文章编号:1671-4598(2025)01-0045-08 DOI:10.16526/j.cnki.11-4762/tp.2025.01.006 中图分类号:TP391 文献标识码:A

基于 FPGA 的高能效纸板缺陷检测系统

陈俊杰，陈哲宇，郑子滨，李 胜
(福州大学 电气工程与自动化学院，福州 350108)

摘要：目前在工业流水线生产过程中主要采用人工检测的方法来剔除不合格纸板，这种方法效率低下，因此在生产过程中实现高能效的、准确的对纸板表面缺陷进行自动检测具有实际意义；依据 YOLO 系列网络在目标检测领域的优异表现和 FPGA 部署网络模型的高能效性，提出了一种基于 FPGA 的高能效纸板缺陷检测系统，通过 YOLOv7-Tiny 网络训练纸板缺陷数据集，并采用 QAT 对网络模型进行再训练和量化，在检测精度仅损失 0.36%前提下，将权重和特征图数据量化为 8 位，降低了硬件资源的消耗；设计了一种复用型多节点可配置架构的硬件加速器，通过多个配置节点实现对不同网络层的推理加速，对各个网络层在硬件层面进行了优化设计，并采用了层内和层间协同的流水线化设计；整个硬件加速系统通过软硬件协同设计实现，合理划分软硬件任务，实现了硬件加速器与软核处理器高度并行工作；最终在 Xilinx VC707 FPGA 评估板上，以 200 MHz 的工作频率实现了 177.96 GOPS 的吞吐量，同时仅消耗了 6.5 W 的功耗，实现了 27.38 GOPS/W 的高能效，分别为 I5-10400F CPU 的 19.7 倍和 GTX 2070S GPU 的 8.6 倍，兼顾了检测速度和功耗，满足了纸板生产的工业环境需求。

关键词：FPGA；表面缺陷检测；硬件加速；YOLO；量化

High-Energy Efficiency Cardboard Defect Detection System Based on FPGA

CHEN Junjie, CHEN Zheyu, ZHENG Zibin, LI Sheng
(School of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China)

Abstract: Currently, manual inspection methods are mainly used to remove defective cardboard in the industrial assembly line production process, which is inefficient. Therefore, it is of practical significance to achieve the high-energy efficiency and accurate automatic detection of surface defects on cardboard during the production process. Based on the excellent performance of the YOLO series network in the field of object detection and the high energy efficiency of FPGA-deployed network models, a high-energy efficiency cardboard defect detection system based on FPGA is proposed. The cardboard defect dataset is trained through the YOLOv7-Tiny network, and the network model is retrained and quantified using quantization aware training (QAT). With a detection accuracy loss of only 0.36%, a quantification of 8 bits for the weights and feature map data is applied to reduce hardware resource consumption. A hardware accelerator with reusable multi-node configurable architecture is designed to achieve inference acceleration for different network layers through multiple configuration nodes. Each network layer is optimized at the hardware level, and an inner and inter layer collaborative pipeline design is adopted. The entire hardware acceleration system is implemented through the collaborative design of software and hardware, with a rational division of software and hardware tasks, achieving the high-speed parallel operation of the hardware accelerator and soft-core processor. Finally, on the Xilinx VC707 FPGA evaluation board, a throughput of 177.96 GOPS is achieved at a working frequency of 200 MHz, while the power only consumes 6.5 W, the high energy efficiency of 27.38 GOPS/W is achieved, which is 19.7 times that of the I5-10400F CPU and 8.6 times that of the GTX 2070S GPU, respectively. It balances detection speed and power consumption, meeting the requirements of industrial environments in cardboard production.

Keywords: FPGA; surface defect detection; hardware acceleration; YOLO; quantization

0 引言

随着包装行业的迅速发展，瓦楞纸板凭借自身优势脱颖而出，成为物品包装的首选材料，被广泛应用于各种工业领域[1-2]。随着纸板生产量越来越大，企业在生产的过程中还要注重纸板的品质。

收稿日期:2023-11-15； **修回日期**:2024-01-03。
基金项目:国家自然科学基金项目(61871133)。
作者简介:陈俊杰(1999-),男,硕士研究生,工程师。
引用格式:陈俊杰,陈哲宇,郑子滨,等.基于 FPGA 的高能效纸板缺陷检测系统[J].计算机测量与控制,2025,33(1):45-52.

基于卷积神经网络（CNN，convolutional neural network）的目标检测技术是计算机视觉领域中的关键技术之一，它能够从图像中获取用户所感兴趣目标的位置信息和类别信息，因此被广泛地应用在各个领域[3-5]。目标检测网络根据执行阶段可以分一阶段检测网络和二阶段检测网络，其中一阶段检测网络的代表为 YOLO（You Only Look Once）[6] 网络，相较于二阶段检测网络，一阶段检测网络的实时性更高，实现了检测精度和检测速度之间的平衡，文献［7－10］采用了 YOLO 系列算法分别对瓶盖封装缺陷、汽车齿轮配件表面缺陷、轮胎缺陷、焊缝缺陷进行检测，然而目前大部分文献中并没有落地的设备来解决实际的工程问题。因此采用 YOLO 目标检测算法代替传统人工检测的方法来剔除带缺陷的纸板，避免残次纸板出厂，保证了出厂纸板的品质，不仅提高了效率，还降低了生成成本。

