DOI:10.19651/j.cnki.emt.2415965

基于改进 YOLOv8 的 SOP 芯片缺陷检测研究*

彭鸿瑞[1] 杨桂华[2]
(1.桂林理工大学机械与控制工程学院 桂林 541006; 2.广西高校先进制造与自动化技术重点实验室 桂林 541006)

摘 要: 针对 SOP 芯片缺陷检测中因缺陷特征相似、缺陷目标小、缺陷尺度差异大造成的检测精度低的问题,本文提出基于改进 YOLOv8 的缺陷检测方法。通过使用 SPD-Conv 模块解决卷积池化过程中的信息丢失问题,并引入 SimAM 注意力机制,使模型学习三维通道中的信息,提高模型对缺陷特征的感知能力;同时使用 BiFPN 代替原特征提取网络,使用双向传递的多尺度特征融合,使模型能更好的区分拥有相似特征和尺度差异大的缺陷;最后增加一个小目标检测头,传递更多的低阶特征信息给高维检测网络,提高对小目标缺陷的检测效果。实验数据表明,该模型相比原模型 mAP@0.5 提高了 5.4%,mAP@0.95 提高了 4.3%,召回率提高了 3%,和其他模型相比有着显著优势。泛化实验中改进算法的 mAP@0.5 相比原模型也提升了 2.7%,并设计了相关系统验证了算法的有效性。

关键词: 芯片缺陷检测;YOLOv8 模型;SPD-Conv;SimAM;BiFPN
中图分类号: TN407 **文献标识码**: A **国家标准学科分类代码**: 510.40

Research on defect detection of SOP chip based on improved YOLOv8

Peng Hongrui[1] Yang Guihua[2]
(1. College of Mechanical and Control Engineering, Guilin University of Technology, Guilin 541006, China;
2. Key Laboratory of Advanced Manufacturing and Automation Technology (Guilin University of Technology), Education Department of Guangxi Zhuang Autonomous Region, Guilin 541006, China)

Abstract: Aiming at the low detection accuracy caused by similar defect features, small defect target and large difference in defect scale in SOP chip defect detection, this paper proposes a defect detection method based on improved YOLOv8. The problem of information loss in the process of convolution pooling is solved by using SPD-Conv module. And introducing the SimAM attention mechanism, the model can learn the information in the 3D channel and improve the model's perception of defect features. At the same time, BiFPN was used to replace the original feature extraction network, and multi-scale feature fusion was used to enable the model to better distinguish the defects with similar features and large-scale differences. Finally, a small target detection header is added to transmit more low-order feature information to the high-dimensional detection network to improve the detection effect of small target defects. Experimental data show that compared with the original model mAP@0.5/% increased by 5.4%, mAP@0.95/% increased by 4.3%, recall rate increased by 3%, has significant advantages compared with other models. In the generalization experiment, the mAP@0.5 of the improved algorithm is also improved by 2.7% compared with the original model, and a relevant system is designed to verify the effectiveness of the algorithm.

Keywords: chip defect detection; YOLOv8 model; SPD-Conv; SimAM; BiFPN

0 引 言

近年来,中国高新技术产业迅速发展,半导体设备市场也在快速扩张。其中半导体生产过程中复杂,微小的变化都可能导致产品缺陷。若是出现划痕或缺脚等各种封装缺陷,对芯片的实际使用会造成很大影响,因此芯片质量检测至关重要[1]。传统的人工目视检测方法效率低下且精度相对较低,而基于机器视觉和深度学习卷积神经网络的检测手段成为新的主流。基于机器视觉的传统目标检测需要手动提取特征,不仅操作复杂,检测速度慢,而且检测精度不高。基于深度学习的目标检测可以通过神经网络学习数据特征,能提取到更复杂、更小的特征,可以有效提升检测准确度。

收稿日期:2024-04-30
* 基金项目:国家自然科学基金地区基金(52065016)项目资助

深度学习自深度卷积神经网络的提出之后开始广泛进入大众的视野，并逐渐应用在工业检测上。其中可以大致分为一阶段模型和两阶段模型。其中一阶段模型具有检测速度快，但精确度较低的特点，其中的代表算法有你只看一次[2](you only look once，YOLO)系列、单步多框检测器[3](single shot multibox detector，SSD)等，两阶段模型在检测精度上要高于一阶段模型，但检测速度较低，以基于区域的卷积神经网络[4](region-based convolutional neural network，R-CNN)系列为代表。罗月童等[5]提出将芯片表面缺陷看作噪音，使用卷积去噪自编码器重构无缺陷图像，同时使用重叠分块策略抑制重构噪音，对弱缺陷实现了很好的检测效果。刘艳菊等[6]提出对深层特征进行反卷积，将深层特征信息与浅层特征信息融合，改善了语义信息不足导致的误检问题。薛阳等[7]提出使用聚合算法确定不同尺度特征图的锚框，提高对不同背景的抗干扰能力，有效提高小目标缺陷的检测效果。但对于特征差异较大的缺陷，其检测准确度较低。周天宇等[8]提出通过增强特征融合的方式优化应用于载波芯片的缺陷多尺度检测，更好地融合了浅层特征图和深层特征图的信息，但这种方式对于特征复杂的缺陷检测效果不佳。Wang 等[9]提出一种将注意力机制与 YOLOv4 相结合的目标检测网络，该网络通过分割模型的校正能突出显示特征图中的缺陷区域，从而更有效的识别微小缺陷。由于只使用了空间注意力机制，对于三维通道中的信息不能得到有效保留，特征提取能力较差。张恒等[10]提出基于面积的边界框融合算法，解决了预测阶段冗余框较多，时间开销大的问题。但对于原网络中卷积池化过程造成的信息丢失问题并没有很好的解决办法。

现有的芯片表面缺陷检测过程中还存在着诸多问题，比如缺陷尺寸差距大、缺陷特征相似、小目标检测效果差等问题。以往的研究人员大多专注于解决芯片缺陷检测中的单一问题，并未给出较为全面的解决方案。本文对上述缺陷进行综合考虑，同时兼顾检测的实时性，在 YOLOv8 模型的基础上提出改进方法。本文的主要改进如下：

1) 针对小目标缺陷难检测和图像分辨率低的问题，本文使用对称正定卷积神经网络[11](symmetric positive definite convolutional neural network，SPD-Conv)结构来代替原始卷积网络中的跨步卷积和池化层，保留了细粒度信息，提升了特征学习效率。同时引入了简单参数自由注意力模块[12](simple parameter-free attention module，SimAM)注意力机制，通过推断特征图中的三维注意力权重工作。最后增加了一个分辨率更低的小目标检测头，加强对小目标的识别能力。

2) 针对芯片缺陷尺寸变化大，特征相似的问题，本文使用加权双向特征金字塔网络[13](bidirectional feature pyramid network，BiFPN)结构。通过引入可学习的权重来学习不同的输入特征，同时重复应用自下而上和自上而下的多尺度特征融合策略，增强对不同尺寸缺陷和相似缺陷的检测能力。

