40745-015-0030-3 [21] R. Nawaratne、D. Alahakoon、D. De Silva 和 X. Yu，“利用深度学习进行实时视频监控的时空异常检测”，《IEEE工业信息学汇刊》，第16卷，第1期，第393–402页，2020年。[22] R. Arroyo、J. J. Yebes、L. M. Bergasa、I. G. Daza 和 J. Almaz´an，“用于实时检测购物中心可疑行为的专家视频监控系统”，《专家系统与应用》，第42卷，第21期，第7991–8005页，2015年。 [在线]。可访问：https://www.sciencedirect.com/science/article/pii/S0957417415004182 [23] J. T. Zhou、J. Du、H. Zhu、X. Peng、Y. Liu 和 R. S. M. Goh，“Anomalynet：视频监控异常检测网络”，《IEEE信息取证与安全汇刊》，第14卷，第10期，第2537–2550页，2019年。[24] J. Pierce、R. Y. Wong 和 N. Merrill，“传感器照明：探索智能摄像头和图像/视频分析的设计特性与伦理影响”，载于《2020年CHI计算系统中的人为因素会议论文集》，丛书：CHI '20。美国纽约：计算机协会，2020年，第1–19页。 [在线]。可访问：https://doi.org/10.1145/3313831.3376347 [25] H. Nissenbaum，“作为情境完整性的隐私”，《华盛顿法律评论》，第79卷，第119页，2004年。[26] Y. E. Appenzeller、P. S. Appelbaum 和 M. Trachsel，“精神科病区视频监控中的伦理与实践问题”，《精神科服务》，第71卷，第5期，第480–486页，2020年。[27] F. Tariq、N. Kanwal、M. S. Ansari、A. Afzaal、M. N. Asghar 和 M. J. Anjum，“面向智慧城市的隐私保护监控方法”，载于《第三届智慧城市研讨会》（SCS 2020），第2020卷，2020年，第450–455页。

[28] W. Hartzog，《隐私 ˜蓝图：争夺新技术设计控制权》。哈佛大学出版社，2018年。[29] J. Daubert、A. Wiesmaier 和 P. Kikiras，“关于物联网中的隐私与信任的观点”，载于《2015年IEEE国际通信会议研讨会》（ICCW）。IEEE，2015年，第2665–2670页。[30] T. Speicher、M. Ali、G. Venkatadri、F. N. Ribeiro、G. Arvanitakis、F. Benevenuto、K. P. Gummadi、P. Loiseau 和 A. Mislove，“在线定向广告中的潜在歧视”，载于《公平性、问责制与透明度会议》。PMLR，2018年，第5–19页。[31] I. D. Raji、T. Gebru、M. Mitchell、J. Buolamwini、J. Lee 和 E. Denton，“挽回颜面：探究人脸识别审计中的伦理问题”，载于《AAAI/ACM人工智能、伦理与社会会议论文集》，2020年，第145–151页。[32] N. Martinez-Martin，“在医疗保健中使用人脸识别技术有哪些重要的伦理影响？”《美国医学会伦理学杂志》，第21卷，第2期，E180页，2019年。[33] L. Introna 和 H. Nissenbaum，“人脸识别技术：政策与实施问题综述”，2010年。[34] E. Selinger 和 B. Leong，“人脸识别技术的伦理”，即将收录于《牛津数字伦理手册》，Carissa V´eliz 编，2021年。[35] F. Angelini、Z. Fu、Y. Long、L. Shao 和 S. M. Naqvi，“基于二维姿态的实时人体动作识别及遮挡处理”，《IEEE多媒体汇刊》，第22卷，第6期，第1433–1446页，2019年。[36] J.-M. Guo、C.-C. Lin、M.-F. Wu、C.-H. Chang 和 H. Lee，“利用基于概率的人脸掩膜预过滤和基于像素的层级特征AdaBoost降低人脸检测复杂度”，IEEE《信号处理快报》，第18卷，第8期，第447–450页，2011年。[37] T. ¨Ozyer、D. S. Ak 和 R. Alhajj，“基于视频数据集的人体动作识别方法综述”，《知识型系统》，第222卷，106995页，2021年。[38] R. Liao、S. Yu、W. An 和 Y. Huang，“结合身体姿态和人体先验知识的基于模型的步态识别方法”，《模式识别》，第98卷，107069页，2020年。[39] X. Jin、T. He、K. Zheng、Z. Yin、X. Shen、Z. Huang、R. Feng、J. Huang、Z. Chen 和 X.-S. Hua，“利用步态预测与正则化，从单幅图像进行换衣行人重识别”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第14 278–14 287页。[40] D. Leslie，“理解人脸识别技术中的偏见”，技术报告，2020 [ ] 4050457。在线。[41] 可访问：https://zenodo.org/record/ G. Jocher、A. Chaurasia、A. Stoken、J. Borovec、NanoCode012、Y. Kwon、K. Michael、TaoXie、J. Fang、imyhxy、Lorna、Z. Yifu、C. Wong、A. V、D. Montes、Z. Wang、C. Fati、J. Nadar、Laughing、UnglvKitDe、V. Sonck、tkia nai、yxNONG、P. Skalski、A. Hogan、D. Nair、M. Strobel 和 M. Jain，“ultralytics/yolov5：v7.0——YOLOv5最先进的实时实例分割”，2022年11月。 [在线]。可访问：https://doi.org/10 5281 7347926 . /zenodo。A. Warsi、M. [42] Abdullah、M. N. Husen、M. Yahya、S. Khan 和 N. Jawaid，“使用YOLOv3的枪支检测系统”，载于《2019年IEEE智能仪器、测量与应用国际会议》（ICSIMA）。IEEE，2019年，第1–4页。[43] S. Narejo、B. Pandey、D. Esenarro Vargas、C. Rodriguez 和 M. R. Anjum，“面向智能监控系统的YOLOv3武器检测”，《工程中的数学问题》，第2021卷，第1–9页，2021年。[44] R. Garg 和 S. Singh，“基于YOLO的智能视频监控：比较研究”，载于《2021年计算、通信与控制进展国际会议》（ICAC3）。IEEE，2021年，第1–6页。[45] Y. Zhang、P. Sun、Y. Jiang、D. Yu、F. Weng、Z. Yuan、P. Luo、W. Liu 和 X. Wang，“ByteTrack：通过关联每个检测框进行多目标跟踪”，2022年。[46] K. Sun、B. Xiao、D. Liu 和 J. Wang，“用于人体姿态估计的深度高分辨率表征学习”，载于CVPR，2019年。[47] W. An、R. Liao、S. Yu、Y. Huang 和 P. C. Yuen，“利用三维姿态估计改进步态识别”，载于《生物特征识别：第十三届中国生物特征识别会议（CCBR 2018），中国乌鲁木齐，2018年8月11–12日，论文集13》。施普林格，2018年，第137–147页。[48] M. M. Hasan 和 H. A. Mustafa，“利用循环神经网络进行多层级特征融合，以实现鲁棒的基于姿态的步态识别”，《国际计算机科学与信息安全期刊》（IJCSIS），第18卷，第1期，2020年。[49] K. Zhou、Y. Yang、A. Cavallaro 和 T. Xiang，“行人重识别的全尺度特征学习”，载于ICCV，2019年。[50] H. Duan、Y. Zhao、K. Chen、D. Lin 和 B. Dai，“重新审视基于骨架的动作识别”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022，第 2969–2978 页。