现场可编程门阵列（FPGA，field programmable gate array）具有丰富的硬件逻辑资源和片内存储资源，可以通过硬件描述语言来重构 FPGA 内部的硬件电路从而实现 CNN 的流水线并行和数据并行，相对于中央处理器（CPU，Central Processing Unit）和图形处理单元（GPU，Graphics Processing Unit）具有较高的能效，因此 FPGA 特别适合作为 CNN 的硬件加速平台。目前，FPGA 已经被广泛地应用于 CNN 硬件加速，文献［11－13］通过将网络模型量化到低位宽精度代替高精度的浮点数网络模型，量化可以节省 FPGA 的硬件资源消耗和片内片外的数据交互量，从而实现更加高效和快速的 FPGA 硬件加速器，但量化会使网络的检测精度下降，如何在量化的同时保证网络的检测精度损失较小是设计中需要关注的问题。同时文献［14－17］中也提出了 YOLO 网络的硬件加速方案，文献［14］通过 Winograd 算法优化了卷积操作，设计了一种低功耗的 YOLOv2 硬件加速器。文献［15］设计了一种片内全流水线化的 YOLOv2-Tiny 硬件加速器实现了较高的吞吐量。文献［16］将输入特征图映射到压缩矩阵乘法的形式，并优化了架构和调度策略，从而实现了低功耗的 YOLOv3-Tiny 硬件加速器。文献［17］提出一种数据块传输策略，并设计两个 14×14 的处理单元矩阵来加速 YOLOv2-Tiny 网络。

本文提出了一种基于 FPGA 的高能效嵌入式纸板缺陷检测系统，利用 YOLOv7-Tiny 网络来对纸板表面缺陷进行准确地、自动化地检测，采用量化感知训练（QAT，quantization aware training）算法来将原本的 32 位浮点网络模型量化为 8 位定点网络模型，在保证网络模型精度仅损失 0.36％的前提下，将权重和特征图数据压缩为原来的 1/4，减少了硬件资源的开销，加快了网络模型的推理速度。并设计了一种复用型多节点可配置的硬件加速器，通过多个可配置节点来实现不同网络层的推理，以这种复用共享硬件加速器的形式来实现 YOLOv7-Tiny 网络，并对各个模块进行了优化设计，节省了硬件资源的消耗和系统的复杂度，最后通过软硬协同设计搭建了整个系统。

1 纸板缺陷检测目标检测网络及优化策略

1.1 纸板缺陷检测网络

YOLOv7[18] 算法于 2022 年提出，相较之前版本的 YOLO 网络，使用了高效聚合的 ELAN 网络模块。YOLOv7-Tiny 算法是 YOLOv7 算法的简化版本，保留了基于级联的模型缩放和高效聚合的网络结构，降低模型的参数量和计算量，提高模型的检测速度和检测精度。

如图 1 所示为 YOLOv7-Tiny 的网络结构图，本文基于 YOLOv7-Tiny 网络训练了人工智能纸板缺陷自动检测模型，网络分为输入端、主干结构和输出端 3 个部分，高效长程距离网络（ELAN，efficient long-range attention network）模块通过 4 个特征计算块进行特征提取并拼接融合，利用不同长度的梯度路径，让深层网络能够获得更多的特征信息。SPPCSPC 结构中的空间金字塔（SPP，spatial pyramid pooling）结构可以通过不同尺度的最大池化来增大图像的感受野。YOLOv7-Tiny 输入图像尺寸为 416×416，输出层可以在 3 个尺度上预测。

1.2 网络量化

通常深度学习框架中训练得到的网络模型以 32 位精度的浮点数为主，但将 32 位精度的网络模型直接部署在 FPGA 上会带来很大计算压力和存储压力，从而阻碍了 YOLO 网络的应用，并且 FPGA 中的数据流都是二进制定点数，这使得在 FPGA 中定点数运算比浮点数更加高效，所以需要对网络模型进行定点量化[19]。

本文使用 QAT 算法来对用于纸板缺陷检测的 YOLOv7-Tiny 网络进行量化，在网络加入伪量化节点来标记参数的数值范围，然后通过在网络训练中模拟量化来调整量化数据的缩放系数和零点，从而减小量化所带来的误差，不同网络层的权重和特征图数据可以通过训练得到的缩放系数和零点来量化为 8 位定点数。QAT 算法能够在压缩网络尺寸的同时保证较小的精度损失，使网络模型在 FPGA 上实现更加高效。

在网络前向推理过程中，特征图数据和权重数据量化为 8 位的位宽，通过 QAT 得到的缩放系数 S 和零点 Z 来将浮点数 X 映射为定点数 q 的公式如下式所示：