1　本文改进的 YOLOv8 网络

本文改进的 YOLOv8 网络模型结构如图 1 所示。网络采用 SPD-Conv 替代卷积步长和池化层，在下采样过程中避免丢失可学习信息。此外在空间金字塔池化之前引入 SimAM 注意力机制，使得图像的空间和通道参数可以被协同关注，提高网络检测精度。同时使用一种特征金字塔网络 BiFPN，对各层次特征进行融合，更好地获取多尺度信息并提高模型抗干扰能力。最后针对小目标缺陷，增加一个分辨率更低的小目标检测头，提高对小目标缺陷的检测能力。

1.1　主干网络 SPD-Conv 模块

在检测小目标缺陷时，小目标缺陷在整个图像中所占的像素比例较低，并且有些图像本身分辨率不高。同时，在卷积神经网络的早期层中使用跨步卷积和池化造成的细粒度信息丢失，能给模型提供的可学习信息更少，导致小目标缺陷往往难以被准确检测。

为了解决这个问题，本文引入 SPD-Conv 模块代替原始卷积网络中的卷积步长和池化层。SPD-Conv 是由一个空间深度层(space-to-depth，SPD)和一个无步长卷积层构成。SPD 推广了一种图像转换技术[14]对特征映射 X 进行下采样，同时最大限度的保留通道维度中的所有信息。此外，在每个 SPD 之后添加了一个无步长卷积操作，此操作可以通过可学习参数减少额外增加的卷积层中的通道数量。

本文对任意大小为 $S\times S\times C_1$ 的中间特征映射 X 的子特征映射序列进行切片：

$$\begin{aligned}
&f_{0,0}=X[0:S:scale,0:S:scale],\\
&f_{1,0}=X[1:S:scale,0:S:scale],\cdots,\\
&f_{scale-1,0}=X[scale-1:S:scale,0:S:scale];\\
&f_{0,1}=X[0:S:scale,1:S:scale],f_{1,1},\cdots,\\
&f_{scale-1,1}=X[scale-1:S:scale,1:S:scale];\\
&\vdots\\
&f_{0,scale-1}=X[0:S:scale,scale-1:S:scale],f_{1,scale-1},\cdots,\\
&f_{scale-1,scale-1}=X[scale-1:S:scale,scale-1:S:scale].
\end{aligned}\tag{1}$$

一般给定任意原始特征映射 X，子映射可以由构成 X 的所有特征映射按比例整除组成。因此每个子图都可以按一个比例对 X 进行下采样。比如当 $scale=2$ 时，可以得到 4 个子映射，每个映射的维度为 $(S/2,S/2,C_1)$。接下来沿着通道维度将子特征映射连接起来得到一个新的特征映射 X'，X' 在空间维度上相比 X 减少了 $scale$ 倍，在通道维度上增加了一个 $scale^2$ 倍，图 2 为当 $scale=2$ 时 SPD-Conv 的示意图。

在 SPD 特征转换层之后，使用一个具有 $C_2(stride=1)$ 过滤器的非跨步卷积层连接以最大限度地保留所有的判别

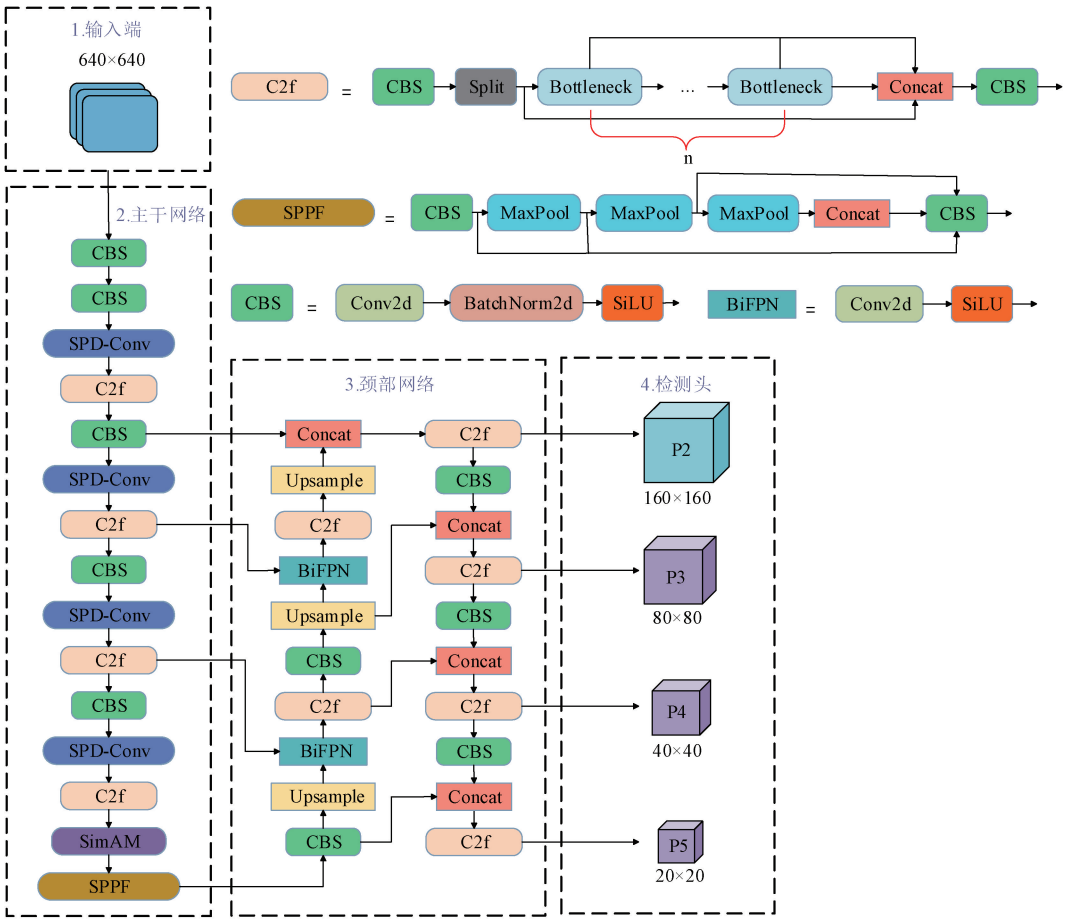


图 1　本文改进的 YOLOv8 网络结构

Fig. 1　The improved YOLOv8 network structure proposed in this paper

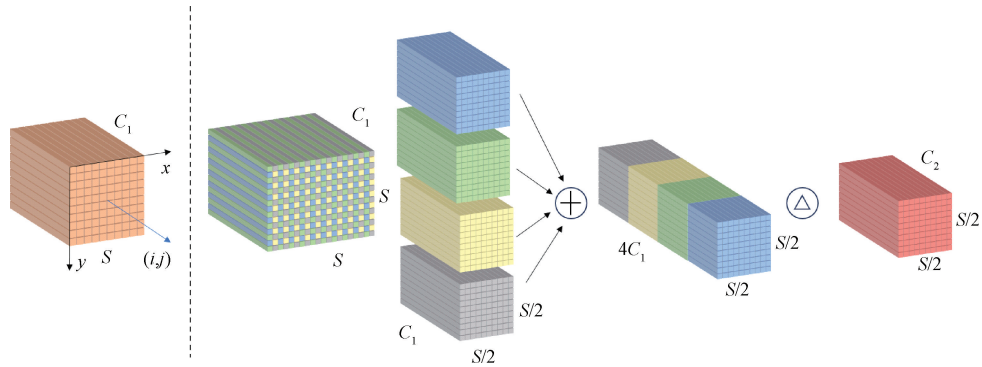


图 2　$scale = 2$ 时 SPD-Conv 的示意图

Fig. 2　Schematic diagram of SPD-Conv at $scale = 2$

特征信息。其中 $C_2 < scale^2 C_1$，并进一步转换：

$$X'\left(\frac{S}{scale},\frac{S}{scale},scale^2C_1\right)\rightarrow X''\left(\frac{S}{scale},\frac{S}{scale},C_2\right) \quad (2)$$

