DOI: 10.19650/j.cnki.cjsi.J2412531

用于 Mini/Micro-LED 芯片缺陷检测的全局特征压缩卷积神经网络*

田心如[1],褚 洁[1],蔡觉平[1],温凯林[1,2],王宇翔[1]
(1. 西安电子科技大学微电子学院 西安 710071; 2. 苏州鸿鹄骐骥电子科技有限公司 苏州 215000)

摘 要:微型发光二极管(Mini/Micro-LED)是下一代显示技术。随着 Mini/Micro-LED 芯片物理尺寸的微小化,制造良品率下降、集成度激增,Mini/Micro LED 芯片的快速、精确检测成为工业生产的关键。然而由于芯片尺寸小、分布密集,单个目标的特征信息占比不足,且工业检测要求检测算法速度快、易部署,Mini/Micro-LED 芯片缺陷检测仍面临巨大挑战。针对这些问题,设计了一种压缩注意力细节-语义互补卷积神经网络(CADSC-CNN)。在特征融合网络加入基于自注意力机制的编码器结构,更容易获取全局信息,对小目标的特征信息进行补充;同时对自注意力进行压缩操作减少模型的参数量,提高检测速率。此外,通过工业相机采集的 Mini/Micro-LED 数据集验证该方法的有效性。实验表明,该方法的平均精度均值(mAP)达到了 95.6%,速度为 100.6 fps。

关键词: 缺陷检测;Mini/Micro-LED;卷积神经网络;自注意力
中图分类号: TH89 TP391 **文献标识码**: A **国家标准学科分类代码**: 510

Global feature compression convolutional neural network for defect detection in Mini/Micro-LED chips

Tian Xinru[1], Chu Jie[1], Cai Jueping[1], Wen Kailin[1,2], Wang Yuxiang[1]
(*1. College of Microelectronics, Xidian University, Xi'an 710071, China*;
2. Suzhou Honghu Qiji Electronic Technology Co., Ltd, Suzhou 215000, China)

Abstract: Mini/Micro-LED represents the next generation of display technology. As the physical size of Mini/Micro-LEDchips becomes smaller, fabrication yields have decreased while the degree of integration has significantly increased. Consequently, fast and accurate inspection of Mini/Micro-LED chips is crucial for industrial production. However, inspecting Mini/Micro-LED chips remains challenging due to their small size and dense distribution. The limited feature information from individual targets and the need for fast, easily deployable inspection algorithms add to these challenges. To address these issues, we designed a compressed attention detail-semantic complementary convolutional neural network (CADSC-CNN). By incorporating an encoder structure based on a self-attention mechanism into the feature fusion network, it becomes easier to acquire global information to complement the features of small targets. Additionally, the compression operation of self-attention reduces the model's parameter count, thereby improving the detection rate. We validated the effectiveness of this method using a Mini/Micro-LED dataset collected by an industrial camera. Experiments demonstrated that this method achieves a mean average precision (mAP) rate of 95.6% and a speed of 100.6 frames per second.

Keywords: defect detection; Mini/Micro-LED; convolutional neural network; self-attention

0 引 言

微型发光二极管(Mini/Micro-LED)是一种新型显示技术,具有亮度高、响应速度快和分辨率高等特点,可广泛应用于可穿戴显示、智慧医疗显示、先进显示等领域[1-2]。然而在 Mini/Micro-LED 芯片的制备过程中,芯片缺陷是不可避免的。随着 LED 芯片尺寸减小、芯片密

收稿日期:2024-02-27 Received Date: 2024-02-27
*基金项目:国家自然科学基金面上项目(62274123)、陕西省自然科学基础研究计划项目(2024JC-YBQN-0615)、中央高校基本科研业务费专项资金项目(XJSJ23054)、陕西省博士后项目(2023BSHEDZZ173)资助

度增加,人工目检难度增大、误判率升高,无法满足工业化的检测要求[3-4]。因此,为提高 Mini/Micro-LED 的生产效率和质量,实现准确快速的 Mini/Micro-LED 芯片缺陷检测成为芯片制造过程中必不可少的环节[5]。

