视联网大模型在城市治理场景中的应用

章伟，张驰，沈阳

（中国电信股份有限公司上海分公司，上海 201203）

摘　要：随着视频监控技术和人工智能技术发展，依托视频流和人工智能算法的城市治理应用已成为了智慧城市建设当中的重要一环。然而城市治理里面长尾场景众多，传统人工智能算法效果不佳，研发成本也非常高，制约了城市治理数字化、智慧化的发展。多模态大模型作为一种整合多种模态信息的大模型技术，和传统小模型相比有着信息更全面、泛化性能更强的优点，在城市治理长尾场景中相较于传统人工智能技术有更好的表现。因此，提出了一种在城市治理长尾场景中基于多模态大模型实现算法研发和落地的方法，在实验中相较于传统视觉人工智能技术在功能和性能指标上有显著的提升。

关键词：城市治理；长尾场景；多模态大模型

中图分类号：TP399

文献标志码：A

doi: 10.11959/j.issn.1000−0801.2025095

0　引言

当前，社会治安形势日趋复杂，传统的治安防控措施已经难以满足现实要求。党的十八大报告中强调“要深化平安建设、完善立体化社会治安防控体系”，旨在构建“视频+云网”的视联网能力，充分发挥视频监控系统作用，推进社会治安防控体系建设。近年来，政府大力推进天网工程、视频全覆盖工程，进一步完善社会治安防控体系。除了平安城市，在工业、教育、零售业等不同行业当中，视频监控也越发普及，发挥了其在安防上的巨大作用，为事前发现、事中监控、事后追溯提供了技术支持。基于视联网能力的视频上云服务极大地降低了视频监控普及的门槛，也为视频监控场景与人工智能结合形成了土壤。

近年来，随着人工智能技术逐步发展，视频监控场景因其场景重要、数据充足、任务明确、门槛较低等特点，成为了人工智能技术大规模落地的试验田，逐步形成了一条“视频监控+人工智能分析”的智能视频应用落地模式。通过人工智能，对视频监控内容进行提炼，自动甄别出大家关心的内容，缩短问题发现周期，解决人工查看投入的大量人工成本。但由于城市治理场景中多为长尾场景，长尾场景具有特异性的特点，需要识别的目标类型种类繁多，业务逻辑规则也不尽相同，在传统以深度学习目标检测为主体的技术框架下，会出现如下几个问题：首先，算法训练所需要的数据通常难以采集，通用数据集或者通过爬虫等手段也难以采集到所需类型的数据，只能从目标场景中采集数据，这会有非常大的工作量；其次，算法模型需要重新训练，难以复用其他现有的模型，需要投入大量研发资源；此外，算法只能在当前场景中应用，复制推广的可能性较低，均摊下来研发成本非常高，性价比很

收稿日期：2024−11−29；**修回日期：**2025−02−26

低，而用户预算普遍有限，难以支付高额的研发费用。这就形成了一个用户有需求但预算有限、能力供应商有心无力的困境，城市治理数字化程度和智能化程度难以进一步深化。因此，如何在数据匮乏的场景下提升模型通用性、减少研发过程中的研发成本投入就成为了亟待解决的难题。

2022 年年底，ChatGPT 横空出世，引领了一波通用人工智能的热潮。以 Transformer[1]和 Mixture of Expert[2]为核心的大语言模型[3]快速风靡全球，涌现出了面向各行各业的海量智慧应用。这给予了视觉领域很多可借鉴的经验，整合文本和图片能力的多模态大模型成为了研究的热点。多模态大模型具备更加完整和丰富的信息表示，能够提升准确性和扩展通用性，而且能够通过文本这一更加自然化的方式作为信息输入。因此，本文提出了一种基于视联网大模型来解决城市治理场景中算法通用性差、数据匮乏、研发投入高问题的方法，并进行了实验论证。