1.2　SimAM 注意力机制

注意力机制一般可以划分为空间注意力机制和通道注意力机制。空间注意力机制通过对图片中的每个像素点进行打分，然后对打分结果加权平均，从而得到关注度更高的区域。通道注意力机制则是对不同通道之间的信息进行关联整合，增强模型表征能力。相比于传统的空间注意力机制和通道注意力机制，SimAM 注意力机制提出了一种统一权值的注意力模块，可以直接计算三维的注意力权重。三维注意力机制的融合过程如图 3 所示。

SimAM 注意力机制结合了视觉神经科学中神经元的概念，每个神经元都有不同的重要性。其中的活跃神经元

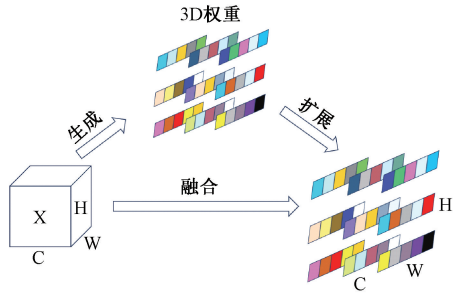


图 3　SimAM 注意力机制结构图

Fig. 3　SimAM attention mechanism structure diagram

可以抑制周围神经元的活动，所以对于显示出明显空间抑制效应的神经元应该被给予更高的优先级。因此，为每个神经元定义了以下能量函数：

$$e_t(w_t,b_t,y,x_i)=(y_t-\hat{t})^2+\frac{1}{M-1}\sum_{i=1}^{M-1}(y_0-\hat{x}_i)^2 \tag{3}$$

式中：e_t 表示目标神经元 t 的能量函数；w_t 与 b_t 表示线性变换的权值与偏置；y 表示神经元的输出；y_t 表示目标神经元的输出；y_0 表示其他神经元的输出；x_i 表示其余神经元；M 代表该通道上神经元的个数。$\hat{t}$ 与 $\hat{x}_i$ 由 t 和 x_i 经线性变化而来，相关的线性变化如式(4)所示。

$$\begin{cases}\hat{x}_i=x_iw_t+b_t\\ \hat{t}=tw_t+b_t\end{cases} \tag{4}$$

求解能量函数的最小值可以找到目标函数 t 与同一通道中其余神经元之间的线性可分性，通常 y_t 和 y_0 采用二值化标签，即−1 和 1。同时将正则化添加到等式中，最终的能量函数由式(5)～(7)给出：

$$e_{t1}(w_t,b_t,y,x_i)=\frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_tx_i+b_t))^2 \tag{5}$$

$$e_{t2}(w_t,b_t,y,x_i)=(1-(w_tt+b_t))^2+\lambda w_t^2 \tag{6}$$

$$e_t(w_t,b_t,y,x_i)=e_{t1}(w_t,b_t,y,x_i)+e_{t2}(w_t,b_t,y,x_i) \tag{7}$$

通过式(8)得到一个关于 w_t 和 b_t 的快速闭合形式解。

$$\begin{cases}w_t=-\dfrac{2(t-\mu_t)}{(t-\mu_t)^2+2\sigma_t^2+2\lambda}\\ b_t=-\dfrac{1}{2}(t+\mu_t)w_t\end{cases} \tag{8}$$

式中：$\mu_t=\dfrac{1}{M-1}\sum_{i=1}^{M-1}x_i$ 和 $\sigma_t^2=\dfrac{1}{M-1}\sum_{i}^{M-1}(x_i-\mu_t)^2$ 是在该通道中除 t 以外的所有神经元上计算的均值和方差。因为式(7)中所示的解是在单个通道上获得的，所以假设单个通道中的所有像素都遵循相同的分布，在此基础上计算所有神经元上的均值和方差，并将其重新运用在该通道上的所有神经元。故最小能量可以用式(9)计算：

$$e_t^*=\frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\sigma}2+2\lambda} \tag{9}$$

式中：$\hat{\mu}=\dfrac{1}{M}\sum_{i=1}^{M}x_i$ 和 $\hat{\sigma}^2=\dfrac{1}{M}\sum_{i=1}^{M}(x_i-\hat{\mu})^2$，表示所有神经元的平均值和方差。式(9)表明能量越低，神经元 t 与周围神经元的区别就越大，其权值就越高。式(10)为整个模块的细化阶段：

$$\tilde{X}=\mathrm{sigmoid}(\frac{1}{E})\odot X \tag{10}$$

式中：sigmoid 是为了限制 $\dfrac{1}{E}$ 中的过大值；E 为跨越通道和空间维度的所有 e_t^* 的总和；X 为图像输入的特征；$\tilde{X}$ 为对输入特征进行加权处理后的结果。

1.3　特征融合网络 BiFPN

使用卷积神经网络提取多层次特征时，对于高层次网络，其感受野大，语义表征能力强，但特征图分辨率较低，缺乏空间几何特征。相反，低层次网络的感受野虽然小，但特征图分辨率大，空间几何特征丰富。所以为了提升检测效果，需要对多层次特征进行特征融合。Tsung-Yi Lin 等于 2016 年提出的特征金字塔网络[15](feature pyramid networks，FPN)被广泛用于多尺度特征融合，随后路径聚合网络[16](path aggregation network，PANet)、神经架构搜索特征金字塔网络[17](neural architecture search feature pyramid networks，NAS-FPN)等更多跨尺度特征融合网络结构研究也被相继提出。FPN 会在升降维度的过程中造成信息损失，且只有自顶向下的过程，很难将低层的信息传递到最后一层。PANet 在 FPN 的基础上多了一个自底向上的传递过程，让图像的底层信息也能得到有效传递。NAS-FPN 在 FPN 的基础上用神经架构搜索来找到一个不规则的特征网络拓扑，然后重复应用，但这样融合输出特征的贡献通常是不均匀的。

为了解决该问题，引入了一种加权双向特征金字塔网络，通过使用可学习的权值来区分不同输入特征的重要性，同时多次使用自顶向下和自底向上的多尺度特征融合。BiFPN 首先切除了网络中对整体的特征网络贡献比较小的结点，其次在处于同一级别的原始输入和输出节点中间添加一条额外的信息传送路径，这样可以在不增加额外成本的情况下融合更多信息。最后将每一个双向路径都视为一个特征网络层，并在特征融合过程中多次调用，以此来达到更好的特征融合效果，改善因缺陷特征相似造成的误检。BiFPN 特征网络结构如图 4 所示。