以摄像系统与视觉检测算法为核心的自动光学检测(automated optical inspection,AOI)技术可以自动定位和检测 LED 芯片[6],能够有效减少人力消耗、降低检测误判率,更适合进行 Mini/Micro-LED 芯片缺陷检测[7]。传统 AOI 技术主要为基于模板匹配的检测算法,Zhong 等[8]提出了一种基于斑点分析的模板匹配法,根据斑点梯度方向特征定位 LED 芯片。然而模板匹配法存在检测速率慢、无法定位缺陷目标、目标图像与模板图像角度不同时检测效果差等不足。目前,基于卷积神经网络(CNN)的模型在视觉领域表现出色,通过多层卷积可提取深层特征信息从而实现芯片的缺陷检测[9]。Lin 等[10]提出了一种基于 AlexNet 网络结构的缺陷检测器,能够精确定位 LED 芯片的缺陷位置。Shu 等[11]提出了一种并行空间金字塔池网络(spatial pyramid pooling network,SPP-Net),用于 LED 芯片表面质量检测。Chen 等[12]开发了一种基于 YOLOv3 网络的缺陷检测器,并将其应用于表面贴装器件发光二极管(surface mounted devices LED,SMD-LED)芯片的检测过程。

卷积神经网络具备强大的特征提取能力,在图像检测任务中取得了良好的性能。然而由于 Mini/Micro-LED 芯片尺寸小密度高的特点以及工业检测对模型参数和检测速率的要求,Mini/Micro-LED 芯片缺陷检测仍存在以下挑战:1)Mini/Micro-LED 尺寸缩减至 50 μm 以下,LED 芯片间的距离小于 5 μm,集成度激增 100 倍,属于微小目标的工业检测。2)Mini/Micro-LED 单个芯片特征信息不足,在卷积神经网络中经过逐层传递容易丢失细节信息[13]。特征信息的丢失对 Mini/Micro-LED 芯片缺陷检测效果影响较大,自注意力机制能够补充全局特征提升芯片缺陷检测效果[14]。3)由于工业检测部署环境的限制,检测方法要求实时性和低计算开销。自注意力模型导致参数量大和复杂度高,并不满足 Mini/Micro-LED 芯片缺陷检测任务速度快、高效率的要求[15]。

针对以上问题,本文提出了一种压缩注意力细节-语义互补卷积神经网络(compressed attention detail-semantic complementary convolutional neural network, CADSC-CNN)。首先芯片图像经过特征提取网络获得不同尺度的特征图。其次通过本文提出的压缩注意力细节-语义互补特征融合网络进行特征融合。与传统的特征金字塔相比,本文构建的压缩注意力细节-语义互补特征融合网络在融合细节信息和语义信息之后引入了基于自注意力机制的编码器结构,并对自注意力机制进行了压缩。自注意力能够捕获全局依赖关系,全局上下文作为额外信息对 Mini/Micro-LED 芯片的特征信息进行补充;对自注意力机制通过卷积操作压缩输入向量的维度,从而减少模型的参数量。最后,通过检测网络对不同尺寸的特征图进行预测目标的分类回归。此外,本文通过建立了 Mini/Micro-LED 光学图像采集系统,构建了 Mini/Micro-LED 数据集进行模型验证。

1　Mini/Micro-LED 芯片微小密集及工业检测速率要求的问题

主流检测网络在自然图像中表现优异,自然图像的目标尺寸大、分布稀疏,而 Mini/Micro-LED 芯片尺寸小、集成度高,与主流检测网络不适配。普遍认为尺寸小于 32×32 的目标为小目标,自然环境下采集的 MS COCO 数据集中小目标仅占全部目标的 1/3。主流检测网络在 MS COCO 数据集上的检测平均精度(average precision,AP)表明,小目标检测和中大目标检测仍存在较大差距,检测平均精度如表 1 所示。采集到 Mini/Micro-LED 芯片图像属于工业图像,包含 100~2 000 个待测芯片,单个芯片的绝对尺寸分布在 10×20~25×35,属于微小目标检测。因此主流目标检测算法对 Mini/Micro-LED 芯片缺陷的检测效果不佳,对芯片缺陷检测带来巨大挑战。

表 1　主流检测网络在 MS COCO 数据集上的检测平均精度

Table 1　Average detection accuracy of leading detection networks on the MS COCO dataset　(%)

网络模型	AP-大	AP-中	AP-小
Faster-RCNN	59.1	48.8	26.5
SSD	48.5	37.6	20.1
YOLOv3	55.3	45.9	23.7
YOLOv5	62.6	51.2	29.4
ViT	64.5	55.2	34.1

