

融合自适应置信度的轻量化 YOLOv8 检测模型

刘梓熙

南加利福尼亚大学

DOI: 10.32629/ems.v8i9.22147

[摘要] 边缘设备的算力与访存带宽双重受限,使检测模型的参数压缩不能仅以浮点运算量为唯一目标。针对深度可分离卷积在低通道数下计算强度不足、实际时延与理论量级背离的问题,构建以部分卷积与跨阶段特征复用为核心的轻量主干,并在特征融合通路上以通道剪裁替代整层删除,保留多尺度语义的完整性。检测头的固定置信度阈值被替换为随目标尺度与局部密度联动的自适应判决机制,阈值由预测分布的统计量在线估计,缓解密集小目标场景下漏检与误检此消彼长的困境。实验结果表明,参数量与推理时延同步下降,平均精度均值未出现明显折损,小目标类别的召回获得改善,模型在嵌入式平台上的实测帧率满足实时约束。

[关键词] 目标检测;轻量化网络;自适应置信度;特征融合;边缘部署

引言:

轻量化设计长期存在一个被指标掩盖的错位。浮点运算量的削减并不等价于时延的等比下降,深度可分离结构将标准卷积拆为逐通道与逐点两步,运算量显著减少,访存次数却因中间特征图的反复读写而上升,在访存带宽受限的嵌入式芯片上,理论加速比往往无法兑现^[1]。以运算量为唯一优化目标的搜索过程,可能收敛到访存密集而计算稀疏的结构。

1 融合自适应置信度的检测模型设计

1.1 目标检测模型基础分析

单阶段检测将定位与分类合并于单次前向推理,摒弃候选框生成环节,速度优势来自结构而非实现优化。主干网络承担层次化特征抽取,浅层保留高分辨率的边缘与纹理,深层获得大感受野的语义表征,二者的分辨率与语义强度呈天然的倒置关系。特征融合模块以自顶向下与自底向上的双向通路缓解该倒置,深层语义经上采样与浅层特征相加或拼接,浅层定位信息再经下采样回流^[2]。

检测头的解耦设计将分类分支与回归分支分离,两类任务的最优特征本不相同,分类关注语义类别的判别性,回归依赖边界处的空间精度,共享参数迫使二者在梯度上相互妥协。无锚框范式进一步移除预设先验框,直接回归目标中心到四边的距离:

$$\hat{b} = (l, t, r, b), \mathcal{L}_{\text{reg}} = 1 - \text{IoU}(\hat{b}, b) + \frac{\rho^2(b, c)}{d^2} + \beta v$$

其中中心距离项与长宽比共同约束回归方向,避免交并比无重叠时梯度消失。

1.2 轻量化网络结构优化设计

标准卷积参数量由输入通道、输出通道、卷积核尺寸乘积决定,通道相关性造成大量冗余^[3]。部分卷积选取部分通道做卷积运算,剩余通道直接传递后拼接融合:

$$Y = \text{Concat}(\text{Conv}(X_{[c_p]}), X_{[c_p:]})$$

该结构降低计算量同时维持高计算强度,访存计算比优于深度可分离卷积,实测时延更贴合理论降幅。跨阶段特征分流架构切断重复梯度传递,提升多层堆叠参数利用率。

主干输出三尺度特征接入融合模块,横向支路采用归一化可学习权重加权融合;增设统计汇聚旁路采集目标尺度、密度特征,独立输入阈值单元;全部下采样使用步长卷积,降低小特征丢失概率。

1.3 自适应置信度检测机制构建

全局固定阈值无法适配疏密差异图像,稀疏样本分数区分明显,密集场景正样本分数下沉易被误筛,构建尺度、密度联动自适应阈值公式:

$$\tau(s, \rho) = \tau_0 + \lambda_s \log \frac{s_{\text{ref}}}{s} + \lambda_\rho \sigma(\rho - \rho_0)$$