特征融合过程中，由于不同的输入特征具有不同的分辨率，因此它们通常对输出特征的贡献不相等。芯片表面缺陷中字符粘连和部分表面划痕缺陷尺寸差距很大，对此 BiFPN 采用快速归一化融合模块来平衡不同特征的权重，减少因特征尺寸差距过大造成的误检和漏检。其输入与输出的关系式为：

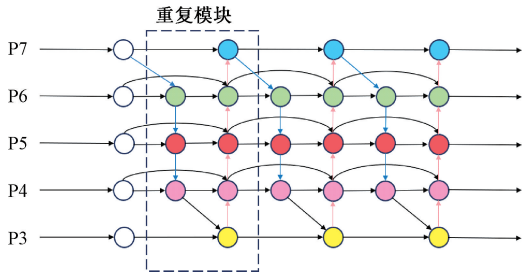


图 4　BiFPN 特征网络结构

Fig. 4　BiFPN feature network architecture

$$O=\sum_i \frac{\omega_i}{\varepsilon+\sum_j \omega_j}\cdot I_i \tag{11}$$

式中：I_i 为输入特征图；ω_i 为输入特征 I_i 对应的学习权重，其中每个 ω_i 后使用 RELU 来确保 $\omega_i \geqslant 0$。同时设置 $\varepsilon=0.0001$，以避免数值不稳定。通过上式使归一化权重的值在 0～1 之间下降。以 BiFPN 的 P_5 特征层为例，P_5 特征层的中间特征层 P_5^{td} 和输出特征层 P_5^{out} 分别为：

$$P_5^{td}=Conv\left(\frac{\omega_1\cdot P_5^{in}+\omega_2\cdot Resize(P_6^{in})}{\omega_1+\omega_2+\varepsilon}\right) \tag{12}$$

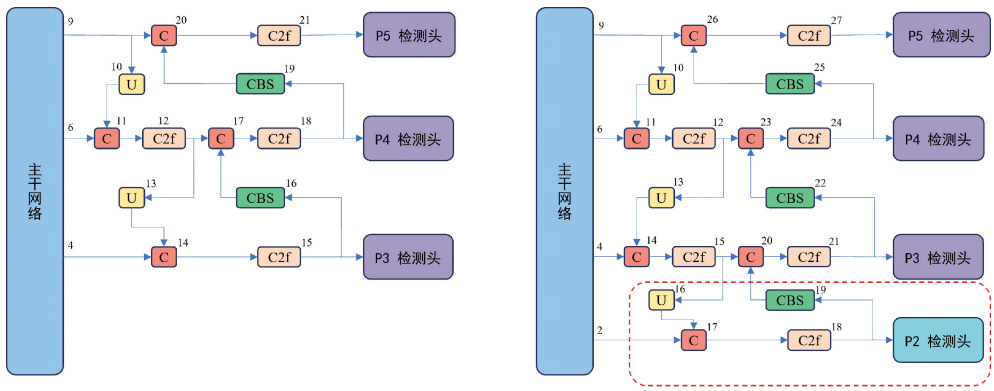
$$P_5^{out}=Conv\left(\frac{\omega'_1\cdot P_5^{in}+\omega'_2\cdot P_5^{td}+\omega'_3\cdot Resize(P_4^{out})}{\omega'_1+\omega'_2+\omega'_3+\varepsilon}\right) \tag{13}$$

式中：P_5^{in} 为第 5 级输入特征，*Resize* 为上采样或下采样操作。

1.4　小目标检测头

YOLOv8 中包含有 3 个检测头，分别对应检测大、中、小尺寸的目标，如图 5(a)所示。然而在检测尺寸过小的目标时会出现检测效果不佳的情况。尺寸过小的目标在模型加深的过程中会丢失部分低阶特征信息，例如纹理、边缘等信息。新增加的检测头 P2 包含丰富的低阶特征信息，其对应的检测特征图大小为 160×160，可以为其他的高维特征补充更多细节信息，优化其他检测头的检测效果。

从图 5(b)可以看出，P2 检测头融合了不同层次的特征图，即 C2f 处理的第 4 层和第 2 层特征图和空间金字塔汇集的第 9 层骨干网络特征图。这些特征图从上到下融合，并通过上采样进行处理，形成深度语义特征层。这个深度语义特征层通过 C2f 模块和一个额外的解耦头输出。接下来，经过一系列 Conv、Concat 和 C2f 操作后，微小目标的低级特征信息将被传输到其他 3 个尺度特征层，从而提高对小缺陷的检测。通过使用新的检测头，提高了检测芯片表面划痕和字符粘附等小缺陷的能力，使模型能够检测更小的目标。



(a) 原始检测网络
(a) Original detection network

(b) 增加了P2检测头
(b) Adding P2 detection head

图 5　检测网络增加 P2 检测头

Fig. 5　Detection network with added P2 detection head

2　实验设计

2.1　数据集构建

数据集使用 500 万像素的相机搭配焦距为 35 mm、分辨率为 500 万的镜头在环形光源下进行拍摄。图像分辨率为 2 592×1 944，共 396 张。

通过 LabelImg 标注工具进行人工标注，数据集包含 6 个类别，分别是正常、引脚上翘、引脚缺失、引脚弯曲、字符粘连、表面划痕。同时采用改变亮度、裁剪、平移、运动模糊、加噪声、翻转的方式对数据集进行扩充增强，最终数

表 1　芯片缺陷数量分布表

Table 1　Chip defect number distribution table

类别	数量
正常	492
引脚上翘	1 071
引脚弯曲	1 332
引脚缺失	896
字符粘连	590
表面划痕	1 718

据集共 2 772 张，再按照 8∶2 的比例将数据集划分为训练集和验证集。每种芯片缺陷类别具体数量如表 1 所示，部分芯片缺陷的图例如图 6 所示。

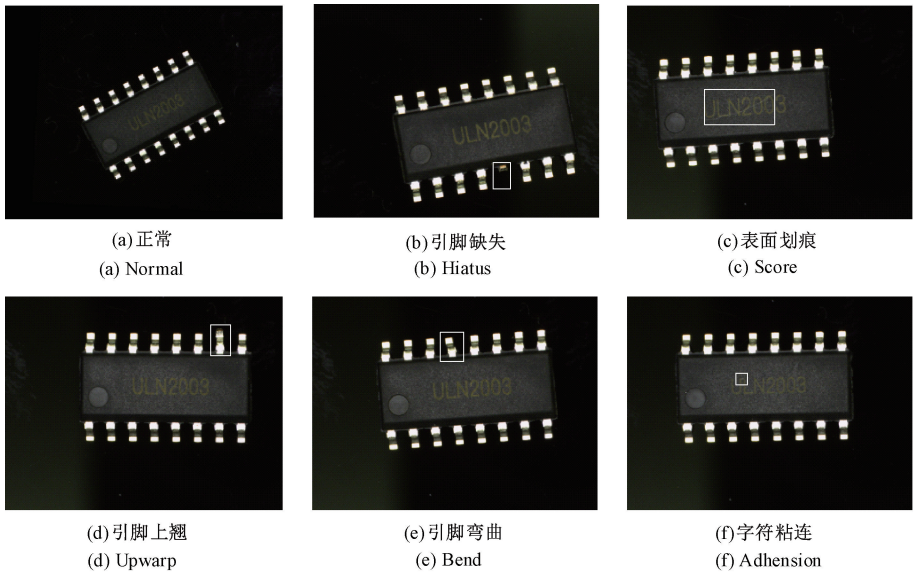


图 6　芯片缺陷样本

Fig. 6　The chip defect samples

2.2　实验环境配置

本实验操作系统为 Windows10，GPU 型号 NVIDIA GeForce RTX 4060 Ti，显存 8 G。CPU 型号 AMD Ryzen 5 5600G with Radeon Graphics。深度学习框架选择 python-3.8.18、cuda-11.7、pytorch-2.0.1，学习率设置为 0.01，批次大小设置为 16，训练周期 300 轮。为了确保实验结果的可靠性与准确性，所有实验均不使用预训练权重。