随着工艺的发展,Mini/Micro-LED 芯片的尺寸缩小至 200 μm 以下,芯片间距缩小至 5 μm 以下,因此芯片具有小尺寸高密度的特点。卷积神经网络卷积层的感受野有限,需要堆叠深度网络才能拥有全局感受野。对于小目标检测,单个目标的特征信息占比小,特征信息在卷积神经网络中经过长距离地移动与融合容易丢失细节信息。这种堆叠下采样的操作对中大型尺寸的目标检测影响较小,因为经过深层网络后仍保留足够的中大型目标特征信息。但由于小目标芯片包含的特征信息不足,少量特征信息的丢失会对小目标芯片检测造成较大的影响,网络模型无法提取足够的特征信息对芯片进行定位和分类;此外对于高密度的芯片,进行多次下采样后,特

征信息反应到深层特征图上将聚合成一个点,导致检测模型无法区分,如图 1 所示。

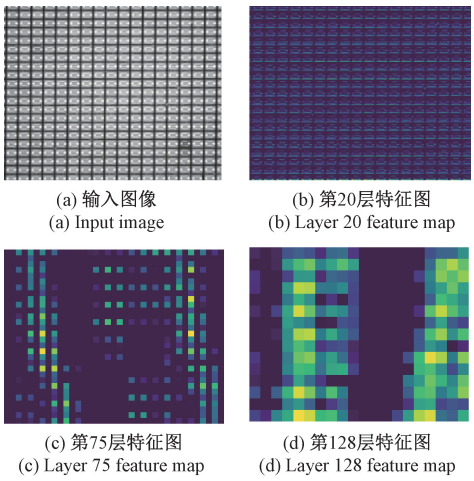


图 1 多层卷积后的 Mini/Micro-LED 芯片特征图分辨率对比

Fig. 1 Comparison of Mini/Micro-LED chip feature map resolution after multi-layer convolution

自注意力机制可以获取芯片图像的全局特征信息[16],对芯片特征信息进行补充,但是自注意力使模型参数量迅速增加,不符合工业检测快速的需求。自注意力机制将芯片图像分成图像块,学习所有输入图像块之间的远程依赖关系,复杂度与输入图像尺寸的平方成正相关,因此基于自注意力的模型计算复杂。卷积神经网络和自注意力模型的参数量对比如表 2 所示。基于自注意力的模型需要获取更多的芯片图像作为训练数据,然而 Mini/Micro-LED 芯片图像为工业数据集,样本数量匮乏且标注成本高昂;并且 Mini/Micro-LED 芯片检测属于工业应用,通常部署在嵌入式平台,嵌入式平台算力有限,不适用于复杂模型,导致检测性能较差,不满足 Mini/Micro-LED 芯片缺陷检测任务速度快、高效率的要求。

表 2 卷积神经网络和自注意力模型在 ImageNet 数据集上的参数量和计算

Table 2 Number of parameters and computational complexity of convolutional neural networks and self-attention models on the ImageNet dataset

类别	网络模型	模型计算力/GB	参数量/$\times 10^6$	输入尺寸
卷积神经网络	ResNet-18	1.8	11.7	224
	ResNet-50	4.1	25.6	224
	EfficientNet-B0	0.4	5.3	224
	EfficientNet-B1	0.7	7.8	240
自注意力模型	ViT-B/16	18.7	86.5	384
	ViT-L/16	65.8	304.33	384
	PVT-M	6.7	44.2	224
	PVT-L	9.8	61.4	224

2 压缩注意力细节-语义互补卷积神经网络

2.1 CADSC-CNN 框架

本文提出的 CADSC-CNN 为一阶段目标检测算法,框架如图 2 所示,CADSC-CNN 框架由特征提取网络、压缩注意力细节-语义互补特征融合网络和检测网络 3 部分组成。

