项目详细方案

基于 FPGA 的铝片表面工业缺陷检测系统

摘 要

缺陷检测通常是指物体表面缺陷、异常的检测，在铝片的工业生产中，由于各方面的影响，会产生诸如折皱、针孔、脏污等异常情况。对于铝合金、钢材等材料，板材与卷材的表面缺陷直接影响到产品最终品质与定价，个别缺陷甚至会影响产品下一阶段的安全可靠性，在工业检测中的缺陷检测一直是非常受重视的环节。

目前硬件市场设备五花八门，且各端对人工智能模型都有不同的要求，本文目标是建立一个基于 FPGA 的铝片表面缺陷检测模型池，实现在高、中、低端 FPGA 设备都能够快速部署与使用相关铝片表面缺陷检测模型。

随着国内半导体产业链的不断完善、芯片设计能力的不断加强，以及工艺的进步以及技术的发展，已经使得国产 FPGA 能够实现一定程度的自我供给。FPGA 以低延时、低成本、高性能等优势成为人工智能模型部署的理想硬件载体。

本文以英特尔 FPGA 中国创新中心提供的数据集作为数据源，提出了适用于高端 FPGA 设备的 Cascade-RCNN-ResNet50-FPN，中端 FPGA 设备的 SSD-MobileNet 以及低端 FPGA 设备的 TOOD-RepVGG-FPN 算法模型，三种模型在设计初期是针对各类硬件场景进行架构的，但考虑到近年来 FPGA 的不断成熟，以及能够加速人工智能模型推理等优良特性，本文针对 FPGA 进行了模型的结构优化，在模型设计之初便避免了使用诸如 deformable convolution、matrix nms 之类的特殊算子，使其能够轻松部署在 FPGA 等各种硬件上，实现基于 FPGA 的铝片表面缺陷检测。

高端 FPGA 设备的主要需求是高精度，本文选用 Cascade RCNN 为基础检测网络，再选用特征提取效果较好的 ResNet50 作为其骨干网络，并且针对铝片表面缺陷尺寸变化较大的问题加入了 FPN，将 ResNet50 输出的具有大量位置信息的浅层特征层与和具有丰富语义信息的深层特征层进行均衡化融合。最后得到的模型体积约为 260M，在实验环境下，其 FPS 为 14.3。而 mAP 和总平均 IoU 高达 68.6%与 0.841，完美实现高端 FPGA 设备的缺陷检测任务。

中端 FPGA 设备的主要需求是模型兼顾检测速度与精度，本文选用 SSD 作为检测网络，MobileNet 为骨干网络。最后得到的模型体积约为 20M，且能够在 FPGA 上完美部署与运行。在实验环境下，推理的 FPS 为 25.9，满足实时检测的同时，mAP 和总平均 IoU 也达到了惊人的 65.6%和 0.831，极为出色的完成中端 FPGA 设备的实时缺陷检测任务。

低端 FPGA 设备的主要需求是模型能够在运算能力较弱、内存较小的条件下执行指定模型。本文选用 TOOD 为基础检测网络，RepVGG 为骨干网络，FPN 解决缺陷多尺度问题。最后得到的模型体积仅 25M，在实验环境下，其 FPS 为 26.1，满足实时检测需求的同时，mAP 和总平均 IoU 达到了 60.9%与 0.722，较好的实现了低端 FPGA 设备的实时缺陷检测任务。

关键词：铝片表面缺陷；FPGA；FPN；实时检测

ABSTRACT

Defect detection usually refers to the detection of surface defects and abnormalities. In the industrial production of aluminum sheet, due to the influence of various aspects, there will be abnormal conditions such as wrinkling, pinhole and dirt. For aluminum alloy, steel and other materials, the surface defects of plate and coil directly affect the final quality and pricing of products, individual defects will even affect the safety and reliability of the next stage of the product, defect detection in industrial testing has been a very important link.

At present, there are various devices in the hardware market, and each end has different requirements for artificial intelligence models. The objective of this paper is to establish a FPGA-based aluminum sheet surface defect detection model pool, so as to achieve rapid deployment and use of relevant aluminum sheet surface defect detection models in high, middle and low-end FPGA devices.

With the continuous improvement of the domestic semiconductor industry chain, the continuous strengthening of chip design capability, as well as the progress of the process and the development of technology, domestic FPGA has been able to achieve a certain degree of self-supply. FPGA has become an ideal hardware carrier for ai model deployment due to its advantages of low latency, low cost and high performance.

Based on the data set provided by Intel FPGA China Innovation Center as the data source, this paper proposes cascade-RCNN-RESNET50-FPN suitable for high-end FPGA devices. Ssd-mobilenet of mid-end FPGA device and TOOD-RepVGG-FPN algorithm model of low-end FPGA device, the three models are designed for various hardware scenarios at the initial stage of design. However, considering the continuous maturity of FPGA in recent years, and the excellent characteristics of accelerating artificial intelligence model reasoning, In this paper, the structure of the model is optimized for FPGA. At the beginning of the model design, special operators such as Deformable Convolution and Matrix NMS are avoided, so that it can be easily deployed on various hardware such as FPGA to realize the detection of aluminum surface defects based on FPGA.

The main requirement of high-end FPGA equipment is high precision. Cascade RCNN is selected as the basic detection network in this paper, and ResNet50 with good feature extraction effect is selected as its backbone network. FPN is added to solve the problem that the size of aluminum surface defect changes greatly. The shallow feature layer with a lot of location information output by ResNet50 and the deep feature layer with rich semantic information are balanced and fused. Finally, the volume of the model is about 260M, and its FPS is 14.3 in the experimental environment. MAP and the total average IoU are as high as 68.6% and 0.841, which perfectly realize the defect detection task of high-end FPGA devices.

The main requirement of mid-end FPGA device is that the model takes into account both detection speed and accuracy. SSD is selected as the detection network and MobileNet as the backbone network in this paper. Finally, the size of the model is about 20M, and it can be perfectly deployed and run on FPGA. In the experimental environment, the FPS of inference is 25.9, meeting the requirements of real-time detection, while mAP and total average IoU are 65.6% and 0.831, excellently completing the real-time defect detection task of mid-range FPGA

equipment.

The main requirement of low-end FPGA devices is that the model can execute the specified model under the condition of weak computing power and small memory. In this paper, TOOD is used as the basic detection network, RepVGG is used as the backbone network, and FPN is used to solve the defect multi-scale problem. Finally, the volume of the model is only 25M, and its FPS is 26.1 in the experimental environment. While meeting the requirements of real-time detection, mAP and total average IoU reach 60.9% and 0.722, which better realize the real-time defect detection task of low-end FPGA devices.

KEY WORDS：Aluminium Surface Defects；FPGA；FPN；Real-Time Detection

目 录

前 言

随着工业化进程的不断推进，我国产业制造的不断转型升级，以及现代计算机技术的飞速发展，现代工业对自动化程度的要求越来越高，工业体系向着智能化的方向发展，人们对工业产品表面质量、精度以及可靠度的要求也越来越高。然而在提高生产效率的同时，产品的表面缺陷问题仍然不可避免。

在缺陷检测领域，传统的人工检测方法不仅效率低下、易误判，而且容易产生疲劳，无法满足稳定的大批量检测需求。随着图像处理与计算机技术的发展，机器视觉检测技术被广泛应用到各种生活场景中。由于铝片表面缺陷在尺寸、形状等方面呈现的复杂性，传统图像处理技术采用的方法，已经难以满足工业检测需求。近年来，随着深度学习技术飞速发展，计算机视觉技术凭着其高精度与稳定的优点，已在目标检测、工业质检等领域被广泛的应用。

本文的出发点是建立一个多硬件设备支持的铝片表面缺陷检测模型池，实现在不同硬件设备条件下都能快速部署与使用相关铝片表面缺陷检测模型。

FPGA 是英文 Field Programmable Gate Array 的缩写，即现场可编程门阵列，它是在 PAL、GAL、EPLD 等可编程器件的基础上进一步发展的产物。它是作为专用集成电路(ASIC)领域中的一种半定制电路而出现的，既解决了定制电路的不足，又克服了原有可编程器件门电路数有限的缺点。

随着国内半导体产业链的不断完善、芯片设计能力的不断加强，以及工艺的进步和技术的发展，已经使得国产 FPGA 能够实现一定程度的自我供给。FPGA 以低延时、低成本、高性能等特性成为人工智能模型部署的理想载体。

本文所有模型均对 FPGA 采取了针对性优化，能够完美部署、运行在高、中、低端 FPGA 上，部分模型能够实现实时检测的要求。除此之外，部分模型还能够部署在服务端、移动端、树莓派等硬件环境。真正实现了多硬件支持的铝片表面缺陷检测模型池。

1 绪论

1.1 研究背景和目的

1.1.1 研究背景

近年来，我国工业化水平开始全方位提高，各行业快速发展。铝片是工业生产过程中非常重要的原材料，其需求量逐年增加。在如此大的需求背景下，铝片生产整体质量并不高，由于多种因素致使铝片表面出现一系列质量问题，如铝片表面出现针孔、擦伤、脏污、折皱等，此类缺陷会严重影响到下游生产线的再加工，降低用户的使用体验。

当前这一过程主要依赖人工进行缺陷检测与质量把控，然而工人们的检测标准很难统一，且成本高、易受环境影响等现实因素。

目前，已有多家企业及高校对铝片表面质量检测方法进入了深入研究。主流的检测方法是基于视觉系统的缺陷检测，不需要接触铝片表面，通过视觉传感器安装于铝片生产线上，搭配专用的检测系统，实现图像采集、处理、分析、归类的功能。但准确率不高、硬件成本高、推理速度慢。

FPGA 具有低延时、低成本、高性能等特性，成为人工智能模型部署的理想载体。若将铝片表面缺陷检测模型部署在 FPGA 上将大大降低经济成本、提高工厂效率，保障产品质量。将缺陷检测模型部署到 FPGA 设备上替代人工进行缺陷检测是不可逆转的时代趋势。

1.1.2 研究目的

对于铝合金、钢材等材料，板材与卷材的表面缺陷直接影响到产品最终品质与定价，个别缺陷甚至会影响下一个阶段产品的安全可靠性，所以对于工业检测中的缺陷检测一直是非常关键的环节。在生产流水线中，板材、卷材的表面缺陷检测目前依然依靠质检员目检完成，工作环境要么是在生产线上，要么以吊装抬升，工人站立于板材下方检测两种方式，目检不仅效率低下，还存在严重的生产安全隐患，亟需利用人工智能技术手段将智能设备部署于生产线上，提高人工智能表面缺陷检测范围，将人工目检作为辅助性检测手段，提高检测效率与质量，保障生产安全。

1.2 国内外研究现状

近些年来，随着深度学习技术的发展，图像分类、目标检测和人脸识别等计算机视觉领域的任务得到了迅速的发展。在 20 世纪 80 年代，卷积神经网络被提出。而在 2012 年，深度卷积神经网络 AlexNet 被首次提出并应用于大型图像分类任务 ImageNet，这标志着图像处理技术的特征提取能力大大提高。深度卷积神经网络通过对原始图像进行卷积与池化运算，从复杂的铝片表面中对缺陷特征抽取出来，并对缺陷特征与表面缺陷检测空间进行映射，有效表达了铝型材表面图像的缺陷信息，有效克服了传统图像处理技术特征提取能力弱。基于上述优势，大量研究者也在尝试将深度学习引入铝片表面缺陷检测。

张磊等利用空域变换对铝型材表面缺陷进行预处理，并对原始图像与预处理图

像进行融合，极大地提高了铝材目标检测精度。李超贤等人采用变形卷积技术提高目标检测网络对铝型材表面缺陷的适应性，提升了铝材表面缺陷检测实时性。Zhang 通过在深度卷积神经网络引入空间注意力机制，提升了对铝型材中部分罕见缺陷的识别精度。

由上述国内外研究现状可知，深度学习理论具有深度特征自提取的特点，能够将铝片表面缺陷特征进行抽象化表达，对提升铝片表面缺陷检测的精度与稳定性具有推动作用。但由于铝片生产环境复杂，在铝片表面缺陷检测过程中仍然存在样本不平衡、噪声干扰以及网络轻量化程度低等突出问题。因此，实现基于 FPGA 的铝片表面缺陷检测具有重要意义。

1.3 主要研究内容

本文主要内容是通过研究铝片表面缺陷检测算法，训练模型，努力提高模型推理准确率。进而可将模型应用于 FPGA 上，实现基于 FPGA 实现的铝片表面缺陷检测。研究内容如下：

（1）通过分析原始数据集，更加深入了解原始数据集中各类缺陷的属性，从而更好的了解数据集的缺陷数量、种类、尺度等信息，并对模型结构的确定有一定的指导意义。

（2）基于深度学习的铝片表面缺陷检测改进算法。针对铝片表面缺陷目标较小且重叠度较高的场景下进行缺陷检测时仍存在漏检的问题，本文使用 FPN 对其进行优化，效果良好。

（3）采用 ResNet50 网络和多尺度特征融合均衡化网络，分别对 Cascade RCNN 算法的骨干网络以及特征融合模块进行改进，提出了改进的 Cascade RCNN 铝片表面缺陷检测算法。

（4）采用 TOOD-RepVGG-FPN、SSD-MobileNet、CascadeRCNN-ResNet50-FPN 三种目标检测算法进行铝片表面缺陷检测，对比分析检测效果，选出特定应用场景最合适的铝片表面缺陷检测算法。

1.4 论文章节结构

本论文共分为十个章节，具体章节架构安排如下：

第 1 章：绪论。主要阐述本课题的研究背景及意义，分析总结铝片表面缺陷检测方法的研究现状，说明本文的研究内容及组织结构。

第 2 章：相关方法与技术。主要对采用深度学习方法进行铝片表面缺陷检测时所涉及的主要理论知识进行了分析。

第 3 章：数据集分析。主要对数据集从各类缺陷尺寸、数量、长宽比等角度进行数据分析。

第 4 章：基于 Cascade RCNN 的缺陷检测改进算法。主要介绍了 Cascade RCNN 算法的改进思路以及采用的各种 Tricks。

第 5 章：基于 SSD 的缺陷检测改进算法。主要介绍了 SSD 算法的改进方法。

第 6 章：基于 TOOD 的缺陷检测改进算法。主要介绍了 TOOD 算法的改进方法

第 7 章:模型实验与结果分析。主要分析了 TOOD-RepVGG-FPN、SSD-MobileNet、CascadeRCNN-ResNet50-FPN 三种目标检测算法结果指标，证明了模型的安全性高，性能卓越等，以及各模型对应部署的 FPGA 类型。