1　多模态大模型

多模态大模型[4]是一种基于深度学习的先进机器学习技术，它能够处理并融合多种媒体数据类型，如文本、图像、音频和视频等。多模态大模型的核心思想在于通过学习和理解不同模态之间的关联，实现更加智能化和全面的信息处理。多模态大模型远比单模态模型要复杂，其复杂性主要体现在数据对齐、数据融合、统一标识等方面。数据对齐能保证不同模态的数据在时间和空间上的一致性；数据融合将多模态数据整合在一起，充分利用各模态的信息；统一标识构建一个统一的表示空间，使得不同模态的数据能够互相理解和结合。

多模态大模型首先将不同模态的数据进行预处理，然后将预处理过的数据输入一个深度神经网络。在这个网络中，模型会进行多层的特征提取和融合，从而学习各种模态之间的共同语义和潜在关系。这种跨模态的学习能力使模型能够在多种任务上表现出色，如跨媒体检索、语义对齐、跨模态生成等。多模态大模型结构示意图如图 1 所示。

传统单一模态模型存在大量的局限性，其具体表现主要为信息不全面和上下文缺失。单一模

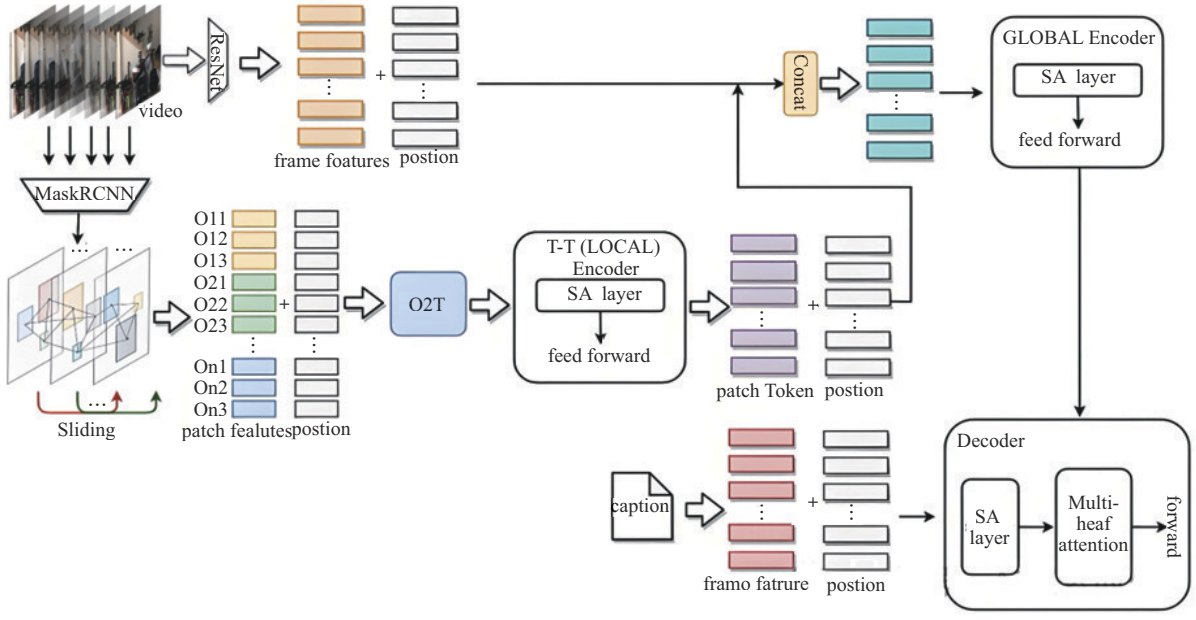

图 1　多模态大模型结构示意图

态的信息往往不够全面，例如，仅依赖文本描述可能无法准确理解一个场景，而只依赖图像可能无法准确获取文字内容和背后的含义。多模态大模型的优点在于其能够充分利用不同模态数据的信息，提取出更加丰富和全面的特征，从而提高模型的性能和泛化能力。通过学习和理解不同模态之间的关联，模型还能够进一步增强其语义理解和表达能力，使自身在处理复杂任务时更加准确和高效。此外，人与人之间的交流就是通过多种形式来表现，包括视觉、听觉、嗅觉、触觉等人类五感，而目前人机交互以指令或者简单的视觉交互如人脸识别等为主，这些方式使得人机交互存在很大的局限性，而多模态大模型有推进人机交互更加自然化和智能化的潜力。

