Journal of Image and Signal Processing 图像与信号处理, 2026, 15(3), 309-318
Published Online July 2026 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2026.153027

Hans汉斯

基于YOLOv10s的织物缺陷检测轻量化研究

肖钦元，逯力红

天津工业大学物理科学与技术学院，天津

收稿日期：2026年6月2日；录用日期：2026年6月24日；发布日期：2026年7月7日

摘　要

针对移动端织物缺陷检测算力受限、推理延迟高的问题，本文基于YOLOv10s提出轻量化模型MR-YOLOv10s。首先，引入MobileNetV4重构骨干网络，利用万能倒残差块(UIB)与移动端多查询注意力(Mobile MQA)剥离计算冗余，降低内存访问成本。其次，在颈部网络引入结构重参数化RCS-OSA模块，通过训练阶段多分支强化与推理阶段模块融合，在不增加时延的前提下有效回补精度损耗。此外，结合SCDown与C2fCIB单元提升特征提纯能力。实验结果显示，MR-YOLOv10s参数量为4.18 M，计算量为12.5 GFLOPs，较基线模型缩减近50%。与Faster R-CNN相比，推理速度提升约8倍(单张仅需3.1 ms)，实现了速度与能效的深度缩减，适配移动端部署需求。

关键词

YOLOv10s，MobileNetV4，RCS-OSA，轻量化

Lightweight Research on Fabric Defect Detection Based on YOLOv10s

Qinyuan Xiao, Lihong Lu

School of Physical Science and Technology, Tiangong University, Tianjin

Received: June 2, 2026; accepted: June 24, 2026; published: July 7, 2026

Abstract

To address the issues of limited computing power and high inference latency in mobile device fabric defect detection, this paper proposes a lightweight model named MR-YOLOv10s based on YOLOv10s. Firstly, MobileNetV4 is introduced to restructure the backbone network, and the universal inverted residual block (UIB) and mobile multi-query attention (Mobile MQA) are utilized to eliminate computational redundancies and reduce memory access costs. Secondly, a structure reparameterization RCS-OSA module is introduced in the neck network. Through multi-branch reinforcement in

文章引用: 肖钦元, 逯力红. 基于 YOLOv10s 的织物缺陷检测轻量化研究[J]. 图像与信号处理, 2026, 15(3): 309-318.
DOI: 10.12677/jisp.2026.153027

the training stage and operator fusion in the inference stage, the accuracy loss is effectively compensated without increasing latency. Additionally, the SCDown and C2fCIB units are combined to enhance the feature purification ability. Experimental results show that MR-YOLOv10s has 4.18 M parameters and 12.5 GFLOPs of computational power, reducing nearly 50% compared to the baseline model. Compared with Faster R-CNN, the inference speed is approximately 8 times faster (only 3.1 ms per single instance), achieving a significant reduction in both speed and energy efficiency, and adapted to the deployment requirements of mobile devices.

Keywords

YOLOv10s, MobileNetV4, RCS-OSA, Lightweight

1. 引言

纺织业作为国民经济的重要产业之一，产品质量的把控是增强企业竞争力的关键。目前，织物缺陷检测正从传统人工向着自动化、智能化发展。然而，实际生产场景中不仅存在固定流水线，还存在有手持质检终端、移动式巡检设备等算力资源受限的检测场景。这些移动端平台通常受限于电池续航与 ARM 架构处理器的算力，对算法的推理延迟、内存占用及硬件能效比提出了极为严苛的要求。因此，开发一种适配于移动检测算力的轻量化检测系统，对于实现便携的低成本的织物缺陷检测具有重要意义。

传统的目标检测算法主要依赖数字图像处理技术，通过人工设计的特征算子，如灰度共生矩阵(Gray-Level Co-occurrence Matrix, GLCM)等统计学方法[1]、Gabor 变换等纹理特征分析[2]以及基于傅里叶变换的频谱分析方法，来量化织物表面规律。但是，传统检测方法对算子预设的依赖程度较高，对背景复杂的微小缺陷的识别精度不高，难以在复杂多变的生产环境中实现高精度的自动识别。随着 2012 年 AlexNet 的提出，深度学习技术掀起了研究热潮，并逐渐取代传统特征技术[3]。目前，基于深度学习的目标检测算法可分为双阶段(Two-stage)和单阶段(One-stage)两大类。双阶段检测以 Faster R-CNN 为代表，通过预生成候选区域再进行分类，虽然精度较高但在速度上难以满足实时需求[4]；单阶段检测以 SSD 和 YOLO 系列为代表[5]，能够在一次前向传播中同时完成目标分类与定位，在追求高效率的工业检测领域表现出了更强的应用前景。

