doi:10.6041/j.issn.1000-1298.2025.05.060

基于深度学习加速模型的杂乱目标实时视觉检测方法

余永维　陈天皓　杜柳青　方　荣
(重庆理工大学机械工程学院,重庆 400054)

摘要:在农业机械自动装配产线上,其嵌入式控制平台片上资源极其有限,而基于卷积神经网络的深度学习检测系统参数量过大,难以直接移植于嵌入式平台,为此,本文提出一种基于改进 ResNet18 - SSD(Single shot multi-box detector)和现场可编程门阵列(Field programmable gate array,FPGA)加速引擎的深度学习实时检测方法。为了降低参数量的同时提高检测模型准确性,提出基于 ResNet18 - SSD 的深度学习快速检测模型,利用优化改进后的 ResNet18 网络替换 SSD 模型的 VGG16 前置网络,引入多分支同构结构和非对称并行残差结构,使其能适应遮挡、光线昏暗等复杂场景;在满足检测精度需求的情况下,采用动态定点量化的方式,对模型数据量进行缩减,以提高检测模型执行效率。针对改进 ResNet18 - SSD 模型中消耗资源严重的卷积层,提出一种基于 Winograd 算法的 FPGA 加速引擎,提高模型检测实时性,通过软硬件协同设计,从硬件加速器与软件网络轻量化两个角度进行联合优化,实现轻量化、加速性能及复杂场景下准确性三者之间的平衡。在 Xilinx FPGA 嵌入式平台的实验结果表明,本文方法检测准确率达到 93.5%,当工作频率为 100 MHz 时,单幅图像检测时间为 80.232 ms,满足实时性需求。

关键词:目标检测;FPGA;动态定点量化;Winograd 算法
中图分类号:TP273⁺.5　　**文献标识码**:A　　**文章编号**:1000-1298(2025)05-0617-08　　**OSID**:

Real Time Visual Detection for Cluttered Targets Based on Deep Learning Acceleration Model

YU Yongwei　CHEN Tianhao　DU Liuqing　FANG Rong
(*College of Mechanical Engineering*, *Chongqing University of Technology*, *Chongqing* 400054, *China*)

Abstract: In the automatic assembly line of agricultural machinery, the on-chip resources of its embedded control platform are extremely limited, and the parameter amount of the convolutional neural network-based deep learning detection system is too large, which is difficult to be directly transplanted to the embedded platform. Therefore, a deep learning real-time detection method based on improved ResNet18 - SSD (single shot multi-box detector) and field programmable gate array (FPGA) acceleration engine was proposed. In order to improve the accuracy of the detection model while reducing the number of parameters, a deep learning fast detection model based on ResNet18 - SSD was proposed, which utilized the optimized and improved ResNet18 network to replace the VGG16 predecessor network of the SSD model, and introduced a multi-branch isomorphic structure and an asymmetric parallel residual structure, so as to adapt to the complex scenes such as occlusion, dim light; and in the case of meeting the detection accuracy requirements, a dynamic fixed-variance network was used to meet the detection accuracy requirements. Under the condition of meeting the requirements of detection accuracy, the dynamic fixed-point quantization was adopted to reduce the model data volume to improve the execution efficiency of the detection model. Aiming at improving the convolutional layer in the ResNet18 - SSD model, which consumed serious resources, an FPGA acceleration engine based on the Winograd algorithm was proposed to improve the real-time performance of the model detection, and through the

收稿日期:2024 - 11 - 19　修回日期:2025 - 01 - 17
基金项目:国家自然科学基金项目(52375083)、重庆英才计划项目(CQYC20220207232/cstc2024ycjh - bgzxm0052)、重庆教委科研重大项目(KJZD - M202401101)、重庆自然科学基金项目(cstc2021jcyj - msxmX0372)和重庆技术创新与应用项目(CSTB2022TIAD - CUX0017)
作者简介:余永维(1973—),男,教授,博士,主要从事智能制造技术研究,E-mail:weiyy@cqut.edu.cn
通信作者:杜柳青(1978—),女,教授,博士,主要从事智能检测技术研究,E-mail:lqdu@cqut.edu.cn

software-hardware co-design, joint optimization was carried out from the perspectives of the hardware gas pedal and the lightweighting of the software network, so as to achieve a balance between the lightweighting, acceleration performance, and accuracy in the complex scene. Experimental results on the Xilinx FPGA embedded platform showed that the detection accuracy of the proposed method reached 93.5%, and the detection time of a single image under the operating frequency of 100 MHz was 80.232 ms, which met the real-time demand.

Key words: object detection; FPGA; dynamic fixed-point quantization; Winograd algorithm

0 引言

近年,深度学习在视觉检测领域不断取得突破[1-3]。传统目标检测采用 Selective Search[4-5]方法,其通过大量候选区域,在候选区域对图像特征进行提取,然后对目标进行识别。Faster R-CNN[6-8]卷积神经网络提高了目标检测准确性。相比 Faster R-CNN 算法,SSD 目标检测算法在图像检测速度上有明显优势[9-11]。

在农业机械自动装配产线上,智能装配机器人大多采用嵌入式控制平台,其片上资源极其有限,而基于卷积神经网络的深度学习检测模型(如 Faster R-CNN、SSD)参数计算量巨大,难以直接移植部署[12-13]。而现有轻量化检测模型如 MobileNet、EfficientNet 和 GhostNet 等,其轻量化方法主要集中在降低模型参数量和计算复杂度上,通过减少浮点运算数来降低模型复杂度,但对检测模型准确影响较大。因此,研究能够部署在嵌入式平台、基于深度学习轻量化加速模型目标零件实时检测方法,解决农业机械智能装配机器人在粘连、堆叠、光照变化及环境因素干扰等复杂条件下零件检测率低、鲁棒性差等问题有重要意义。