在实际应用中，多模态大模型在多个场景中展现出了巨大的潜力。在对于个人的应用中，文生图或图生文都是多模态大模型非常典型的应用，广泛应用在短视频、计算机辅助设计等多种软件中。在面向行业场景中，多模态大模型也有广泛的应用场景。例如，在医疗影像诊断领域，多模态大模型可以处理包括医学图像在内的多种数据，提高诊断的准确性和效率。在城市治理中，通过整合和分析来自不同模态的数据，提高城市治理的效率和准确性，比如利用多模态大模型进行城市事件“一屏统览”，通过机器视觉分类任务对视频进行分析和研判，实现一次预训练即可覆盖大部分城市综合治理监管要求的场景识别，通过自主学习和泛化能力，提高城市治理的智能化水平。随着技术的不断发展和应用的深入拓展，多模态大模型将在未来发挥更加重要的作用。

2 应用方法

在城市治理场景中，本文聚焦于利用多模态大模型减少长尾算法研发的时间和人工投入，结合城市治理场景业务特点，形成了“两步走”的方法论。第一步，对于用户提出的长尾算法需求，通过多模态大模型结合提示词快速形成可用的目标检测模型；第二步，通过该模型进行数据积累和持续优化，再通过模型蒸馏或者基于数据重新训练小模型等方式，形成最终为用户提供的高性价比的技术服务。

2.1 基于多模态大模型的长尾算法快速研发

在长尾场景中用户提出需求阶段，用户想要验证的是需求是否能被现有AI能力满足，取得的效果是什么样子，如果效果好，用户才有继续推进直到付费的意愿。传统的深度学习方法在这种情况下只能从零或者有限的基础开始训练一版模型，但训练过程中又会遇到数据匮乏的问题，在数据采集和增强上要付出大量的精力。传统视觉小模型[5]只能以图片数据作为信息的输入，需要从海量图片数据和标注信息中学习到所要检测目标特点，数据不充分和不全面必然会导致训练出来的模型效果不佳，呈现于用户的效果难以保证，能力提供方在很多场景也会因为工作量和不确定性望而却步。而多模态大模型具备非常强的通用识别能力，结合提示词这种通过文本形式进行精确信息输入的优势，便可以在无须开发的情况下快速形成能力。

多模态模型带来的另一个优势是可以完成许多复杂任务，比如对场景内容的理解，包括判断一个场景是不是街道、是不是厨房等。传统做法基于检测或者分割技术，通常只能根据特定标志物来判断，比如识别道路，或者灶台等，这需要大量数据训练，识别效果也难以保证。而通过多模态大模型，结合少量数据微调以及提示词，即可实现相对比较准确的场景理解，工作量相比传统方法大幅度降低。

2.2 利用多模态大模型进一步孵化高性价比小模型

多模态大模型的一个主要劣势在于模型参数量大，推理所需算力成本高，性价比不如传统小

模型。从大模型转换到小模型有两种方式，一种是通过模型蒸馏[6]，另一种是基于大模型产生的数据重新训练小模型。模型蒸馏是一种模型压缩技术，其核心思想是将一个复杂的大型模型（通常被称为教师模型）的知识转移到另一个更小、更简单的模型（通常被称为学生模型）中。模型蒸馏可以使模型减小并提升推理速度，但模型蒸馏也存在训练时间较长、数据需求大、模型性能可能下降、对特定问题适用性差等缺点。和模型蒸馏相比，通过大模型积累数据然后重新训练小模型多数情况下是更稳妥的选择，一般都能在长尾场景中取得比较高的性价比。

多模态大模型积累数据主要有两种来源，一种是通过多模态大模型配合提示词先行试用以后在真实场景中积累数据，另一种是从公开数据集或者互联网数据中提取相关的数据，同时也可以获取到数据的标注信息。当然这个标注信息可能不够精确，还需要结合人工进行调整，但相对于完全人工进行数据收集和标注来说，可以极大地降低人工工作量。当积累到足够的数据以后，再训练小模型，小模型便是一个针对于长尾场景的专用模型，效果和性价比相较于大模型便会都有明显的提升。