在目标检测任务的推进中，Redmon 等人提出的 YOLOv1 将检测流程统一为单阶段网络，大幅提升了检测效率[6]。后续的 YOLOv3 引入多尺度预测与残差主干网络，增强了对织物缺陷尺度变化的适应能力[7]。Ultralytics 团队开发的 YOLOv5 采用 CSPDarknet 与多分支融合结构，已在纺织工业实时检测中广泛应用[8]。而进一步优化而来的 YOLOv8 则通过解耦检测头与改进特征融合策略，在多种瑕疵的识别任务中实现了更均衡的精度与速度表现。

然而，这些传统模型在推理阶段仍依赖非极大值抑制(Non-Maximum Suppression, NMS)后处理环节，用以滤除重叠的预测框。在算力受限的嵌入式或移动端设备中，NMS 往往涉及大量的跨硬件调度与串行计算，极易成为全流程推理的实时性瓶颈。相比之下，YOLOv10s 通过引入双重标签分配策略，彻底消除了对 NMS 的依赖，实现了真正意义上的端到端(End-to-End)检测，在理论上具备更高的推理效率上限[9]。但对于手持质检终端等极低功耗场景，原生 YOLOv10s 的骨干网络仍存在一定的计算冗余，且在进

行大幅度轻量化压缩后，极易导致对织物微小缺陷的特征捕捉能力下降。如何在维持端到端高效检测逻辑的前提下，实现算法架构在移动端硬件上的性能突破，成为当前研究的热点。

2. MR-YOLOv10s 模型

为实现全流程的轻量化减负增效，本文从两个方面进行改进：

(1) 骨干网络轻量化重构

引入 MobileNetV4 架构对原生骨干网络进行重构。MobileNetV4 是 Google 针对移动终端全平台(包括 CPU、DSP 及 NPU)设计的最新一代轻量级视觉骨干模型。其设计精髓在于通过硬件感知(Hardware-Aware)的神经架构搜索技术，构建了核心的万能倒残差块(Universal Inverted Bottleneck, UIB)，以及使用针对移动加速器优化的移动端多查询注意力机制(Mobile Multi-Query Attention, Mobile MQA)注意力模块。UIB 模块打破了传统轻量化算子的固定设计定式，通过在特征升维与降维层之间灵活配置深度可分离卷积，实现了对硬件计算效率的自适应匹配。Mobile MQA 通过优化算术运算与内存访问的比率，显著提高了移动加速器上的推理速度，这是决定移动端推理性能的关键因素。MobileNetV4 还引入了改进的神经网络架构搜索(Neural Architecture Search, NAS)策略，通过粗粒度和细粒度搜索相结合的方法，显著提高搜索效率并改善模型质量。

选择 MobileNetV4 的核心原因在于其能够从底层逻辑上剥离织物图像处理中的计算冗余，通过秩引导的设计有效锁定最具判别性的纹理通道。这种设计不仅极大地压缩了模型的参数规模，更关键的是通过优化算子组合降低了内存访问成本(Memory Access Cost, MAC)，对于缓解移动端手持设备在长时间检测过程中的发热与降频问题具有重要的工程意义。

(2) 颈部特征聚合网络优化

针对骨干网络轻量化后特征表征能力会减弱的客观事实，本文在颈部网络(Neck)引入了基于结构重参数化卷积与单次聚合结构的基于单次聚合的实时卷积(Real-time Convolutional Stem with One-Shot Aggregation, RCS-OSA)模块。其技术优势在于利用结构重参数化(Reparameterization)技术，在训练阶段构建复杂的多分支通路，增强模型对断经、星跳等微小缺陷弱信号的拟合能力；在推理阶段则通过数学变换将多分支合并为单路卷积，实现在不产生额外推理时延的前提下精度无损提升。选择 RCS-OSA 替换原生 C2f 模块，是因为其单次聚合(One-Shot Aggregation, OSA)机制能够显著缩减特征在多层传递中的冗余计算链路，有效防止因骨干轻量化而导致的缺陷语义特征被背景噪声所稀释，从而在轻量化框架内最大程度地回补精度损耗。