第 8 章：铝片表面缺陷检测系统设计与实现，主要说明了系统的架构方式与运行流程。

第 9 章：介绍了模型在 FPGA 上的部署流程。

第 10 章：总结与展望。总结论文已完成的工作，并分析未来工作的方向。

2 相关方法与技术

本章主要介绍相关深度学习算法模型，包括 One Stage 算法模型和 Two Stage 算法模型，通过对比分析各算法模型的运行流程、优缺点以及运算量，选取本文所研究的模型。

2.1 缺陷检测概述

铝片表面缺陷检测属于图像目标检测问题。目标检测技术是通过运用特定的计算机视觉算法，训练经过预处理和标注好的训练集，检测出目标类别和位置信息的一个过程。基于深度学习的算法具有比传统目标检测更优秀的检测能力，深度学习网络可以提取目标更深层次的特征信息，能更精确的识别图像中目标位置与类别信息，能更好的完成目标检测任务，用于图像缺陷检测相关深度学习算法可分为两大类，One-Stage 算法模型和 Two-Stage 算法模型，包括 SSD、YOLO 等系列算法模型和 Cascade RCNN、Faster RCNN 等系列算法模型。

2.2 TOOD-RepVGG-FPN

TOOD-RepVGG-FPN 即 PP-YOLOE，它是基于 PP-YOLOv2 的卓越的单阶段 Anchor-Free 模型，超越了多种流行的 yolo 模型。PP-YOLOE 是一系列的模型简称，拥有 s/m/l/x 等不同规格的模型，可以通过 width multiplier 和 depth multiplier 配置灵活调节模型，具体模型结构如图 2-1 所示。

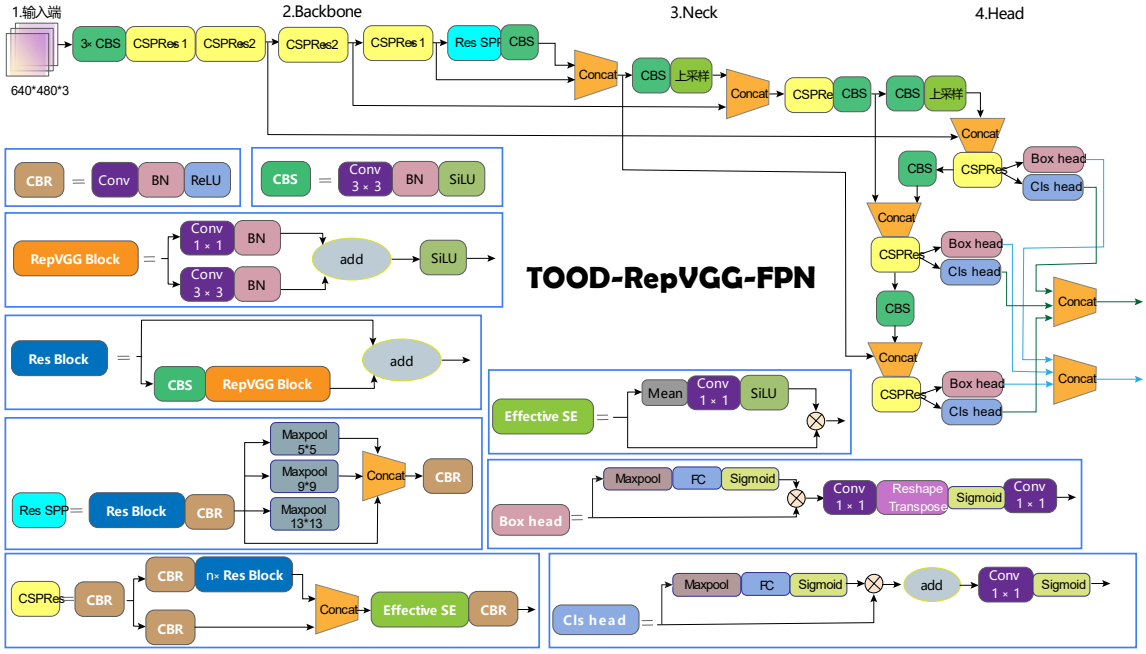


图 2-1 TOOD-RepVGG-FPN 网络设计模型图

由图 2-1 可知，PP-YOLOE 由以下方法组成：可扩展的 backbone 和 neck、Task Alignment Learning、Efficient Task-aligned head with DFL 和 VFL、SiLU 激活函数。

其主要组成部分如下：

（1）骨干网络：PPYOLOE 使用的 Backbone 主要是使用 RepVGG 模块以及 CSP 的模型思想对 ResNet 的改进，同时使用了 SiLU 激活函数、Effitive SE Attention 等模块。为了防止 Backbone 的替换会影响其检测性能，本文用一些可变形的卷积层来替换卷积层，可变卷积网络（DCN）在很多模型中已经得到了很好地效果，并且 DCN 本身不会显著增加模型中参数和运算量，但实际应用中，过多的 DCN 会增加推理时间，为了平衡效率和精度，只在最后阶段用 DCN 替换 3×3 的卷积层，将修改后的主干网络表示为 RepVGG-DCN，输出为 C3、C4、C5。

（2）颈部网络：PP-YOLO 中也使用了 FPN，并在特征图之间添加了横向连接。主干网络的输出特征 C3、C4、C5 作为 FPN 结构的输入，得到对应的输出。

（3）检测网络：PP-YOLO 的检测头很简单，通过一个 3×3 卷积和一个 1×1 卷积来获得最后的输出预测。每个输出的通道数为 3×(K+5)，其中 K 是类别数(本文中 K=4)，3 代表了不同尺寸的 anchors，5 代表预测的坐标偏移量（4 个偏移值和一个置信度）。

2.3 SSD-MobileNet

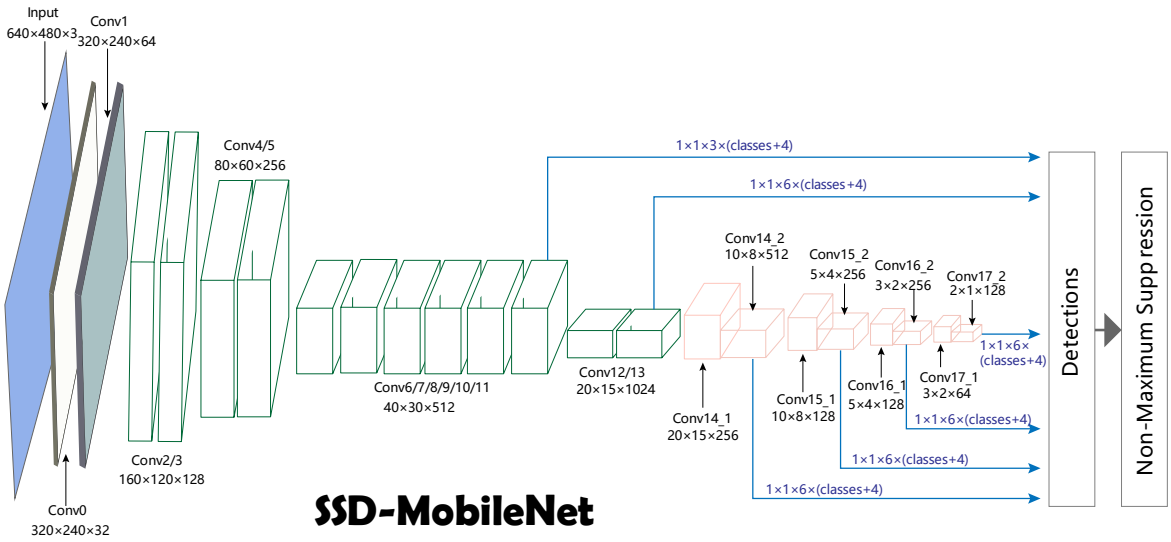


图 2-2 SSD- MobileNet 网络架构图

SSD-MobileNet 的网络结构如图 2-2 所示。

SSD 算法以 VGG16 作为基本特征提取网络，并且添加了几层额外的特征提取层。网络包含 13 层卷积层和 3 层全连接层。然而，强大的特征表示能力，也意味着网络模型更复杂、参数量更多。这样的模型虽然能够提取到更好的视觉特征，但是并不适合部署到移动设备、嵌入式设备、自动驾驶等计算能力受到限制的场景中。为了在特征表示能力和模型复杂度之间找到一个更好的平衡，Google 大脑实验室提出了专门应用于移动和嵌入式设备的 MobileNet 卷积神经网络，网络的核心组成部分是深度可分离卷积（Depthwise separable convolution）如图 2-3 所示。

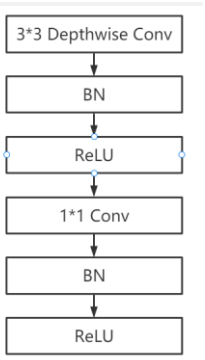


图 2-3 深度可分离卷积

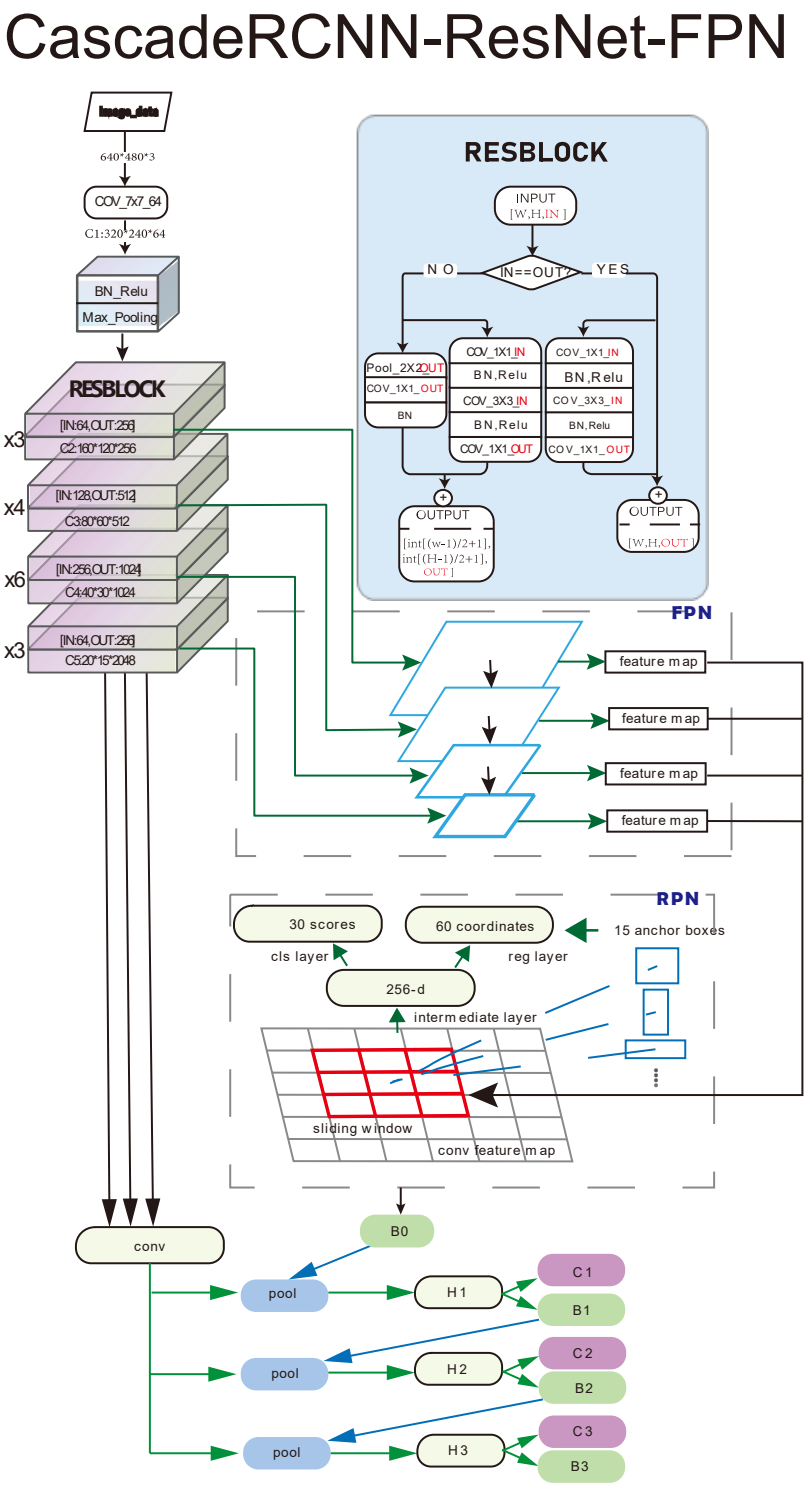
由图可知，深度可分离卷积由逐通道卷积（Depthwise Conv）和逐点卷积（Pointwise Conv）两个卷积层组成。逐通道卷积对输入卷积层的每个通道应用 1 个单一的卷积核，输出和输入的通道数目相等，逐点卷积对逐通道卷积的输出应用 1×1 的卷积核产生新的特征图。当输入卷积特征图的通道数目为 M，输出通道数目为 N 时，标准卷积的参数量为 3×3×M×N，逐通道卷积的参数量为 3×3×1×M，逐点卷积的参数量为 1×1×M×N，则深度可分离卷积与标准卷积参数量之比 β 用公式 2-1 表示为。

$$\beta = \frac{3 \times 3 \times \mathrm{M} + \mathrm{M} \times \mathrm{N}}{3 \times 3 \times \mathrm{M} \times \mathrm{N}} = \frac{1}{9} + \frac{1}{N} \qquad \textbf{公式 2-1}$$

由公式 2-1 可知，随着网络模型输出通道数目的加深，深度可分离卷积的参数量仅占标准卷积参数量的 1/9，因此使用了深度可分离卷积的网络结构在参数量上往往占有优势。

与此同时，MobileNet 由 1 个标准卷积、13 个深度可分离卷积及 1 个平均池化层堆叠而成，深度可分离卷积的应用增加了网络深度的同时，显著降低了 MobileNet 的模型参数量。MobileNet 及其变种不仅在速度上占有优势，在特征提取能力上也能取得和标准卷积神经网络相似的性能。

2.4 Cascade RCNN-ResNet-FPN



图表 2-4 Cascade RCNN-ResNet-FPN

Cascade RCNN-ResNet-FPN 的网络架构如图 2-4 所示。

Cascade RCNN 整体检测流程是通过骨干网络提取特征，再通过 FPN 组合特征图，生成更加易于分类与定位的特征图组，RPN 结构按尺度在对应的特征图上生成

