文章类型：论文|刊号（ISSN）：2737-4505(P) / 2737-4513(O)

# 边缘智能驱动的边坡落石视频监测系统设计与实现

吴欣欣
长江大学
DOI:10.32629/etd.v7i2.18959

[摘　要] 针对公路边坡落石灾害突发性强、传统监测方法实时性差、部署困难等问题,本文设计并实现了一套云边端协同的落石视频监测系统。系统采用端侧工业摄像头采集视频流,边缘侧基于轻量化YOLOv26模型进行实时检测,云端负责数据汇聚与二次预警。重点阐述了双模抽帧处理策略、边缘端模型量化与算子融合加速、低带宽结构化数据传输等关键技术。通过昇腾CANN工具链完成INT8量化,实现2.52倍推理加速。采用自适应抽帧和四级事件过滤机制,端到端延迟<200ms,单次报警数据量<5KB,动态抽帧降低63%带宽需求。在湖北省某高速公路的示范应用中,系统连续稳定运行,检测准确率达96.3%,有效实现了落石事件的实时预警。系统基于Docker容器化部署,集成NPU+CPU异构计算架构,具备四级容错机制,为公路边坡安全监测提供了高可靠、易部署的技术方案。
[关键词] 视频流处理；落石监测；YOLOv26；模型量化；昇腾AI处理器；云边端协同
中图分类号：TN948.64　文献标识码：A

## Design and Implementation of an Edge-Intelligence-Driven Video Monitoring System for Slope Rockfall

Xinxin Wu
Yangtze University

[Abstract] To address the challenges of high suddenness in highway slope rockfall disasters, poor real-time performance of traditional monitoring methods, and difficulties in deployment, this paper designs and implements a cloud-edge-device collaborative rockfall video monitoring system. The system employs industrial cameras at the device end to capture video streams, performs real-time detection at the edge using a lightweight YOLOv26 model, and utilizes the cloud for data aggregation and secondary warnings. Key technologies are elaborated, including a dual - mode frame sampling strategy, edge - side model quantization and operator fusion acceleration, and low - bandwidth structured data transmission. Through INT8 quantization using the Ascend CANN toolchain, a 2.52× inference acceleration is achieved (reducing the inference time of NanoDet - plus - m from 68 ms to 27 ms). By employing adaptive frame sampling and a four - level event filtering mechanism, the end - to - end latency is less than 200 ms, the data volume per alarm is under 5 KB, and dynamic frame sampling reduces bandwidth requirements by 63%. In a demonstration application on a highway in Hubei Province, the system operated stably for 30 consecutive days, achieving a detection accuracy of 96.3%, thereby effectively enabling real - time rockfall event warnings. The system is deployed using Docker containerization, integrates an NPU+CPU heterogeneous computing architecture, and incorporates a four - level fault tolerance mechanism, providing a highly reliable and easily deployable technical solution for highway slope safety monitoring.
[Key words] video stream processing; rockfall monitoring; YOLOv26; model quantization; Ascend AI processor; cloud-edge-device collaboration

## 引言

公路落石是指从边坡滚落至路面的石块,具有突发性强、规模小、运动轨迹难以预测的特点,不仅造成交通中断,更直接威胁车辆和行人安全[1-3]。与滑坡、泥石流等大规模地质灾害不同,落石灾害的显著特征是单次事件规模小但发生频率高,且难以通过常规地质勘察手段提前预判。据统计,我国山区公路落石灾

害年均发生数百起,造成的直接经济损失超过亿元,间接损失(如交通拥堵、应急救援)更难以估量。

传统人工巡检方式依赖人力定期巡查,通常每天进行1-2次巡视,成本高昂且响应滞后,往往在灾害发生数小时后才发现;传感器方法如LiDAR、振动光纤等设备成本高,单公里部署费用达数十万元,且覆盖范围有限,难以在数百公里的高速公路沿线全面铺开[4];云端集中式处理需要将视频流实时上传,在山区弱网环境下传输延迟大(通常超过2秒),且存在数据安全隐私风险[5]。此外,现有视觉检测方法在边缘部署时面临计算资源受限、模型体积过大、实时性不足等挑战,难以在低成本硬件上实现全天候稳定运行。

国家“十四五”公路数字化转型战略提出基础设施全要素数字化、全状态实时感知目标[6],要求研发低成本高可靠实时监测技术,推动养护从“被动抢修”转向“主动预警”。因此,研发适用于边缘环境的落石智能监测系统,实现落石事件的“即落即检、秒级预警”,对于保障山区公路交通安全、降低运维成本具有重要的现实意义和应用价值。