通过将深度适配移动端的 MobileNetV4 骨干网络与具备特征增强能力的 RCS-OSA 颈部进行协同集成，本研究构建了一种基于 YOLOv10s 改进的轻量化检测算法，将其命名为 MR-YOLOv10s。该模型并非盲目追求单一维度的参数缩减，而是致力于在移动端硬件约束下寻找精度、速度与功耗的最佳平衡点。通过这种跨层级的轻量化重构，旨在构建一个既能灵敏识别织物主要疵点，又能在移动端平台实现快速响应的视觉检测方案。

2.1. 轻量化改进机理

2.1.1. UIB 单元与自适应神经架构搜索

MobileNetV4 的技术核心在于通过 UIB 构筑了一个具有高灵活度的算子计算空间。UIB 的设计初衷是解决传统轻量化算子在不同计算设备(如 GPU 的并行计算能力与 CPU 的标量处理能力)上的表现差异。其拓扑结构包含两个灵活的深度卷积(Depthwise Convolution, DW)插入位置，起始位置的 Extra DW 能够有效增加感受野，而中间位置的 Middle DW 则侧重于局部空间的非线性映射。其结构如图 1 所示。

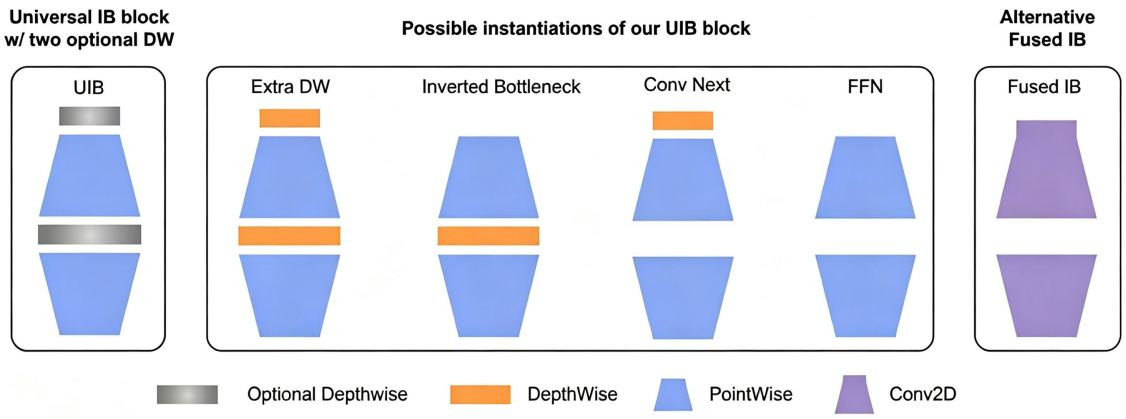

Figure 1. UIB structure
图 1. UIB 结构

在织物缺陷检测任务中，UIB 的万能属性体现在对纹理特征的自适应选择上。通过改进的 NAS 策略，模型在浅层网络中趋向于类 ConvNext 结构，利用前置的大核深度卷积捕获织物表面经纬交织的全局频率分布；在计算密集的深层网络中，模型会自动切换为前馈网络或标准倒残差模式，利用点卷积在极低延迟下完成通道间的语义整合。这种基于秩引导的动态算子分配，使得骨干网络能够精准剥离背景中的低秩冗余信息，仅保留对缺陷判定至关重要的异质信号，从而在算法底层实现了计算强度与推理能效的统一。

2.1.2. Mobile MQA 机制

Mobile MQA 通过采用非对称的查询(Query)与键值(Key Value)投影方式，对注意力计算的过程进行简化。在处理织物缺陷信号时，MQA 允许模型在保持查询的表达能力的同时，对键和值的维度进行压缩来降低缓存占用[10]。这种设计在识别长距离连续且有方向性的缺陷时，使模型在一定程度保持检测精度。同时，使模型能够在移动端有限的内存带宽内，仍然保持良好的全局感受度，确保在移除 NMS 这种后处理环节后，依然对特征信号有着较好的识别能力，有效缓解了极致轻量化导致的特征退化问题。

Mobile MQA 采用了多头注意力共享，其表达式为：

$$attention_j = softmax\left(\frac{\left(XW^{Q_j}\right)\left(SR(X)W^K\right)^T}{\sqrt{d_k}}\right)\left(SR(X)W^V\right)$$
$$\text{Mobile\_MQA}(X) = Concat\left(attention_1,\cdots,attention_n\right)W^O \tag{1}$$