候选框，最终由多个 ROI Pooling 结构再对候选框进行多次框回归。

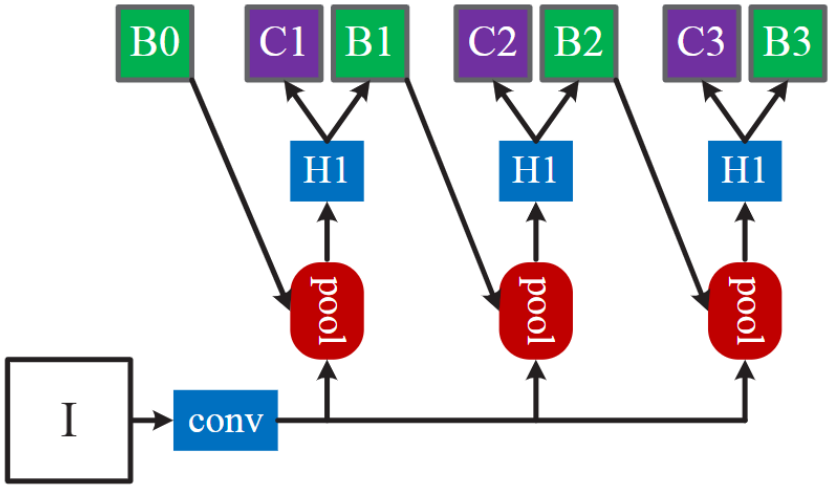


图 2-5 Cascade RCNN 网络结构

通过将 RPN 输出的检测框输入一个框回归模块(如图 2-6 所示)，再将框回归模块的输出作为下一层的框回归模块的输入，如此重复提高 IOU 阈值，将使得预测框的定位效果获得提升。Cascade RCNN 可以提升检测效果，其关键点在于其采用了 4 次框回归。这些框回归模块通过生成d_x、d_y调整矩形框的位置上，生成 d_w、d_h调整矩形框的长宽。但这些调整只具有线性调整能力，一旦生成的矩形框与真实框的差距过大，就很难通过框回归调整为真实框。因此 Cascade RCNN 通过将多个框回归模块级联来增强线性回归的能力。

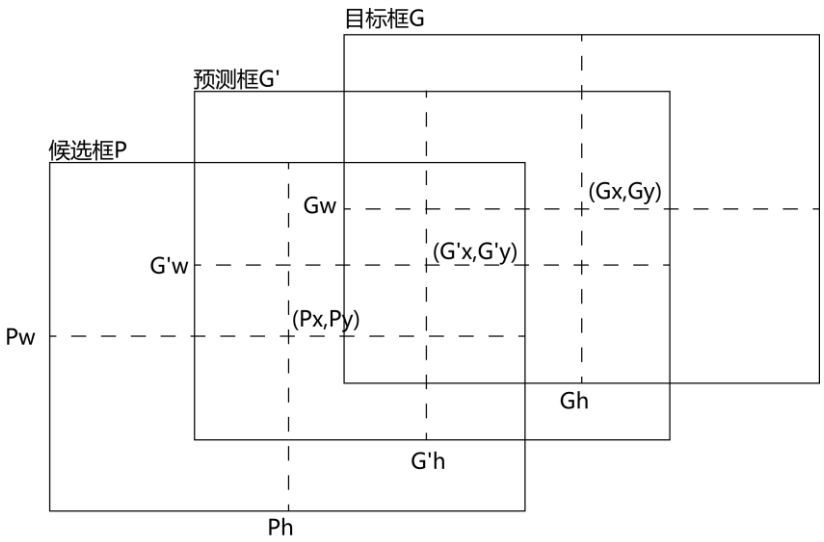


图 2-6 边框回归

骨干网络部分采用 ResNet50，赋予模型强大的特征提取能力。

采用 FPN 结构，使得模型对小尺寸缺陷变得更加敏感。

总而言之，Cascade RCNN 这一检测方法的优点是通过级联多个渐增阈值的框回

归模块使其将原本低质量的检测框可以优化为高质量的检测框，提升了检测的准确率和召回率。

2.5 本章小结

本章主要对采用深度学习技术进行铝片表面缺陷检测时所需的基础理论知识进行了详细的分析，为后文铝片表面缺陷检测算法的研究提供了理论基础，详细分析了 TOOD-RepVGG-FPN、SSD-MobileNet 和 Cascade RCNN-ResNet-FPN 模型架构。

3 数据集分析

3.1 数据集介绍

本文的数据集由英特尔 FPGA 中国创新中心提供，原始数据集中共分为训练集和验证集，其中训练集包含 400 张图片、1062 个缺陷，验证集包含 354 张图片、941 个缺陷，共包括针孔、擦伤、脏污和折皱四种缺陷，具体如图 3-1 所示。

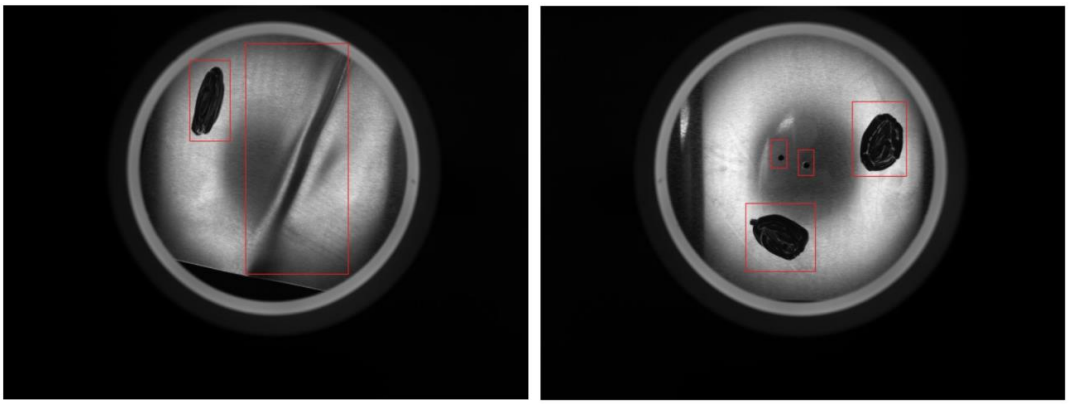


图 3-1 铝片表面缺陷数据集展示

数据集的环境稳定，在之后的图片预处理阶段，不需要使用数据增广策略，但存在数据集较小、目标缺陷特征大小不均衡等问题。

3.2 各类缺陷的数量分析

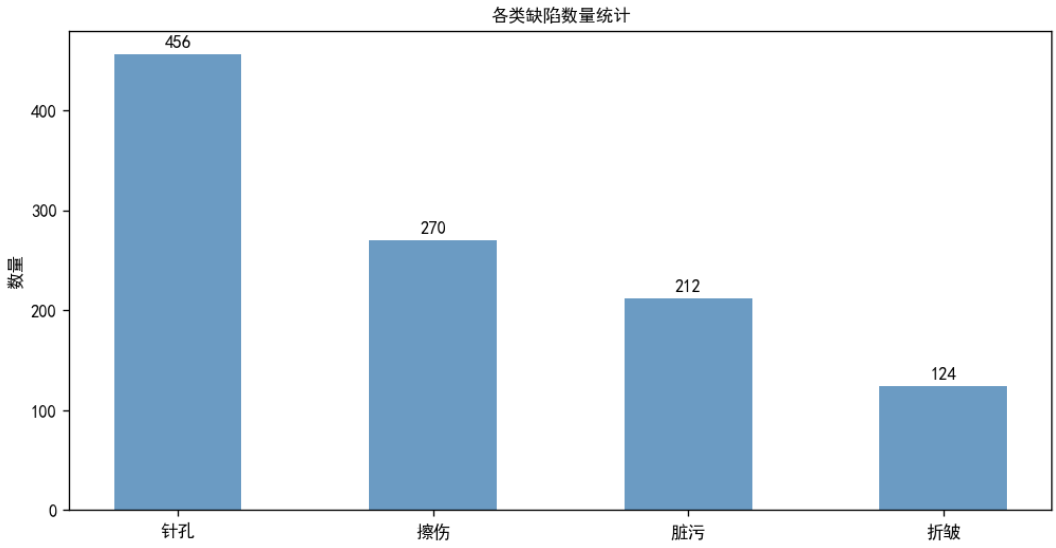


图 3-2 各类瑕疵数量统计

训练集中共出现 1062 个缺陷，各类瑕疵数量统计如图 3-2 所示。从图中可以看出针孔的数量最多，出现折皱的数量最少，各类瑕疵数量不平均。其比例为 3.7: 2.1: 1.7: 1，可以在预处理阶段对擦伤、脏污、折皱增加采样比例，使之训练数据数量均

衡。

3.3 长宽比分析

3.3.1 缺陷长宽比分析

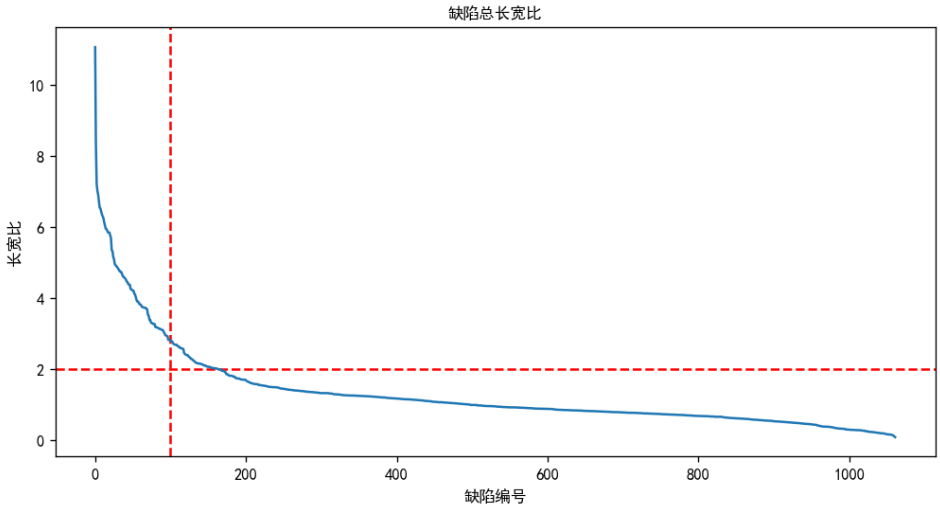


图 3-3 缺陷总长宽比

通过分析数据集各缺陷的总长宽比，从而了解数据集中所有缺陷的总长宽比范围。如图 3-3 所示。长宽比主要在 0~2 之间。

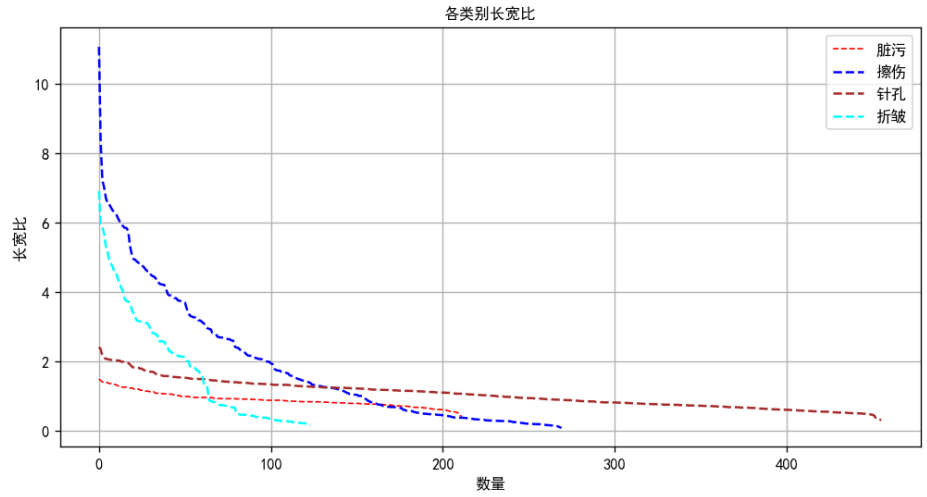


图 3-4 各类别长宽比

分析各类别长宽比，如图 3-4 所示，其针孔与脏污的长宽比较为稳定，约在 1 左右，而擦伤与折皱的长宽比变化较大，主要分布在 1-2 之间。

通过计算得到：

- 缺陷平均大小：擦伤 > 折皱 > 脏污 > 针孔

- 缺陷平均长宽比：针孔: 擦伤: 脏污: 折皱 =0.90: 1.91: 1.06: 1.82

由以上两条结论可在训练过程中确定预选框的比例。

3.3.2 各类缺陷长宽分析

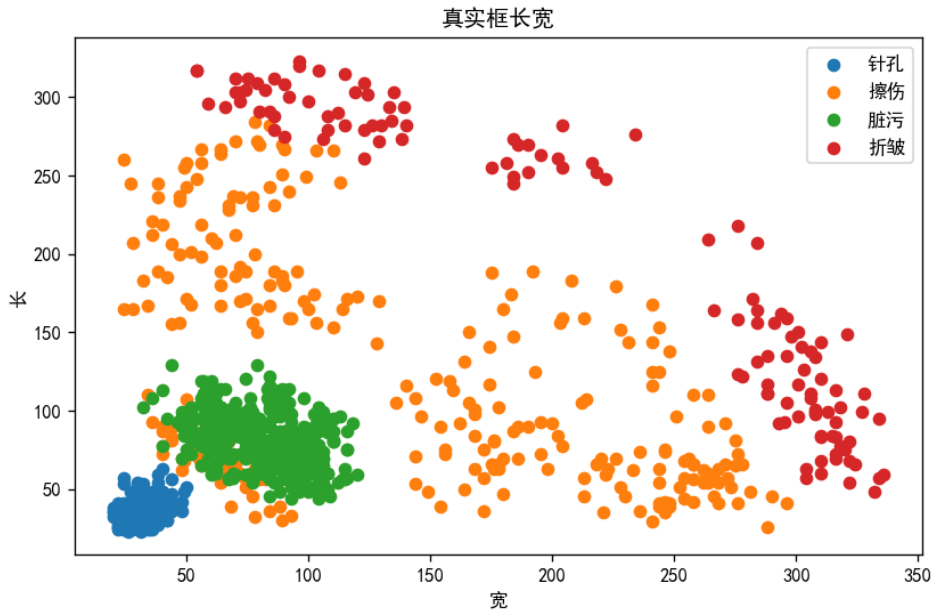


图 3-5 各类缺陷长宽

为了更加深入了解原始数据集中四类缺陷的真实框长宽，通过对各类缺陷长宽进行可视化分析，从而更好地调整各类缺陷的锚框，如图 3-5 所示。

由图可得，针孔与脏污的真实长宽较为稳定，擦伤与折皱长宽不稳定，存在较大波动。

通过计算得到各缺陷平均长宽：

- 针孔:脏污:折皱:擦伤= (31.9,36.1):(80.3,81.0):(218.0,196.6):(140.2,121.6)