$$X = S(q - Z) \tag{1}$$

卷积层的作用是提取特征信息，计算公式如公式（2）所示：

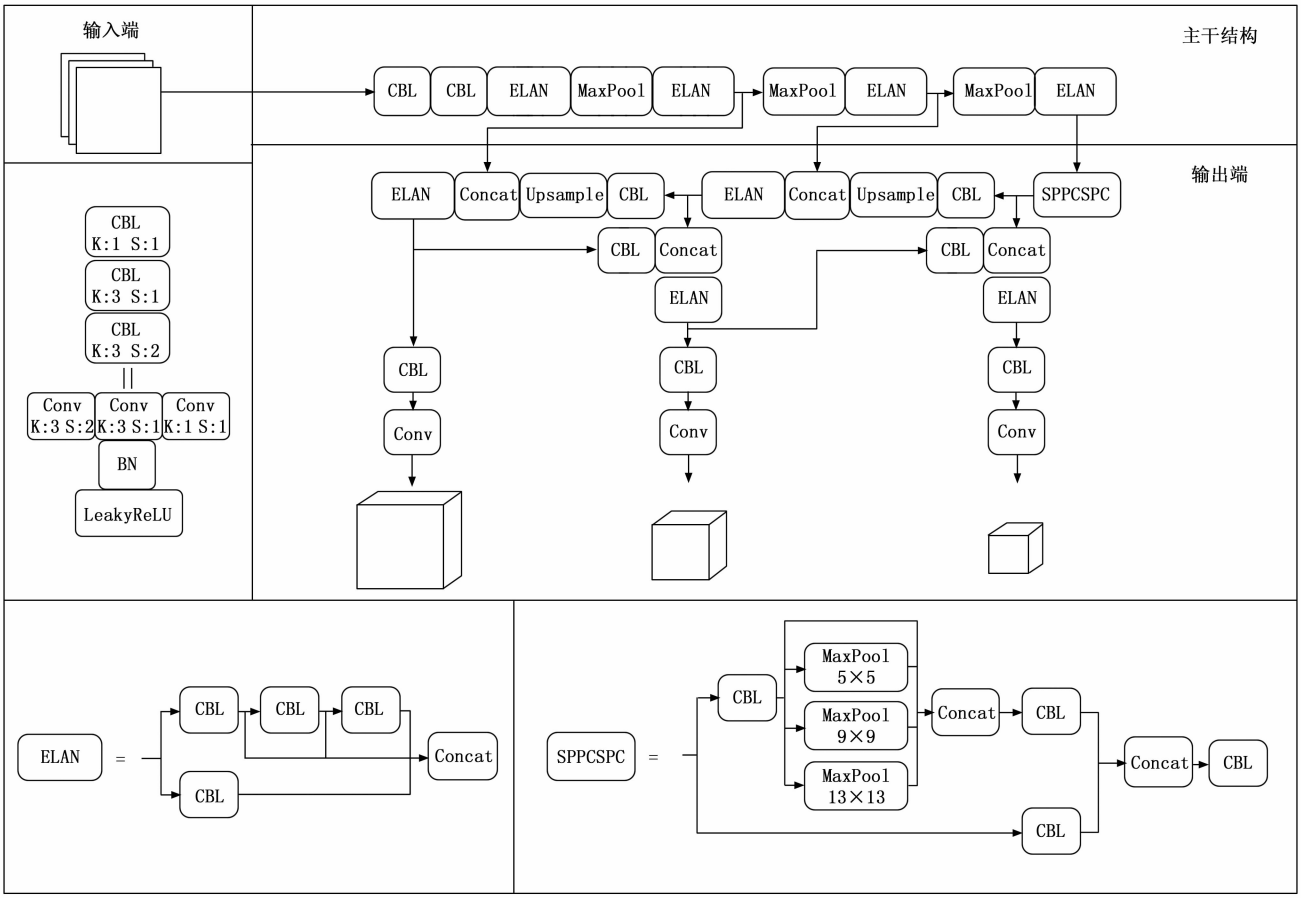

图 1　YOLOv7-Tiny 网络结构图

$$Y_{\text{conv}} = W \otimes X_{\text{conv}} + b \tag{2}$$

其中：Y_{conv} 和 X_{conv} 分别表示卷积层的计算结果和输入数据，W 和 b 分别为权重和偏置。将式（1）代入式（2）可得下式：

$$q_Y = \frac{S_W S_X}{S_Y}(q_W - Z_W) \otimes (q_X - Z_X) + \frac{S_b}{S_Y}(q_b - Z_b) + Z_Y \tag{3}$$

由于权重和偏置采用对称量化，所以 Z_W 和 Z_b 都为 0。同时为了使算法在 FPGA 中计算更加高效，在大部分工程实践中，会取 $S_b = S_W S_X$，故可得下式：

$$q_Y = \frac{S_W S_X}{S_Y}[q_W \otimes (q_X - Z_X) + q_b] + Z_Y \tag{4}$$

S_W 和 S_X 的缩放比例为 8 位，可知 $S_W S_X$ 的缩放比例为 16 位，在 QAT 中，偏置以 32 位的数据位宽来进行计算，放缩系用 $S_W S_X$ 来代替 S_b 作为偏置的数会损失一半的信息量。误差在网络层中不断传播将会影响最终的检测结果，因此本文提出公式（5）来解决该问题：

$$q_Y = 2^{-n}\left[2^n \frac{S_W S_X}{S_Y} q_W \otimes (q_X - Z_X) + \frac{1}{S_Y} 2^n r_b\right] + Z_Y \tag{5}$$

由于 $\frac{S_W S_X}{S_Y}$ 是浮点数，需要找到一个合适的 n 通过移位量化得到 $2^n \frac{S_W S_X}{S_Y}$ 将公式转换为全定点计算，$2^n r_b$ 相当于对偏置的浮点数 r_b 进行了 n 位的移位量化，通过这种方式可以使得偏置的缩放比例达到甚至超过 32 位，从而避免了偏置的信息损失。

YOLOv7-Tiny 中使用的激活函数为 LeakyReLU，LeakyReLU 的负激活值设定为 0.125，即 2^{-3}。在硬件设计中可以用算术右移三位来取代浮点运算，从而减少 FPGA 上的硬件资源消耗。因此，LeakyReLU 在 QAT 训练后得到的量化计算公式如下式所示：

$$q_{\text{act}} = \begin{cases} \frac{S_{\text{Conv}}}{S_{\text{act}}}(q_{\text{Conv}} - Z_{\text{Conv}}) + Z_{\text{act}}, q_{\text{Conv}} > Z_{\text{Conv}} \\ 2^{-3}\left[\frac{S_{\text{Conv}}}{S_{\text{act}}}(q_{\text{Conv}} - Z_{\text{Conv}}) + Z_{\text{act}}\right], q_{\text{Conv}} \leqslant Z_{\text{Conv}} \end{cases} \tag{6}$$