式中，X 为输入特征张量，j 为并行注意力头的索引，每个头 j 都有自己独立的权重 W^{Q_j}，所有的注意力头共享同一组 Key 和 Value，与标准 MHA (每个头都有独立的 K 和 V)相比，MQA 大幅减少了参数量。同时引入空间缩减技术($SR(X)$)，大幅降低了自注意力计算的复杂度。

2.1.3. RCS-OSA 的特征补强

颈部网络中引入的 RCS-OSA 模块的目的在于解决特征融合过程中的内存墙问题[11]。传统的 C2f 模块虽然梯度丰富，但频繁的跨层连接会导致大量的中间特征图积压在内存中，增加了移动端的功耗。RCS-OSA 通过 OSA 策略，将多层卷积的输出整合，从而一次性输出。

该结构的优势在于利用结构重参数化(Reparameterization)的数学特性。模型在训练过程中，通过 1 ×

1 与 3 × 3 卷积的并行路径，增强了特征信息表达的准确性，能够更敏锐地对微小缺陷的信号进行提取。在推理阶段，这些分支通过算子融合(Fused Kernel)转换为类似于单路卷积的模式，在不消耗额外计算资源的前提下，有效回补了骨干极致轻量化带来的精度损失。

2.2. 模型网络架构

2.2.1. 轻量化骨干网络

骨干网络是模型提取织物缺陷特征的关键结构，MR-YOLOv10s 采用了集成化的 MobileNetV4ConvSmall 作为轻量化骨干[12]。该骨干网络被展开为五个连续的特征提取阶段(Stage)，分别对应为第 0 至 4 层。这种编排方式为骨干网络建立了清晰的层级结构，为后续的跨尺度融合提供了准确的层级引导。

具体的层级分布中，第 0 和 1 层对应浅层纹理提取阶段，进入第 2 到 4 层后，特征图的语义表达能力有了明显的增强，这一部分为颈部网络获取多尺度信息提供了保障。其中第 2 层(P3)输出原始分辨率 1/8 的特征图，由于它保留了很高的空间响应密度，所以被看作是模型识别细小缺陷的关键层。层级 3 (P4)输出 1/16 尺度的特征图，用于平衡语义深度与空间定位精度。层级 4 (P5)作为骨干末端层级输出具备最高抽象维度的 1/32 尺度的特征图。

在骨干网络的输出末端，设置了特征值全局强化模块。第 5 层的快速空间金字塔池化(Spatial Pyramid Pooling-Fast, SPPF)模块利用 5 × 5 卷积和的连续池化操作，在不增加参数的前提下扩大感受野，实现缺陷特征与复杂织物背景的初步解耦。随后的第 6 层 PSA 模块通过对深层张量执行部分自注意力建模，强化模型对疵点全局几何分布的感知深度，确保进入颈部网络前的特征流具备足够的判别性。

2.2.2. 颈部网络

颈部网络具有多尺度特征对齐聚合与信息耦合增强的功能，本研究利用 RCS-OSA 模块重构了 FPN-PAN 的聚合节点。

在自上而下的采样路径中，深层语义信息经上采样层后，分别与主干网络层级 3 和层级 2 传来的中间特征进行维度拼接。在此过程中，RCS-OSA 模块取代了传统的 C2f 结构，作为融合阶段的主控算子。该模块利用重参数化技术对汇合后的特征流执行单次聚合处理，在有效识别特征信号的同时，大幅减少了参数在移动端内存中的驻留时间，降低了内存占用与功耗。

在自下而上的下采样路径中，MR-YOLOv10s 采取了差异化的增强策略。第 13 层利用步长为 2 的传统卷积完成空间压缩。针对深层特征容易在多级融合中产生信息混叠的问题，第 16 层通过了 SCDown 模块进行去耦下采样，利用先调整通道后压缩空间的技术原理，识别高纬度的特征信号。在进入检测头前的最后阶段，模型通过使用 C2fCIB 单元，利用它紧凑型倒残差路径对多尺度融合后的最高维特征进行提纯，确保了模型对横档类大尺度缺陷判定精确度。

2.2.3. 检测头输出

MR-YOLOv10s 的网络结构在第 19 层完成了对三个尺度分支(P3, P4, P5)的最终集成。这三个分支分别承载着不同空间分辨率的特征映射，通过输入 640 × 640 分辨率的待测图，检测头分别在 80 × 80、40 × 40 和 20 × 20 三个维度上进行目标预测。这种设计确保了算法对从像素级到全幅面级的织物疵点均具备极高的覆盖率。