通过如上数据，在可在预训练阶段得到各类聚类的长宽大小，确定各预选框大小、长宽比。

3.4 数据分析总结

通过以上分析，本文可以得知数据集中共出现四种缺陷种类，分别为针孔、脏污、折皱及擦伤，各类缺陷差异如表 3-1 所示。：

表 3-1 各类缺陷差异

	针孔	脏污	折皱	擦伤
数量比	3.7	2.1	1.7	1

平均大小	最小	小	大	大
平均值	（31.9,36.1）	（80.3,81）	（218,196.6）	（140.2,121.6）
平均长宽比	0.9	1.91	1.06	1.82

通过统计分析以上数据，对本文模型结构选择具有一定的指导意义，具体如下：

（1） 本文的主要评判指标为模型精度，因此预测网络优先选择精度较高的网络；

（2） 数据集的环境稳定，不需要使用数据增广策略；

（3） 缺陷种类少，分类、预测难度不大，骨干网络可选择较为简单的网络；

（4） 缺陷数量比有差异，需要调整对折皱、脏污和擦伤的训练比例；

（5） 得到各缺陷的大小、比例等数据，可调整模型训练时锚框的大小与比例。

（6） 缺陷种类平均大小有差异，需要采用 FPN 策略。

3.5 本章小结

本章主要对数据集进行了详细的分析，对各类缺陷的数量、长宽、长宽比方向进行分析，并提出了一系列结论，对后续模型的结构有一定的指导意义。

4 基于 Cascade RCNN 的缺陷检测改进算法

本文针对 Cascade RCNN 算法仍存在对小目标和重叠度高的缺陷识别率低的问题，提出了一种基于 Cascade RCNN 的铝片改进算法，并完美部署与运行在高端 FPGA 设备上。通过采用 ResNet50 替换原骨干网络作为 Cascade RCNN 算法的特征提取网络，以及提出基于 FPN 网络改进的多尺度特征融合均衡化网络，提高铝片表面缺陷检测的精度。

4.1 改进的 Cascade RCNN 检测算法

针对 Cascade RCNN 算法在进行缺陷检测时仍存在对小目标缺陷漏检的问题，提出了一种改进的 Cascade RCNN 算法。改进方式主要有：将特征提取网络 VGG16 替换为分类精度更高的 ResNet50，提取更多的缺陷特征信息；借鉴 FPN 网络的思想，提出了多尺度特征融合均衡化网络，均衡化融合浅层特征的位置信息和深层特征的语义信息，提高小目标的识别率。改进后的 Cascade RCNN 网络结构如图 4-1 所示。

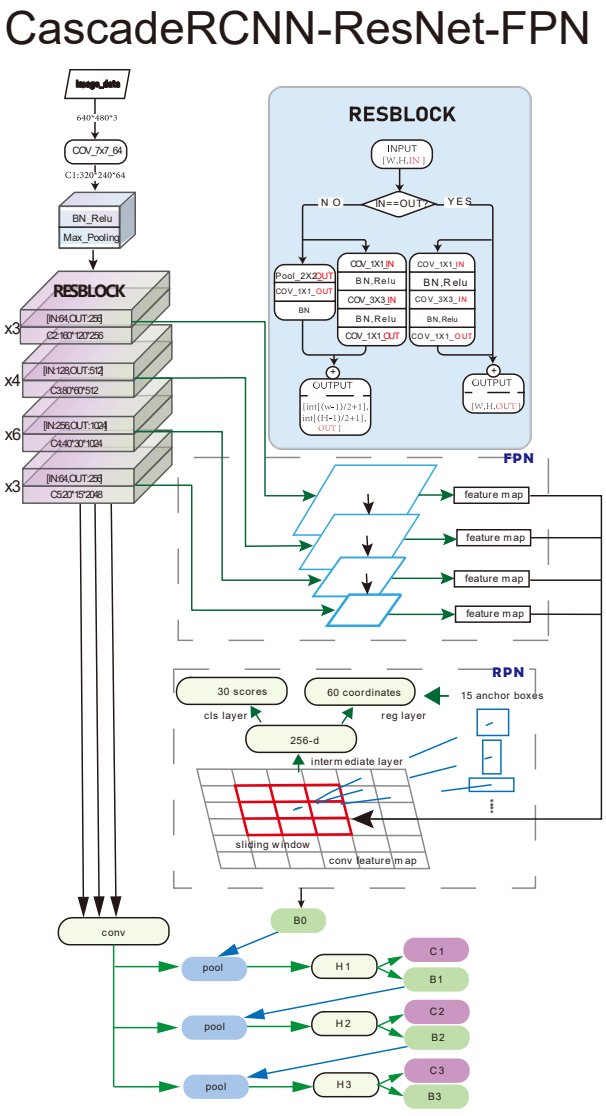


图 4-1 改进后 Cascade RCNN 网络结构

4.1.1 ResNet50 模块

ResNet50 是经典的运用残差单元结构组成的网络，是在 2015 年由何凯明等人共同提出，旨在解决因卷积神经网络不断加深而可能导致的梯度消失和网络退化的问题。其网络层次如图 4-2 所示。ResNet50 包括一个步长为 2 的 7×7 卷积层、四个残差单元级联卷积层和一个全连接层。其中残差单元是由一个 3×3 的卷积核、两个 1×1 的卷积核和一个残差连接分支组成，其结构如图 4-3 所示。其中：x_1 为输入，其期望输出为 h(x)，但加入一个恒等映射 x_1 后原始学习特征变成 $h(x)=(f(x)+x_1)$，因此残差是指 $f(x)=h(x)-x_1$，使得网络拟合更容易，直接映射的加入也使得网络的下一层一定比前一层的图像信息多。ResNet50 不仅计算量比 VGG16 更少，而且能提取更多铝片特征信息。

Layer name	Output size	50-Layer
Conv1	112×112	7×7, 64, stride=2
Conv2_x	56×56	3×3 max pool, stride=2
		$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Conv3_x	28×28	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
Conv4_x	14×14	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$
Conv5_x	7×7	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	Average pool, 1000-d fc, softmax
FLOPs		3.8×10^9

图 4-2 ResNet50 网络结构

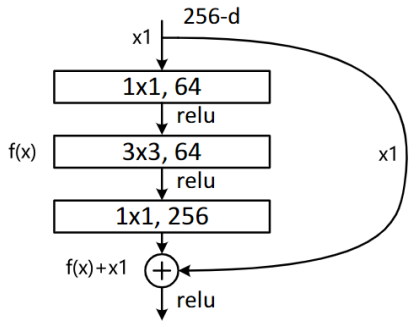


图 4-3 ResNet50 网络的残差单元结构图

4.1.2 多尺度特征融合均衡化网络

当前，FPN 网络是目标检测算法中最常用的特征融合网络，具体实现方式如图 4-4 所示。FPN 是在特征提取网络生成的多尺度特征图的基础上，将最后一层特征图做 2 倍上采样后与前一层特征图相加生成融合后的特征图，然后将新生成的特征图以同样的方式进行特征融合，逐层重复以上操作，形成特征金字塔。但是特征金字塔网络长距离信息流动会导致缺陷信息的流失和更多的关注相邻分辨率，这种融合方式会产生缺陷特征信息不平衡问题，所以本文基于 FPN 提出多尺度特征融合均衡化网络，利用相同深度融合的平衡语义特征来增强多层次的特征。首先将通过 FPN 网络形成的每层特征图通过上采样操作变成相同尺度，然后进行像素级的聚合均值，再将聚合均值化的特征图通过与之前相反的采样操作形成不同尺度的特征图，从而得到缺陷特征信息更平衡的多尺度特征图，实现缺陷特征的多层复用和融合均衡化，提高小目标缺陷的敏感度。

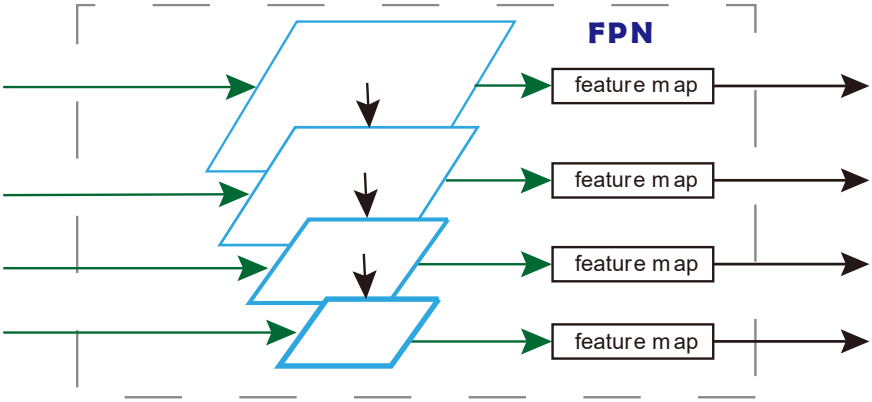


图 4-4 多尺度特征融合均衡化网络结构

4.1.3 Cascade RCNN-ResNet50-FPN 组合

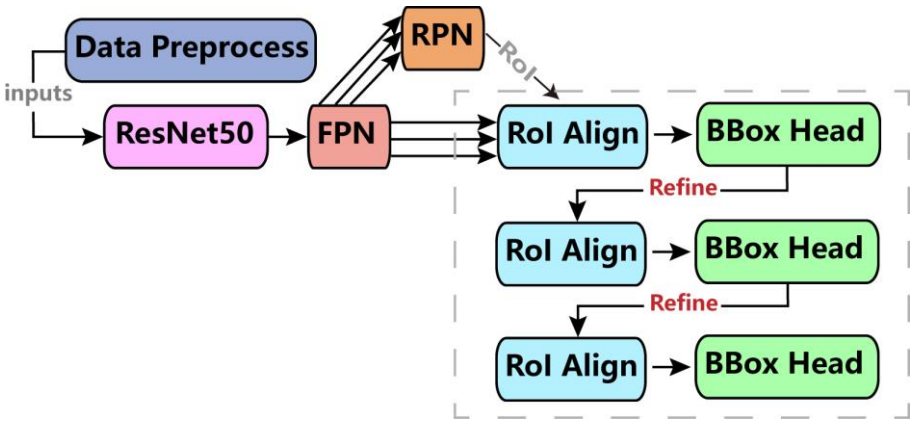


图 4-5 Cascade RCNN-ResNet50-FPN 结构

ResNet50 的最初作用是用来做分类网络的，要使其运用在目标检测中充当骨干网络，只需要将最后一层的全连接层删除，并将 4 个大块的特征图输入 FPN 中，最后从 RPN 中将预测框与 FPN 得到的特征图传入 Cascade RCNN 即可完成铝片表面缺陷检测网络的搭建。

4.1.4 采用的 Tricks

(其他两个网络采用的 Tricks 相同，此后不再赘述)

事实上，对于 Cascade RCNN-ResNet50-FPN 的优化运用了很多 tricks，由于相对简单，本文只将采用的 tricks 与介绍列为下表：

- RandomResize：对图片进行 resize 用不同的制度训练可以提高鲁棒性；
- RandomFlip：图片执行水平翻转；
- NormalizeImage：归一化；
- Batch_transforms：因为需要生成不同框，大小不一，因此要跟随不同规格的图片填充 stride；
- Shuffle：打乱图片；
- Drop_last：抛弃一些数据；

4.2 缺陷检测实验步骤

(其他两个网络的训练步骤类似，此后不再赘述)

缺陷检测实验中，改进的 Cascade RCNN 缺陷检测算法从训练到测试获取检测结果的具体实现步骤如下：

（1）将训练集作为网络的输入，先通过骨干网络 ResNet50 进行特征提取，将 ResNet50 的四个残差单元级联卷积层的输出作为多尺度特征图输出；

（2）将多尺度特征图通过多尺度特征融合均衡化网络进行特征融合，输出信息更丰富且均衡的多尺度特征图；

（3）将多尺度特征图输入至共享的 RPN 网络，产生存在缺陷的预测框；

（4）把产生的缺陷建议区域与 ResNet50 网络最终输出的特征图相结合，经过池化运算后得到各类缺陷的候选区域的特征图；

（5）将输出特征图送入 Cascade RCNN Head 后得到各类缺陷的分类信息和位置信息，通过非极大值抑制后选出最佳的预测框；

（6）重复以上步骤直至模型收敛，则训练结束；

（7）提取出缺陷检测模型的权重参数；

（8）将测试集通过缺陷检测权重模型进行检测，得到各类缺陷的检测结果和定位结果；

4.3 本章小结

本章主要介绍了改进后的 Cascade RCNN 检测算法，并将其应用在高端 FPGA 设备上进行铝片表面工业缺陷检测。首先，对 Cascade RCNN 算法的两个改进部分的改进目的及改进方式进行了详细的分析；然后，介绍了该模型训练到测试获取检测结果的具体实现步骤与相关 Tricks。

5 基于 SSD 的缺陷检测改进算法

为了提升目标检测的 SSD 模型的检测准确率及检测速度，满足其在中端硬件资源配置条件下也能达到较高的检测速度要求，本文根据基本 SSD 模型，结合轻量级的深层神经网络 MobileNet 的思想，构建一种结合特征金字塔的多尺度卷积神经网络结构。结合反卷积操作、特征提取和区域映射、正负样本处理等方法改善模型对中小型占比目标的检测效果。完整结构如图 5-1 所示。