拼接层用于将不同网络层的特征图信息融合，由于不同网络层的特征图数据量化后的缩放因子和零点参数不一致，特征图数据需要重新量化来减小误差，量化后的拼接层的计算公式如下式所示：

$$q_{\text{concat}} = \frac{S_{\text{Concat}}}{S_{\text{act}}}(q_{\text{act}} - Z_{\text{act}}) + Z_{\text{Concat}} \tag{7}$$

2 纸板缺陷检测硬件加速器框架设计

2.1 复用型多节点可配置的硬件架构

如图 2 所示为本文所设计的硬件加速器架构图，硬件加速器为复用型多节点可配置的硬件架构，硬件加速器内置高级可扩展接口（AXI，advanced extensible interface）总线中的 AXI-Lite 接口，该接口会将接收到的信息放在寄存器中，利用寄存器的配置信息来配置各个选择器让硬件加速器实现不同的功能，从而实现 YOLOv7-Tiny 纸板缺陷检测网络中不同的网络层，通过不断复用该硬件加速器就可以实现网络的前向推理。控制模块首先会根据配置信息分析出网络层类型，并生成其他硬件模块的使能控制信号。YOLOv7-Tiny 纸板缺陷检测网络在推理过程中需要缓存大量的特征图和参数数据，由于 FPGA 的片内存储资源有限，不足以将所有数据缓存在片内，因此需要双倍速率（DDR，double data rate）存储器和片内的存储资源协同缓存，而将 DDR 存储器的数据读取到片内缓存时存在访存时延，所以本文采用双输入缓存的机制，通过缓存选择器使两块输入特征图缓存空间交替进行数据写入和计算，降低了数据传输时延，减小了访存时延对加速器的影响。填充模块主要是在特征图外围填充数据来保证卷积后能够保留更多的特征信息，由控制模块生成的填充使能信号来控制，当信号拉高时，输出特征图数据量化后的零点值，而当信号拉低时，读取输入特征图缓存空间的数据输出。特征图数据需要减去量化后的零点值然后根据卷积类型的配置信息选择数据通路，卷积核为 1×1 的特征图数据直接输入卷积模块计算，而卷积核为 3×3 的特征图数据需要通过行缓冲器形成 3×3 的滑动窗口数据再送入卷积模块进行卷积运算，行缓冲器可以对输入数据进行预先缓存，避免了数据搬移和复制所产生的时延。卷积模块会通过卷积类型选择器来读取不同排列的权重和特征图数据进行卷积运算，卷积模块会输出两种计算结果，一种为卷积计算输出结果，另一种为经过卷积、批归一化和激活函数计算后的结果。

YOLOv7-Tiny 纸板缺陷检测网络模型的主干网络中用于图像下采样的最大池化层使用的是 2×2 的滑动核，因此在将卷积模块的计算后的数据输入下采样模块之前，需要用行缓冲器生成 2×2 的滑动窗口数据。同时为了减少片外数据传输所带来的时延，本文设计了卷积-上采样、卷积-下采样的层间流水线化结构，各模块层内的完全流水线化设计和层间流水线化设计提高了硬件加速器的吞吐量，同时为了能够保存层间流水线结构中卷积层的输出数据，增设了一个中段输出特征图缓存空间。SPP 模块用于匹配不同尺寸滑动核的最大池化，不同的尺寸的滑动核需要的图像填充圈数不同，用来保证池化后的图像保持原来的尺寸。YOLOv7-Tiny 纸板缺陷检测网络通过拼接层将不同通道的特征图拼接在一起，使网络能够学习到更多的特征信息，提高网络性能，从硬件实现的层面上来看，需要进行量化计算后将数据传输到 DDR 存储器预设的偏移地址上就可以实现特征图数据的拼接，最后通过加速类型选择器来选择写入输出特征图缓存的数据流。

2.2 卷积模块架构概述

如图 3 所示为卷积模块架构的概述图，卷积模块采用 16 输入通道和 16 输出通道并行计算。YOLOv7-Tiny 纸板缺陷检测网络模型中最大的卷积核尺寸为 3×3，因此乘法计算阵列的大小设计为 3×3 可向下兼容。3×3 卷积和 1×1 的卷积可以共用 3×3 的乘法阵列，只需将不同排列的权重数据和特征图数据输送进卷积乘法阵列即可完成相应的计算，3×3 的卷积需要在一个周期内同时输入特征图的 9 个像素点数据和对应的 9 个权重数据，而 1×1 的卷积在一个周期内输入特征图的 1 个像素点数据和对应的 1 个权重数据。对于不同步长的卷积运算，需要控制输出数据使能信号的来选择性地输出