2.3　模型评价标准

为了验证模型对芯片缺陷的检测性能，本研究采用了精确度（precision，P）、召回率（recall，R）、平均精度（average presion，AP）、平均精度均值（mean average presion，mAP）、浮点运算量（floating-point operation，FLOPs）、参数量（Parameters）、检测速度（frames per second，FPS）等评价指标。

精确度的含义是在全部预测为正的结果中，被预测正确的正样本所占的比例，即有没有误检。召回率的含义是预测正确的正样本在所有正样本中所占的比例，即有没有漏检。平均精度 AP 是基于精确度和召回率进行计算的，是模型在所有置信度阈值（intersection over union，IoU）下的值。平均检测均值 mAP@0.5 指 IoU 设为 0.5 时的 mAP；mAP@0.5∶0.95 表示 IoU 在 0.5 到 0.95 上的 mAP。IoU 表示交并比，即预测框和真实框相交区域面积和两者合并区域面积的比值。FLOPs 用于测量模型运行的时间，值越低代表模型越小，模型计算所需的时间也越少。计算公式如下：

$$P = \frac{T_P}{T_P + F_P} \tag{14}$$

$$R = \frac{T_P}{T_P + F_N} \tag{15}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{16}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP \tag{17}$$

式中：T 为实际结果为正样本；F 为实际结果为负样本；P 为推断为正样本，N 为推断为负样本；故 T_P 为正确推断为正样本的数量；F_P 为错误推断为正样本的数量；F_N 为错误推断为负样本的数量；n 为总类别数。

3　实验结果与分析

3.1　YOLOv8 基线模型对比

YOLOv8 目前有 5 个版本的模型，分别是 n、s、m、l、x。不同的模型对应不同的模型复杂度和参数量。为了确定哪个模型更适合芯片数据集，分别对 5 个模型进行测试。测试结果如表 2 所示。

表 2　基线模型对比

Table 2　Comparison of baseline models

模型	mAP@0.5/%	FLOPs/M	Parameters/M
YOLOv8n	92.4	**8.2**	**3.01**
YOLOv8s	91.6	28.7	11.14
YOLOv8m	92.3	79.1	25.86
YOLOv8l	**92.5**	165.4	43.63
YOLOv8x	91.5	258.1	68.16

可以发现对于芯片数据集来说，更高的参数量和更复

杂的模型结构不一定具有更显著的检测效果。相反,由于模型复杂度的提高,模型计算时间也会大幅增加。综合考虑下,虽然 YOLOv8l 拥有最高的检测精度,但参数量高了约 40 M,计算成本过大,故选择 YOLOv8n 作为本研究的基线模型。

3.2 不同注意力机制分析对比

为了验证验证注意力机制对模型检测性能的影响,本研究横向对比分析了 SE、CBAM、ECA、SimAM 四种注意力机制。实验时注意力机制添加位置均为 SPPF 模块上方。实验结果如表 3 所示。

表 3 不同注意力机制的对比

Table 3 Comparison of different attention mechanisms

注意力模块	P/%	mAP@0.5/%	Parameters/M
Baseline	93.4	92.4	**3.01**
CBAM	**94.6**	93.6	3.07
SE	91.9	92.0	3.02
ECA	92.8	92.6	**3.01**
SimAM	94.0	**93.9**	**3.01**

由表 3 结果分析可知,4 种注意力机制均未对模型参数大小产生影响。相比于基线模型和其余 3 种注意力机制,SimAM 注意力机制所得 mAP 值最高,精确度也达到了 94.0%。虽然 CBAM 注意力机制的精确度最高,但其 mAP 值仅为 93.6%。在添加 SE 模块后,模型的 mAP 值出现了下降,模型精确度也是 4 种注意力机制中最低的。ECA 模块对比基线模型,其 mAP 值仅仅提升了 0.2%。以上实验初步证明了引入 SimAM 注意力机制对于芯片缺陷检测有一定优势。此外,为排除注意力模块添加位置对实验结果的影响,分别将 SimAM 注意力模块添加到网络的不同位置实验,实验结果如表 4 所示。

表 4 不同 SimAM 注意力机制添加位置

Table 4 Different SimAM attention mechanisms add positions

SimAM 添加位置	P/%	mAP@0.5/%	Parameters/M
Baseline	93.4	92.4	**3.01**
SPPF	**94.9**	**93.9**	**3.01**
P3	94.7	93.8	**3.01**
Head	94.0	93.4	**3.01**
Backbone	94.6	93.5	**3.01**

SPPF 表示将注意力机制加在 SPPF 层上方,P3 表示将注意力机制加在 P3 检测头上,Head 表示将注意力机制加在每一个检测头上,Backbone 表示将主干网络的所有 C2f 模块替换为 C2f_SimAM 模块。从表 4 中可以看出,SimAM 注意力机制加在 SPPF 层上的表现最好。在参数量相同的情况下,P 和 mAP 值分别提升了 1.5 个百分点。实验结果表明,使用 SimAM 注意力机制并将其添加在 SPPF 层上方对芯片缺陷的检测能达到最好的效果。

3.3 消融实验

为了验证本研究中的主干网络 SPD-Conv 模块、SimAM 注意力机制、BiFPN 特征融合网络和 P2 小目标检测层对模型性能的影响,对上述改进后的模型进行消融实验,结果如表 5 所示。