相比 CPU、GPU 等硬件处理器,FPGA 可实现硬件灵活定制,能够高效地实现算法加速、数据处理,从而提高系统性能,其在性能、开发周期、资源利用率等方面均具有优势,因此利用 FPGA 特点来加速卷积神经网络是解决传统深度学习模型效率低、速度慢的一种新思路[14-16],该领域代表性研究进展有:徐欣等[17]采用 FPGA 对多卷积核并行计算方法,使不同的卷积核同时对输入图像进行卷积,以实现卷积核层面的并行加速计算。QIU 等[18]提出多卷积计算单元并行计算的方式,但其以损失片上数据传速率为代价,整体系统稳定性受到影响。文献[19-20]采用端到端的方式将所有卷积层完全在片上进行布置,这种方式仅适用于小型神经网络,当网络模型复杂度增加时需要的片上资源以及带宽会明显增加。陈朋等[21]引入分割参数,设计了一种基于改进动态配置的 FPGA 卷积神经网络优化方法。谢坤鹏等[22]针对片上资源受限、算子操作类型复杂等问题,提出一种面向嵌入式 FPGA 的卷积神经网络稀疏化加速框架。李天阳等[23]根据卷积和注意力机制的计算特征,提出一种面向 FPGA 的非线性与归一化加速单元。以上在 FPGA 上实现卷积神经网络加速方法主要采用提高计算并行性、以及降低片上存储资源占用等方法。如何合理设计神经网络架构的同时充分发挥 FPGA 的并行计算能力以实现加速的目的是目前亟需解决的难题。

针对采用嵌入式平台的农业机械智能装配机器人在粘连、堆叠、光照变化及图像模糊等复杂场景下零件准确率低、鲁棒性差等问题,目标检测对精度、实时性、小尺寸多目标检测的要求,本文通过软硬件协同设计,从硬件加速器与软件网络轻量化两个角度进行联合优化,以实现轻量化、加速性能及复杂场景下准确性三者之间的平衡。提出一种基于改进 ResNet18-SSD 的深度学习快速检测模型,引入多分支同构结构和非对称并行残差结构优化 ResNet18 网络,用改进后 ResNet18 网络替换 SSD 模型 VGG16 前置网络,以适应遮挡、光线昏暗等复杂场景;采用动态定点量化策略,提出一种基于 Winograd 算法的 FPGA 加速引擎,以提高检测实时性。

1 基于改进 ResNet18-SSD 的深度学习加速检测模型

1.1 深度学习加速检测模型构建

SSD 网络模型采用 Selective Search 方法,在大量候选区域对图像特征进行提取分类。SSD 目标检测算法在检测速度上较 Faster R-CNN 系列检测模型有明显优势。

提出引入 ResNet18 模块来优化 SSD 原型网络,并对 ResNet18 进行通道裁剪。基于改进 ResNet18-SSD 的深度学习加速检测模型如图 1 所示,其采用改进 ResNet18 作为加速检测模型基础网络,使新模型大幅降低计算参数量,同时提高模型检测精度。为适应光线灰暗、图像对比度低、图像模糊、目标重叠等复杂场景,提出多分支同构结构和非对称并行残差结构来提高网络模型对不同尺度图像敏感性和

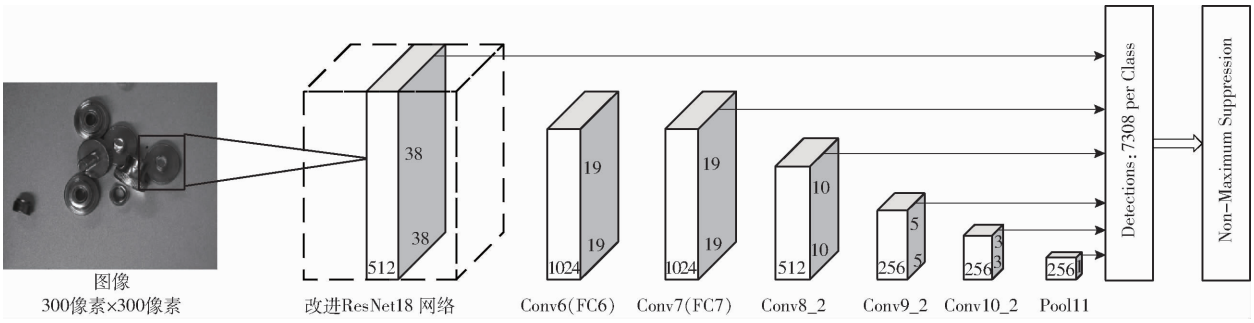

图 1　基于改进 ResNet18 – SSD 的深度学习加速检测模型

Fig. 1　Deep learning accelerated detection model based on improved ResNet18 – SSD

特征提取能力，增强模型鲁棒性。

1.2　ResNet18 网络优化改进

ResNet18 包括带有权重的卷积层和全连接层共 18 层，骨干结构为 1 个 7 ×7 卷基层和由 4 个残差块组成的模块。针对目标检测复杂场景对准确性和实时性的要求，对 ResNet18 网络结构进行优化，首先采用多分支同构结构替换掉 ResNet18 网络的原有卷积层，池化层保留，然后将非对称并行残差卷积引入残差结构，以增强网络多样性，改进 ResNet18 网络如图 2 所示。