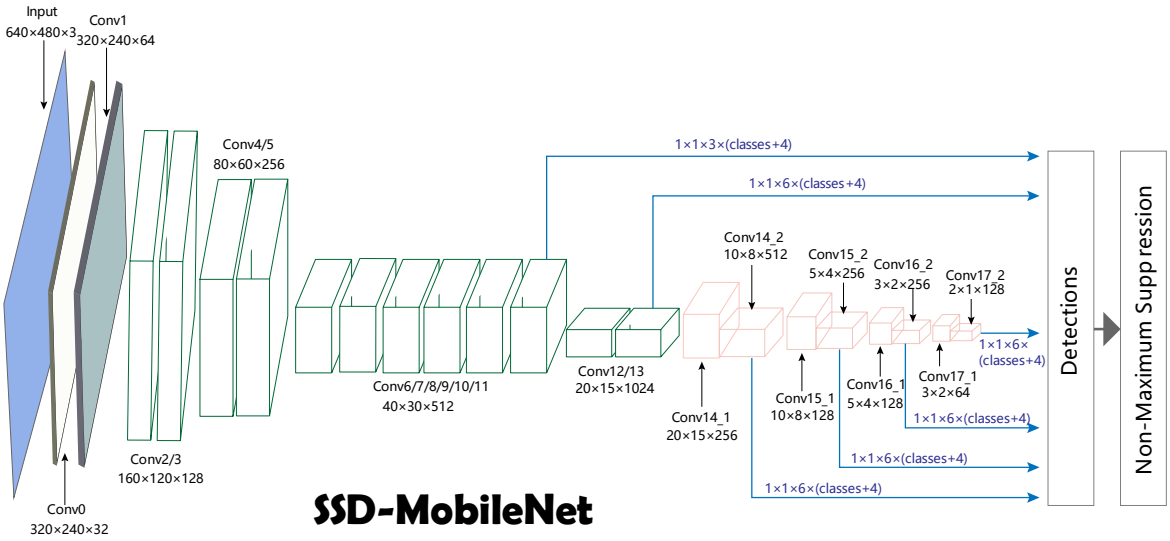


图 5-1 SSD-MobileNet 结构

5.1 改进的 SSD 检测算法

SSD 模型中结合特征金字塔的多尺度卷积神经网络的思想主要体现在获取不同卷积层、不同尺度的特征图数据来进行目标检测。而在 SSD 模型中特征的提取采用的是逐层提取并抽象化的思想，低层的特征主要对应于占比较小的目标，高层的特征主要对应于占比较大的目标的抽象化的信息。即如果待检测的目标在图中占比越小，特征图在经过层层卷积、池化的操作之后，在高层特征层可能出现信息丢失的情况，以致于检测不到占比较小的目标。

针对基本的 SSD 模型对于图像多目标识别的缺陷，无法很好地检测中小型目标这一问题，本文采用改进的多尺度卷积神经网络结构进行目标检测，对 SSD 模型中的低特征层和高特征层采用不同的改进策略提高模型的检测能力；同时融合 MobileNet 的基本思想，提高网络结构的检测速度，提高模型的准确性、实时性和鲁棒性。

为了提高 SSD 模型对于小型目标的检测能力，对低特征层采取特征提取、反卷积操作保留特征图中更多的特征信息，进而对于小型目标的特征区域，通过特征映射在保留有更多小型目标特征信息的特征图上进行特征提取，对于提高模型的小型占比目标的检测能力是十分有必要的。

低特征层包含有更多的细粒度的特征信息，保留更多细节的特征信息可以更加准确地检测目标。反卷积操作扩大了卷积运算之后的特征图的大小，同时也保留了更

多的细节特征信息，提高了模型的特征表达能力。但是既然保留了更多的特征信息，那么在运算时必然会使得运算时间增加，降低检测速度，影响一些模型整体的检测效果。其基本思想如图 5-2 所示。

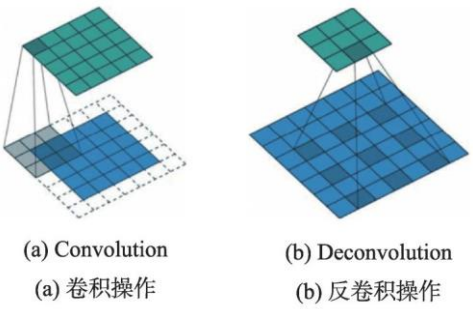


图 5-2 卷积和反卷积示意图

SSD 模型中主要结构有多个池化层，因此本文通过设置扩张率的参数，通过反卷积操作将特征区域放大，保留更多的小型占比目标的特征信息，便于提升对小型目标的检测能力。由于 SSD 模型是对于 640×480 的输入图像进行输出的区域建议，因此本文将反卷积操作之后的图像增大至 640×480，一方面保证了充足的特征信息，另一方面也可以获得更加准确的目标区域建议。

在进行反卷积操作之后，对获取到的特征图进行特征区域映射。从而根据卷积核的大小或图像与特征图大小的比例关系建立起输入图像的目标区域与特征图的某一点的对应关系，将用于进行目标检测的特征图的每个位置映射到原图像中相应的位置，并在对应的位置生成不同比例大小的预测框。

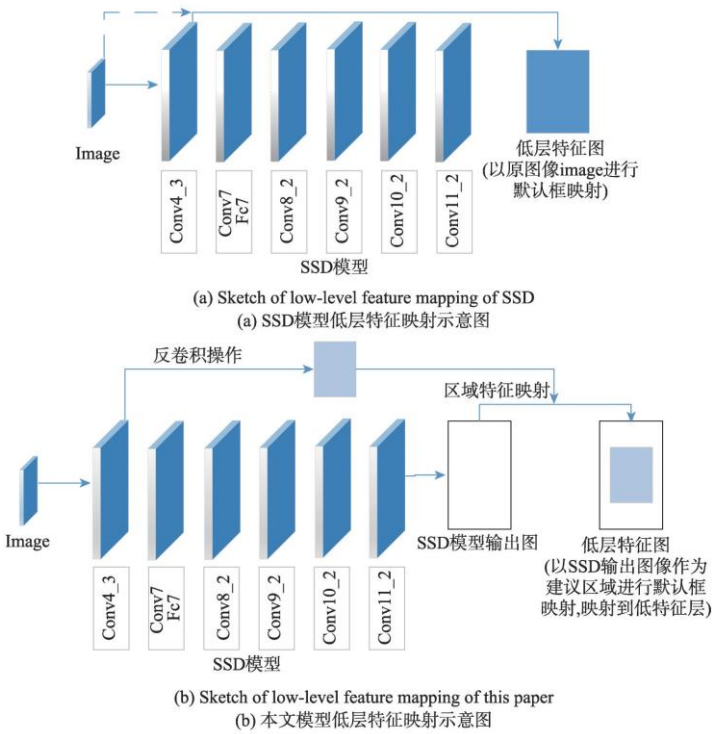


图 5-3 低层特征映射示意图

基本的 SSD 模型在预测框映射时是默认针对整个图像数据进行的，并且对低层特征层的特征图进行区域映射是针对输入图像进行的。但是这里低层特征层在执行反卷积操作后，是将 SSD 模型输出后的图像作为特征图数据进行区域映射的模板。具体思想如图 5-3 所示。

在本文模型中，通过将模型输出图像作为低层特征经反卷积操作之后的区域映射的模板，默认框映射到原图像的对应位置。

MobileNet 的网络结构是基于深度可分离卷积块的堆叠设计。通过权衡延迟时间和精度要求，基于宽度因子和分辨率因子构建合适规模、合适速度的 MobileNet 结构。其网络结构的基本思想是将通道间的相关性和空间相关性完全分离出来，同时大大减少计算量和参数量。

由于 SSD 模型是一种不需要产生候选区域，直接产生物体的类别概率和位置坐标，经过单次检测即可得到最终的检测结果。而 MobileNet 是使用这种算法的具体网络结构，用于进行特征提取。这使得二者可以结合。通过 SSD 模型融合了 MobileNet 的网络思想，结合了二者的优势，保留原有 SSD 模型的网络结构，使用 3×3 的卷积核进行特征处理，保证了模型的准确率。在此基础上，本文模型将原本的大量冗余的参数变成小型参数结构，减少网络计算量的同时，降低了对于硬件资源的消耗，有利于加快模型的收敛速度，改善模型的基本性能。

通过该 SSD-MobileNet 网络结构，有效减少了模型的计算量和参数量。

5.2 本章小结

本章主要介绍了改进后的 SSD 检测算法。首先，取 SSD Head 作为检测网络，其骨干网络选择参数量、运算量较小的轻量级网络 MobileNet。将两个结构进行组合，

为了提升目标检测的 SSD 模型的检测准确率及检测速度，满足其在较低的硬件资源配置条件下也能达到较高的检测速度要求，本文根据基本 SSD 模型，结合轻量级的深层神经网络 MobileNe 的思想，构建一种结合特征金字塔的多尺度卷积神经网络结构。结合反卷积操作、特征提取和区域映射、正负样本处理等方法改善模型对中小型占比目标的检测效果。并使其完美部署与运行在中端 FPGA 设备上。

6 基于 TOOD 的缺陷检测改进算法

TOOD 是 Chengjian Feng 等人在 2021 年 8 月提出的专用于检测网络，其具有精度高、参数量小、可使分类与预测任务进行对齐的优良特性。本文使用 TOOD Head 作为预测网络，骨干网络使用 RepVGG，并使用 FPN 解决铝片表面缺陷尺度不均衡问题。本文避免使用诸如 deformable convolution 或者 matrix nms 之类的特殊算子，以使其能轻松地部署在多种多样的低端硬件上，完整结构如图 6-1 所示。

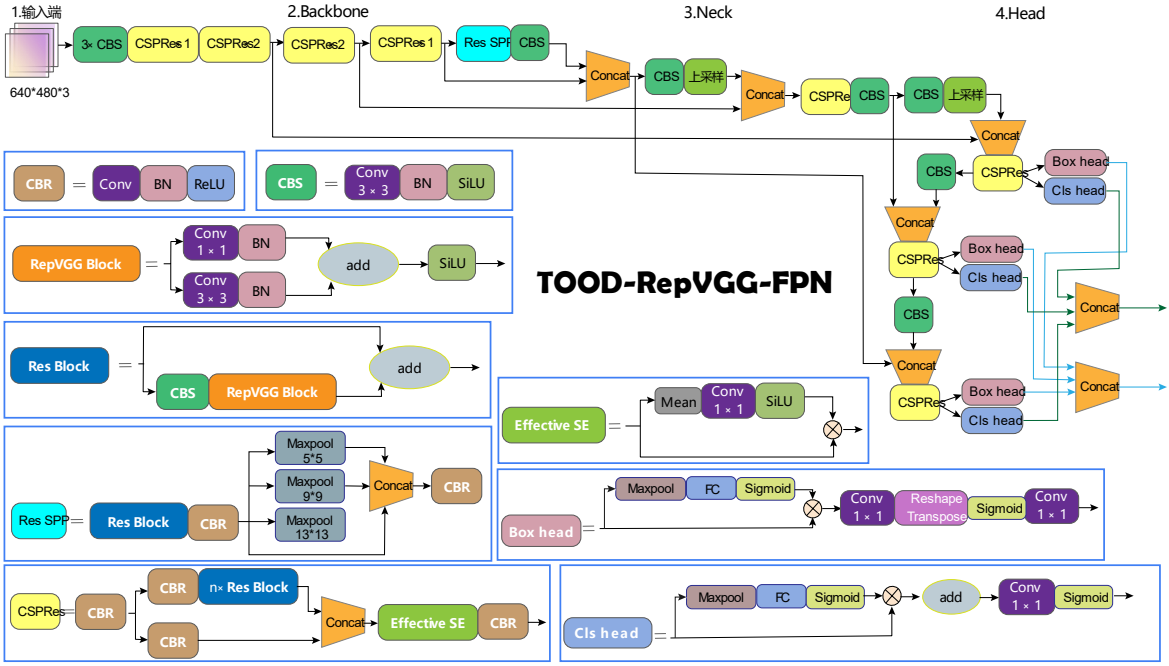


图 6-1 TOOD-RepVGG-FPN 结构

6.1 改进的 TOOD 检测算法

本网络的骨干网络采用了极简模型 VGG 的变种 RepVGG，其是在 VGG 的基础上进行了改进，主要思想包括:

1. 在 VGG 网络的 Block 块中加入了 Identity 和残差分支，相当于把 ResNet 网络中的精华应用到 VGG 网络中；
2. 模型推理阶段，通过 Op 融合策略将所有的网络层都转换为 3×3 卷积，便于网络的部署和加速。

图 6-2 模型展示了推理阶段的重参数化过程，其实就是一个 OP 融合和 OP 替换的过程。图 A 从结构化的角度展示了整个重参数化流程，图 B 从模型参数的角度展示了整个重参数化流程。

RepVGG 的优势如下：

1. 3×3 卷积非常快。在 GPU 上，3×3 卷积的计算密度可达 1×1 和 5×5 卷积的 4 倍。

2. 并行度高。同样的计算量，“大而整”的运算效率远超“小而碎”的运算。

3. RepVGG 主体部分只有一种算子：3×3 卷积接 ReLU。在设计 FPGA 芯片时，给定芯片尺寸或造价可以集成海量的 3×3 卷积+ReLU 计算单元来达到很高的效率。

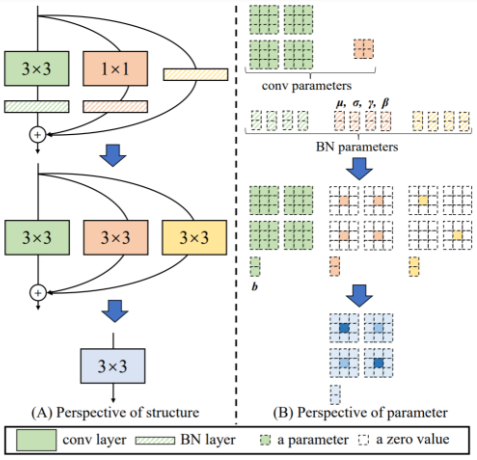


图 6-2 RepVGG 过程图

基于 TOOD 模型改进的颈部网络依然为 FPN 结构，这里不做赘述。

最后的预测网络为 TOOD 的 head，简称为 T-Head，主要是包括了 Cls Head 和 Loc Head，T-Head 简图如图 6-3，具体来说，T-head 首先在 FPN 特征基础上进行分类与定位预测；然后 TAL 基于所提任务对齐测度计算任务对齐信息；最后 T-head 根据从 TAL 传回的信息自动调整分类概率与定位预测。

