无人机巡检中的边缘AI识别应用研究

王燚，杨博文，唐双江
四川通信科研规划设计有限责任公司，四川成都 610041

摘　要：随着无人机技术的快速发展，无人机在巡检领域的应用日益广泛，特别是在电力、建筑等行业。无人机巡检可以大幅提高巡检效率，减少人力成本，并能在复杂环境中进行高效作业。然而，传统的无人机巡检方法依赖于人工识别，存在效率低、易出错等问题。无人机技术结合深度学习目标识别的应用，可以显著提高巡检效率和准确性。本文旨在开发一种基于深度学习的无人机巡检边缘嵌入式目标识别模型，将其部署在无人机边缘设备上，实时地从航拍视频流中检测和识别目标，为巡检提供高效、精准的解决方案。

关键词：无人机巡检；边缘嵌入；目标检测
中图分类号：TN929.5　　**文献标志码：**A　　**文章编号：**1672-0164（2024）06-0113-05

1 引言

随着无人机技术的进步，结合图像处理技术，为应急通信[1]、野外救援[2]、巡检巡航[3]和遥感测绘[4]等领域的广泛应用带来新的机遇。然而，随着应用场景的扩展，无人机图像检测方法的重要性日益凸显。如何提升图像检测的精确度并降低对软硬件资源的需求，已成为当前无人机智能应用系统研究的关键方向。

2 相关工作

无人机智能巡检技术采用了多种方式提升检测效率，除视觉技术外，还包括雷达探测、声学探测和射频探测等，但由于技术成本及功耗体积等限制，对于中小型无人机，基于机器视觉的目标检测与跟踪技术仍是低空无人机智能巡检应用的主要发展方向[5-6]。MCUNet[7]和NanoDet[8]主要设计用于生产低功耗微控制器，提高边缘CPU推理速度。YOLO系列算法具有更好的精度和更快的推理速度，在国内广泛应用于工业领域。YOLOv1[9]是一种典型的单级目标探测器，在此基础上进行了一系列改进，得到了YOLOv2[10]和YOLOv3[11]，具有更快的检测速度和更高的检测精度。YOLOv4[12]重新设计主干、颈部和头部的三个独立结构，使它们在单一GPU上更好地训练。目前，YOLOv5[13]、YOLOX[14]、PPYOLOE[15]等在实时检测和部署方面具有极大的竞争力。

虽然这些流行的YOLO类检测模型运行效果较好，但复杂的网络结构增加了计算代价。因此，如何平衡兼顾视觉AI的功耗和性能，将模型计算载体前端化、轻量化，成为无人机巡检任务中一个亟待解决的问题。人们提出了不同的实时检测模型，适用于不同的边缘器件，主要集中在简单、高效的实时目标探测器结构设计上，以便在无人机巡检任务中可以实现实时检测效果[16-17]。Tijtgat等人[18]设计了一个基于运行YOLOv2的NVIDIA Jetson TX2边缘计算设备的系统，以实现无人机的实时目标检测。Abdulghafoor等人[19]提出了一种将边缘计算设备与深度流软件开发工具包（DS-SDK）4.0.2[5]相结合的方法，以实现一种可以处理高性能视频流的卷积网络模型。为了提高实时视频流检测系统的实用性。Haq等人[6]在NVIDIA Jetson单板计算机上部署了深度流框架来运行深度学习算法，特别是YOLO算法。该研究还验证了深度流框架可以在虚拟机中运行得很好，这可以进一步提高模型的性能和部署的可移植性。

针对以上不足之处开展了基于无人机载人工智能轻量化的研究、应用和验证。通过无人机机载轻量化AI技术将视频图像进行实时动态识别、目标检测，以供地面指挥所及时对突发情况做出应对。通过机载端运行的轻量化AI与后端神经网络处理器集中运行AI对航拍视频的处理结果对比，多方面体现了轻量化AI在部署灵活性和功耗比上的可行性和优势。