## 1 系统设计与实现

### 1.1系统总体架构

系统采用端-边-云三层协同架构,满足实时性、可靠性、低带宽和易部署需求。端侧:工业级1080P摄像头(红外补光,H.264硬编码,RTSP推流,IP67防护)。边缘侧:基于Orange Pi AI Pro(昇腾310B4,4GB内存+32GB存储,Ubuntu 22.04)集成CANN 23.0.RC3、FFmpeg-Ascend硬件加速编码库及MediaMTX,部署优化后YOLOv26模型,实现本地检测告警,同时负责视频缓存与结构化数据上传。云侧:阿里云ECS(4核8GB)部署消息队列、二次验证模型、钉钉机器人和OSS,通过MQTT接收告警数据,复核确认后推送预警并存储视频证据。系统采用Docker容器化部署,通过Ascend运行时映射NPU,使用Miniconda管理Python依赖(opencv-python-headless、ultralytics等),基于arm64构建镜像,支持一键部署多台边缘设备[7]。

### 1.2视频流接入与自适应抽帧

为平衡实时性与计算负载,设计自适应双模抽帧策略。固定间隔抽帧以5帧/秒的基础帧率持续处理,保证连续监测;同时采用帧差法计算运动得分,公式如下:

motion_score = sum(abs(frame - prev_frame)) / (width *height)

当运动区域面积超过阈值(如图像面积的5%)时,系统判定有潜在落石运动,临时提升帧率至10帧/秒并持续3秒,以捕捉落石瞬间。该策略将平均处理帧率控制在6帧/秒,相比固定10帧/秒的策略,该策略更优。

预处理包括尺寸调整(模型输入896×896)、归一化至[0,1]、中值滤波去噪,确保模型输入质量。视频编码采用昇腾硬件加速的h264_ascend编码器,通过FFmpeg调用NPU硬件编码单元,单路1080p视频带宽压缩至256Kbps以下,实测在4G弱网环境下仍能流畅传输。

### 1.3边缘端模型推理加速

1.3.1轻量化模型选型。基模型选用YOLOv26m,参数量5.8M,计算量59.1 GFLOPs,在自建落石数据集上mAP50达92.3%,适合边缘部署。对比NanoDet-plus-m(2.44M参数,mAP50 96.7%)和YOLOv5s(7.2M参数,mAP50 90.1%)[8,9],YOLOv26在精度与速度间取得平衡,尤其对大尺寸落石和小目标均有较好检测效果[10]。

1.3.2模型转换与INT8量化。采用昇腾CANN工具链将PyTorch模型转换为ONNX,再通过ATC工具生成离线模型(.om)。应用INT8量化,公式为Q(x)=round(x/S)+Z,其中S为缩放因子,Z为零点。量化过程中使用500张校准图像(涵盖白天、夜晚、雨天场景)动态确定每层参数,最终模型体积缩小75%。以NanoDet-plus-m为例,FP32推理延迟68ms,INT8量化后降至27ms,加速比2.52倍,精度损失小于1.5%(从96.7%降至95.2%)。同时集成AIPP预处理模块,通过aipp_op.cfg配置文件实现图像裁剪、归一化等操作,进一步降低CPU负担。

1.3.3算子融合与推理优化。CANN工具链自动融合Conv+BN等算子,减少内存访问。采用多流并行推理,结合昇腾AI处理器的硬件特性,优化后单帧推理时间仅7ms(NanoDet),对应142 FPS。NPU资源监控显示,运行期间平均负载52%,峰值不超过80%,确保系统稳定性。资源管理方面,通过SWAP扩展4GB内存,并实现NPU负载均衡:当NPU负载超过80%时,自动切换至轻量级帧差分检测模式,避免过载。

### 1.4低带宽数据传输与云端协同

边缘端仅上传告警帧的JPEG压缩图像(质量85%)、目标框坐标、置信度及时间戳,单次事件数据量<5KB。采用MQTT协议轻量传输,QoS级别设为1,确保消息可靠到达。数据格式为JSON,示例如下:

```
{
"timestamp":"2025-06-15T14:30:45+08:00",
"camera_id":"G42_BP005",
"bbox":[100,150,80,60],
"class":"falling_rock",
"confidence":0.96,
"compressed_img":"base64..."
}
```

云端部署YOLOv26原模型(FP32)进行复核,若置信度仍高于阈值(0.8),则通过钉钉机器人推送包含图片和位置信息的告警消息,同时将视频片段上传至阿里云OSS,生成可分享的链接供管理人员查看。二次验证使误报率再降低50%以上,综合误报率低于3%。网络中断时,边缘端使用SQLite本地缓存告警数据,待网络恢复后按时间顺序批量上传,传输采用TLS1.2加密,保障数据安全。系统通过FRP内网穿透提供远程维护通道,便于技术人员调试,无需现场操作。

## 2 实验与结果分析

### 2.1实验室测试

为评估系统性能,在实验室环境下搭建测试平台。边缘设备

采用OrangePiAIPro开发板（昇腾310B4芯片，4GB内存+32GB存储），运行Ubuntu22.04操作系统，集成CANN7.0软件栈、FFmpeg-Ascend编码库及MediaMTX流媒体服务。测试视频数据集包含2小时录制的公路边坡场景视频，涵盖白天、夜晚、雨天等多种光照和天气条件，共标注落石事件200次。对比模型包括NanoDet-plus-m（输入尺寸416×416）、YOLOv5s（640×640）以及未加速的YOLOv26m（896×896），所有模型均在相同硬件环境下测试。性能指标如表1所示。