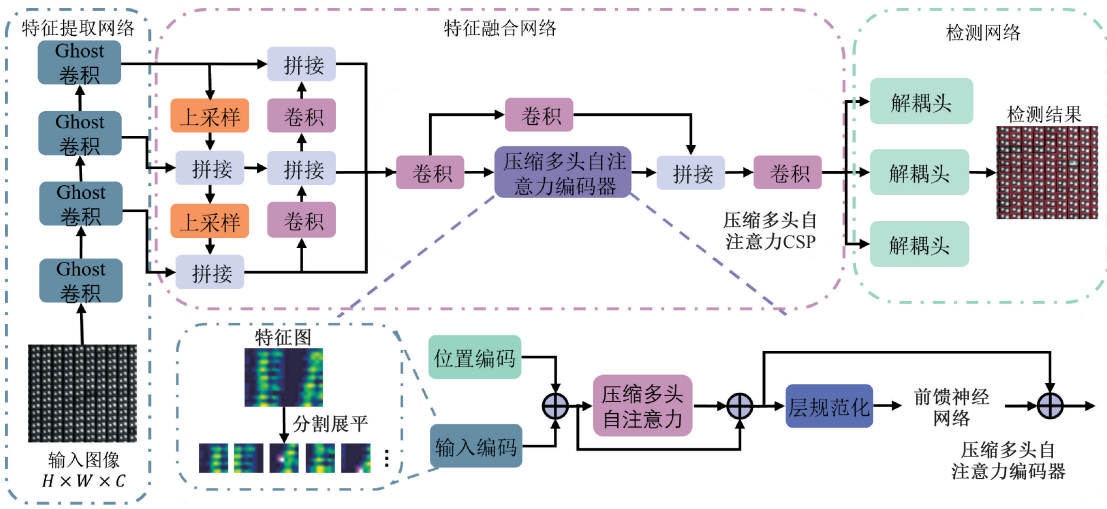


图 2 压缩注意力细节-语义互补卷积神经网络模型框架

Fig. 2 Framework of the compressed attention detail-semantic complementary convolutional neural network model

1)特征提取网络

特征提取阶段微小目标局部特征提取不足、特征图的冗余性造成多余计算消耗,本文提出了多尺度幻影卷积特征提取网络。

多尺度网络能够利用高分辨率、小感受野的低层特征图,高分辨率保留更多细节、小感受野聚焦局部区域特征,从而提高局部特征的提取能力。幻影卷积(GhostConv)能够高效地生成特征图,将部分常规的卷积替换为线性变换,从而减少不必要的计算消耗。GhostNet[17]作为主干网络,由一个 3×3 的卷积模块和 4 个幻影卷积特征层组成。卷积模块对芯片图像进行通道数调整,幻影卷积特征层对 Mini/Micro-LED 图像进行特征提取,获得 4 个不同尺度的特征图。幻影卷积模块的卷积步长为 1,能够加深网络的深度。幻影卷积下采样模块的卷积步长为 2,可以将特征图高和宽进行压缩,特征提取网络的具体参数选择参考文献[17]。

2)压缩注意力细节-语义互补特征融合网络

高底层特征无法兼顾细节信息和语义信息、单个芯片可利用特征量少,为此,设计了自注意力细节-语义互补特征融合网络。该网络由细节-语义互补融合模块和融入压缩多头自注意力模块的跨阶段局部瓶颈结构(compressed multi-head self-attention cross stage partial, CMSA-CSP)组成。细节-语义互补模块采用特征金字塔(feature pyramid networks, FPN)结构和路径聚合网络(path aggregation network, PAN)结构,融合了高层特征的语义信息和低层特征的细节信息。FPN 由两个上采样模块和拼接模块的组合构成,将高层特征图自上而下传递,先通过上采样改变尺寸大小,再与低层特征图进行特征拼接,从而使高层特征图中包含的语义信息传递到低层特征图。PAN 由两个卷积和拼接模块的组合构成,融合后的低层特征图自下而上传递,先通过卷积减小特征图尺寸,再通过拼接进行融合,从而使高级特征图中保留更多的细节信息。

CMSA-CSP 结构对互补后的特征数据提取全局特征信息。自注意力提取的全局信息能够对图像的全部芯片进行关联和比较,从而补充了单个目标的特征信息。过早采用自注意力机制强调全局依赖性,容易忽视重要的局部信息,因此为了更好的提取局部信息,我们将自注意力模块放在细节-语义互补模块后。

细节-语义互补融合模块中卷积的卷积核为 3,3×3 卷积核能够捕捉低层特征图的细节信息;压缩自注意力模块采用 1×1 的卷积,1×1 卷积可以在不损失特征信息的基础上,改变特征图的通道数,1×1 卷积的计算成本远低于相同深度的 3×3 卷积,能够减少模型计算复杂度。特征融合网络参数如图 3 所示。