3 方法

3.1 轻量化NPU计算平台

NPU（Neural Processing Unit，神经处理器）与CPU（Central Processes Unit，中央处理器）和GPU（Graphics

收稿日期：2024年6月14日；**修回日期：**2024年7月29日

Processing Unit，图形处理器）等通用处理器设计思路不同，NPU是一种专门针对神经网络中的运算逻辑进行设计的SOC解决方案，无需考虑神经网络并不需要的一些计算单元，同时，通过突触权重实现存储和计算一体化，能够运行多个并行线程，其大部分时间集中在低精度的算法，因此，NPU的这种对AI计算的折中优化使得NPU相比CPU和GPU在算效上更加高效，在相比于CPU能获取更高处理帧率的同时，其功耗开销却不足GPU的1%。这凸显了嵌入式NPU的小型化、低功耗和低成本优势，能充分满足无人机AI轻量化应用的硬件需求。

考虑要实时进行目标识别的同时部署到无人机边缘设备，YOLOv8在网络模型、训练策略、损失函数、模型大小、推理速度等方面相比于其他YOLO系列的方法，非常适合轻量化AI的部署。

3.2 改进的轻量化目标检测算法

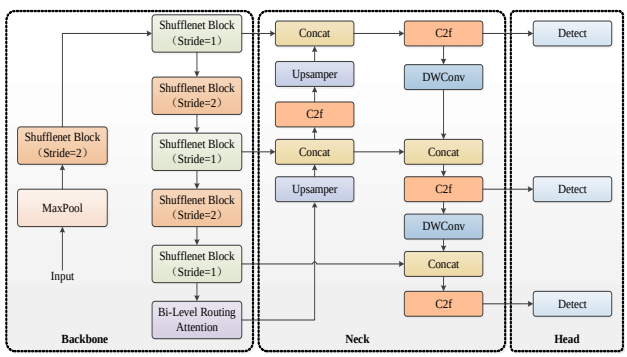

图1 检测模型结构图

准确、及时地跟踪和分析目标的位置、状态对于无人机巡检系统的决策和部署至关重要。以往的识别算法往往存在大量的计算量，需要较高的计算内存和设备的计算资源。针对资源有限的边缘设备，本项目提出了一种轻量的目标识别方法，如图1所示，可轻松实现机载边缘设备的部署。选取的基线模型YOLOv8是主流的实时目标探测器，主要由主干、颈部和头部组成。首先，使用轻量级网络Shufflenetv2[20]重建YOLOv8的主干。其次，利用深度可分卷积来代替颈部层的标准卷积。然后，利用BiFormer中的注意力（Bi-Level Routing Attention，BRA）[21]，在骨干网络输出特征之前，提高模型的定位能力和特征提取能力。

BRA是一种动态的、查询感知的稀疏注意机制，在粗粒度区域级别上过滤掉最不相关的键向量和值向量，这样就只保留一小部分路由生成细粒度区域。然后，在这些细粒度区域并集中的每个查询，通过其键值所指向的最相关的查询进行加权求合，完成注意力计算，该方法可以缓解多头自注意力的可伸缩性问题。具体来说，首先对Q和K应用每个区域平均推导出区域级查询向量和键向量，$Q^r,K^r \in R^{s^2 \times C}$。然后，通过$Q^r$和转置的$K^r$之间的矩阵乘法，推导出区域到区域亲和图的邻接矩阵，$A^r \in R^{s^2 \times s^2}$：

$$A^r = Q^r(K^r)^T \tag{1}$$

邻接矩阵A^r中的条目度量了两个区域在语义上的关联程度。然后通过为每个区域只保留top-k个连接来修剪亲和图。推导一个路由索引矩阵，$I_r \in N^{s^2 \times k}$，使用行向topk算子：