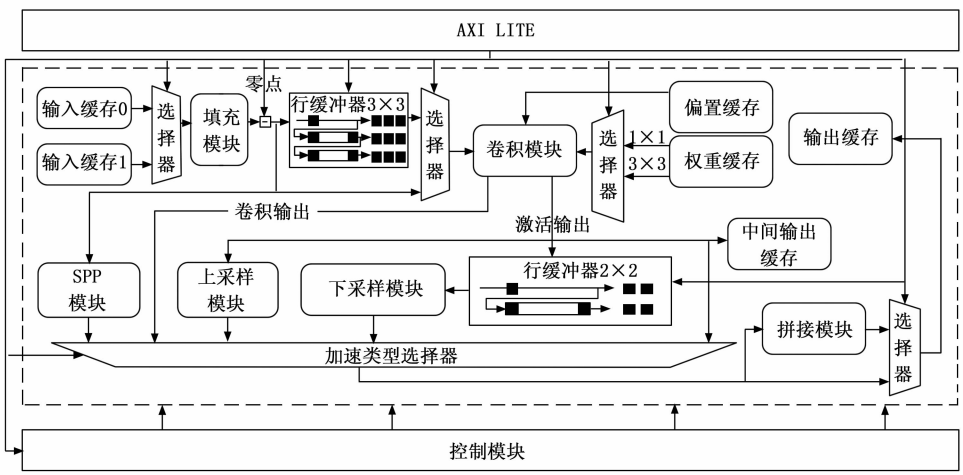

图 2 硬件加速器总体架构

数据。乘法阵列计算后的数据通过加法树分多周期计算 16 个输入通道乘法阵列累加和并且保存到通道累加缓存器，并在下一次迭代中，会读取通道累加缓存器中的数据进行相加后再写入通道累加缓存器，循环迭代直到计算完所有输入通道。再根据公式（5）进行量化计算得出量化的卷积层计算结果，减去卷积层特征图数据的零点之后，判断该数与零点的关系，根据公式（6）进行激活层的量化计算计算并输出结果，至此完成 16 个输出通道的特征图数据计算，并如此往复完成所有输出通道的特征图数据计算。

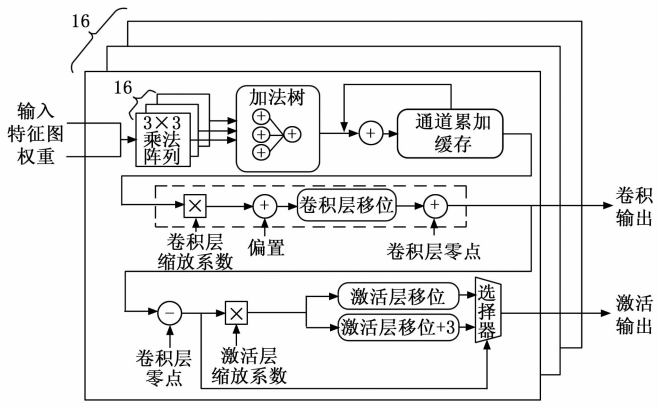

图 3　卷积模块架构图

2.3　卷积-下采样层间流水线化架构概述

YOLOv7-Tiny 纸板缺陷检测网络中采用步长为 2、滑动核大小为 2×2 的最大池化层来对图像进行下采样，将特征图尺寸缩小 1/4。如图 4 所示为本文所设计的卷积—下采样层间流水线化的设计，在网络模型中卷积—下采样层间流水线中卷积的卷积核尺寸为 1×1，特征图数据逐行逐通道批次输入卷积模块，在卷积模块计算最后批次的输入通道数据时，卷积模块计算后的数据会直接输入行缓冲器中来生成 2×2 的滑动窗口，最后通过两周期比较器计算出最大值。在开始池化运算之后，计数器会对输出的行列值进行计数，由于最大池化层步长为 2，所以只有当行计数器和列计数器的值同时为偶数时，才将数据写入输出特征图缓存空间。卷积—下采样的层间流水线化设计避免了使用额外的存储空间来缓存卷积模块计算完的数据，降低了系统的时延。

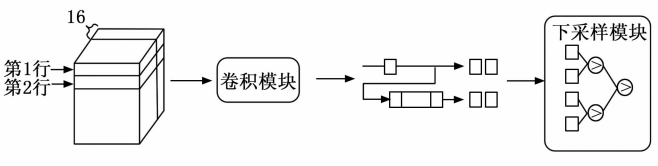

图 4　卷积—下采样层间流水线化架构图

2.4　多节点可配置 SPP 模块架构概述

SPP 模块通过使用不同尺寸滑动核的最大池化来增强网络的感受野。如图 5 所示为多节点可配置 SPP 模块的硬件架构图，为了节省硬件资源的消耗，本文所使用的 3 个尺度的滑动核分别为 3×3、5×5、7×7，行缓存器依据最大滑动核的尺寸来设计，最大可同时输出 7 行的数据，由于不同尺寸滑动核的图像填充值不同，所以需要根据图像的填充值来将行缓冲器配置成不同的长度。行缓冲器可以向下兼容，利用不同数量的行缓冲器可以输出不同大小的滑动窗口数据，再通过 SPP 选择器来选择输出滑动窗口数据，最后输出的滑动窗口数据通过比较器分多周期筛选出最大值。

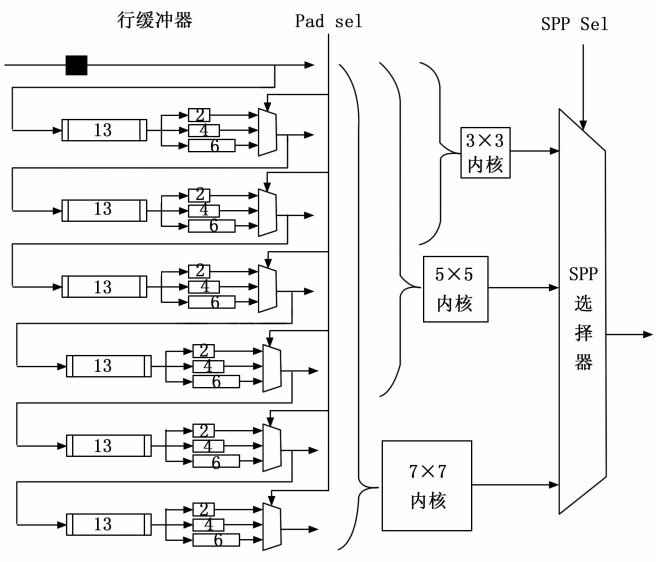

图 5　多节点可配置 SPP 模块架构图

2.5　地址映射型上采样模块架构概述

YOLOv7-Tiny 纸板缺陷检测网络中有两个上采样层，分别将 13×13 的特征图扩充为 26×26、26×26 的特征图扩充为 52×52，再与主干网络中的特征图进行拼接融合，有效地减少了网络的信息损失。如图 6 所示为地址映射型上采样模块硬件设计的架构图，卷积模块计算得到的数据流首先写入上采样模块中的数据缓存空间，通过控制数据缓存空间的读取地址来实现图像上采