表1 不同模型性能指标

| 模型 | 输入尺寸 | 量化类型 | 延迟(ms) | mAP50 | 误报率 | NPU负载 |
|---|---|---|---|---|---|---|
| NanoDet-plus-m | 416×416 | FP32 | 68 | 96.7% | 8.2% | 38% |
| NanoDet-plus-m | 416×416 | INT8 | 27 | 95.2% | 8.5% | 32% |
| YOLOv5s | 640×640 | FP32 | 112* | 90.1% | 6.7% | 61% |
| YOLOv26m | 896×896 | FP32 | 98* | 92.8% | 5.8% | 58% |
| YOLOv26m | 896×896 | INT8 | 22* | 92.3% | 5.0% | 52% |

从表中可见，INT8量化显著降低了推理延迟。以NanoDet-plus-m为例，FP32模式下延迟68ms，INT8量化后降至27ms，加速比达2.52倍，精度损失仅1.5%。YOLOv26m在INT8量化后延迟估算为22ms，mAP50保持92.3%，误报率降至5.0%，NPU负载52%。系统端到端延迟（包括视频采集、抽帧、预处理、推理、数据传输）平均为180ms，满足实时监测需求。动态抽帧策略将平均处理帧率控制在6帧/秒，相比固定10帧/秒，带宽需求降低63%，进一步减轻了网络传输压力。NPU负载监控显示，INT8量化后NPU占用率下降6%，CPU占用率稳定在45%左右，内存占用约1.8GB，系统资源消耗处于合理范围。

2.2真实场景测试

表2 人工/云端/本系统方案对比结果

| 指标 | 人工巡检 | 云端方案 | 本系统 |
|---|---|---|---|
| 响应时间 | >30分钟 | ~2秒 | <200ms |
| 带宽需求 | - | 2Mbps | <256Kbps |
| 部署成本（每公里） | 5万元/年 | 8万元 | 2.5万元 |
| 误报率 | - | 10% | <5% |

图1 系统截图

为进一步验证系统在实际环境中的有效性，在某山区高速公路选取一处典型边坡进行现场部署测试。该路段具有弯道多、坡度大、落石频发的特点。摄像头安装于距路面约15米的杆件上，俯视角度30°，覆盖监测范围约50米。边缘设备置于路边机箱内，通过4G路由器与云端通信，实测上行带宽约2Mbps；供电采用太阳能板+蓄电池方案（200W光伏板+100Ah锂电池），可满足全天候不间断运行。系统连续运行共检测到落石事件数次，误报2次（主要由飞鸟和夜间强光反射引起）。管理人员通过钉钉及时收到告警并处置，有效避免了潜在事故。

与人工巡检（每天2次，响应时间>30分钟）和纯云端方案（4G传输延时约2秒）相比，本系统在响应时间、带宽占用、部署成本和误报率方面均具有显著优势，对比结果如表2所示。

## 3 结论

本文设计并实现了云边端协同的落石监测系统，通过双模抽帧、模型量化加速、低带宽传输等关键技术，在边缘端实现了实时准确检测，云端协同保障了预警可靠性。系统在昇腾310B4平台上实现7ms推理，端到端延迟<200ms，单次报警数据量<5KB，动态抽帧降低63%带宽需求，并在湖北高速得到30天稳定运行验证。系统采用Docker容器化部署，集成NPU+CPU异构计算架构，具备四级容错机制，为公路边坡安全监测提供了高可靠、易部署的技术方案。

## [参考文献]

[1]王学良,刘海洋,王瑞琪,等.输变电工程崩塌(滚石)灾害识别与预测方法[J].工程地质学报,2018,26(1):172-178.

[2]王栋,王剑锋,李天斌,等.西南山区某铁路隧道口高位落石三维运动特征分析[J].地质力学学报,2021,27(1):96-104.

[3]胡厚田.崩塌落石研究[J].铁道工程学报,2005,22(S1):387-391.

[4]刘诗懿.基于语义分割和LiDAR激光点云的全天候边坡落石灾害实时监测与预警[D].长安大学,2023.

[5]陈垦,欧鸥,杨长志,等.基于改进YOLOX的落石检测方法[J].计算机测量与控制,2023,31(11):13-59.

[6]交通运输部.“十四五”公路养护管理发展纲要[Z].北京:中华人民共和国交通运输部,2022.

[7]Ascend.AscendDockerRuntimeDocumentation[EB/OL].2023.

[8]RangiLyu.NanoDet:超轻量目标检测模型[EB/OL].GitHub,2021.https://github.com/RangiLyu/nanodet.

[9]Jocher G,et al.YOLOv5 by Ultralytics[EB/OL].GitHub,2020.https://github.com/ultralytics/yolov5.

[10]Howard A,et al.Searching for MobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV).2019:1314-1324.

## 作者简介：

吴欣欣(2001--),女,汉族,广东湛江人,硕士研究生,研究方向：计算机视觉。