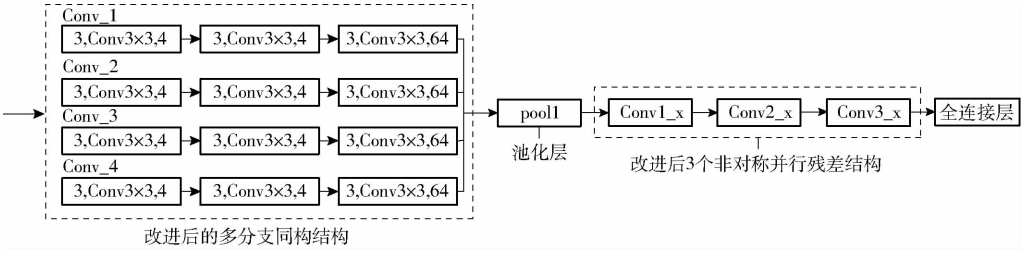

图 2　改进 ResNet – 18 网络结构

Fig. 2　Improved ResNet – 18 network structure

(1)多分支同构结构

为了在不增加通道数条件下提取输入图像更多特征信息，ResNet18 原型网络采用尺寸为 7 ×7 大卷积结构。但在复杂场景下，目标图像存在对比度低、图像模糊、清晰度低等问题，7 ×7 大卷积对复杂目标图像表达能力不强，存在特征提取不完整的情况。因此，提出引入多分支同构结构，共 4 个同构分支，每个分支由尺寸为 3 ×3 的卷积层叠加而成，能够增加网络多样性，提高网络模型对不同尺度图像的敏感性，增强网络表达能力和鲁棒性。

(2)非对称并行残差结构

ResNet18 原型结构中主要有 2 种残差块结构：①面向 64 维图像，采用 2 个堆叠的尺寸为 3 ×3 卷积层实现特征提取，然后将提取到的图像特征与输入特征融合得到输入图像整体特征。②面向 256 维图像，通过一个尺寸为 1 ×1 卷积层对输入图像进行降维，采用尺寸为 3 ×3 卷积层实现特征提取，利用 1 个尺寸为 1 ×1 卷积层对图像维度进行恢复。这 2 种传统残差块结构单一，特征挖掘能力有限，为了进一步提高 ResNet18 网络提取特征能力，本文设计了非对称并行残差结构，如图 3 所示。改进后残差块在原始残差块上增加 2 条并行支路，1 条并行支路由尺寸为 3 ×3 的卷积层叠加批标准化层(Batch normalization，BN)组成，另 1 条并行支路由 1 个尺寸为 3 ×3 卷积层和 2 个尺寸为 1 ×3 卷积叠加组成。2 条并行支路采用非对称卷积操作对图像特征进行提取，提高网络模型对不同尺度图像特征挖掘提取能力，这种方式在提高网络模型精度的同时，可以有效抑制模型过拟合现象，增加模型鲁棒性。

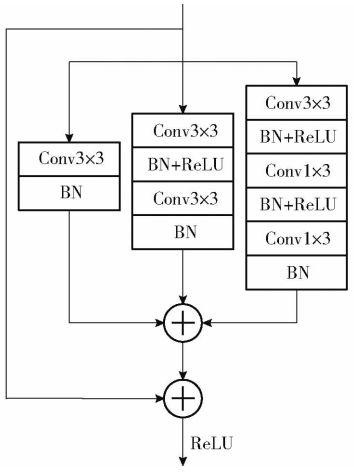

图 3　改进残差块示意图

Fig. 3　Schematic of improved residual block

2　基于 FPGA 的深度学习模型加速方法设计

针对改进 ResNet18 – SSD 模型，在保证精度前

提下,采用 8 位整型动态量化方法来减少参数量,提升卷积神经网络训练和推理效率;针对 ResNet18 - SSD 的部分卷积计算,提出基于 Winograd 算法的卷积硬件加速引擎,大幅更高模型并行计算能力。

2.1 基于动态定点数据量化的加速方法

主流的深度学习框架大多采用 32 bit 或 64 bit 浮点数进行推理计算,在满足精度要求前提下,适当减少参数量可有效缓解 FPGA 因片上资源不足导致的计算缓慢问题。因此,提出采用动态定点数整形量化方法以缩小数据位宽,从而提高 FPGA 神经网络加速器性能。采用动态定点数将所有数据分成若干部分,同一部分数据共享指数。在存储数据时,类似于整型存储操作,将指数部分与其余部分各自单独存储。具体表示方式如图 4 所示。

图 4 动态定点数存储格式

Fig. 4 Dynamic fixed-point number storage format

其计算式为

$$V=(-1)^{S}2^{f_l}\sum 2^{i}F_i \quad (i\in[0,b_w-2]) \tag{1}$$

式中 f_l——共享指数 S——数据位数

b_w——数值部分存储位宽

F_i——尾数位置 i 的 0 或 1

在动态定点数量化时,为了保证量化后数据有效性,加入舍入操作至关重要[18],采用舍入操作有两种:①当模型对层内可训练参数进行训练时,如出现梯度消失时可进行随机舍入。②在模型前向传播过程中为使量化数据尽可能接近真实值,当被舍入部分大于真实值 0.5 则向上取整,反之向下取整。

2.2 基于 FPGA 和 Winograd 算法的卷积加速引擎

典型卷积加速算法有 FFT 和 Winograd 算法[24-25],FFT 算法仅在大卷积核情况下才具有计算优势,而且 FFT 算法中包含的复数运算会加大 FPGA 硬件计算开销。在硬件计算时,乘法器要耗费计算资源相较于加法器会成倍增长,而二维 Winograd 算法可以实现用更少的乘法器来完成卷积计算,以此来提高卷积计算速度。采用 Winograd 算法进行卷积计算引擎的设计。二维 Winograd 算法计算式为

$$\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}[(\boldsymbol{GgG}^{\mathrm{T}})\odot(\boldsymbol{BdB}^{\mathrm{T}})]\boldsymbol{A} \tag{2}$$