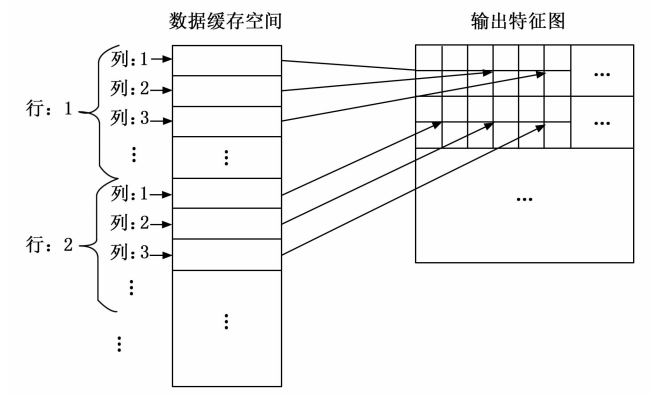

图 6　地址映射型上采样模块架构图

样。上采样模块开始工作后，一个列数据输出两个周期的数据后才会读取下一个列数据，当输出完一行的有效数据后读取地址返回该行首部，再以相同的方式再输出一行与上一行相同的数据，至此，完成了输入空间缓存中第一行输入数据的上采样操作，重复此操作即可完成完整的特征图数据的上采样操作。

3 纸板缺陷检测硬件加速系统

纸板缺陷检测硬件加速系统的框架如图 7 所示，采用软硬协同的方式来实现高能效 YOLOv7-Tiny 纸板缺陷检测网络的推理，在片内搭载 Microblaze 软核 CPU，通过 CPU 来对系统进行整体调度、执行控制任务，硬件加速器用于执行网络前向推理的计算任务。通过工业 CameraLink 摄像头采集图像数据传入 FPGA 板卡中，经过硬件加速器的加速推理后将预测结果显示在显示屏上。

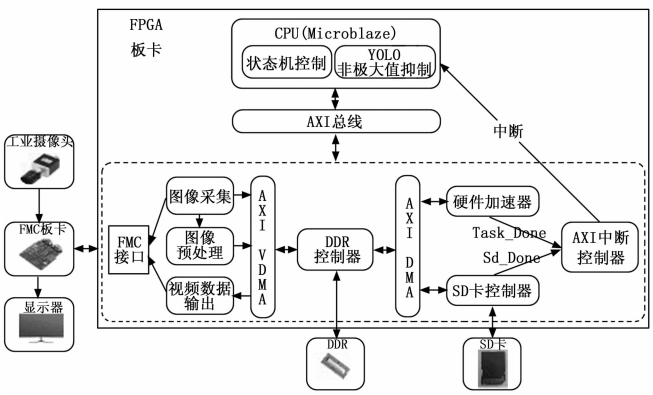

图 7 纸板缺陷检测硬件加速系统框架

如图 7 所示，在系统上电后，CPU 软件端首先会通过 FMC 板卡的 CameraLink 接口和 IIC（内部集成电路，inter-integrated circuit）接口分别配置工业摄像头和显示器的初始化参数。使用纸板缺陷数据集训练后 YOLOv7-Tiny 网络模型的权重、偏置和量化因子等参数数据进行重新排列后预先以二进制文件的格式保存在安全数码卡（SD 卡，secure digital card）卡中，在进行纸板表面缺陷检测之前，SD 卡控制器会通过 SD 卡槽的 SPI（串行外设接口，serial peripheral interface）硬件接口将权重、偏置、缩放因子和零点这些参数数据从 SD 卡读取到预设的双倍速率同步动态随机存储器（DDR SDRAM，double data rate synchronous dynamic random access memory）的偏移地址上，单个扇区读取完毕 SD 卡会向 CPU 发送 SD _ Done 中断信号，并开启下一个扇区读取。

CPU 软件端设置各个网络层的配置参数，CPU 根据这些参数通过 AXI 接口来配置硬件加速器，从而实现纸板缺陷检测网络中不同的网络层前向推理。硬件加速器通过 AXI Stream 接口来读写数据，只要在软件端向 AXI DMA 设置传递 DDR 存储器的地址指令信息和数据传输长度就可以实现加速器和 DDR 存储器的数据交互。硬件加速器执行完一次就会向 CPU 发送 Task _ Done 中断信号，CPU 会重新配置硬件加速器的参数，如此反复直到最终完成整个网络的推理。CPU 根据网络推理的结果，通过非极大值抑制（NMS，non-maximum suppression）算法求得纸板表面缺陷的坐标位置。

如图 8 所示为系统软硬协同设计的运行方式，通过 CPU 软件端设置的状态机和中断信号来控制实现硬件加速器进行不同的配置。系统在执行网络层推理时，在第一阶段中，CPU 软件端会先计算数据传输参数，在下一个阶段中，CPU 软件端将硬件加速器配置为接收模式同时计算的传输参数，硬件加速器则接收从 DDR 存储器传输来的数据，在阶段三中硬件加速器的输入特征图的两个缓存空间分别用于推理和接收数据，直到执行到阶段四中完成所有输入通道的计算后进入阶段五，硬件加速器配置为发送模式，将输出特征图发送至 DDR 存储器，如此循环直到完成所有输出通道特征图数据的计算。CPU 和加速器并行工作的设计，能够充分地减少系统时延，充分发挥了 CPU 调度控制和加速器加速推理的优势。