最终的检测头仍保留了 YOLOv10s 的 v10Detect，检测头接收这三个分支的输出，通过解耦的分类与回归路径生成检测结果。整个网络结构通过对骨干网络使用 MobileNetV4 的底层重构，以及在颈部灵活运用 RCS-OSA 和保留 SCDown，实现了一套针对移动端硬件深度优化的闭环体系。这种全流程的轻量化重构使得模型不仅消除了计算过程中的非必要开销，还通过无 NMS 依赖的端到端快速推理，满足了移动

终端在实际工业工况下对缺陷实时监控的性能要求。

MR-YOLOv10s 的整体结构如图 2 所示。

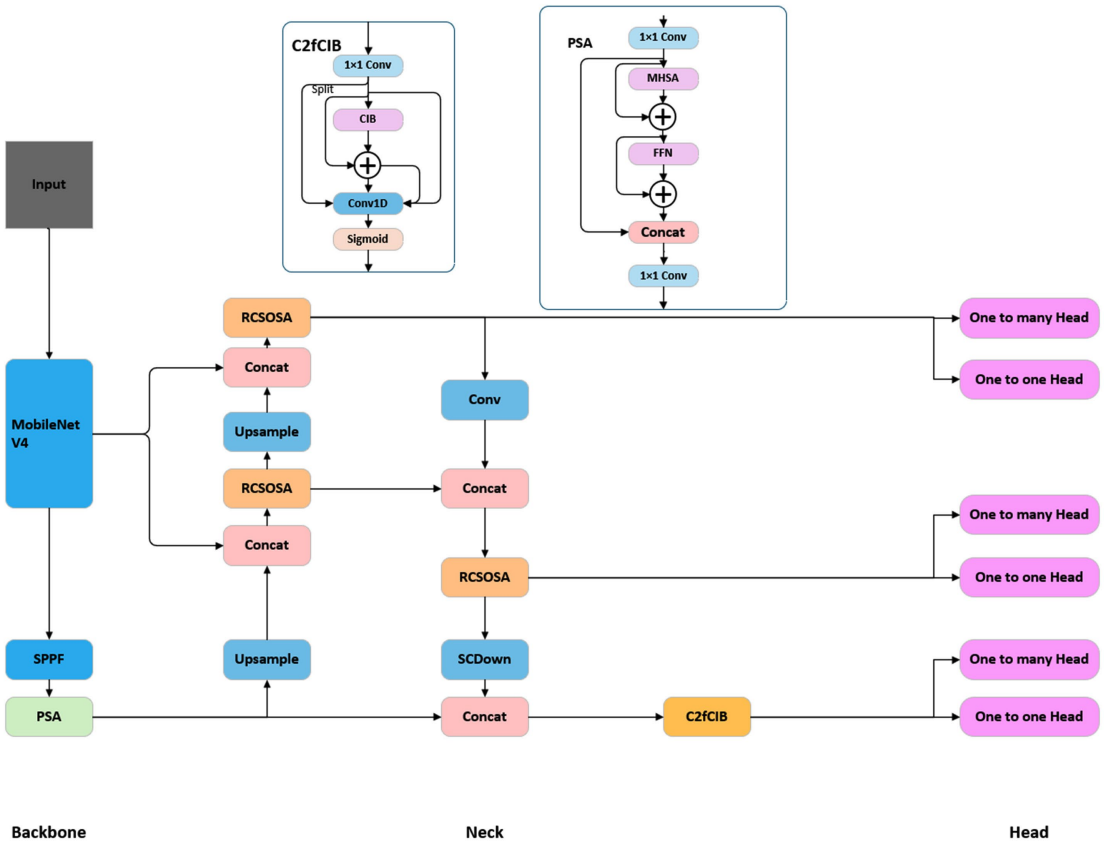

Figure 2. MR-YOLOv10s
图 2. MR-YOLOv10s

3. 实验结果与分析

3.1. 实验环境

所有实验均在统一完善且稳定的软硬件环境下完成。软件配置方面，操作系统为 Windows11，使用 Python 3.9.19、torch-2.0.1、cu118 构建深度学习框架。硬件配置方面更为重要，本文使用的硬件具体配置参数如表 1 所示。