尺度系数下调小目标判定门限,密度系数抬高拥挤区域阈值,超参通过验证集网格搜索获取。

尺度系数下调小目标判定门限,密度系数抬高拥挤区域阈值,超参通过验证集网格搜索获取。

阈值由特征统计实时计算,设置双重约束:阈值取值上下限防止区域无检测结果;采用高斯低通平滑阈值分布图,消除相邻区域阈值突变带来的目标检测闪烁问题。

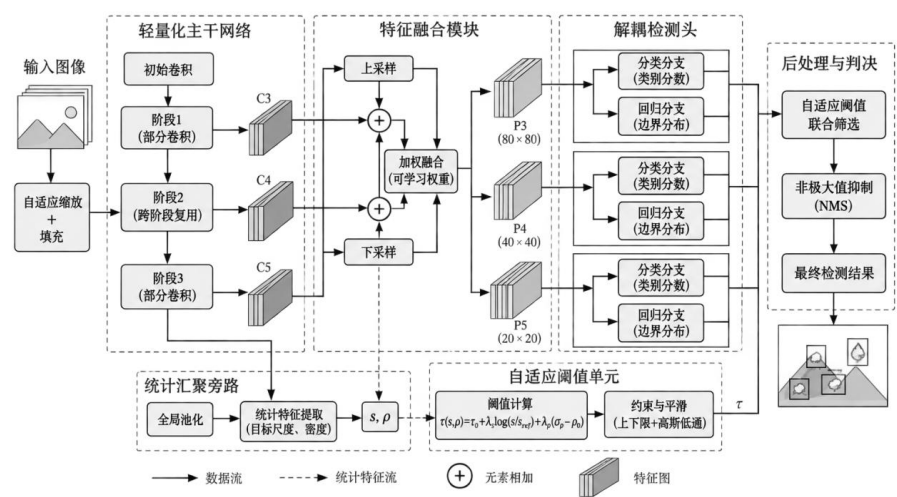


图 1 轻量化主干与自适应置信度检测头的整体架构图

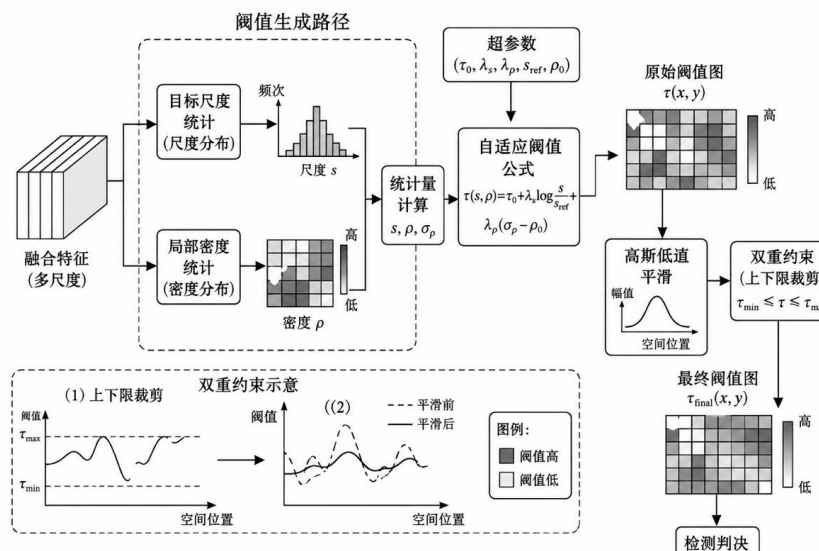


图2 自适应置信度阈值的生成路径与约束示意图

图 2 自适应置信度阈值的生成路径与约束示意图

2 融合自适应置信度的模型实现

2.1 改进模型整体架构设计

输入图像经自适应缩放至统一分辨率，缩放采用保持长宽比的填充方式，直接拉伸会改变目标形状先验，使回归分支学到与部署阶段不一致的分布^[4]。特征提取阶段逐级降采样并扩展通道，各阶段末端引出旁路至融合模块。融合完成后的三尺度特征分别送入解耦检测头，分类分支输出类别分

数，回归分支输出边界分布，二者在后处理阶段由自适应阈值联合筛选。

2.2 特征提取与融合模块改进

小目标的困境在于其有效像素在深层几乎耗尽。单纯增加浅层分支的权重会引入大量背景噪声，改进方式转向在融合前对浅层特征施加空间注意力，以深层语义生成的掩码调制浅层响应，抑制背景而保留目标区域的高分辨率细节。

表 1 检测模型各改进模块的作用对象与适用条件对照

改进模块	作用对象	引入的结构代价	适用条件
部分卷积主干	主干各阶段的通道冗余	通道拼接带来的访存开销	部署平台访存带宽受限
跨阶段特征复用	深层堆叠的梯度路径重复	分流合流处的特征图缓存	网络深度超过若干阶段

语义调制的横向连接	浅层特征中的背景响应	掩码生成的额外逐点卷积	目标与背景纹理相近
可学习加权融合	多尺度分支的数值幅度失衡	每条连接一个标量参数	各尺度目标数量分布不均
自适应置信度判决	固定阈值的场景适应性	统计汇聚支路的推理时延	目标密度在图像内变化剧烈