式中 $\boldsymbol{A}$、$\boldsymbol{G}$、$\boldsymbol{B}$ 为变换矩阵,为固定值。⊙为点乘操作,对于一个 4×4 特征图和 3×3 卷积核,其变换矩阵为

$$\boldsymbol{G}=\begin{bmatrix}1 & 0 & 0\\ 1/2 & 1/2 & 1/2\\ 1/2 & -1/2 & 1/2\\ 0 & 0 & 1\end{bmatrix}$$

$$\boldsymbol{B}=\begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 1 & 0\\ 0 & -1 & 1 & 0\\ 0 & 1 & 0 & -1\end{bmatrix}\quad \boldsymbol{A}=\begin{bmatrix}1 & 0\\ 1 & 1\\ 1 & -1\\ 0 & -1\end{bmatrix}$$

以模型中 3×3、步长为 1 的卷积为例进行卷积计算引擎设计,当输入特征图为 4×1,卷积核为 3×1,输出结果为 2×1 时(简称 $F(2,3)$),其一维计算式为

$$\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}[(\boldsymbol{Gg})\odot(\boldsymbol{Bd})] \tag{3}$$

式(2)为由一维计算式(3)变换而来。在 FPGA 硬件上实现一维计算公式 $F(2,3)$,在此基础上实现 3×3、步长为 1 卷积计算引擎。将式(3)代入数值并展开可得

$$\boldsymbol{d}=[d_0\ d_1\ d_2\ d_3]^{\mathrm{T}} \tag{4}$$

$$\boldsymbol{g}=[g_0\ g_1\ g_2]^{\mathrm{T}} \tag{5}$$

$$m_1=(d_0-d_2)g_0 \tag{6}$$

$$m_2=(d_1+d_2)\frac{g_0+g_1+g_2}{2} \tag{7}$$

$$m_3=(d_2-d_1)\frac{g_0-g_1+g_2}{2} \tag{8}$$

$$m_4=(d_1-d_3)g_2 \tag{9}$$

$$\boldsymbol{Y}=\begin{bmatrix}m_1+m_2+m_3\\ m_2-m_3-m_4\end{bmatrix} \tag{10}$$

图 5 为 $F(2,3)$ 的权值预处理 FPGA 硬件结构图,其通过 4 个加法器对一维特征图进行计算($\boldsymbol{Bd}$),图 5 通过 3 个加法器和两次移位操作完成对权重与计算过程($\boldsymbol{Gg}$)。经过预处理后的一维特征图与权值通过 4 个乘法器进行点乘(($\boldsymbol{Gg}$)⊙($\boldsymbol{Bd}$)),共包含 8 个加速器,负责将前向传播结果进行后计算($\boldsymbol{A}^{\mathrm{T}}[(\boldsymbol{Gg})\odot(\boldsymbol{Bd})]$)。

在此基础上,式(2)中 $\boldsymbol{g}$ 可以看作是 1×3 的行向量,在行向量每一列上存在 3 个元素,$\boldsymbol{d}$ 同理。每个列向量方向预算即为 $F(2,3)$ 的计算过程。二维与一维公式区别还在于行向量方向有更多的预处理运算($\boldsymbol{gG}^{\mathrm{T}}$,$\boldsymbol{dB}^{\mathrm{T}}$),以及在后处理阶段后再进行一次变换($\times\boldsymbol{A}$)。图 6 为 $F(2\times2,3\times3)$ 硬件总体设计结构。