$$I^r = \text{topkIndex}(A^r) \tag{2}$$

利用区域到区域的路由索引矩阵I^r，可以应用细粒度的标记注意。对于区域i中的每个查询标记，它将关注所有位于k个路由区域的并集中的键值对，它们以$I^r_{(i,1)},I^r_{(i,2)},\cdots I^r_{(i,k)}$为索引。收集的键和值张量如下表示：

$$K^g = \text{gather}(K,A^r),V^g = \text{gather}(V,I^r), \tag{3}$$

然后将注意力应用到收集到的键值对上：

$$O = \text{Attention}(Q,K^g,V^g),V^g + \text{LCE}(V) \tag{4}$$

LCE是一个局部上下文增强模块[21]，函数LCE（·）通过深度卷积进行参数化，将核大小设置为5。

3.3 无人机载AI模型部署

检测算法在pytorch框架下生成的pt模型权重文件与NPU的适配度较低，直接运行检测效率较低，在Rockchip的NPU开发套件RKNN-Toolkit的基础上，将权重pt转换为适配NPURKNN模型，实现了检测模型在RK3399NPU平台上的加载和推理计算。YOLO模型转换流程如图2所示：

图2 YOLO模型转换流程

3.3.1 模型算子转换

Pytorch、ONNX和RKNN同为深度学习网络框架，其模型算子的分类与功能具有相似性。算子可以根据其功能和用途被分为不同的类别，在模型转换时，前向模型的算子函数会被解释转换为后向模型使用的对应算子结构，保证了两者网络框架在逻辑上的一致性。

3.3.2 PT模型至ONNX模型

在转换pt模型时，一方面，PyTorch会用跟踪法执行前向推理，把遇到的算子整合成计算图；另一方面，PyTorch还会把遇到的每个算子翻译成ONNX中定义的算子。在这个翻译过程中，可能会碰到以下两种情况：

①该算子可以一对一地翻译成一个ONNX算子。

②该算子在ONNX中没有直接对应的算子，会翻译成一至多个ONNX算子。

使用netron（深度学习网络可视化工具）导入pt模型网络结构和ONNX模型网络结构的局部直观展示图如图3与图4所示：

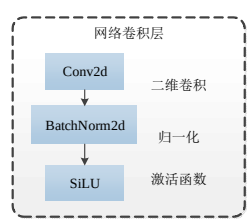

图3 pt模型局部网络结构

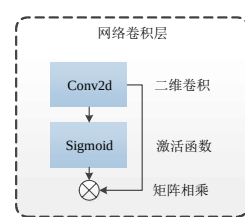

图4 ONNX模型局部网络结构

由于PyTorch算子是向ONNX对齐的，故主要算子转换关系如表1所示：

表1　PyTorch与ONNX算子对应关系

原pt算子结构	对应ONNX算子结构	主要功能
Conv2d	Conv、Mul	由多个输入平面组成的输入图像上应用2D卷积
BatchNorm 2d		对数据做归一化处理，使其分布一致
SiLU	Sigmoid	激活函数

整体来看，ONNX模型在保留pt模型网络基本逻辑结构的基础上，将诸如Focus等集成化模块拆分成了对边缘计算处理器更加易读的形式，输入输出参数与原模型完全对齐，保证了模型的可靠性，但代价是网络结构更加冗杂，同时量化过程中会不可避免地裁剪掉一些参数，导致分割细节丢失。

3.3.3　ONNX模型至RKNN模型

ONNX作为一种通用的开放格式，充当过渡模型来存储网络结构，最终目的是转换得到适合在边缘AI设备上部署的RKNN模型。RKNN是由Rockchip提出的一种针对其硬件平台的模型部署框架，其核心思想是将传统神经网络中的全连接层替换为局部连接层，从而实现对数据的高效计算。在RKNN中，每个神经元只与输入数据的一个局部区域相连，这大大减少了模型的参数数量，提高了模型的泛化能力。