表 5 模型消融实验结果

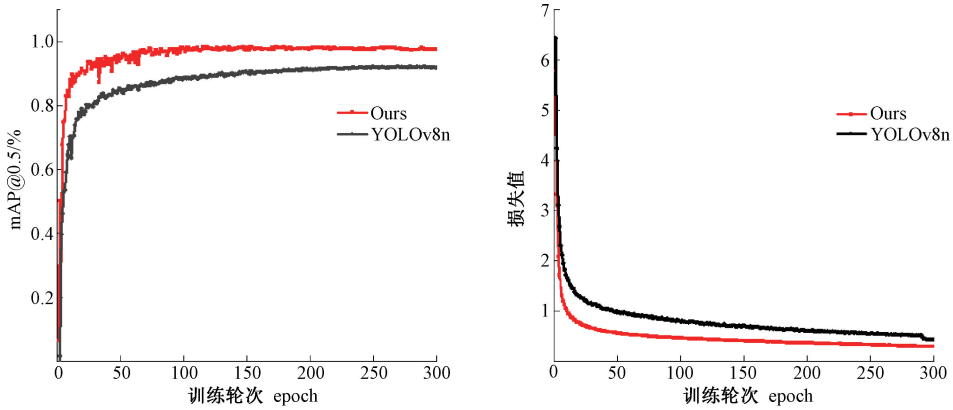
Table 5 Results of ablation experiment

YOLOv8n	SPD	SimAM	BiFPN	P2	Parameters/M	R/%	mAP@0.5/%	mAP@0.95/%
√					**3.01**	91.9	92.4	70.5
√	√				3.27	92.8	93.9	72.4
√		√			**3.01**	93.0	93.9	72
√			√		3.02	93.0	93.6	71.3
√				√	3.21	93.1	94.0	71.9
√		√	√		3.02	92.7	94.2	71.1
√	√			√	3.20	91.7	94.9	72.8
√	√		√	√	3.17	94.4	96.5	70.8
√	√	√	√	√	3.17	**94.9**	**97.8**	**74.8**

从表 5 中可以看出每个模块的加入都能有效提高检测能力。其中 SPD 模块由于在主干网络多加了一层结构,不可避免地增加了参数量,但 mAP@0.5 提高了 1.5%,mAP@0.95 提高了 1.9%。SimAM 模块提升的检测效果和 SPD 相差不大,但无参数量上的增加,说明直接计算三维的注意力权重能在几乎不影响模型大小的情况下提高模型的检测效果。当只使用 BiFPN 模块时 mAP@0.5 提高 1.2%,mAP@0.95 提高 0.8%。说明添加 BiFPN 模块能更好地检测出具有相似特征的缺陷。但显然单独使用 BiFPN 模块时 mAP 的提升值没有其他模块高,因为特征融合的过程中会不可避免地导致一些细节信息丢失,从而导致误检、漏检。在同时使用 BiFPN 模块和 SimAM 注意

力机制之后，可以看到情况得到了改善。最后，本文所得的模型相比基础模型，能够在仅增加少量参数量的情况下，使 mAP@0.5 提高 5.4%，mAP@0.95 提高 4.3%，R 提高 3.0%。对于小外形封装（small outline package，SOP）芯片缺陷数据集中的微小缺陷，特征相似缺陷及不同尺寸缺陷都有较好的检测效果。

改进前后损失值与 mAP@0.5/%对比曲线如图 7 所示。从图 7 可以看出改进前模型的收敛速度较慢且最后仍有少许波动。改进后的模型损失值曲线较为平滑且收敛速度更快，最终的损失值也低于原模型，没有出现过拟合、欠拟合的情况，说明改进后模型的整体性能明显优于原模型。



(a) 平均精度均值@0.5/%对比曲线
(a) mAP@0.5/% compari son curve

(b) 损失值对比曲线
(b) Loss comparison curve

图 7　改进前后损失值与 mAP@0.5/%对比曲线

Fig. 7　Comparison of training loss values and mAP@0.5/% comparison curve

3.4　不同检测模型对比实验

为验证改进模型的性能，将改进模型与双阶段检测的经典算法 Faster-RCNN 和单阶段检测的经典算法 SSD 以及 YOLO 系列的其余算法进行横向对比，结果如表 6 所示。

表 6　不同检测模型对比实验

Table 6　Comparative experiment of different detection models

模型	mAP@0.5/%	mAP@0.95/%	Parameters/M	FLOPS/M
SSD	89.5	55.8	26.29	62.7
Faster-RCNN	90.7	49.4	137.10	370.2
YOLOv5	91.3	69.0	7.08	16.5
YOLOv6	91.5	68.5	4.50	13.1
YOLOv7	92.6	67.7	37.22	105.2
YOLOv8	92.4	70.5	3.01	8.2
YOLOv9	93.8	70.3	2.62	10.7
YOLOv10	93.4	69.1	2.70	8.2
文献[18]	93.5	69.2	2.63	11.7
文献[19]	90.2	63.6	**0.64**	**4.3**
Ours	**97.8**	**70.8**	3.18	15.8

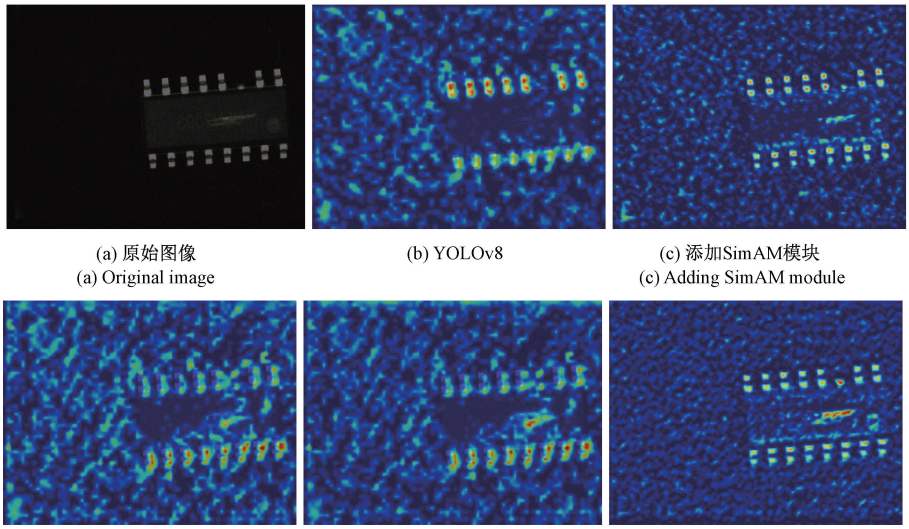
从表 6 中可看出改进后的模型相比其他模型拥有最高的 mAP 值，同时参数量也相对较少，计算成本较小。虽然 YOLOv7 相比本文基线模型，mAP@0.5 提高了 0.2%，但代价是参数量增大，计算成本过大。最新的 SOTA 模型 YOLOv9 和 YOLOv10，虽然相比 YOLOv8，两个算法都做到了降低参数量的同时提升准确度，但提升幅度都很小，不如本文的改进算法。同时本文还对比了两个最新的 YOLOv8 改进算法，虽然二者在参数量方面都比本文改进算法少，但在检测准确度方面与本文算法还是有一定的差距。可以看出改进后的算法参数量和浮点计算量不是最低，但也满足了实时检测的要求，且具有最高的检测精度，兼顾了检测速度和检测准确度，在 SOP 芯片缺陷检测上具有优越性。

3.5　可视化分析

经过上述实验，可以确定改进的 YOLOv8 模型提供了最佳的综合性能，满足了实际 SOP 芯片生产工艺的要求。

为了更直观地观察不同模块在芯片缺陷检测方面的改进情况,本研究使用 Grad CAM[20]绘制了热力图,可以更直接地观察到不同网络模型检测性能的差异。Grad CAM 利用训练权重反向传播,在空间维度上对获得的梯度矩阵进行全局平均池化,并通过加权激活特征层的每个通道,以获得目标的热力图。热力图中某个区域的亮度大小可以反映图像中对模型输出产生影响的大小。