3 实验

3.1 实验数据集

为验证基于改进 ResNet18 - SSD 和 FPGA 硬件加速的深度学习加速检测模型的有效性,以 Xilinx

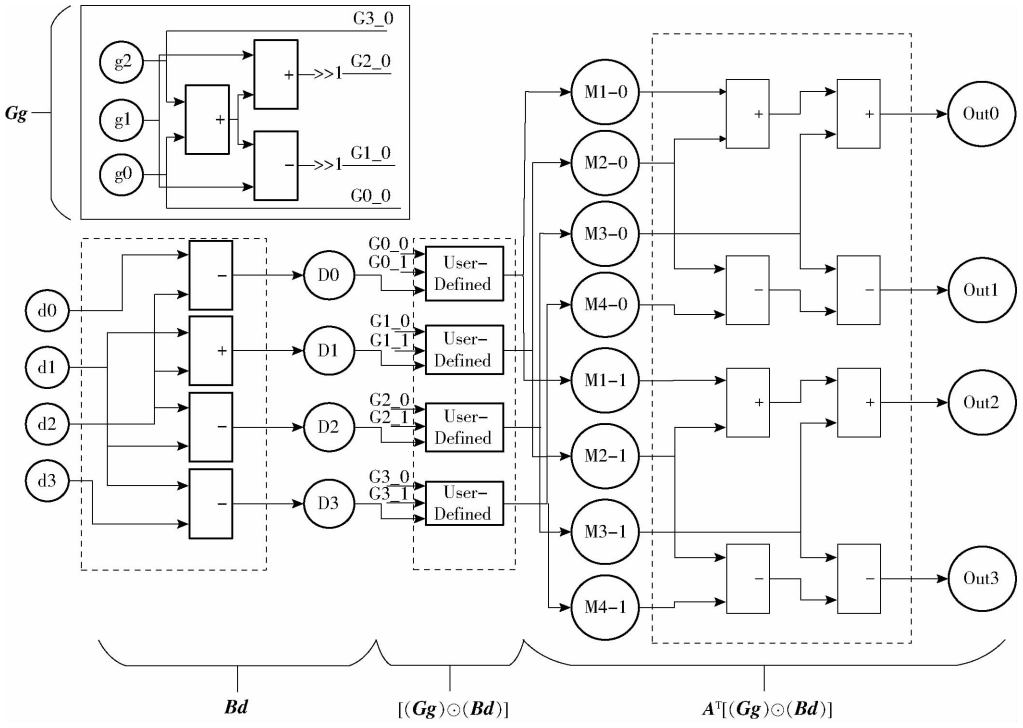

图 5　$F(2,3)$硬件结构设计

Fig. 5　$F(2,3)$ hardware architecture design

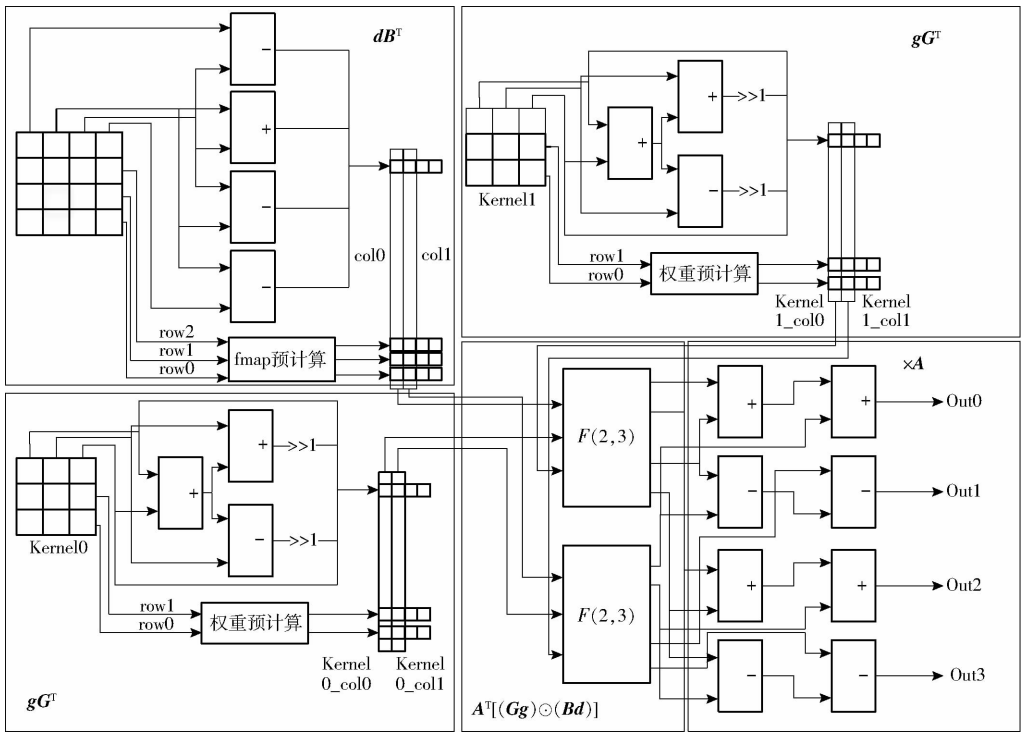

图 6　$F(2\times2,3\times3)$硬件结构设计

Fig. 6　Hardware structure design of $F(2\times2,3\times3)$

公司 XCZU3EG－1SFVC784I 型 FPGA 为嵌入式平台，Xilinx 端采用 Ubuntu 16.04LTS 操作系统，采用 Caffe 深度学习框架。以螺母、螺钉等 5 类零件为检测对象，在零件遮挡、异物干扰、多类堆叠、光照昏暗、图像模糊、对比度低等场景下采集零件图像，共 6 000 幅，随机选取 4 500 幅图像作为训练集，剩余 1 500 幅图像作为测试集。数据集样本种类如图 7 所示，测试集和训练集所含零件种类与数量如表 1 所示。

在数据集制作时，分别模拟农业机械智能装配产线上的无序摆放、零件堆叠、杂物干扰、曝光过度、光线昏暗等复杂场景。采用 LabelImg 工具制作数据集，标注每个零件坐标及类别信息。然后保存每幅零件标注后对应生成的xml信息文件，以记录图

图 7 数据集样本种类

Fig. 7 Types of data set samples

表 1 测试集和训练集所含零件种类与数量

Tab. 1 Types and number of parts included in test set and training set

种类	训练集	测试集
螺母	3 560	828
螺钉	3 523	834
滚花大头螺母	4 212	1 139
滚花平头螺母	3 529	1 348
压铆螺母	3 473	984

像标注信息。

3.2 零件检测效果

(1)单类零件检测效果

单类零件检测效果如图 8 所示,本文方法对单类零件识别准确率超过 96%,在光线不足、光照不均、过量曝光等复杂条件下,仍有很好的检测效果。

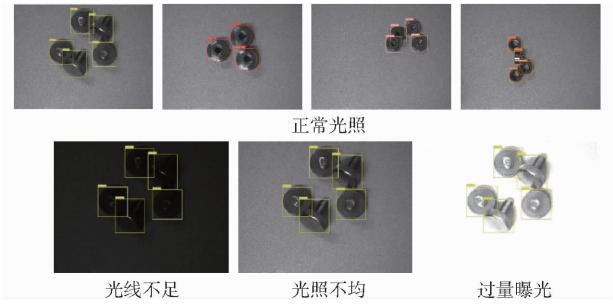

图 8 单类零件识别效果

Fig. 8 Effect of recognizing single type of parts

(2)多类零件检测效果

多类零件检测效果如图 9 所示,在光照昏暗、过度曝光、多目标零件、零件堆叠遮挡、异物干扰等复杂条件下,本文方法依然取得较好的检测效果。零件堆叠遮挡场景下识别率准确不低于 92.5%,过曝/光线不足场景下识别准确率不低于 92.3%,异物干扰场景下识别准确率不低于 89.4%,平均准确率达到 93.5%。每类零件在不同场景下识别准确率如表 2 所示。