[51] Y. Chen、Z. Zhang、C. Yuan、B. Li、Y. Deng 和 W. Hu，“面向基于骨架的动作识别的通道级拓扑细化图卷积”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第13 359–13 368页。[52] A. Markovitz、G. Sharir、I. Friedman、L. Zelnik-Manor 和 S. Avidan，“用于异常检测的图嵌入姿态聚类”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第10 539–10 547页。[53] R. Morais、V. Le、T. Tran、B. Saha、M. Mansour 和 S. Venkatesh，“学习视频中骨架轨迹的规律性以进行异常检测”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2019年，第11 996–12 004页。[54] J. Carreira 和 A. Zisserman，“动作识别，何去何从？一种新模型与 Kinetics数据集”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第6299–6308页。[55] H. Fan、B. Xiong、K. Mangalam、Y. Li、Z. Yan、J. Malik 和 C. Feichten-hofer，“多尺度视觉Transformer”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第6824–6835页。[56] J. Liang、H. Zhu、E. Zhang 和 J. Zhang，“Stargazer：一种面向智能交通的基于Transformer的驾驶员动作检测系统”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第3160–3167页。[57] W. Liu、G. Kang、P.-Y. Huang、X. Chang、Y. Qian、J. Liang、L. Gui、J. Wen 和 P. Chen，“Argus：用于扩展视频分析的高效活动检测系统”，载于《IEEE/CVF冬季计算机视觉应用会议研讨会论文集》，2020年，第126–133页。[58] K. Corona、K. Osterdahl、R. Collins 和 A. Hoogs，“Meva：用于活动检测的大规模多视角、多模态视频数据集”，载于《IEEE/CVF冬季计算机视觉应用会议论文集》，2021年，第1060–1068页。[59] T.-Y. Lin、M. Maire、S. Belongie、L. Bourdev、R. Girshick、J. Hays、P. Perona、D. Ramanan、C. L. Zitnick 和 P. Doll´ar，“Microsoft COCO：上下文中的常见物体”，2014年。[60] P. Dendorfer、H. Rezatofighi、A. Milan、J. Shi、D. Cremers、I. Reid、S. Roth、K. Schindler 和 L. Leal-Taix´e，“MOT20：拥挤场景中的多目标跟踪基准”，2020年。 [在线]。可访问：https://arxiv.org/abs/2003 09003 。E. Ristani、F. Solera、R. Zou、R. [61] Cucchiara 和 C. Tomasi，“多目标、多摄像头跟踪的性能度量与数据集”，载于《欧洲计算机视觉会议》多目标跟踪基准测试研讨会，2016年。[62] W. Wang、J. Dai、Z. Chen、Z. Huang、Z. Li、X. Zhu、X. Hu、T. Lu、L. Lu、H. Li 等，“InternImage：利用可变形卷积探索大规模视觉基础模型”，arXiv预印本 arXiv:2211 05778，2022年。。[63] L. Zheng、M. Tang、Y. Chen、G. Zhu、J. Wang 和 H. Lu，“利用单目标跟踪改进多目标跟踪”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2021年6月，第2453–2462页。[64] Y. Xu、J. Zhang、Q. Zhang 和 D. Tao，“ViTPose：用于人体姿态估计的简单视觉Transformer基线”，载于《神经信息处理系统进展》，A. H. Oh、A. Agarwal、D. Belgrave 和 K. Cho 编，2022年。 [在线]。可访问：https: //openreview.net/forum?id=6 H2pBoPtm0s [65] M. Wieczorek、B. Rychalska 和 J. Dabrowski，“图像检索中质心非凡有效性的探讨”，载于《神经信息处理：第28届国际会议》，ICONIP 2021，印度尼西亚巴厘岛沙努尔，2021年12月8日至12日，《论文集》第IV部分。柏林、海德堡：施普林格出版社，2021年，第212–223页。 [在线]。可访问：https://doi.org/10 1007 978 3 030 92273 3 18 。/ - - - - J. Liu、A. [66] Shahroudy、M. Perez、G. Wang、L.-Y. Duan 和 A. C. Kot，“NTU RGB+D 120：用于三维人体活动理解的大规模基准”，–《IEEE模式分析与机器智能汇刊》，第42卷，第10期，第2684–2701页，2019年。[67] H. Duan、J. Wang、K. Chen 和 D. Lin，“PySkl：迈向良好的骨架动作识别实践”，2022年。 [在线]。可访问：https://arxiv.org/abs/2205 09443 。W. Liu、D. L. W. Luo 和 S. Gao，“未 [68] 来帧预测用于异常检测——一种新的基线”，载于《2018 IEEE计算机视觉与模式识别会议》（CVPR），2018年。[69] L. Wang、D. Q. Huynh 和 P. Koniusz，“近期基于Kinect的动作识别算法比较综述”，《IEEE图像处理汇刊》，第29卷，第15–28页，2020年。