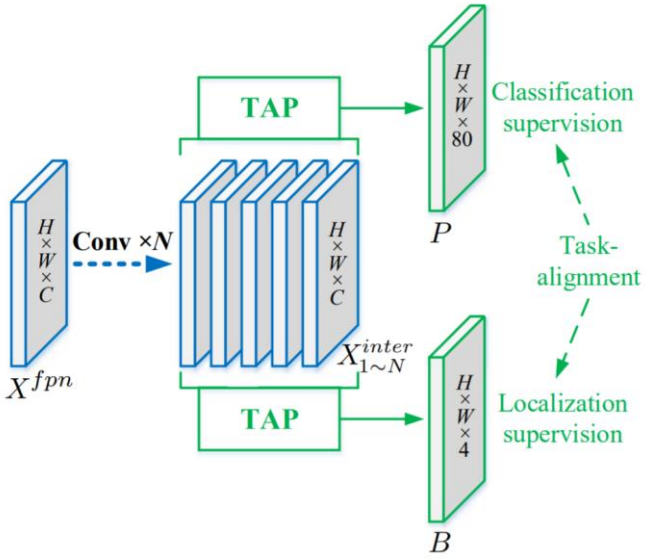


图 6-3 TOOD-Head 结构简图

6.2 本章小结

本文通过避免使用诸如 deformable convolution 或者 matrix nms 之类的特殊算子，骨干网络采用较为简单的 RepVGG 网络，使用 FPN 解决小尺寸缺陷的问题，实现了部署方便、运算参数小、检测速度快的功能，使其能轻松地部署在低端 FPGA 硬件上。

7 模型实验与结果分析

7.1 实验环境配置

本文相关环境的具体配置方案如表 7-1 所示。

表 7-1 嵌入式设备环境

环境	CPU	内核模块/GPU	操作系统
FPGA 环境	ARMv7 Processor	nnadrv	Ubuntu 4.9.78
检测系统环境	Intel Core i7-7700K 32G	GTX 1080Ti 11G	Ubuntu 22.04
模型训练环境	Intel Gold 6271C 32G	Tesla V100 32G	Ubuntu 16.04.6

7.2 评价指标 mAP

缺陷检测的目标是预测框与真实框的距离尽量重合，同时所有的真实框都应该被预测出来。但想二者均取得满意的结果是比较困难的，如果需要提高检测框与真实框的 IOU，那么就需要设置一个较高的 IOU 阈值，将一些 IOU 较小的检测框舍去，这一过程可能导致一些缺陷未被检测到。同样如果希望尽可能的多检测到缺陷，设置的 IOU 阈值则较低，其必然会导致检测框的质量下降。这其中前者被称为准确率，后者被称为召回率。反映了算法的准确性与可靠性。因此衡量一个目标检测算法需要其可以同时计算准确率与召回率。目标检测的评价标准采用 mAP 来衡量高准确率与高召回率。计算 mAP，先要计算构成 PR 曲线的准确率与召回率。

一条 PR 曲线对应一个 IoU 阈值。本文选择阈值为 0.5 对样本进行划分，IoU 阈值大于 0.5 的就认为是正例，小于 0.5 的就认为是负例，从而计算相应的精准率和召回率，具体定义如表 7-所示。

表 7-2 测试结果混淆矩阵

	与真实框 IOU 大于 0.5	**与真实框 IOU 小于 0.5**
预测框置信度大于阈值	TP	FP
预测框置信度小于阈值	FN	TN

准确度（Precision）通过真正的正例样本除以预测为正例样本得到，其定义如公式 7-1 所示。

$$Precision = \frac{TP}{TP + FP} \qquad \textbf{公式 7-1}$$

TP 为检测为正样本的正样本检测框，FP 为检测为正样本的负样本检测框。

召回率（Recall）通过真正的正例样本除以正确预测的样本，其定义如公式 7-1 所示。

$$Recall = \frac{TP}{TP + FN} \quad \text{公式 7-2}$$

TP 为检测为正样本的检测框，FN 为检测为负样本的检测框。

随后计算 PR 曲线，将所有生成的检测框按照置信度从大到小进行排列，随后从第一个置信度最高的样本开始计算以它为阈值的准确率与召回率，这时规定大于等于阈值的检测框为正例，小于阈值的检测框为负例，并以此作为标准计算这时的准确率与召回率，对每一个生成的检测框都进行这一操作。这时将以上计算的所有召回率为横轴坐标，对应的所有准确率为纵轴坐标，即可得 PR 曲线。

随后计算 AP(average precision)即平均精准度。先对 PR 曲线进行平滑处理即对 PR 曲线上的每个点，Precision 的值取该点右侧最大的 Precision 的值。再对其进行积分， 积分公式定义如下：

$$AP = \int_0^1 p(r)dr \quad \text{公式 7-3}$$

对所有类别计算各自的 AP，再取其所有类别 AP 的均值即为 mAP。mAP 可以很好的衡量一个目标检测算法的可靠性与检测性能。

7.3 实验结果及分析

为了验证三种模型的性能与拟合程度，本文对其 PR 曲线、mAP、FPS、总平均 IoU 以及训练的 Epochs 进行比较。

7.3.1 实验数据介绍

原始数据集中共分为训练集和验证集，其中训练集包含 400 张图片、1062 个缺陷数量，验证集包含 354 张图片、941 个缺陷数量，共包括针孔、擦伤、脏污和折皱四种缺陷。

7.3.2 Epoch-mAP 分析

此步骤的主要目的是确定模型是否拟合完成，即查看 Epoch-mAP 图，随着训练的 Epoch 增加，mAP 是否逐渐平缓，若逐渐平缓，则证明模型接近训练完成。具体如图 7-1 所示。

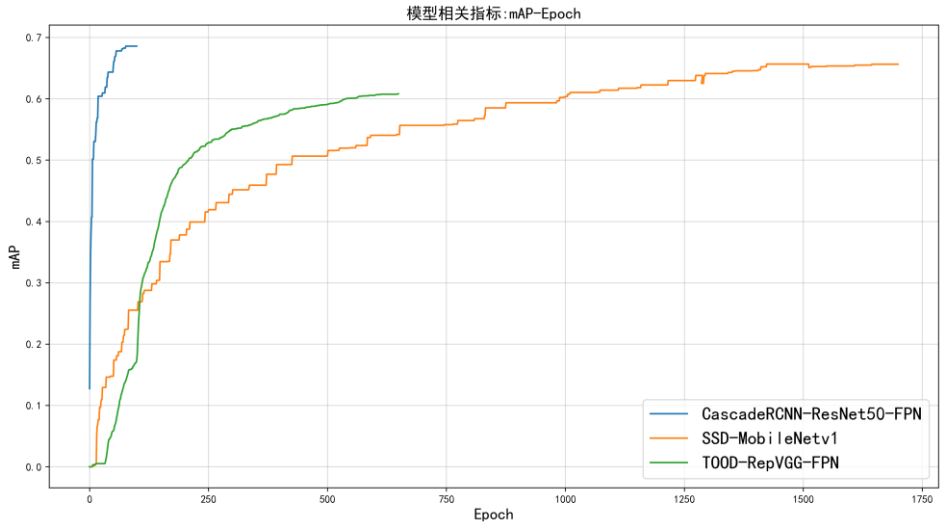


图 7-1 各模型训练过程分析

通过监测模型在整个训练过程中 mAP 的变化情况，可以直观地反映训练效果，如图 7-1 所示。其中，水平轴代表训练过程中的 Epoch，垂直轴代表 mAP 值。从图中可以看出：

(a) Cascade RCNN-ResNet50-FPN 在训练到 100 epochs 时 mAP 趋于平缓，最终 mAP 为 68.6%；

(b) SSD-MobileNet 在训练到 1700 epochs 时 mAP 趋于平缓，最终 mAP 为 65.6%。

(c) TOOD-RepVGG-FPN 在训练到 650 epochs 时 mAP 趋于平缓，最终 mAP 为 60.8%；

各模型训练结果的 mAP 均高于 60%，证明其预测效果优异；

7.3.3 各模型 PR 曲线分析

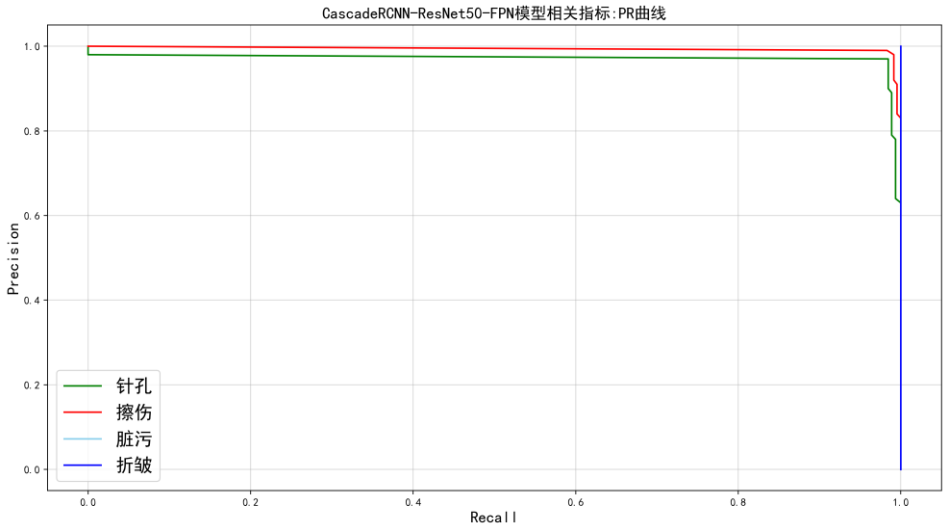


图 7-2 CascadeRCNN-ResNet50-FPN 模型 PR 曲线

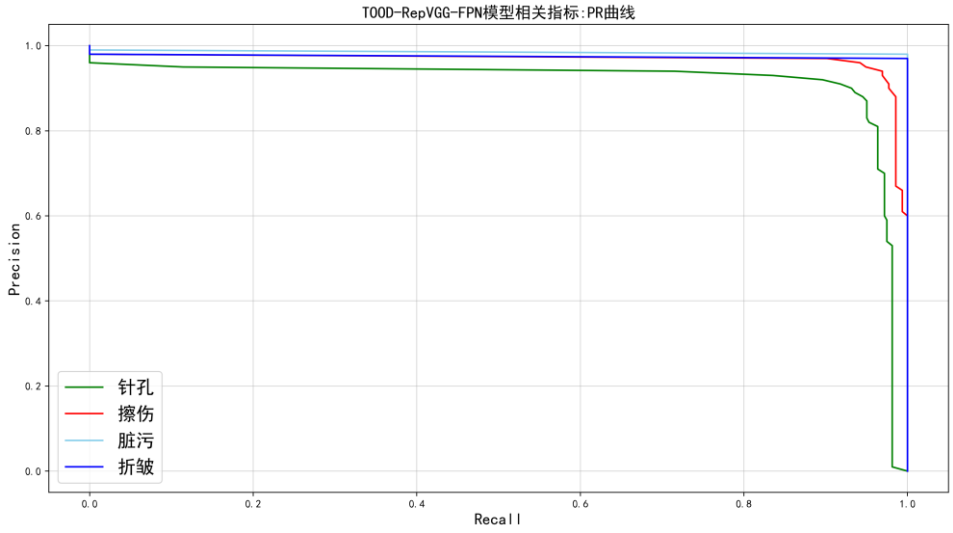


图 7-3 TOOD-RepVGG-FPN 模型 PR 曲线

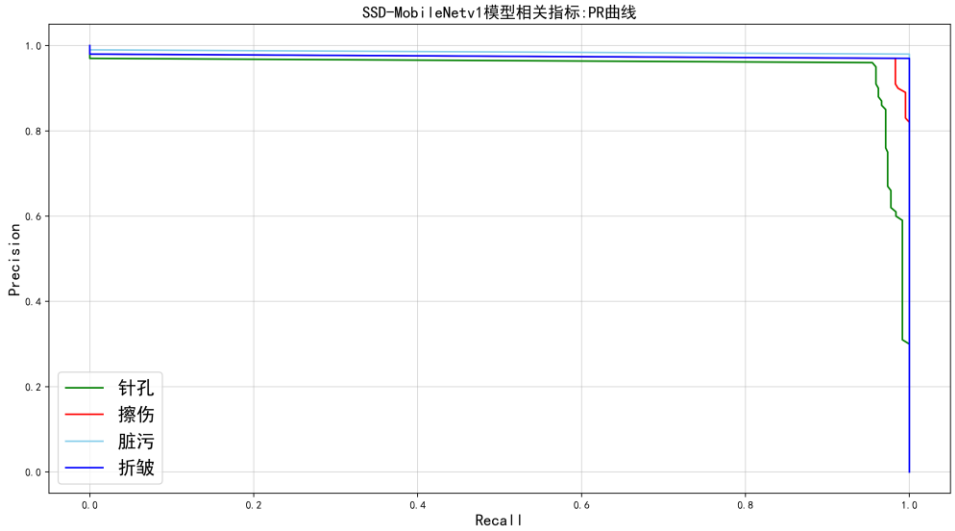


图 7-4 SSD-MobileNetv1 模型 PR 曲线

各模型的 PR 曲线如图 7-2、图 7-3 和图 7-4 所示，从 PR 曲线来看，通过对比分析数据集中的四种缺陷，无论是 CascadeRCNN-ResNet50-FPN、TOOD-RepVGG-FPN 或 SSD-MobileNet，他们各自的准确率（P）和召回率（R）效果都非常好。结合 PR 曲线可证明模型结果可靠，安全性高。

7.3.4 各模型 IoU 对比

IoU 是目标检测任务中进行预测框分类时常用的指标，也可用作边界框回归的损失函数，其计算公式如公式 7-4 所示。

$$IoU = \frac{|A \cap B|}{|A \cup B|} \qquad \textbf{公式 7-4}$$

表 7-3 为各模型不同缺陷种类 IoU 对比表，除 TOOD-RepVGG-FPN 总 IoU 为

0.722 外，其余二者总 IoU 都高于 0.83，此时模型预测框与真实框基本重合，证明其卓越的检测能力。

表 7-3 各模型不同缺陷种类 IOU 对比表

模型类名	总 IoU	针孔	擦伤	脏污	折皱
CascadeRCNN	0.841	0.793	0.840	0.857	0.866
SSD	0.831	0.776	0.830	0.852	0.859
TOOD	0.722	0.678	0.670	0.804	0.768

从表 7-3 的各模型不同缺陷种类 IoU 可知，相对于折皱与脏污，针孔与擦伤更难被模型检测到。这可能与缺陷的大小与特征不明显有关。

7.3.5 各模型 FPS-mAP-体积对比

表 7-4 各模型 FPS-mAP-模型体积对比表

模型类名	mAP	FPS	模型体积
Cascade RCNN	68.6%	14.3	264M
SSD	65.6%	25.9	25M
TOOD	60.9%	26.1	20M