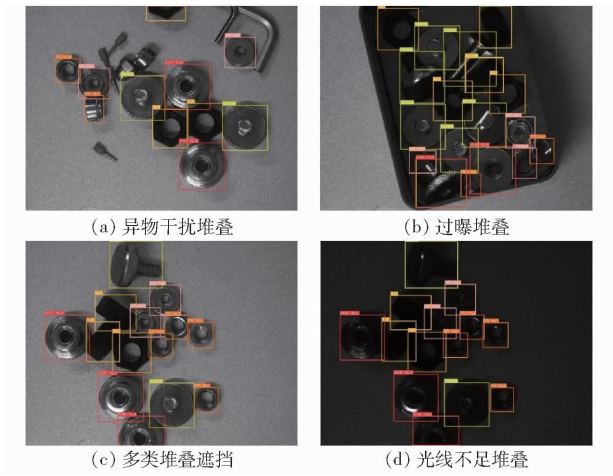

图 9 多类零件混放堆叠和异物干扰检测效果

Fig. 9 Detection effect of mixed stacking of multiple types of parts and foreign object interference

表 2 各种干扰情况下多类零件识别准确率

Tab. 2 Recognition accuracy of multi-class parts in various interference situations %

零件	多类堆叠遮挡	过曝/光线不足堆叠	异物干扰堆叠
螺母	94.7	92.3	94.3
螺钉	93.5	93.6	89.4
滚花大头螺母	93.6	94.5	94.5
滚花平头螺母	92.5	93.3	93.3
压铆螺母	93.3	93.5	93.5

对包括图 9 在内的所有测试样本在不同复杂场景下漏检率、误检率进行统计,最大漏检率为异物干扰堆叠场景下螺钉零件漏检率为 3.84%,最大误检率为异物干扰堆叠场景下螺钉零件误检率为 6.35%,如表 3 所示。

表 3 各种干扰情况下多类零件漏检率与误检率结果

Tab. 3 Results of leakage rate and false detection rate of multi-category parts under various interference situations

种类	测试集	多类堆叠遮挡场景				过曝/光线不足堆叠场景				异物干扰堆叠场景			
		漏检数	误检数	漏检率/%	误检率/%	漏检数	误检数	漏检率/%	误检率/%	漏检数	误检数	漏检率/%	误检率/%
螺母	828	16	28	1.92	3.38	20	44	2.42	5.31	16	31	1.93	3.74
螺钉	834	21	33	2.52	3.96	19	34	2.28	4.08	32	53	3.84	6.35
滚花大头螺母	1 139	32	41	2.81	3.60	22	41	1.93	3.60	23	40	2.02	3.51
滚花平头螺母	1 348	36	65	2.67	4.82	35	55	2.60	4.08	35	55	2.60	4.08
压铆螺母	984	28	38	2.85	3.86	21	43	2.13	4.37	19	45	1.93	4.57

3.3　模型检测实时性

在工作频率 100 MHz 下，以检测速度及传输帧率评估本文模型检测实时性。1 500 幅测试集零件图像（尺寸为 512 像素 ×768 像素）识别稳定平均时间为 80.232 ms，稳定平均帧率为 12.46 f/s，满足工厂实际检测速率要求。模型实时检测界面如图 10 所示。

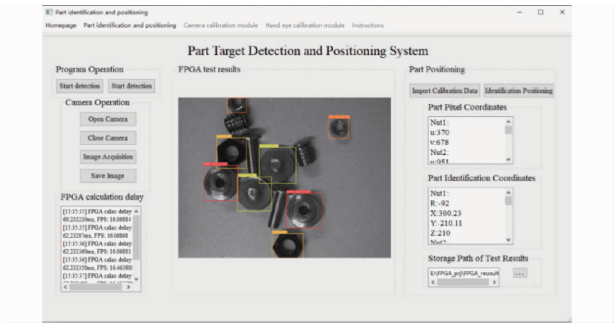

图 10　模型实时检测界面

Fig. 10　Model real-time inspection interface

3.4　不同模型性能评估

评估 VGG－SSD 与本文提出的 ResNet18－SSD 及改进 ResNet18－SSD 模型复杂度，3 种模型计算量和参数量对比如表 4 所示。以零件堆叠遮挡场景下评估 3 种模型检测准确率及平均准确率，如表 5 所示。与 VGG－SSD 模型相比，本文提出的改进 ResNet18－SSD 方法在计算量和参数量分别减少 61.83% 和 25.82%，大幅降低了网络复杂度，而检测平均准确率从 90.8% 提高到 93.5%，体现了本文方法优势。与改进前 ResNet18－SSD 模型相比，改进 ResNet18－SSD 方法在计算量和参数量略有增加，但检测平均准确率从90.8% 提高到 93.5%，对复杂场景有更好的适应性。

表 4　不同模型计算量和参数量对比

Tab. 4　Comparison of computational and parametric quantities for different models