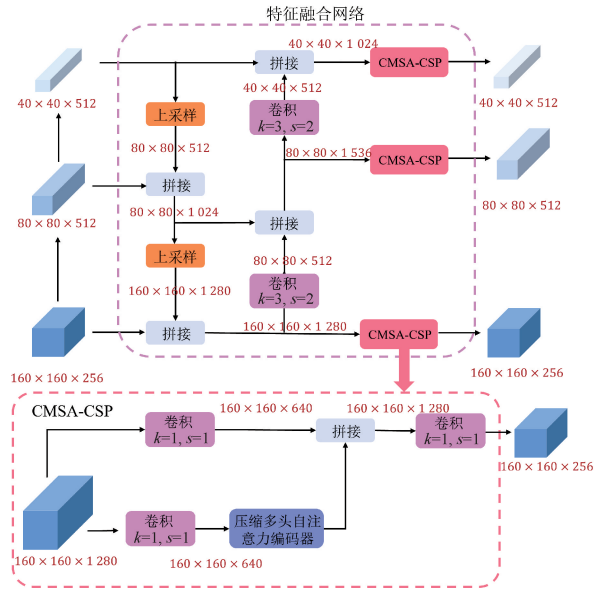


图 3　特征融合网络参数

Fig. 3　Parameters of feature fusion network

3)检测网络

分类和回归任务对网络权重的需求不同,耦合头同时预测类别和边界框,使不同任务之间关系复杂、相互影响大,因此本文采用并行解耦头。解耦头能够针对不同任务分别处理特征信息,并采用相应的损失函数,从而防止两个任务的误差相互影响。每个解耦头分为两个分支,分别提取类别特征和位置特征。解耦头每个分支由一个 3×3 的卷积模块和全连接层组成,在 3 个不同尺寸的特征图上进行目标的分类与详细坐标的回归。

2.2　基于长程依赖的融合全局特征

本文提出了一种压缩多头自注意力跨阶段局部瓶颈结构,CMSA-CSP 将卷积神经网络与基于多头自注意力机制的网络以残差结构相结合。

自注意力机制提取全局信息,全局信息能够综合图像中所有位置的特征信息,总结了微小目标的形状模式和排列规律,全局信息还能够提示网络在相似条件下寻找相似特征,扩大可利用特征的数量,全局信息作为一个额外信息对特征量不足的小目标进行了信息补充。由于自注意力计算复杂导致内存成本上升、自注意力提取全局信息时容易忽略局部信息,因此采用跨阶段局部瓶颈结构(cross stage partial, CSP)结构。CSP 结构将输入通道分为两部分,一部分关注全局信息,另一部分关注局部信息,最后将两部分拼接,从而同时关注全局信息和局部信息。两部分输入通道经过两分支进行并行计算,每一层的计算量平均分配,有效提高了每一个计算单元的利用率,从而减少内存成本。

CSMA-CSP 结构如图 4 所示,分为两个并行结构,一

个分支对特征图只进行卷积操作调整通道数，另一分支包括卷积层和压缩多头自注意力编码器（compressed multi-head self-attention encoder，CSMA-Encoder），两分支进行特征融合后调整通道数。压缩多头自注意力编码器结构类似于 Transformer 解码器，包括位置编码（position encoding）、压缩多头自注意力（compress multi-head self attention，CMSA）、层规范化（layer normalization，LN）、前馈神经网络（feed forward）和残差结构[18-19]。自注意力机制可以提高模型获取 Mini/Micro-LED 图像全局上下文信息的能力，前馈神经网络做进一步特征提取。