表 7-3 为各模型 FPS-mAP-IoU 对比图，图 7-5 为其三维图。能够更好的观察各模型相关性能。

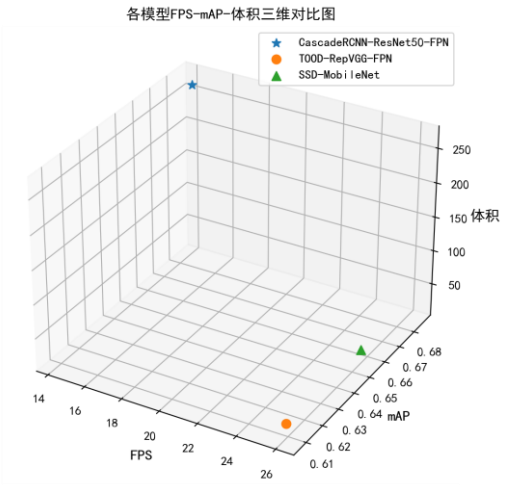


图 7-5 模型 FPS-mAP-IOU 类比图

由图 7-3 与表 7-4 可得如下结论：

(a) Cascade RCNN-ResNet50-FPN 在 FPS 为 14.3 的情况下，mAP 高达 68.6%，检测效果异常优异。但模型体积较大，适合部署在有高精度要求的高端 FPGA 设备上。

(b) SSD-MobileNet 的 FPS 为 25.9，满足实时检测的条件，其 mAP 相较于 CasRCNN-ResNet50-FPN 差别不大，但 FPS 提升了约 2 倍。非常适合部署在中端 FPGA 设备上。

(c) TOOD-RepVGG-FPN 的 FPS 高达 26.1，满足实时监测条件，而 mAP 为 60.9%，模型体积为 25M，适合部署在有实时检测需求的低端 FPGA 设备上。

7.4 本章小结

本章通过分析模型的各种指标得到了很多结论。

(a) 通过 Epoch-mAP 曲线可确定三种模型均已完成训练。

(b) 通过分析各模型 PR 曲线可证明各类检测模型的可靠性与稳定性。

(c) 通过分析各模型 IoU 值可证明模型卓越的检测能力。

通过如上三步分析，本文证明了三种模型的可靠性，稳定性与模型卓越的检测能力。

通过分析各模型 FPS-mAP-体积可选出特定场景与通用场景的最佳模型。

(A) 通用场景选用 SSD-MobileNet，其能够实时检测、精度较高且硬件成本低。

(B) 高端 FPGA 选用 Cascade RCNN-ResNet50-FPN，其检测速度理想，精度最高。

(C) 中端 FPGA 选用 SSD-MobileNet，其运算量较小，能够实时检测的同时精度较高。

(D) 低端 FPGA 选用 TOOD-RepVGG-FPN，其检测速度最快，精度理想。

8 铝片表面工业缺陷检测平台设计与实现

8.1 系统总体方案设计

本系统平台分为两部分。一部分是前台演示主页网站，主要功能在于将深度学习模型训练结果进行可视化展示以及项目概况介绍。其主要功能模块分为三部分：项目介绍模块、文件上传模块、可视化展示模块。另一部分则为后台管理系统模块，主要功能为管理检测记录和与 FPGA 通信并调用指定模型进行缺陷检测作业。具体软件平台设计架构图如图 8-1 所示。

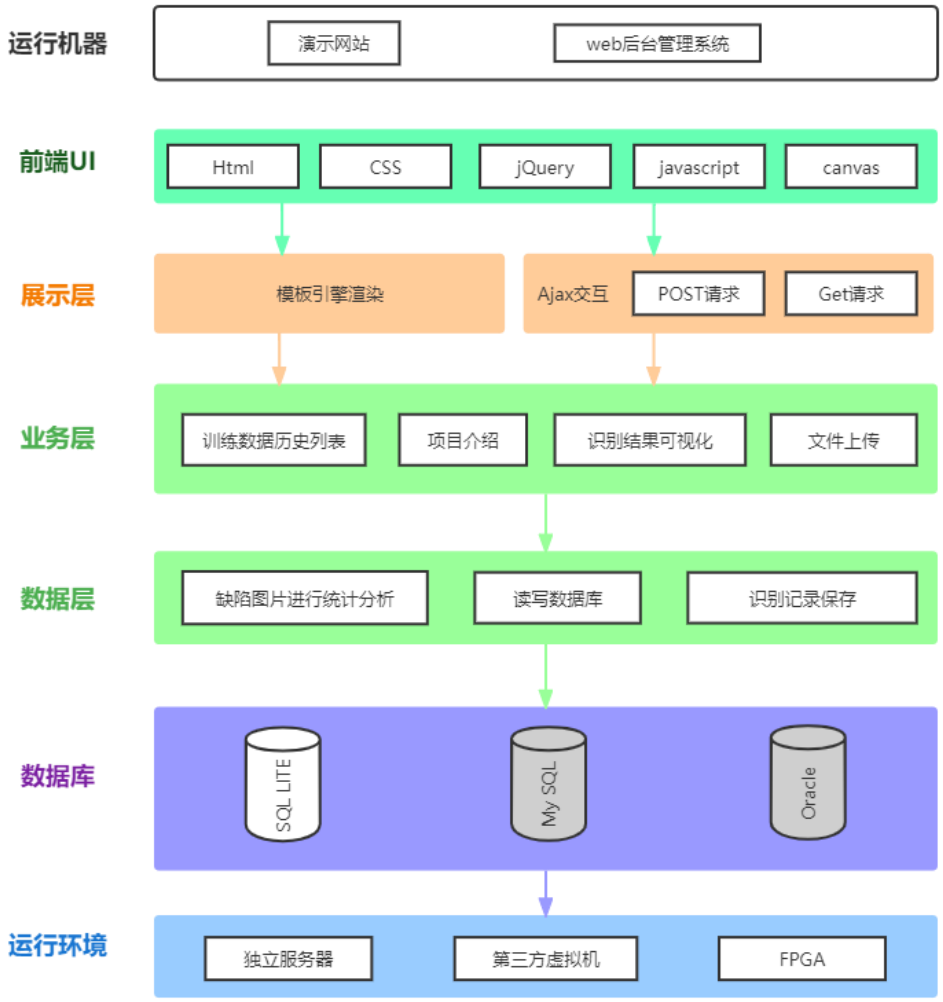


图 8-1 软件平台设计架构图

8.2 系统工作流程介绍

缺陷检测系统的流程如下，流程图如图 8-2 所示：

(a) 后台启动服务，监听指定端口，等待前端发送请求；

(b) 前端被调用，发送图片与指定模型名至后台；

(c) 后台与 FPGA 设备进行通讯，发送图片与运行指定模型进行缺陷预测；

(d) FPGA 设备完成预测，返回 JSON 格式的预测结果至后台；

(e) 后台接收预测结果，保存至数据库并转发至前端；

(f) 前端取得指定模型检测结果，并将结果渲染到前端页面。

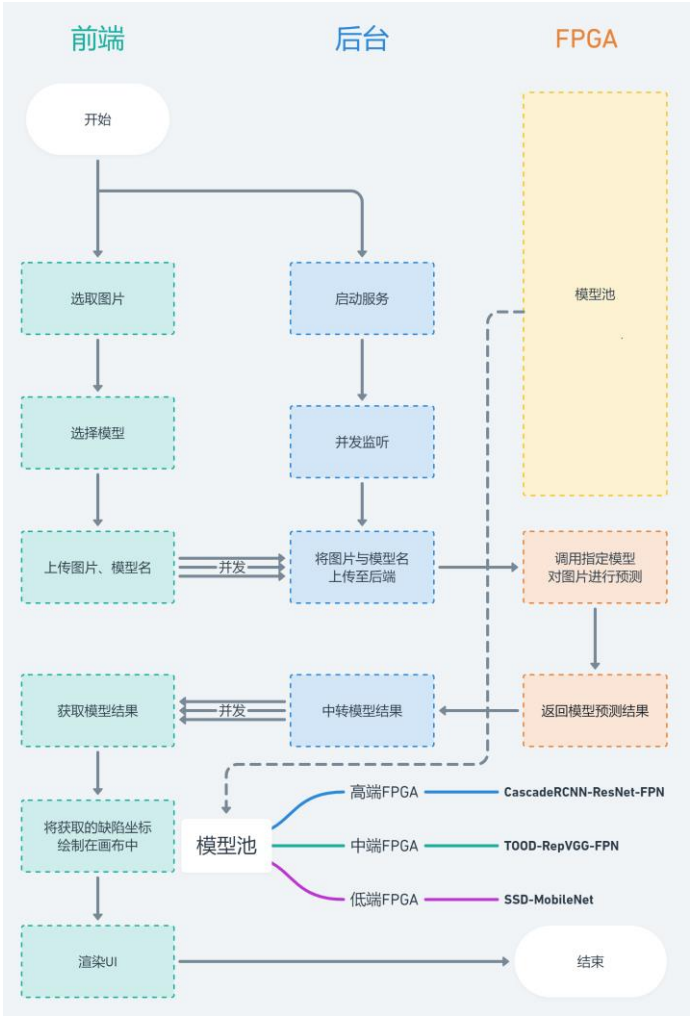


图 8-2 检测平台工作流程图

8.3 整体功能模块设计

检测平台主要分为五大功能模块，下面介绍每个功能模块的实现过程。

（1）项目介绍模块主要功能是将本系统进行基本介绍，利用 HTML, CSS 进行样式排版。系统介绍模块分为项目简介，模型介绍。利用 HTML5 网页视频在线播放演示视频，在线 PDF 查看技术在线展示项目文档。使得用户可以直接通过本网站了解本系统基本用途与功能。部分页面如图 8-3 所示。

图 8-3 检测平台上传模块

（2）文件上传模块主要功能是上传铝片缺陷图片到服务器，进行缺陷检测。本模块主要使用 Ajax 异步网络传输，文件 FromData 格式,利用 JS 内置 Blob 方法将图片格式转换为 blob 二进制进行前后端通信进行文件上传，与此同时提交指定模型名，告知后台调用指定模型进行缺陷检测作业。如图 8-4 所示。

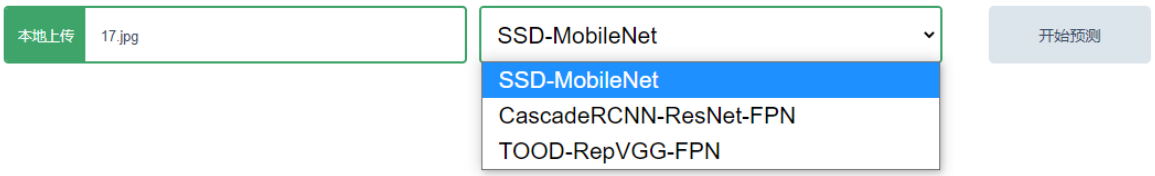


图 8-4 检测平台上传模块

（3）预测功能可视化模块主要功能是预测缺陷位置，可视化模块主要功能是利用模型训练后返回的缺陷相关数据，包括缺陷名称、坐标等，将缺陷标识在前台画板上，添加鼠标交互与动画。将训练结果可视化，可互动化，使得用户可以更加直观的了解缺陷情况。其主要使用 canvas 画板技术将识别结果坐标绘制成框，标记在前台图片中，并且利用 JS，CSS3 Animation 进行鼠标移动交互动画。最终编译出具有交互功能的缺陷识别结果可视化画板，其效果如图 8-5 所示。

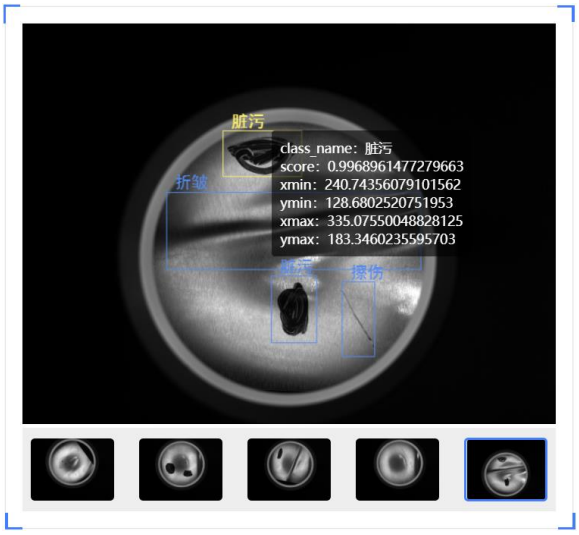


图 8-5 识别功能可视化

（4）检测历史记录模块主要功能是使用 GET 请求接口 http://127.0.0.1:8000/detect?size=10&page=1 时（size 控制每次需要的数据量，默认为 10。Page 为页面数，默认为 1），会得到 JSON 格式的相关检测历史记录，其渲染后效果如图所示如图 8-6 所示。

图 8-6 检测历史记录模块

（5）图片检测模块主要功能是使用 POST 请求发送表单（图片与模型名）请求接口 http://127.0.0.1:8000/detect 时，得到 JSON 格式的相关检测结果。其 JSON 结构如图 8-7 所示。

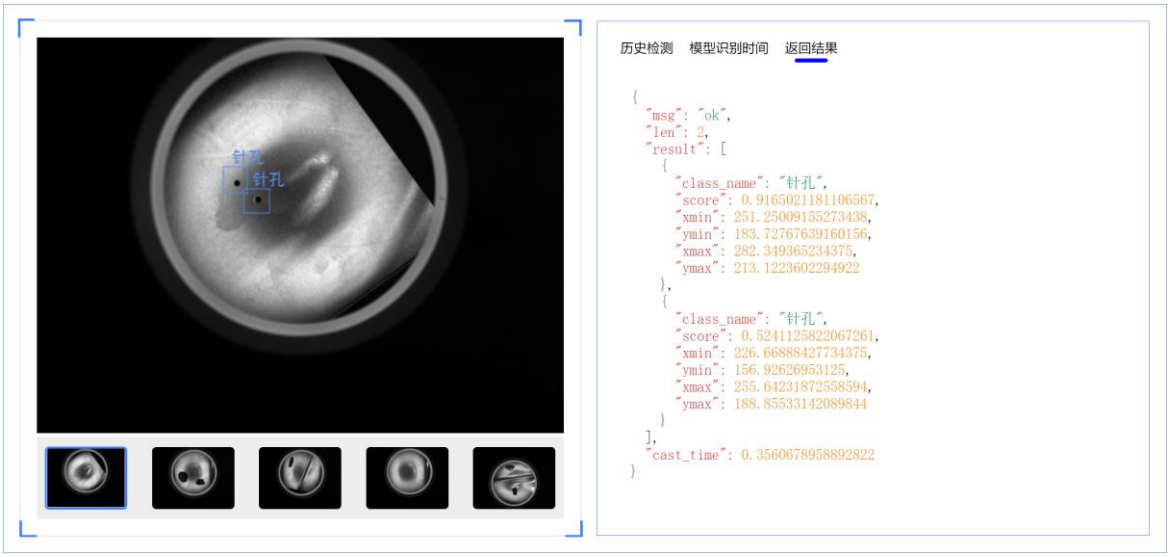


图 8-7 图片检测模块

8.4 本章小结

本章主要介绍了铝片表面缺陷检测平台的展示页面、铝片缺陷检测功能流程，以及前端渲染结果。通过本系统上传待预测的图片与指定模型名，可实现快速调用指定模型、得到并在前端渲染预测的缺陷位置。完美做到了设计之初的所有功能。