参数	ResNet18－SSD	本文方法	VGG－SSD
计算量	4.536×10^{10}	6.768×10^{10}	1.773×10^{11}
参数量	1.43×10^{7}	1.58×10^{7}	2.13×10^{7}

表 5　不同模型检测准确率

Tab. 5　Detection accuracy of different models　%

模型	螺母	螺钉	滚花大头螺母	滚花平头螺母	压铆螺母	平均准确率
VGG－SSD	91.5	88.2	92.6	85.1	86.2	88.7
ResNet18－SSD	94.6	90.1	93.5	87.8	88.0	90.8
本文方法	94.7	93.5	93.6	92.5	93.3	93.5

本文方法与目前能应用于边缘部署的 3 种轻量化模型 MobileNet、EfficientNet 和 GhostNet 测试结果如表 6 所示，本文方法不管在对单类零件还是多类堆叠零件的检测中均优于其他 3 种。原因为现有轻量化检测模型如 MobileNet、EfficientNet 和 GhostNet 等，其轻量化方法主要集中在降低模型参数量和计算复杂度上，通过减少浮点运算数来降低模型复杂度，但这对检测模型准确性影响较大。而本文方法采取的策略通过软硬件协同，从软件网络轻量化与硬件加速器两个角度进行联合优化。本文检测网络参数量在 4 个轻量化模型中相对较小，计算量远高于其他 3 种模型，但通过本文设计的 FPGA 硬件加速引擎得以解决，保证了模型轻量化、加速性能及复杂场景下准确性三者之间的平衡。

表 6　不同轻量化模型复杂度与准确率

Tab. 6　Complexity and accuracy of different lightweight models

模型	参数量	计算量	单类零件平均准确率/%	多类堆叠零件平均准确率/%
MobileNet V4 Hybrid－L	3.59×10^{7}	7.20×10^{9}	91.1	83.2
GhostNet V2 1.6X	1.23×10^{7}	3.99×10^{8}	85.8	80.6
EfficientNet v2－s	2.40×10^{7}	8.80×10^{9}	91.6	83.8
本文方法	1.58×10^{7}	6.77×10^{10}	96.5	93.5

4　结论

（1）提出了多分支同构结构和非对称并行残差结构来改进 ResNet18 网络，提高网络对不同尺度图像特征的挖掘能力和鲁棒性。通过引入改进 ResNet18 替换 SSD 模型的基础网络即 VGG16 前置网络，大幅降低了参数量，同时提高了检测精度。

（2）采用动态定点量化的方式降低对 FPGA 片上资源的要求，针对 3 × 3 的卷积层设计了其 Winograd 快速卷积算法的硬件加速引擎，实现了改进 ResNet18－SSD 模型移植于 FPGA，执行效率大幅提高。

（3）FPGA 实验结果表明，当工作频率为 100 MHz，单幅图像检测时间为 80.232 ms，复杂情况下平均准确率能够达到 93.5%，满足实际场景应用需求。

参　考　文　献

[1]　亢洁，刘港，郭国法，等. 基于多尺度融合模块和特征增强的杂草检测方法[J]. 农业机械学报，2022，53(4)：254－260.
KANG Jie, LIU Gang, GUO Guofa. Weed detection based on multi-scale fusion module and feature enhancement[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(4): 254－260. (in Chinese)

[2]　朱红春，李旭，孟炀，等. 基于 Faster R－CNN 网络的茶叶嫩芽检测[J]. 农业机械学报，2022，53(5)：217－224.
ZHU Hongchun, LI Xu, MENG Yang, et al. Tea bud detection based on Faster R－CNN network[J]. Transactions of the