4 结果分析

4.1 网络量化

在 Pytorch 深度学习框架中分别对 YOLOv7-Tiny 网络进行初始训练和 QAT，本文以百度飞桨上开源的纸板缺陷数据集进行实验测试，来自于纸板生产厂的 1 057 张真实场景的图片，收集了纸板在生产过程中产生的破损、起泡、划痕等一些缺陷。图 9 列出了全精度网络训练、QAT 所得到的网络的平均精度（mAP，mean

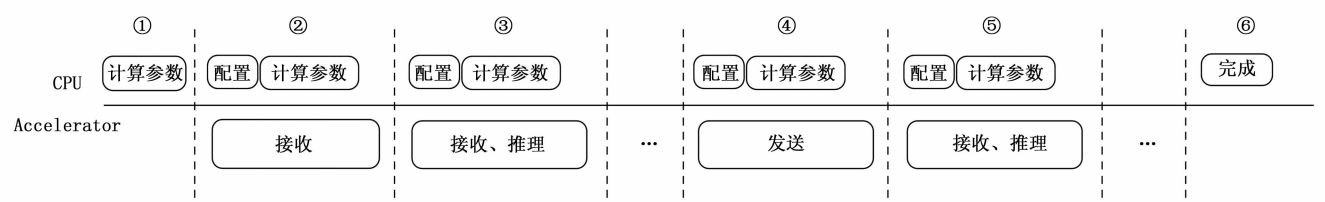

图 8 软硬协同运行方式

average precision)。

训练收敛稳定后全精度网络训练的精度为 89.03%，QAT 的精度为 88.67%。如表 1 所示，在 *mAP* 值仅下降了 0.36%的情况下，将 32 位浮点模型压缩为 8 位定点数模型，QAT 模型的参数尺寸约为全精度网络模型的 1/4，极大地减少了硬件资源消耗，使系统能够实现更大的并行度。

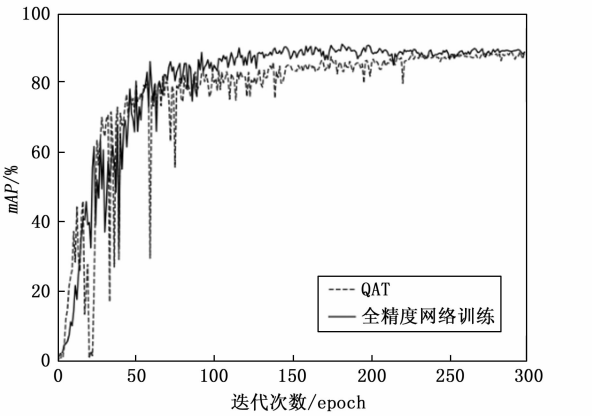

图 9　网络训练 *mAP* 变化曲线

表 1　全精度网络训练模型与 QAT 模型对比

训练类型	量化位宽/bit	模型参数尺寸/M
全精度网络训练	32	23.9
QAT	8	6.003

4.2　实现结果

本文采用 Xilinx Virtex-7 VC707 作为纸板缺陷检测硬件加速系统的部署平台，在表 2 中分析了硬件加速器核心资源的使用情况。数字信号处理器（DSP，digital signal processor）为主要的硬件乘法器资源，主要消耗在卷积模块中乘法阵列设计。块随机存储器（BRAM，block random access memory）为主要的片内存储器资源，主要用于缓存空间模块的实现。查找表资源（LUT，look up table）主要用于硬件设计中各部分组合逻辑的实现。触发器资源（FF，flip flop）主要用于硬件中各部分时序逻辑的实现。

表 2　详细硬件资源占用

资源	消耗/个	总量/个	利用率/%
LUT	110 664	303 600	36.45
FF	172 964	607 200	28.49
BRAM	229	1 030	22.23
DSP	1 257	2 800	44.89

本文分别在 FPGA、CPU 和 GPU 三种平台上进行测试，表 3 列出了本系统与 CPU、GPU 平台性能的对比，CPU 的测试平台为 I5-10400F，GPU 为 GTX 2070S，CPU 与 GPU 在 Pytorch 框架下进行网络推理，并记录运行时间和实时功耗。本文所设计的系统在能效为 27.38 GOPS/W 分别为 CPU 的 19.7 倍和 GPU 的 8.6 倍，具有优越性。

表 4 列出了本文与近年来其他文献中的硬件加速系统性能的对比，本文所设计的 YOLOv7-Tiny 纸板缺陷检测硬件加速系统，由于模型与数据集的不同，各系统部署的模型尺寸也不同，吞吐量为每秒硬件加速系统所能够处理的数据量，可通过吞吐量与功耗之比来衡量不同系统的性能，本文能效优于文献［16，17，20，21］的硬件加速系统，实现了高能效的系统架构设计，达到了计算性能和功耗两者的平衡，更加适合实际应用。

表 3　不同平台性能比较

平台	I5-10400F CPU	GTX 2070S	Virtex-7 Vc707
延迟/ms	81.8	24	31.3
吞吐量/GOPS	67.97	231.29	177.96
功耗/W	48.58	72.18	6.5
能效/(GOPS/W)	1.39	3.20	27.38