Table 1. Hardware configuration and parameters
表 1. 硬件配置及参数

硬件	型号参数
CPU	Intel i7-14700 k
GPU	4070
GPU 显存	12 G

网络模型训练过程中，需要对网络进行一些参数配置，部分参数对于训练结果有着很大的影响。其

中模型训练轮数设置为 200 个 epoch，批大小为 16，初始学习率为 0.01。

本研究采用阿里云天池平台开源的织物缺陷竞赛数据集。该数据集使用工业相机拍摄于真实的工业生产现场，包含了多种复杂纹理背景下的织物图像，是国内织物缺陷检测领域权威的公开数据集，具有较高的应用价值。根据需求筛选出 8 大类典型缺陷，经过数据增强扩充得到本研究所用的实验数据集。

3.2. 消融实验

为了验证每个改进模块对轻量化模型性能的独立贡献度，本研究设计了严格的单变量消融实验，以原始 YOLOv10s 为基线模型，在完全一致的实验环境下，逐步引入 MobileNetV4 主干与 RCS-OSA 模块。消融实验结果如表 2 所示。

Table 2. Ablation experiment
表 2. 消融实验

Backbone	Neck	mAP@0.5	参数量/M	GFLOPs
原始	原始	0.79	8.09	24.8
MobileNetV4	原始	0.602	3.95	11.2
MobileNetV4	RCS-OSA	0.618	4.18	12.5

从消融实验结果可以发现：

(1) 当模型主干替换为极简的 MobileNetV4ConvSmall 后，由于该模块序列大幅精简了卷积路径并脱离了参数密集的 CSP 结构，模型实现了实质性的减重。参数量由 8.08 M 锐减至 3.95 M，缩减比例达 51.1%；GFLOPs 下降了 54.8%。但受骨干轻量化影响，模型的特征表征能力有所下降，mAP 降低到了 0.602；

(2) 在 Neck 引入 RCS-OSA 模块替换 C2f，在仅微幅增加 0.23 M 参数量的前提下，利用结构重参数化技术成功将精度回补至 0.618。1.8%的增益证明了通过增强颈部的信息复用深度有效补偿了骨干轻量化带来的语义特征缺失，在移动端硬件约束下找到了精度与效率的最优平衡点；

(3) MR-YOLOv10s 模型相较于基线 YOLOv10s，参数量缩减了 48.3%，GFLOPs 降幅约 49.6%，推理速度提升了约 35%，在大幅轻量化的同时，保留了织物缺陷检测的核心效能，完全适配移动端算力受限场景的需求。

上述两种改进方案的结果曲线如图 3 和图 4 所示，通过曲线对比可知，使用两项改进的网络结构，在训练 mAP 的上升速度上明显高于仅使用 MobileNetV4 的网络，同时它的曲线更加的平滑，这也表明它的训练更加的稳定且高效。

3.3. 主流算法对比实验

本研究通过将 MR-YOLOv10s 与经典的 Faster R-CNN 进行对比实验，验证其在工业部署中的价值。实验旨在通过分析在检测精度水平基本相似的前提下，不同网络架构在检测效率上的差异，对比结果如表 3 所示。

Table 3. Controlled experiment
表 3. 对比实验

网络模型	mAP@0.5	参数量/M	GFLOPs	检测时间/ms
Faster R-CNN	0.629	41.6	91.2	25
MR-YOLOv10s	0.618	4.18	12.5	3.1