使用netron（深度学习网络可视化工具）导入检测模型的RKNN网络结构的局部直观展示如图5所示：

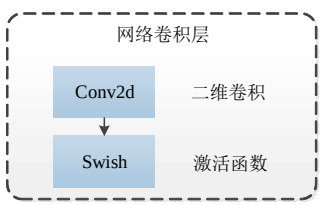

图5　RKNN模型局部网络结构

与上一节中的ONNX网络结构对比，主要算子转换如表2所示：

表2　ONNX算子与RKNN算子对应

原ONNX算子结构	对应rknn算子结构	主要功能
CONV	CONV2D	对输入图像进行2D卷积
Sigmoid	SWISH	激活函数

整体来看，RKNN与ONNX网络架构基本一致，主要是引入了SWISH非线性激活函数，其最大的优点是可以使用更少的参数让神经网络获得更低的偏差和更好的泛化能力，此外，其较低的计算开销也有助于边缘AI设备更好的性能表现。

3.3.4　算法函数结构轻量化

检测算法中的一些函数结构对于边缘计算设备，特别是需要量化操作的NPU芯片来说，在速度和精度上的影响显著，其中影响明显的是激活函数结构。卷积之后的激活操作采用了SiLU函数，这个结构本身在NPU上的效率不高，同时会在转换时被ONNX拆分，对检测结果的稳定性和可靠性有较大影响。因此，在转换为ONNX模型时，主要针对图5中的sigmod + mul组合，将hardSwish模块合并，提升在NPU上的运行效率。

4　AI智能感知模型检测结果与分析

为验证边缘部署的AI智能检测模型的有效性，首先通过VisDrone2019数据集[21]对YOLO模型进行训练。其中用于训练、验证以及测试的数据一共包含10209张静态图像(6471张用于训练，548张用于验证，3190张用于测试)。测试视频采用无人机传回的航拍影像，包含了行人、卡车、货车、小轿车等多类目标。为了验证分析模型轻量化对推理性能的影响，分别采用轻量化目标检测算法的PT模型和轻量化RKNN模型对上述测试视频进行预测。实验时分别将PT模型和RKNN模型加载到同一台PC机的虚拟机上，保证推理NPU型号、核数量以及其余配置完全相同。检测如图6和图7所示，检测结果如图8和图9所示：

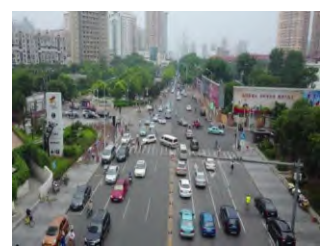

图6　VisDrone2019数据集图例1

图7　巡检现场检测图例

图8　PT模型检测结果图例

图9　RKNN模型检测结果图例

图10　不同检测方法的结果对比

表3 基线模型与改进模型性能指标对比

Method	P ↑	R ↑	mAP_{50} ↑	mAP_{50-95} ↑	F1 ↑	Latency (ms) ↓	FPS (bs =1) ↓
YOLOv8n	45.6	35.4	35.6	20.9	39.9	9.1	91.7
Improved-YOLO	47.3	35.2	36.0	20.8	40.4	5.5	125
Improved-YOLO+BRA	48.0	35.8	36.8	21.4	41.0	6.8	116.3

不同流行的检测模型检测结果如图10所示，可以看出，在VisDrone数据集上，文中的检测模型与基线模型YOLOv8和流行的YOLOv6相比，目标检测的数量和置信度方面均有显著提升，表3显示了不同方法定量的对比结果，可以看出文中方法不仅在精度上有提升，在推理速度上也有明显的改善，这将有助于部署到边缘设备，供无人机巡检过程中实时处理智能感知任务，对后续的决策和部署提供重要的技术支持。

实验统计得出模型混淆矩阵见表4。

表4 模型混淆矩阵等效一维表

PT模型样本分类值	00	01	10	11
PT模型样本数	728	356	1203	18807
RKNN模型样本分类值	00	01	10	11
RKNN模型样本数	693	391	1358	18652