表 4　硬件加速系统性能比较

	文献[16]	文献[17]	文献[20]	文献[21]	本文
网络模型	Tiny YOLO	YOLOv3-Tiny	YOLOv2-Tiny	YOLOv3-Tiny	YOLOv7-Tiny
平台	XCZU3EG	Ultra96 V2	ZC706	Zedboard	VC707
时钟频率/MHz	100	250	100	100	200
图像尺寸	—	448×448	416×416	416×416	416×416
数据拉宽/bit	1-8	8	16-32	16	8
延迟/ms	63	121	128.74	532	31.3
吞吐量/GOPS	71.04	31.5	41.99	10.45	177.96
功耗/W	6	4.26	7.50	3.36	6.5
能效/(GOPS/W)	11.84	7.40	5.6	3.11	27.38

如图 10 所示为纸板缺陷检测硬件加速系统的检测效果示意图，将程序烧录进 FPGA，由工业摄像头采集图像数据经由系统处理后，将检测结果显示在左侧显示屏上，图中效果显示对纸板表面的三处缺陷进行了检测。

图 10　检测效果示意图

5 结束语

本文提出了一种基于 Xilinx VC707 FPGA 的高能效纸板缺陷检测系统，通过 QAT 对网络进行量化，设计了一种复用型多节点可配置架构的硬件加速器，通过配置多个不同的节点实现不同的加速器功能，复用硬件加速器来实现网络的前向推理。对各个硬件模块进行了优化，并设计了层内、层间的双流水线化的结构，最后通过软硬件协同设计最终实现了整个检测系统，实现了 88.67%的平均检测精度和 177.96 GOPS 的吞吐量，功耗仅为 6.5 W，能效高达到 27.38 GOPS/W，优于 CPU、GPU 和其它的 FPGA 硬件加速系统，实现了检测速度、精度和功耗的平衡，满足了工业环境中高能效的需求。

参考文献：

[1] FADIJI T, AMBAW A, COETZEE C J, et al. Application of finite element analysis to predict the mechanical strength of ventilated corrugated paperboard packaging for handling fresh produce [J]. Biosystems Engineering, 2018, 174: 260 - 281.

[2] KELLICUTT K, LANDT E. Development of design data for corrugated fiberboard shipping containers [J]. Tappi J, 1952, 35: 398 - 402.

[3] ZHENG Z, ZHAO J, LI Y. Research on Detecting Bearing-Cover Defects Based on Improved YOLOv3 [J]. IEEE Access, 2021, 9: 10304 - 10315.

[4] CHEN S H, TSAI C C. SMD LED chips defect detection using a YOLOv3-dense model [J]. Advanced Engineering Informatics, 2021, 47: 101255.

[5] 王 宸，张秀峰，刘 超，等. 改进 YOLOv3 的轮毂焊缝缺陷检测 [J]. 光学精密工程，2021，29 (8)：1942 - 1954.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779 - 788.

[7] 赵 磊，矫立宽，翟 冉，等. 基于 YOLOv5 的瓶盖封装缺陷轻量化检测算法 [J]. 激光与光电子学进展，2023，60 (22)：139 - 148.

[8] 朱德平，程 光，姚景丽. 改进 YOLOv5 的汽车齿轮配件表面缺陷检测 [J]. 计算机工程与应用：1 - 9.

[9] PENG C, LI X, WANG Y. TD-YOLOA: An efficient YOLO network with attention mechanism for tire defect detection [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1 - 11.

[10] WANG G Q, ZHANG C Z, CHEN M S, et al. Yolo-MSAPF: multiscale alignment fusion with parallel feature filtering model for high accuracy weld defect detection [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1 - 14.

[11] FARABET C, LECUN Y, KAVUKCUOGLU K, et al. Large-scale FPGA-based convolutional networks [J]. Scaling up Machine Learning: Parallel and Distributed Approaches, 2011, 13 (3): 399 - 419.

[12] 张丽丽，陈 真，刘雨轩，等. 基于 ZYNQ 的 Yolo v3-SPP 实时目标检测系统 [J]. 光学精密工程，2023，31 (4)：543 - 551.

[13] LV P, LIU W, LI J. A FPGA-based accelerator implementaion for YOLOv2 object detection using Winograd algorithm [C] // 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020: 1894 - 1898.

[14] BAO C, XIE T, FENG W, et al. A power-efficient optimizing framework fpga accelerator based on winograd for yolo [J]. IEEE Access, 2020, 8: 94307 - 94317.

[15] NGUYEN D T, NGUYEN T N, KIM H, et al. A high-throughput and power-efficient FPGA implementation of YOLO CNN for object detection [J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2019, 27 (8): 1861 - 1873.

[16] ADIONO T, PUTRA A, SUTISNA N, et al. Low latency YOLOv3-tiny accelerator for low-cost FPGA using general matrix multiplication principle [J]. IEEE Access, 2021, 9: 141890 - 141913.

[17] HUANG H, LIU Z, CHEN T, et al. Design space exploration for yolo neural network accelerator [J]. Electronics, 2020, 9 (11): 1921.

[18] WANG C Y, BOCHKOVSKIY A, LIAO H-Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 7464 - 7475.

[19] GONÇALVES A, PERES T, VÉSTIAS M. Exploring data size to run convolutional neural networks in low density fpgas [C] // International Symposium on Applied Reconfigurable Computing, 2019: 387 - 401.

[20] PREUBER T B, GAMBARDELLA G, FRASER N, et al. Inference of quantized neural networks on heterogeneous all-programmable devices [C] // 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018: 833 - 838.

[21] YU Z, BOUGANIS C-S. A parameterisable FPGA-tailored architecture for YOLOv3-tiny [C] // Applied Reconfigurable Computing. Architectures, Tools, and Applications: 16th International Symposium, ARC 2020, Toledo, Spain, April 1 - 3, 2020, Proceedings 16, 2020: 330 - 344.