3　实验

本文对多模态大模型在城市治理长尾场景的效果进行了实验，并与传统目标检测小模型进行了对比。实验中使用多模态大模型基于 DINOv2[7] 实现，使用传统目标检测模型为 Yolov5s[8]。测试中多模态大模型在预训练基础上使用 50 张场景图片数据进行了微调，而 Yolov5s 模型基于 1 000 张图片进行训练。数据比较少是模拟数据匮乏的长尾场景。

实验场景分为两个，一个是明火场景，明火具有形态不定的特点，而且容易和一些地面反光或者车灯等物体混淆。第二个是后厨场景分类，识别哪些场景是厨房后厨场景，需求普遍存在于明厨亮灶等业务场景中，可以作为判断开通应该配置什么算法、摄像机角度是否正确等问题的重要参考依据。测试数据集每个场景采用 100 张图片进行测试，其中明火数据集包含明火图片和非明火图片，后厨场景图片包含后厨场景以及走廊等非后厨场景和墙壁等模拟摄像机偏转等场景。功能评价指标为准确率（accuracy），性能评价指标为每秒推理帧数（frame per second，FPS）。实验结果见表 1。

表 1　实验结果数据对比

模型	明火场景准确率	后厨场景准确率	推理速度（FPS）
小模型	60%	28%	836
大模型	92%	93%	24

可以看到，多模态大模型在明火场景和后厨场景都获得了 90% 以上的准确率，而小模型在这些场景表现都不理想，主要原因是数据比较匮乏，小模型训练不充分导致模型泛化性能不足。明火场景中小模型存在明显对于红光如汽车尾灯、地面反光等亮斑的误识别，在后厨场景中小模型几乎不能对后厨区域进行检测，准确率低于随机猜测，没有足量的数据根本无法把足够的信息输入到模型中，让模型去理解后厨有什么样的特点。而多模态大模型因为有充足的先验知识，结合少量的数据微调，在以上场景中表现良好，提示词可以有效地描绘要检测的物体，给予模型清晰的数据输入。当然，在性能上小模型远胜于大模型，因为小模型参数更少，架构简单且量化到 INT8 等精度进行计算，而基于 Vision Transformer 的大模型参数多、架构复杂，计算速度远低于小模型。

4　结束语

多模态大模型在长尾场景中表现出了传统小

模型难以比拟的性能，也极大地简化了长尾算法的研发过程，随着时间发展，未来可能成为推进城市治理数字化、智能化的重要抓手，对产业发展有着深远的影响。然而，多模态大模型的性能提升带来了更大的算力开销，算力现在作为重要的基础设施也面临着价格昂贵等问题。如何将多模态大模型技术真正与产业结合，形成物美价廉、安全可靠的新一代城市治理智慧产品还有待业界进一步研究和探索。

参考文献：

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017(30): 5998-6008.

[2] CAI W, JIANG J, WANG F, et al. A survey on mixture of experts[J]. arXiv preprint, arXiv: 2407.06204, 2024.

[3] 郭全中, 杨元昭. 大模型发展回顾与展望[J]. 中国传媒科技, 2024(2): 159-160.

[4] LI C Y, GAN Z, YANG Z Y, et al. Multimodal foundation models: from specialists to general-purpose assistants[J]. Foundations and Trends® in Computer Graphics and Vision, 2024, 16 (1-2): 1-214.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint, arXiv: 1503.02531, 2015.

[7] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: learning robust visual features without supervision[J]. arXiv preprint, arXiv: 2304.07193, 2023.

[8] Glenn Jocher. Yolov5[EB]. 2025.

[作者简介]

章伟（1982–），男，中国电信上海公司工程师，云中台/数字集成部副总经理，主要从事大视频和人工智能等领域顶层设计工作。

张驰（1990–），男，中国电信上海公司工程师，云中台/数字集成部系统架构师，主要从事大视频和人工智能等相关架构设计和研发工作。

沈阳（1995–），男，中国电信上海公司云中台/数字集成部软件开发工程师，主要从事大视频和人工智能等相关技术研发工作。