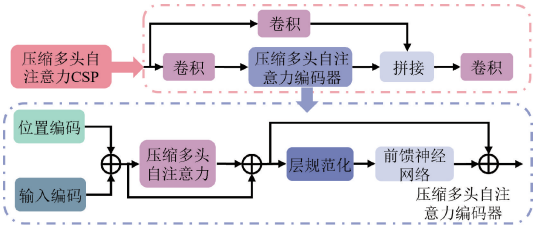


图 4 压缩多头自注意力 CSP 结构

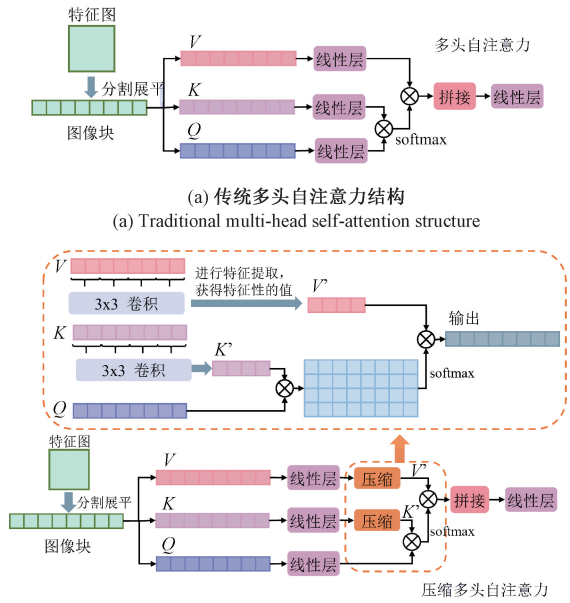
Fig. 4 Compressed multi-head self-attention CSP module

Mini/Micro-LED 芯片具有小尺寸高密度的特点，包含特征信息较少，并且部分特征信息经过长距离地移动与融合容易丢失。而自注意力机制具有全局感受野，提取到的全局信息能够对图像的全部芯片进行关联和比较，从而补充了单个目标的特征信息。因此在特征融合网络加入基于自注意力机制的 CMSA-CSP 结构，可以充分利用局部特征和全局特征从而提高模型对小目标信息的敏感度，对小尺寸、高密度的 Mini/Micro-LED 芯片目标的检测能力提升。

2.3 基于卷积特征稀疏的压缩多头自注意力

自注意力机制会计算所有输入数据的相关性，从而导致计算复杂度上升。因此，为了降低自注意力机制的计算复杂度，提出了压缩多头自注意力结构。传统的多头自注意力结构如图 5（a）所示。Mini/Micro-LED 芯片的特征图作为自注意力层的输入向量，每个输入向量之间都进行了相关性计算，因此模型的参数量大大增加[20-21]，无法满足工业检测快速检测和低计算开销的要求。本文提出的压缩多头自注意力模块通过对自注意力机制进行压缩改进从而减少自注意力机制的复杂度，如图 5（b）所示。

压缩多头自注意力模块在进行注意力操作前，将大小为 $h \times w \times c$ 的特征图划分为固定大小的图像块（patch），并在图像块中嵌入位置信息，图像块大小为 $p^2 \times \frac{hw}{p^2} \times c$，其中 (h,w) 为输入特征图的分辨率，c 为输入特征图的通道数，(p,p) 为图像块的分辨率，$\frac{hw}{p^2}$ 为图像块数量，即压缩多头自注意力输入向量的有效序列长度。对展平后的图像块进行线性投影得到 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵。



(a) 传统多头自注意力结构
(a) Traditional multi-head self-attention structure

(b) 压缩多头自注意力模块
(b) Compressed multi-head self attention module

图 5 传统多头自注意力结构和压缩多头自注意力模块

Fig. 5 Traditional multi-head self-attention structure and compressed multi-head self-attention module

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} = linear(\boldsymbol{patch}) \tag{1}$$

与传统的自注意力机制相比，压缩多头自注意力模块对 $\boldsymbol{K}$、$\boldsymbol{V}$ 采用卷积核为 3，步长为 2 的卷积压缩 $\boldsymbol{K}$、$\boldsymbol{V}$ 的维度，$\boldsymbol{Q}$ 的维度为输出维度，因此不能进行压缩，得到 $\boldsymbol{K}'$、$\boldsymbol{V}'$，每个输入向量的 $\boldsymbol{Q}$ 矩阵分别与所有输入向量的 $\boldsymbol{K}'$ 进行向量点积，得到注意力分数矩阵再乘以其对应的 $\boldsymbol{V}'$ 矩阵，从而得到输出向量：

$$Attention(\boldsymbol{Q},\boldsymbol{K}',\boldsymbol{V}') = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}'^{\mathrm{T}}}{\sqrt{d_{k'}}}\right)\boldsymbol{V}' \tag{2}$$