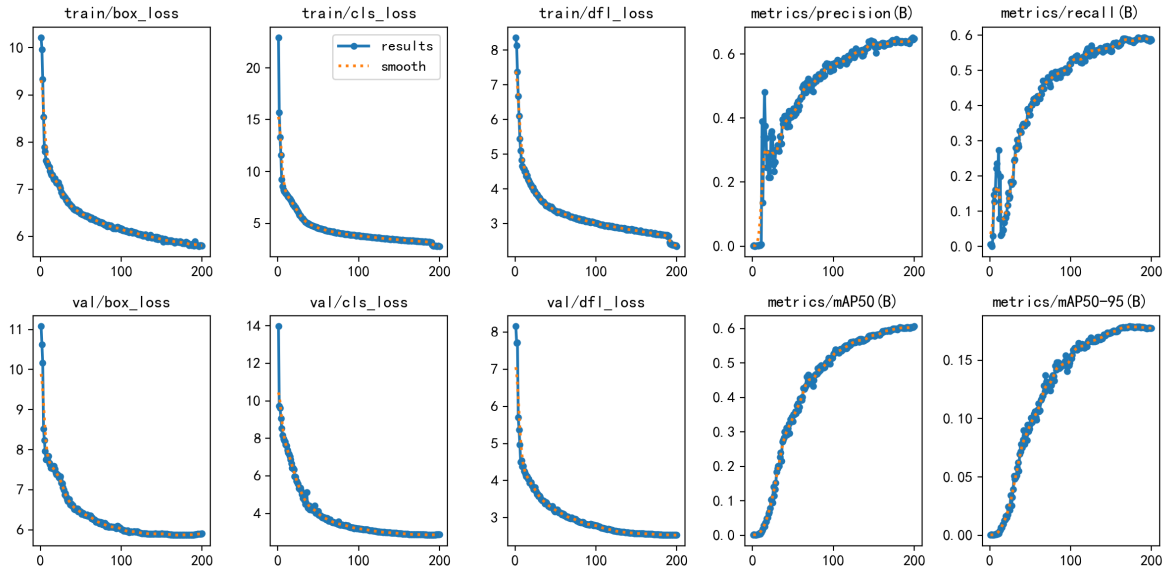

Figure 3. The result using only MobileNetV4
图 3. 仅使用 MobileNetV4 的结果

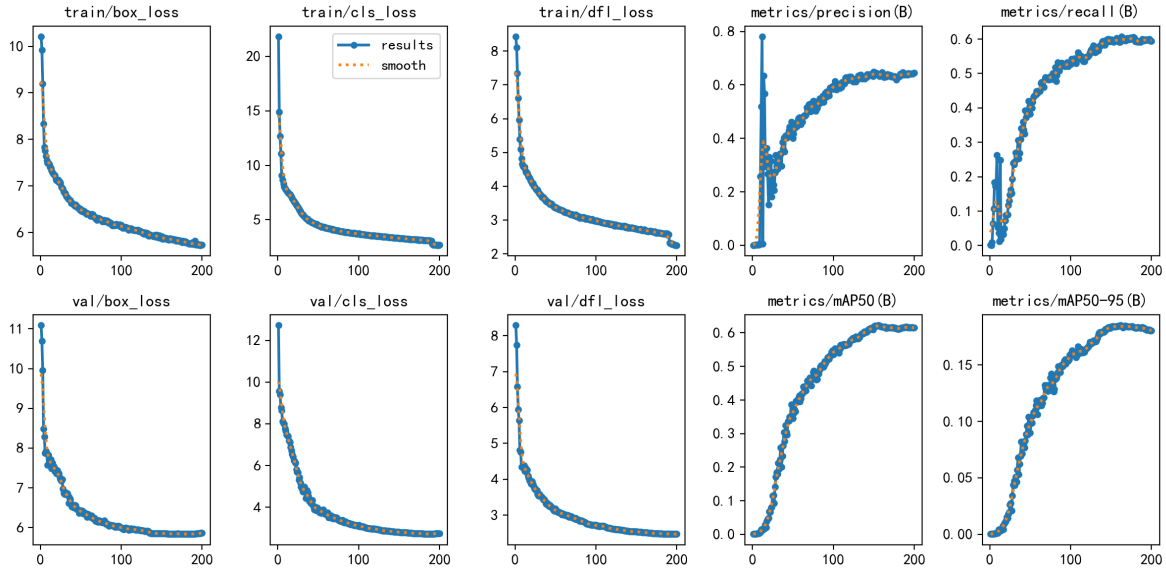

Figure 4. The results of MR-YOLOv10s
图 4. MR-YOLOv10s 的结果

通过实验可以看出，MR-YOLOv10s 在基本保持了与经典双阶段算法 Faster R-CNN 相当的检测精度的同时，在轻量化指标与检测效率上展现出了其轻量化的性能优势。在检测精度与模型规模的权衡上，虽然 Faster R-CNN 的 mAP@0.5 指标比本文模型高出约 1.2%，但它需要以高达 41.6 M 的参数量以及 91.2 GFLOPs 的运算负荷作为代价。相比之下，MR-YOLOv10s 的参数量仅为 4.18 M，约为前者的十分之一，而且计算开销也大幅缩减 12.5 GFLOPs。这种参数水平的差距意味着 Faster R-CNN 必须依赖高性能服务器显卡才能运行，但是 MR-YOLOv10s 则能轻量化地部署在算力受限的移动终端或手持检测设备中，大幅降低了硬件部署成本。