图 8 所示为 SOP 芯片不同模型的热力图,在添加 SimAM 注意力后,图像中间的划痕被突出显示,表明模型已经注意到了缺陷。继续添加 SPD-Conv 后,模型聚焦的缺陷区域变大,但引脚处的缺陷没有被注意到,并且图像中的噪点太多。添加 BiFPN 模块后,图像中的杂质开始减少,模型开始注意到引脚上的缺陷。最终检测效果表明,改进后的模型同时注意到了引脚缺陷和芯片表面上的划痕缺陷,提高了整体检测精度,验证了改进模型的有效性。



(a) 原始图像
(a) Original image

(b) YOLOv8

(c) 添加SimAM模块
(c) Adding SimAM module

(d) 添加SPD-Conv模块
(d) Adding SPD-Conv module

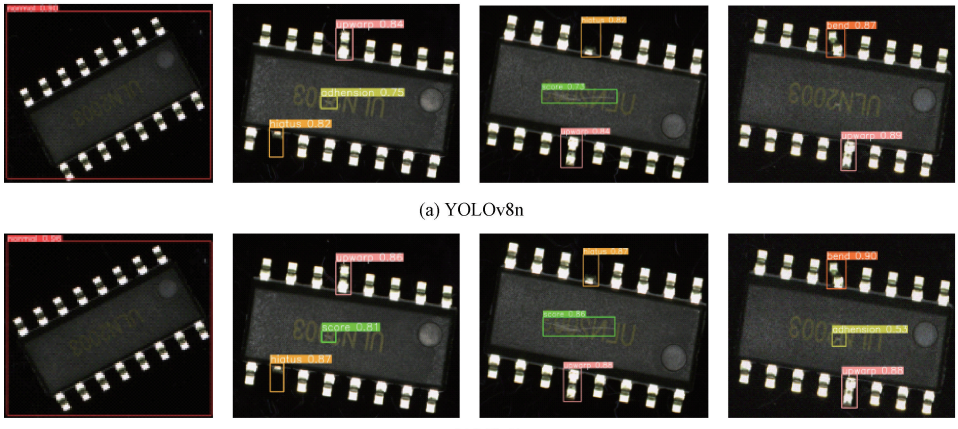
(e) 添加BiFPN模块
(e) Adding BiFPN module

(f) 改进模型
(f) The improved model

图 8　SOP 芯片不同模型的热力图

Fig. 8　SOP chip heat maps for different models

图 9 为 YOLOv8n 模型和改进后的模型对 SOP 芯片缺陷的检测效果对比图。从图 9(b)中可以看到改进后的模型对微小缺陷更为敏感,原模型在芯片表面存在的部分错检情况在改进后也得到了解决。同时由于融合了不同尺寸缺陷的特征信息,对引脚缺陷的检测效果也有了明显改善。综上所述,改进后的模型对 SOP 芯片缺陷有较好的识别效果。



(a) YOLOv8n

(b) 改进模型
(b) The improved model

图 9　SOP 芯片缺陷检测效果对比

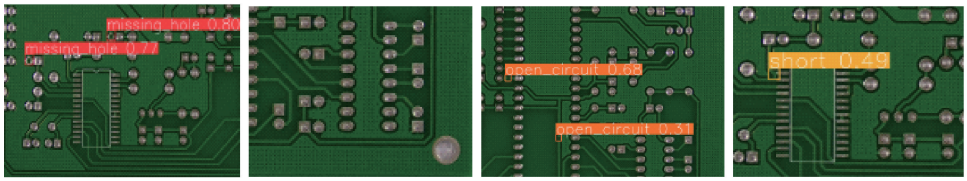
Fig. 9　Comparison of SOP chip defect detection effects

3.6　泛化实验

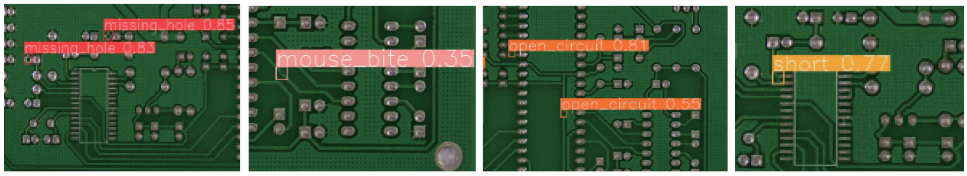
最后本文使用 YOLOv8n 和改进后的算法在北京大学智能机器人开放实验室发布的 PCB_DATASET 公共数据集进行泛化实验。PCB_DATASE 公共数据集有漏孔

(missing hole)、缺口(mouse bite)、开路(open circuit)、短路(short)、毛刺(spur)、余铜(spurious copper)6 类缺陷。由于 SOP 芯片缺陷数据集和 PCB 数据集的缺陷都是拥有相似特征的小目标缺陷,故进行泛化实验验证本文算法的有效性和通用性。

泛化数据实例如图 10 所示。由图 10 可以看出,改进后的模型对于 PCB 缺陷数据集检测效果也有一定的提升,原本算法的漏检、错检得到了有效改善。表 7 显示出模型的整体检测精度和召回率也得到了提高,证明了本文算法具有较好的检测能力和通用性。



(a) YOLOv8n



(b) 改进模型

(b) The improved model

图 10　泛化数据实例展示

Fig. 10　Comparison of detection effects in generalization experiments

表 7　泛化实验

Table 7　Cross-dataset generalization　%

模型	mAP	P	R
YOLOv8n	90.2	93.6	84.5
改进模型	92.9	94.2	88.4

3.7　芯片缺陷检测系统设计

本文基于 PySide6 设计了一个系统,用于对芯片缺陷进行实时检测,系统界面如图 11 所示。系统主要功能是选择训练好的模型,加载图片后点击检测,系统会显示出缺陷类别数量,缺陷的目标数量以及每秒的帧数,并支持使用摄像头功能进行实时检测。

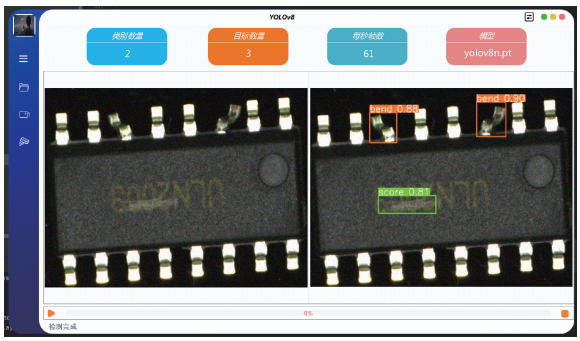


图 11　芯片缺陷检测系统界面

Fig. 11　Chip defect detection system interface

3.8　边缘设备部署

为了验证本文改进算法在边缘设备上的部署情况,将算法部署在嵌入式设备上并进行测试。由于大多数嵌入式设备的算力有限,故采用 TensorRT 框架对算法进行加速。

首先将电脑上训练好的.pt 文件转换为.onnx 格式的中间文件。ONNX 格式能够很好的兼容各种类型的文件格式,用于存储训练好的模型,方便在不同框架之间迁移。接下来使用 TenSorRT 对模型进行加速,会生成一个对应的.engine 文件。将.engine 文件在嵌入式设备上运行即可得到检测效果。在嵌入式设备上的检测情况如表 8 所示。