其中，0代表反例，即不应该被分类器识别到的目标；1代表正例，即应该被分类器识别到的目标。因此，混淆矩阵中的参数含义如下：

11，即TP(True positives)：被正确地划分为正例的个数，即实际为正例且被分类器划分为正例的实例数（样本数）；

01，即FP(False positives)：被错误地划分为正例的个数，即实际为负例但被分类器划分为正例的实例数；

10，即FN(False negatives)：被错误地划分为负例的个数，即实际为正例但被分类器划分为负例的实例数；

00，即TN(True negatives)：被正确地划分为负例的个数，即实际为负例且被分类器划分为负例的实例数。

在此基础上，进一步将视频图像经过PC端的PT模型和NPU开发板的RKNN进行检测识别，由此计算得出模型评价指标整合相关性能参数对比如表5所示：

表5 模型综合性能参数对比

模型 \ 参数	准确度(Accuracy)	精确度(Precision)	召回率(Recall)	检测速率（speed）	功耗（consumption）
PT	92.61%	98.14%	93.99%	28fps	23W
RKNN	91.71%	97.95%	93.21%	25fps	1.2W

从算法评价指标上看，轻量化RKNN模型相比传统PT模型在准确度、精确度和召回率上都有一定程度的下降，但处于可接受的范围内。检测速率上得益于NPU对推理的加速，两者基本一致。但是轻量化的框架带来了显著的功耗降低，同一目标下RKNN模型推理功耗只有PT模型的5.2%。

总体来看，完成从PT模型到RKNN模型的轻量化改造以少量的检测精度为代价，显著降低了算法功耗，实现了对前端设备边缘AI的适配支持，降低了传输开销的约束，在无人机管道巡检行业中有广阔的前景。

5 结论

针对无人机巡检应用中的现状与问题进行分析，围绕边缘AI设备的特点，创新性地提出轻量化AI模型保证性能的同时降低功耗，以集成前端无人机和后端AI边缘平台，解决无人机在巡检过程中前后造成的检测结果实时性差的问题。最终通过实验结果对照验证了观点的可行性，并针对现有问题进一步提出了可优化的方向。

参 考 文 献

[1] 林尚静,田锦,马冀,等.应急通信中异构无人机中继性能仿真评估[J].计算机仿真,2023,40(3):453-459.

[2] 王蓉,吕祖盛,孙嘉,等.基于人像分割的智能搜救无人机系统设计[J].计算机技术与发展,2020,30(8):147-151+156.

[3] 何杏宇,付冲,杨桂松,等.基于任务与巡航方向相关性分析 的无人机任务分配[J].计算机应用研究,2022,39(10):2989-2995+3007.

[4] 荆文龙,周成虎,李勇,等.基于无人机智能基站的空地协同 低空无人机遥感网构建及应用[J].遥感学报,2023(2):209-223.

[5] NVIDIA DeepStream SDK 4.0.2 Release [EB/OL].(2019-12-20)[2024-01-22]. https://docs. nvidia. com/ metropolis/deepstream/4.0.2/dev-guide/index.html.

[6] Haq M A, Fahriani N. Improving YOLO Object Detection Per - formance on Single-Board Computer using Virtual Machine. Emerg[J]. Emerging Information Science and Technology, 2024, 5:36-45.

[7] 张呈宇,李红五,屈阳,等.面向工业互联网的5G边缘计算发展与应用[J].电信科学,2021,37(01):129-136.

[8] 张天魁, 徐瑜, 刘元玮, 等. 无人机辅助MEC系统:架构、关键技术与未来挑战[J]. 电信科学, 2022, 38(8): 3-16.

[9] Lin J, Chen W M, Lin Y, et al. MCUNet: Tiny deep learning on IoT devices[J]. Adv. Neural Inf. Process. Syst. (NeurIPS) 2020, 33:11711 - 11722.