Chinese Society for Agricultural Machinery, 2022,53(5):217 - 224. (in Chinese)
[3] 修春波,孙乐乐. 基于改进 YOLO v4 网络的马铃薯自动育苗叶芽检测方法[J]. 农业机械学报,2022,53(6):265 - 273.
XIU Chunbo, SUN Lele. Potato leaf bud detection method based on improved YOLO v4 network[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022,53(6):265 - 273. (in Chinese)
[4] JRR U, KEAVD S, GEVER T, et al. Selective search for object recognition[J]. International Joural of Computer Vison,2013, 103(2):154 - 171.
[5] REN S, HE K, GIRSHICK R, et al. Faster R - CNN: towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems,2015:91 - 99.
[6] 李翠明,杨柯,申涛,等. 基于改进 Faster R - CNN 的苹果采摘视觉定位与检测方法[J]. 农业机械学报,2024,55(1):47 - 54.
LI Cuiming, YANG Ke, SHEN Tao, et al. Vision detection method for picking robots based on improved Faster R - CNN[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024,55(1):47 - 54. (in Chinese)
[7] 刘毅君,何亚凯,吴晓媚,等. 基于改进 Faster R - CNN 的马铃薯发芽与表面损伤检测方法[J]. 农业机械学报,2024, 55(1):371 - 378.
LIU Yijun, HE Yakai, WU Xiaomei, et al. Potato sprouting and surface damage detection method based on improved Faster R - CNN[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024,55(1):371 - 378. (in Chinese)
[8] 黄成龙,张忠福,华向东,等. 基于改进 Faster R - CNN 和 Deep Sort 的棉铃跟踪计数[J]. 农业机械学报,2023,54(6):205 - 213.
HUANG Chenglong, ZHANG Zhongfu, HUA Xiangdong, et al. Cotton boll tracking and counting based on improved Faster R - CNN and Deep Sort[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023,54(6):205 - 213. (in Chinese)
[9] 余永维,韩鑫,杜柳青. 基于 Inception - SSD 算法的零件识别[J]. 光学精密工程,2020,28(8):1799 - 1809.
YU Yongwei, HAN Xin, DU Liuqing. Target part recognition based Inception - SSD algorithm[J]. Optics and Precision Engineering, 2020,28(8):1799 - 1809. (in Chinese)
[10] 张立杰,周舒骅,李娜,等. 基于改进 SSD 卷积神经网络的苹果定位与分级方法[J]. 农业机械学报,2023,54(6):223 - 232.
ZHANG Lijie, ZHOU Shuhua, LI Na, et al. Apple location and classification based on improved SSD convolutional neural network[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023,54(6):223 - 232. (in Chinese)
[11] 屈志坚,张博语,杨行,等. 基于 MFDC - SSD 网络的接触网定位线夹缺陷识别[J]. 铁道学报,2024,46(5):48 - 57.
QU Zhijian, ZHANG Boyu, YANG Hang, et al. Fault identification of overhead contact system steady ears based on MFDC - SSD network[J]. Journal of the China Railway Society, 2024,46(5):48 - 57. (in Chinese)
[12] 吴启航,丁晓喜,何清波,等. 齿轮箱故障边缘智能诊断方法及应用研究[J]. 仪器仪表学报,2024,45(1):70 - 80.
WU Qihang, DING Xiaoxi, HE Qingbo, et al. Edge intelligent fault diagnosis method in the application of gearbox[J]. Chinese Journal of Scientific Instrument, 2024,45(1):70 - 80. (in Chinese)
[13] WANG Y, YAN J, SUN Q, et al. Bearing intelligent fault diagnosis in the industrial internet of things context: a lightweight convolutional neural network[J]. IEEE Access, 2020,8:87329 - 87340.
[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] HUANG Y, SHEN J, WANG Z, et al. A high-efficiency FPGA-based accelerator for convolutional neural networks using winograd algorithm[J]. Journal of Physics Conference Series, 2018, 1026:012019.
[16] 杨宁,程巍,张钊源,等. 基于 FPGA 加速的 Mask R - CNN 稻瘟病高通量自适应识别模型研究[J]. 农业机械学报,2024, 55(7):298 - 304,314.
YANG Ning, CHENG Wei, ZHANG Zhaoyuan, et al. Research on high-througput adaptive recognition Mask R - CNN model for rice blast disease based on FPGA acceleration[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(7):298 - 304,314. (in Chinese)
[17] 徐欣,刘强,王少军. 一种高度并行的卷积神经网络加速器设计方法[J]. 哈尔滨工业大学学报,2020,52(4):31 - 37.
XU Xin, LIU Qiang, WANG Shaojun. A highly parallel design method for convolutional neural networks accelerator[J]. Journal of Harbin Institute of Technology, 2020,52(4):31 - 37. (in Chinese)
[18] QIU J T, WANG J, YAO S, et al. Going deeper with embedded FPGA platform for convolutional neural network[C]// Proceedings of 2016 ACM/SIGDA International Symposium on Field Programmable Gate Arrays, 2016.
[19] LI H M, FAN X T, JIAO L, et al. A high performance FPGA-based accelerator for large-scale convolutional neural networks [C]// Proceedings of the 26th International Conference on Field Programmable Logic and Applications, 2016.
[20] HUANG C, NI S Y, CHEN G S. A layerbased structured design of CNN on FPGA[C]// Proceedings of the 12th International Conference on ASIC. IEEE, 2017.
[21] 陈朋,陈庆清,王海霞,等. 基于改进动态配置的 FPGA 卷积神经网络加速器的优化方法[J]. 高技术通讯,2020,30(3): 240 - 247.
CHEN Peng, CHEN Qingqing, WANG Haixia, et al. Optimization method of FPGA convolutional neural network accelerator based on improved dynamic configuration[J]. Chinese High Technology Letters, 2020,30(3):240 - 247. (in Chinese)
[22] 谢坤鹏,仪德智,刘义情,等. SAF - CNN:面向嵌入式 FPGA 的卷积神经网络稀疏化加速框架[J]. 计算机研究与发展, 2023,60(5):1053 - 1072.
XIE Kunpeng, YI Dezhi, LIU Yiqing, et al. SAF - CNN: a sparse acceleration framework of convolutional neural network for embedded FPGAs[J]. Journal of Computer Research and Development, 2023,60(5):1053 - 1072. (in Chinese)
[23] 李天阳,张帆,王松,等. 基于 FPGA 的卷积神经网络和视觉 Transformer 通用加速器[J]. 电子与信息学报,2024,46(6): 240 - 247.
LI Tianyang, ZHANG Fan, WANG Song, et al. FPGA-based unified accelerator for convolutional neural network and vision Transformer[J]. Journal of Electronics & Information Technology, 2024,46(6):240 - 247. (in Chinese)
[24] 童敢,黄立波,吕雅帅. 面向现代 GPU 的 Winograd 卷积加速研究[J]. 电子学报,2024,52(1):244 - 257.
TONG Gan, HUANG Libo, LÜ Yashuai, et al. Research on Winograd convolution acceleration for modern GPU[J]. Acta Electronica Sinica, 2024,52(1):244 - 257. (in Chinese)
[25] 梅冰笑,滕文彬,张弛,等. FPGA 平台上动态硬件重构的 Winograd 神经网络加速器[J]. 计算机工程与应用,2024, 60(22):323 - 334.
MEI Bingxiao, TENG Wenbin, ZHANG Chi, et al. Winograd neural network accelerator using dynamic hardware reconfiguration on FPGA platform[J]. Computer Engineering and Applications, 2024,60(22):323 - 334. (in Chinese)