在检测实时性方面，推理速度的提升尤为显著。Faster R-CNN 单张图像的检测时间基本为 25 ms，在工业高频相机的实时采集环境下，极易造成数据处理的积压与延迟。而本研究改进后的 MR-YOLOv10s 仅需 3.1 ms 即可完成一次端到端的推理过程，检测速度提升了约 8 倍。极低的检测时间不仅确保了模型能够敏锐捕捉高速运动下的织物缺陷，更为移动设备提供了充沛的算力冗余，有助于缓解设备在高负荷运行下的发热与功耗问题，延长了移动端连续作业的时间。

4. 结论

本文针对手持质检终端及移动巡检设备对织物缺陷检测算法的轻量化需求，在 YOLOv10s 框架下通过主干重构与颈部优化构建了 MR-YOLOv10s 模型。通过引入具备硬件感知特性的 MobileNetV4 网络重构主干，利用万能倒残差块与 Mobile MQA 机制剥离冗余计算，使得模型参数量较基线 YOLOv10s 下降约 48.3%；针对轻量化导致的特征弱化问题，在颈部引入 RCS-OSA 模块并结合 SCDown 模块，利用结构重参数化技术在推理阶段实现精度的无损回补。实验结果表明，在检测精度与经典双阶段算法 Faster R-CNN 基本持平的前提下，MR-YOLOv10s 的参数量仅为后者的 1/10 左右，计算开销大幅降至 12.5 GFLOPs；同时，其单帧推理时间仅需 3.1 ms，检测速度较 Faster R-CNN 提升了约 8 倍。该模型有效降低了能效与检测速度，为算力受限的移动端环境提供了一种高效、低功耗的织物缺陷在线检测方案。

本文虽然通过模型主干网络压缩降低了计算复杂度，但没有在特定的 ARM 架构芯片或 NPU 上执行指令的深度优化。未来将探索基于 TensorRT 或 OpenVINO 的模型量化部署方案，来进一步提升算法在移动端上的运行能力。同时骨干替换为 MobileNetV4 导致了精度下降，虽然通过引入 RCS-OSA 模块实现部分提升，但仍需探索更好的方法进一步提升精度。

参考文献

[1] Khwakhali, U.S., Tra, N.T., Tin, H.V., Khai, T.D., Tin, C.Q. and Hoe, L.I. (2022) Fabric Defect Detection Using Gray Level Co-Occurence Matrix and Local Binary Pattern. 2022 *RIVF International Conference on Computing and Communication Technologies* (*RIVF*), Ho Chi Minh City, 20-22 December 2022, 226-231. https://doi.org/10.1109/rivf55975.2022.10013920

[2] Kumar, A. and Pang, G.K.H. (2002) Defect Detection in Textured Materials Using Gabor Filters. *IEEE Transactions on Industry Applications*, **38**, 425-440. https://doi.org/10.1109/28.993164

[3] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90.

[4] Ren, S.Q., He, K.M., Girshick, R. and Su, J. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149.

[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., *et al*. (2016) SSD: Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[6] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/cvpr.2016.91

[7] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv: 1804.02767.

[8] Jocher, G., Chaurasia, A., Stoken, A., *et al*. (2022) Ultralytics/YOLOv5: v7. 0-YOLOv5 Sota Realtime Instance Segmentation. Zenodo.

[9] Chen, H., Chen, K., Ding, G., Han, J., Lin, Z., Liu, L., *et al*. (2024) YOLOv10: Real-Time End-To-End Object Detection. *Advances in Neural Information Processing Systems* 37, Vancouver, 10-15 December 2024, 107984-108011. https://doi.org/10.52202/079017-3429

[10] Shazeer, N. (0219) Fast Transformer Decoding: One Write-Head Is All You Need. arXiv: 1911.02150.

[11] Kang, M., Ting, C., Ting, F.F. and Phan, R.C. (2023) RCS-YOLO: A Fast and High-Accuracy Object Detector for Brain Tumor Detection. In: Greenspan, H., *et al*., Eds., *Medical Image Computing and Computer Assisted Intervention—*

MICCAI 2023, Springer, 600-610. https://doi.org/10.1007/978-3-031-43901-8_57

[12] Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., *et al*. (2024) MobileNetV4: Universal Models for the Mobile Ecosystem. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G., Eds., *Computer Vision—ECCV* 2024, Springer, 78-96. https://doi.org/10.1007/978-3-031-73661-2_5