[10] Lyu, NanoDet R. 2021. [EB/OL].(2021-12-20). https://github.com/RangiLyu/nanodet.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016:779 - 788.

[12] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger[C]. IEEE Conference on Computer Vision and Pattern Recognition, Hono - lulu, HI, USA,2017:7263 - 7271.

[13] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv:1804.02767.

[14] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy ofobject detection[J]. arXiv:2004.10934.

[15] YOLOv5. [EB/OL].(2022-11-22)[2024-01-22]. https://github.com/ultralytics/yolov5

[16] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv:2107.08430.

（下转第128页）

Underwater IOT perception analysis platform based on cross-domain unmanned system

ZENG Linxi，SHEN Xu，ZHANG Wei，WANG Jiajia
China Telecom Company Limited Chengdu branch, Chengdu 610000, China

Abstract: Sichuan Province has a good climate and rich water resources, and is a key area for biodiversity protection in the world. The suitable ecological environment, water temperature, water quality and flow velocity provide favorable conditions for the development of aquaculture industry. Our research objectives are mainly focused on two aspects: monitoring the aquaculture environment, especially the water quality, and monitoring the aquaculture objects and aquatic living resources. Based on the land and water cross-domain unmanned system, it is equipped with multiple types of water quality sensors and cameras according to different scenarios, and transmits the collected data back to the data center by various communication methods, so as to realize data analysis and generate analysis reports, and finally present them to the production and management subjects and regulatory departments in the form of the cockpit.

Keywords: Smart fishery, Cross-land unmanned system, Sensor, Underwater communication, Data management and analysis platform

（上接第116页）

[17] Xu S, Wang X, Lv X, et al. PP-YOLOE: An evolved version of yolo[J]. arXiv:2203.16250.

[18] Tijtgat N, Ranst W V, Volckaert B, et al. Embedded real-time object detection for a UAV warning system[C]. IEEE International Conference on Computer Vision Workshops, Venice, Italy, 2017: 2110 - 2118.

[19] Abdulghafoor N H, Abdullah H N. Real-time moving objects detection and tracking using deep-stream technology[J]. Journal of Engineering Science and Technology,2021, 16:194 - 208.

[20] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]. European conference on computer vision (ECCV), 2018:116-131.

[21] Zhu L, Wang X, Ke Z, et al. Biformer. Vision transformer with bi-level routing attention[C].In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023:10323-10333.

[22] Sucheng Ren, Daquan Zhou, Shengfeng He, et al. Shunted self-attention via multi-scale token aggregation[C].arXiv:2111.15193, 2022: 10853 - 10862.

作者简介

王燚（1975—），男，工学博士，高级工程师，主要研究方向为智慧城市中的泛在移动通信系统和智能网络技术研究。

杨博文（1999—），男，工学学士，咨询设计师，主要研究方向为人工神经网络和图像检测。

唐双江（1987—），男，工学学士，高级工程师，主要从事通信网络规划设计与业务咨询研究工作。

Application of edged AI recognition in UAV inspection

WANG Yi, YANG Bowen, TANG Shuangjiang
Sichuan Communication Research Planning&Designing Co.Ltd,Chengdu 610041,China

Abstract: With the rapid development of unmanned aerial vehicle (UAV) technology, the application of UAV in the field of inspection is increasingly extensive, especially in the power, construction, and other industries. UAV inspection can greatly improve inspection efficiency, reduce labor costs, and carry out efficient operations in complex environments. However, the traditional UAV inspection method relies on manual recognition, which has some problems such as low efficiency and easy error. UAV technology combined with the application of deep learning object recognition can significantly improve the efficiency and accuracy of inspection. This paper aims to develop a deep learning-based embedded target recognition model for UAV inspection edge, deploy it on UAV edge equipment, and detect and identify targets from aerial video streams in real time, to provide an efficient and accurate solution for inspection.

Keywords: UAV inspection, Edge embedding, Object detection