个人简介



Armin Danesh Pazho （S’22）目前是美国北卡罗来纳州夏洛特市北卡罗来纳大学的博士生。他专注于人工智能、计算机视觉和深度学习，研究致力于开发面向实际真实世界应用的人工智能，并应对这些领域固有的挑战与需求。具体而言，他的研究涵盖动作识别、异常检测、行人重识别、人体姿态估计和路径预测。



Christopher Neff （S’18）是美国国家科学基金会研究生研究员，也是北卡罗来纳大学夏洛特分校的博士候选人。他的博士论文着眼于如何应对将以人为中心的计算机视觉应用于真实世界场景所面临的挑战。他此前的研究主要涉及行人重识别、人体姿态估计、动作识别、实时系统开发、轻量级算法、噪声数据、域偏移及真实世界应用。



Ghazal Alinezhad Noghre （S’22）目前正在美国北卡罗来纳州夏洛特的北卡罗来纳大学夏洛特分校攻读电气与计算机工程博士学位。她的研究重点是人工智能、机器学习和计算机视觉。她尤其关注异常检测、动作识别和路径预测在真实世界环境中的应用，以及这些领域所面临的挑战。



Babak Rahimi Ardabili 是美国北卡罗来纳大学夏洛特分校公共政策分析项目的博士生。他的主要研究领域是新兴技术政策制定，重点从隐私角度研究人工智能与政策的交叉领域，以及如何将这项技术引入社区所面临的挑战。



Shanle Yao 是北卡罗来纳大学夏洛特分校电气工程专业的研究生。他的论文主要研究计算机视觉流水线性能和吞吐量的优化与应用。他的研究兴趣包括目标检测、多目标跟踪、人体姿态估计、语义分割以及现实世界中的应用。



Mohammedreza Baharani 是 ForesightCares 的机器学习研究员和边缘系统部署工程师。他于 2021 年在美国北卡罗来纳大学夏洛特分校获得计算机工程博士学位，并曾在 TeCSAR实验室担任博士后研究员。他的研究聚焦于计算机体系结构工程与机器学习的交叉领域，旨在让边缘设备上的人工智能算法能够在医疗保健等领域产生积极影响。



Hamed Tabkhi （S’07–M’14）是美国北卡罗来纳大学夏洛特分校电气与计算机工程系副教授。他曾在东北大学担任博士后研究助理。Hamed Tabkhi 于 2014 年在 Gunar Schirner 教授的指导下获得东北大学博士学位。总体而言，他的研究聚焦于面向信息物理系统、实时流处理和新兴机器学习应用的变革性计算系统与架构。