9 模型在 FPGA 上的部署

9.1 FPGA 部署的优势

(1)延迟低

在推理部署的过程中，FPGA 的行为具有确定性，FPGA 可作为硬件加速器，没有时间片、线程或者资源冲突等问题，与 CPU 相比，单个 FPGA 的并行度高。CPU 更适合完成串行任务，而 FPGA 并行可以始终以完全相同的速度执行一件事情。同时 FPGA 适合并行的进行矩阵运算，而神经网络中很多卷积运算本质是矩阵运算。因此若有低延迟要求，FPGA 是人工智能模型的理想载体。

(2)可编程性

FPGA 可通过编程重组电路，直接生成专用电路的方式，专注于某个特定运算(如信号处理，图像处理)当中的计算架构做出相应调整，并进行优化。比如过程中需要大量的乘法运算，可以增加更多的乘法器。而对于 CPU 比较，一旦设计完成便无法更改，只能改造模型从而提升速度。基于 FPGA 灵活改变逻辑的特性，也可反复利用，节约成本。

9.2 FPGA 的部署与工作流程

FPGA 的部署与工作流程如图 9-1 所示：

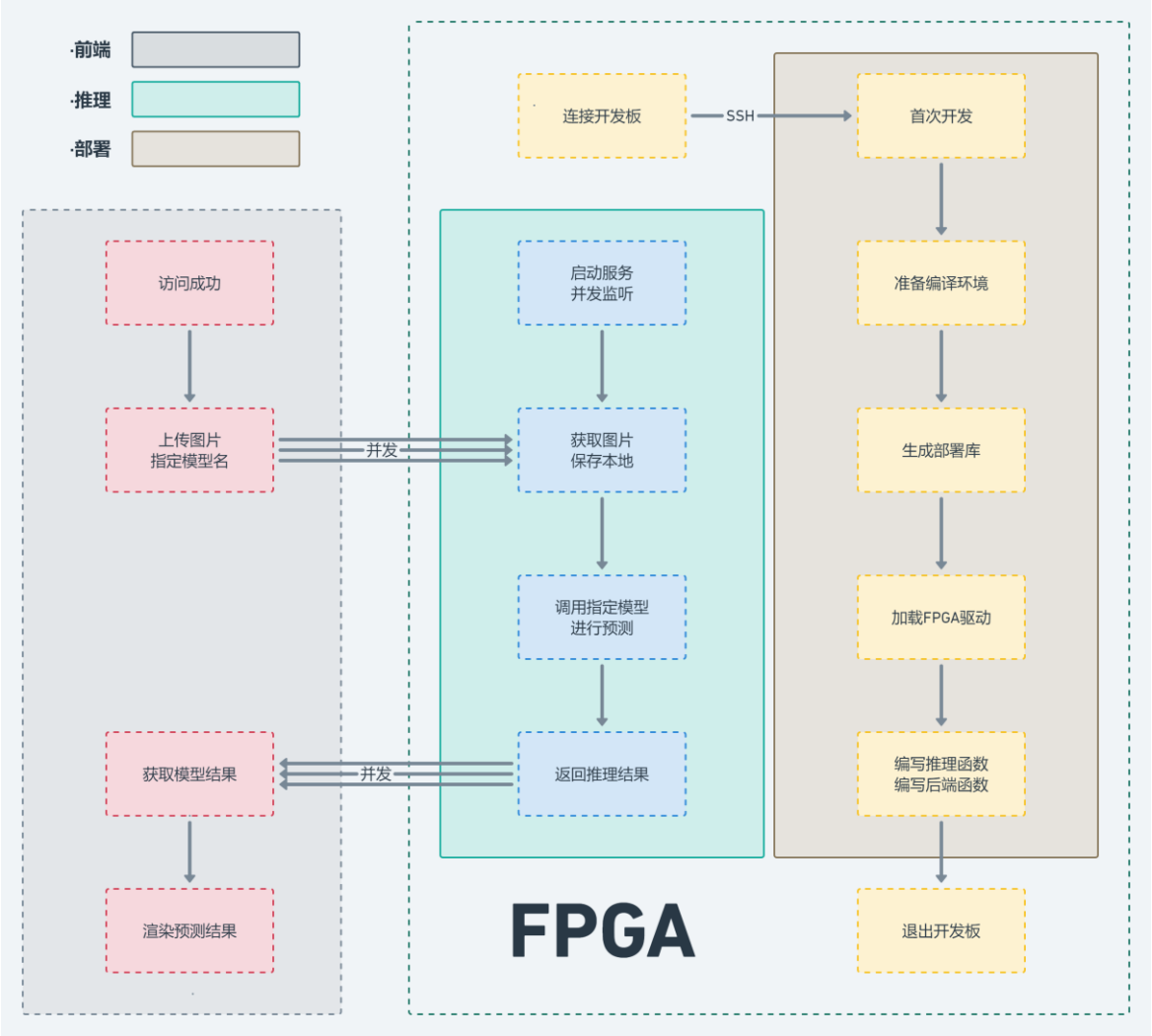


图 9-1 模型在 FPGA 的部署与工作流程

FPGA 的部署与工作流程如下：

（1）开启 VPN，通过 SSH 连接 FPGA；

（2）首次部署：

- 准备编译所需要的环境、如 gcc、cmake、git 等；

- 下载 PaddleLite 源码以及 FPGA 的 SDK，编译生成 PaddleLite 与 Intel FPGA 部署库，添加生成的库文件；

- 加载 FPGA 驱动，开启 FPGA 推理加速；

（3）开启推理：

- 针对需要部署的模型，编写推理程序，以及编写后台程序，开启 HTTP 服务，实现与前端进行交互；

- 启动服务，等待前端发送 POST 请求，传递图片并保存本地；

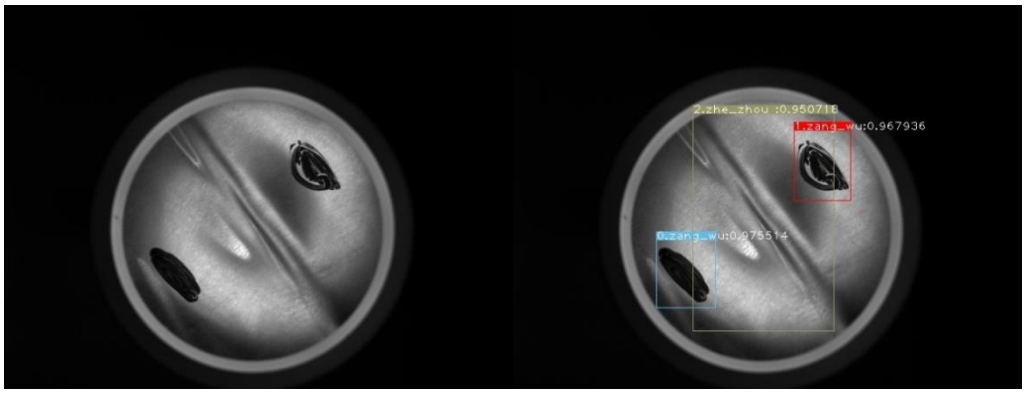
- 调用部署好的模型，开启推理，对图片进行预测；
- 返回 JSON 格式预测数据到后台；
- 后台作为中转站，保存并转发给前端进行预测框渲染；

9.3 模型运行结果

以下是 FPGA 预测返回的结果：（模型为 SSD-MobileNet）

```
root@awcloud:/opt/plite-vi#
root@awcloud:/opt/plite-vi# ./plite-test
iter 0 cost: 1542.493042 ms
iter 1 cost: 1444.418945 ms
iter 2 cost: 1041.850952 ms
iter 3 cost: 1540.166016 ms
iter 4 cost: 1391.167969 ms
warmup: 1 repeat: 5, average: 1392.019385 ms, max: 1542.493042 ms, min: 1041.850952 ms
results: 3
[0] zang_wu - 0.975514 0.281945,0.610528,0.399696,0.789927
[1] zang_wu - 0.967936 0.552598,0.326316,0.664143,0.510868
[2] zhe_zhou  - 0.950718 0.354622,0.282596,0.631859,0.850660
Preprocess time: 28.083000 ms
Prediction time: 1392.019385 ms
Postprocess time: 0.833000 ms
```

预测结果前后对比：



10 总结与展望

10.1 本文总结

本文针对 FPGA 的高、中、低端设备对人工智能模型的主要需求不同，设计了能够部署在不同 FPGA 设备上的铝片缺陷检测模型池，并对当前各类算法与铝片数据集的特点进行针对性优化，其主要工作如下：

i. 查阅相关资料，了解铝片表面缺陷检测方法的国内外现状；

ii. 分析 FPGA 的高、中、低端设备的优势、局限性，并针对其确定了以 Cascade RCNN、SSD、TOOD 为基础的检测网络结构；

iii. 分析英特尔 FPGA 中国创新中心提供的铝片缺陷数据集，为之后模型结构与部分模型训练 Tricks 指明了方向与参数；

iv. 基于 Cascade RCNN、SSD、TOOD 三类算法进行针对铝片数据集的改进与优化；

v. 三种模型的训练，并使用 Epoch-mAP 曲线证明模型已完成训练；

vi. 对比各模型 PR 曲线与召回率，对模型的可靠性进行了证明；

vii. 对比各模型 mAP、FPS、体积，确定模型对应部署的 FPGA 设备；

1. 针对于高端 FPGA 设备的 Cascade RCNN-ResNet50-FPN 模型；
2. 针对于中端 FPGA 设备的 SSD-MobileNet 模型；
3. 针对于低端 FPGA 设备的 TOOD-RepVGG-FPN 模型；

viii. 铝片表面缺陷检测系统的工作流程与功能介绍；

ix. 介绍了 FPGA 的部署流程与推理实现；

10.2 展望未来

本文三种模型其中 SSD-MobileNet 与 TOOD-RepVGG-FPN 均可实现实时检测，但 Cascade RCNN-ResNet50-FPN 的 mAP 只有 14.3，暂且不能满足实时检测需求。

本文三种模型即使加入 FPN 结构，仍然有极小概率出现小目标漏检的情况。

下一步计划：

i. 对 Cascade RCNN-ResNet50-FPN 使用 Slim 对模型进行压缩，使其能够实现高精度的实时瑕疵检测。

ii. 对三类模型继续针对小目标检测进行优化。

参考文献

[1]A Coarse-to-Fine Model for Rail Surface Defect Detection. YU H M,LI Q Y,TAN Y Q,et al. IEEE Transactions on Instrumentation and Measurement . 2019

[2]An adaptive extraction method for rail crack acoustic emission signal under strong wheel-rail rolling noise of high-speed railway. HAO Q S,SHEN Y,WANG Y,et al. Journal of Mechanical Systems . 2021

[3]基于多任务深度学习的铝材表面缺陷检测[J]. 沈晓海,栗泽昊,李敏,徐晓龙,张学武. 激光与光电子学进展. 2020(10)

[4]Rail crack detection and evaluation at high speed based on differential ECTsystem. XU P,ZHU C L,ZENG H M,et al. Measurement . 2020

[5]基于 Faster R-CNN 的零件表面缺陷检测算法[J]. 黄凤荣,李杨,郭兰申,钱法,朱雨晨. 计算机辅助设计与图形学学报. 2020(06)

[6]深度卷积神经网络在目标检测中的研究进展[J]. 姚群力,胡显,雷宏. 计算机工程与应用. 2018(17)

[7]Spatial Pyramid Pooling Deep Convolutional Networks for Visual Recognition. He K,Zhang X,Ren S,et al. IEEE Transactions on Pattern Analysis&Machine intelligence . 2014

[8]基于卷积神经网络的轻量化目标检测网络[J]. 程叶群,王艳,范裕莹,李宝清. 激光与光电子学进展. 2021(16)

[9]Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In ECCV, pages 354-370, 2016. 2, 3

[10]X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by ex- plicit shape regression. In CVPR, pages 2887-2894, 2012. 5

[11]J. Dai, K. He, and J. Sun. Instance-aware semantic segmenta- tion via multi-task network cascades. In CVPR, pages 3150- 3158, 2016. 3

[12]基于深度学习的多视窗 SSD 目标检测方法[J]. 唐聪,凌永顺,郑科栋,杨星,郑超,杨华,金伟. 红外与激光工程. 2018(01)

[13]多尺度卷积特征融合的 SSD 目标检测算法[J]. 陈幻杰,王琦琦,杨国威,韩佳林,尹成娟,陈隽,王以忠. 计算机科学与探索. 2019(06)

[14]钢轨缺陷无损检测与评估技术综述[J]. 张辉,宋雅男,王耀南,梁志聪,赵淼. 仪器仪表学报. 2019(02)

[15]J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. In NIPS, pages 379-387, 2016. 2, 3, 5, 8

[16]P. Dollár, P. Welinder, and P. Perona. Cascaded pose regres- sion. In CVPR, pages 1078-

1085, 2010. 5

[17]C. Elkan. The foundations of cost-sensitive learning. In IJ- CAI, pages 973-978, 2001. 2

[18]P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell., 32(9):1627-1645, 2010. 2, 3

[19]S. Gidaris and N. Komodakis. Object detection via a multi- region and semantic segmentation-aware CNN model. In ICCV, pages 1134-1142, 2015. 3

[20]R. B. Girshick. Fast R-CNN. In ICCV, pages 1440-1448, 2015. 1, 2, 3, 5

[21]S. Han, J. Pool, J. Tran, and W. J. Dally. Learning both weights and connections for efficient neural network. In NIPS, pages 1135-1143, 2015. 6

[22]K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In ICCV, 2017. 3, 6, 8

[23]J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. CoRR, abs/1611.10012, 2016. 8

[24]Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolu- tional architecture for fast feature embedding. In MM, pages 675-678, 2014. 5

[25]H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolu- tional neural network cascade for face detection. In CVPR, pages 5325-5334, 2015. 3

[26]T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Dollár, and C. L. Zitnick. Microsoft COCO: com- mon objects in context. In ECCV, pages 740-755, 2014. 2, 5