表 8　设备部署检测效果对比

Table 8　Comparison of device deployment detection results

fps

模型	台式计算机	嵌入式设备	TenSorRT
YOLOv8	60.4	9.2	—
Ours	66.1	10.5	18.3

可以看出,在经过 TenSorRT 加速后,模型的检测速度大幅提高,基本满足实时检测的帧率要求。并且经过测试,在算力较高的嵌入式设备上部署时,帧率可达 50 fps,完全满足实时检测的要求。

4　结　　论

本文提出了一种改进的 YOLOv8 算法,通过使用 SPD-Conv 模块避免模型在下采样过程中丢失可学习信息,优化对低分辨率和小目标的检测效果。然后在主干网络中加入 SimAM 注意力机制,使模型能学习三维通道中

的信息,提高模型的特征提取能力。同时使用 BiFPN 代替原特征提取网络,多次使用双向路径来进行特征融合,并在此基础上添加一条额外信息传输路径,使低层的信息也可以得到有效传递。最后增加一个小目标检测层,采集丰富的低阶特征信息,同时为其他高维特征补充更多细节信息,优化检测效果。实验结果表明,与原模型比较,改进后的模型 mAP@0.5 提高了 5.4%,mAP@0.95 提高了 4.3%,召回率提高了 3%,和其他模型相比有着显著优势。同时使用 PCB_DATASET 公共数据集进行泛化测试,结果表明本文改进后的算法在不同数据集上也拥有较好的检测能力。最后使用 PySide6 设计了一个芯片缺陷检测系统,并将算法部署在嵌入式设备上,验证了算法的可行性。未来将研究如何提高模型的 FPS,并进行轻量化,优化其在低性能嵌入式设备上的表现,扩展应用场景。

参考文献

[1] 王新宇,蒋三新.芯片缺陷检测综述[J].现代制造技术与装备,2022,58(5):94-98.
WANG X Y, JIANG S X. Overview of chip defect detection[J]. Modern Manufacturing Technology and Equipment,2022,58(5):94-98.

[2] 李大华,徐傲,王笋,等.基于改进 YOLOv5 的印刷电路板缺陷检测[J].电子测量技术,2023,46(23):112-119.
LI D H, XU A, WANG S, et al. Printed circuit board defect detection based on improved YOLOv5[J]. Electronic Measurement Technology, 2023, 46(23): 112-119.

[3] 赵友章,吕进.基于改进 SSD 的交通标志检测算法[J].电子测量技术,2023,46(7):151-158.
ZHAO Y ZH, LYU J. Traffic sign detection algorithm based on improved SSD[J]. Electronic Measurement Technology, 2023, 46(7): 151-158.

[4] 刘应桃,郭世伟,付孟新,等.基于改进 Faster RCNN 的轮对踏面缺陷检测[J].电子测量技术,2023,46(12):34-41.
LIU Y T, GUO SH W, FU M X, et al. Defect detection of wheelset tread based on improved Faster RCNN[J]. Electronic Measurement Technology, 2023, 46(12): 34-41.

[5] 罗月童,卞景帅,张蒙,等.基于卷积去噪自编码器的芯片表面弱缺陷检测方法[J].计算机科学,2020,47(2):118-125.
LUO Y T, BIAN J SH, ZHANG M, et al. Detection method of chip surface weak defect based on convolution denoising auto-encoders[J]. Computer Science, 2020, 47(2): 118-125.

[6] 刘艳菊,王秋霁,张惠玉,等.基于改进 SSD 的工件表面缺陷检测[J].热加工工艺,2024,53(2):134-139.
LIU Y J, WANG Q J, ZHANG H Y, et al. Defect detection of workpiece surface based on improved SSD[J]. Hot Working Technology, 2024, 53(2): 134-139.

[7] 薛阳,叶晓康,孙越,等.基于 Faster-RCNN 的汽车漆面缺陷部位检测[J].计算机应用与软件,2023,40(8):193-200.
XUE Y, YE X K, SUN Y, et al. Method for detection and location of automobile vehicle paint defects area based on Faster-RCNN[J]. Computer Applications and Software, 2023, 40(8): 193-200.

[8] 周天宇,朱启兵,黄敏,等.基于改进 YOLOV3 的载波芯片缺陷检测[J].激光与光电子学进展,2021,58(12):86-93.
ZHOU T Y, ZHU Q B, HUANG M, et al. Defect detection of chip on carrier based on improved YOLOv3[J]. Laser Optoelectronics Progress, 2021, 58(12): 86-93.

[9] WANG SH, WANG H Y, YANG F, et al. Attention-based deep learning for chip-surface-defect detection[J]. The International Journal of Advanced Manufacturing Technology, 2022, 121(3): 1957-1971.

[10] 张恒,程成,袁彪,等.基于 YOLOv5-EA-FPNs 的芯片缺陷检测方法研究[J].电子测量与仪器学报,2023,37(5):36-45.
ZHANG H, CHENG CH, YUAN B, et al. Research on chip defect detection method based on YOLOv5-EA-FPNs[J]. Journal of Electronic Measurement and Instrument, 2023, 37(5):36-45.

[11] SUNKARA R, LUO T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland, 2022: 443-459.

[12] YANG L X, ZHANG R Y, LI L D, et al. SimAM: A simple, parameter-free attention module for convolutional neural networks[C]. International Conference on Machine Learning, 2021: 11863-11874.

[13] TAN M X, PANG R M, LE Q V. Efficientdet: Scalable and efficient object detection[C]. Seattle: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781-10790.

[14] SAJJADI M S M, VEMULAPALLI R, BROWN M. Frame-recurrent video super-resolution[C]. Salt Lake City: IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6626-6634.

[15] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature

pyramid networks for object detection[C]. Honolulu: IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[16] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]. Salt Lake City: IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.

[17] GHIASI G, LIN T Y, LE Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection[C]. Long Beach: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7036-7045.

[18] 王艳,罗健,陶金,等. 基于 YOLOv8-PCB 的印制电路板缺陷检测[J/OL]. 激光与光电子学进展,2024, 1-18[2024-07-25]. http://kns.cnki.net/kcms/detail/31.1690.TN.20240624.2254.060.html.
WANG Y, LUO J, TAO J, et al. Defect detection of printed circuit boards based on YOLOv8-PCB[J/OL]. Laser Optoelectronics Progress, 2024, 1-18[2024-07-25]. http://kns.cnki.net/kcms/detail/31.1690.TN.20240624.2254.060.html.

[19] 李忠科,刘小芳. 基于轻量级 YOLOv8n 网络的 PCB 缺陷检测算法[J]. 电子测量技术,2024,47(4): 120-126.
LI ZH K, LIU X F. PCB defect detection algorithm based on lightweight YOLOv8n network [J]. Electronic Measurement Technology, 2024, 47(4): 120-126.

[20] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. Venice: IEEE International Conference on Computer Vision, 2017: 618-626.

作者简介

彭鸿瑞,硕士研究生,主要研究方向为目标检测、人工智能算法。

E-mail:872533830@qq.com

杨桂华(通信作者),硕士,副教授,主要研究方向为计算机检测与控制技术等。

E-mail:954991219@qq.com