各模块的代价形态互不相同，选择需依部署环境而定。部分卷积与跨阶段复用同属结构级压缩，二者叠加时通道拼接次数累加，访存开销可能反超单独使用任一模块的收益，在带宽极度受限的芯片上宜择一而行。语义调制与可学习加权均作用于融合通路，前者代价随特征图分辨率增长，后者近乎可忽略，在小目标占比高的任务中前者收益明显，反之其代价难以回收。自适应判决的代价集中于后处理而非主干，与前四者在计算资源上不冲突，适合与任意压缩方案组合。

2.3 检测算法训练流程实现

数据增强以马赛克拼接与随机仿射为主，拼接使单张训练图内目标尺度分布拓宽，缓解小目标样本不足。训练末期需关闭拼接增强，其构造的图像与真实分布存在系统性差异，持续开启会使模型在部署数据上产生偏移^[5]。

标签分配采用动态匹配策略，正样本的选取依据分类与回归的联合代价而非静态的中心距离，代价矩阵经最优传输近似求解，使高质量预测优先获得正标签。损失由分类项、回归项与分布焦点项加权构成，各项权重的比例需与自适应阈值机制协同调整，阈值下调会增加低分预测参与后处理的比例，若分类损失权重不同步提高，误检率上升。

3 模型性能测试与结果分析

3.1 实验环境与评价指标

训练在图形处理器工作站完成，部署测试在嵌入式计算模组上进行，两端环境的分离是必要的，仅以工作站时延评估轻量化效果将系统性高估收益。评价指标涵盖精确率、召回率、不同交并比阈值下的平均精度均值、参数量与单帧推理时延。参数量与时延须并列报告，二者在部分卷积等结构上并不同步变化。

3.2 改进模型检测效果分析

对比实验以原始模型及若干通用轻量化变体为基线。改进模型的参数量降至基线的一个明显更小的量级，推理时延同步下降且下降比例接近参数量的压缩比例，验证了计算强度保持策略的有效性。平均精度均值相对原始模型仅有轻微回落，相对同等压缩率的通用变体则明显更高。

自适应置信度的贡献在密集场景中最为集中，小目标类

别的召回获得可观提升，同时误检并未同比例增长，说明阈值的尺度项与密度项确实实现了解耦而非简单的整体平移。消融实验中移除密度项后，拥挤区域的重复框数量回升，移除尺度项后小目标召回下滑，两项的作用域彼此独立^[6]。

3.3 模型应用性能综合评价

嵌入式平台的实测帧率满足实时约束，在持续推理下的温度与功耗处于模组标称范围之内，具备长时运行的可行性。局限同样清晰。自适应阈值的系数依赖验证集搜索，跨场景迁移时需重新标定，尚未实现完全的免调参。统计汇聚支路在目标极度稀疏的图像上估计方差偏大，阈值出现轻微抖动，后续可通过跨帧平滑加以抑制。

结语

轻量化的有效性取决于压缩目标与部署瓶颈是否对齐，以运算量为唯一指标的优化在访存受限平台上难以兑现预期加速。将部分卷积与跨阶段复用组合，压缩比与实测时延降幅得以接近同步。置信度判决从全局固定转为随尺度与密度联动，漏检与误检之间的零和交换被打破，密集小目标场景的召回改善并未以误检上升为代价。自适应系数的跨场景标定仍需人工介入，无监督的在线校准是值得推进的方向。

【参考文献】

[1] 绳鹏飞, 杨军. 采用轻量化主干网络的遥感影像目标检测[J]. 测绘科学, 2023, 48(5): 58-67.

[2] 杨玉敏, 廖育荣, 林存宝, 等. 轻量化卷积神经网络目标检测算法综述[J]. 舰船电子工程, 2021, 41(4): 31-36.

[3] 权宇, 李志欣, 张灿龙, 等. 融合深度扩张网络和轻量化网络的目标检测模型[J]. 电子学报, 2020, 48(2): 390-397.

[4] 初文婷, 刘菁, 张稚萱, 黄峻恺, 陈宇, 袁媛, 魏祥龙. 基于改进 YOLOv8n 的水面漂浮物轻量化检测方法[J]. 水电能源科学, 1-12.

[5] 王骏涛, 刘竞阳, 冯勋伟, 杨加宇, 杨彪, 龙萍. 基于稀疏注意力增强与轻量特征融合的小目标车辆检测方法[J]. 计算机测量与控制, 1-12.

[6] 刘正, 王利众, 梁骁. 基于改进 YOLOv8n 的 PCB 缺陷检测算法[J]. 现代电子技术, 2026, 49(11): 150-158+164.