式中：$d_{k'} = \frac{d_Q}{h}$，d_Q 为 $\boldsymbol{Q}$ 矩阵的列数，即向量维度，h 为指定的常数。

多头自注意力结构为：

$$MultiHead(\boldsymbol{Q},\boldsymbol{K}',\boldsymbol{V}') = Concat(head1,head2...headh)\boldsymbol{W}^O \tag{3}$$

其中，$head_i$ 为：

$$head_i = Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^{K'},\boldsymbol{V}\boldsymbol{W}_i^{V'}) \tag{4}$$

式中：$\boldsymbol{W}^Q$ 为 $\boldsymbol{Q}$ 的权重矩阵；$\boldsymbol{W}^{K'}$ 为 $\boldsymbol{K}'$ 的权重矩阵；$\boldsymbol{W}^{V'}$ 为 $\boldsymbol{V}'$ 的权重矩阵；$\boldsymbol{W}^O$ 为线性投影的权重矩阵。

由于 Mini/Micro-LED 芯片尺寸小，本文选用卷积的方法进行下采样，减少序列长度的同时可以提取有效信息，避免下采样过程中可能将有用的细节信息过滤掉，小

目标的特征信息在全局信息中所占比例较小,信息损失会导致检测效果下降,从而不利于小目标的缺陷检测。特征图作为输入会导致复杂度大大增加,压缩得到 $\boldsymbol{K}'$、$\boldsymbol{V}'$,使 $\boldsymbol{K}'$ 和 $\boldsymbol{Q}$ 进行向量点积时以及注意力分数矩阵乘以其对应的 $\boldsymbol{V}'$ 矩阵时的运算量下降,从而提高模型性能。

因此,可以计算出多头自注意力模块以及压缩多头自注意力模块的复杂度。输入向量矩阵大小为 $n \times d$,其中,$n = \dfrac{hw}{p^2}$,$d = p^2 \times c$。

输入矩阵与权重矩阵相乘分别得到 $\boldsymbol{K}$、$\boldsymbol{Q}$ 和 $\boldsymbol{V}$ 3 个新的矩阵的复杂度为:

$$\Omega = 3nd^2 \tag{5}$$

$\boldsymbol{Q}$ 矩阵与 $\boldsymbol{K}$ 矩阵进行点积计算得到注意力分数矩阵的复杂度为:

$$\Omega = n^2 d \tag{6}$$

注意力分数矩阵与 $\boldsymbol{V}$ 矩阵相乘得到输出向量的复杂度为:

$$\Omega = n^2 d \tag{7}$$

最后将拼接后 h 组注意力汇聚的输出通过线性投影,复杂度为:

$$\Omega = nd^2 \tag{8}$$

可得多头自注意力模块的复杂度为:

$$\Omega = 4nd^2 + 2n^2 d \tag{9}$$

压缩多头自注意力模块中 $\boldsymbol{K}'$、$\boldsymbol{V}'$ 的序列长度减少为原来的 1/2,因此 $\boldsymbol{K}'$ 和 $\boldsymbol{Q}$ 进行向量点积和注意力分数矩阵乘以 $\boldsymbol{V}'$ 矩阵两个过程的复杂度下降。

$$n_{K'} = \frac{n_K}{2}, \quad n_{V'} = \frac{n_V}{2} \tag{10}$$

$\boldsymbol{Q}$ 矩阵与 $\boldsymbol{K}'$矩阵的转置进行点积计算得到注意力分数矩阵的复杂度为:

$$\Omega = \frac{n^2 d}{2} \tag{11}$$

注意力分数矩阵与 V 矩阵相乘得到输出向量的复杂度为:

$$\Omega = \frac{n^2 d}{2} \tag{12}$$

压缩多头自注意力模块的复杂度为:

$$\Omega = 4nd^2 + n^2 d \tag{13}$$

显然,本文提出的压缩多头自注意力模块可以减少模型的复杂度,更加适合 Mini/Micro-LED 芯片缺陷检测。

3　实验与结果分析

3.1　实验设置及数据集

本文搭建了一个 Mini/Micro-LED 光学图像采集系统,该平台由环形光源、工业相机、晶圆托盘组成。如图 6 所示。采集的 Mini/Micro-LED 芯片图像作为数据集进行实验验证,如图 7 所示。数据集共有 3 830 张光学图像,其中健康样本、缺陷样本比例分布为 0∶78∶0∶22,实验使用 4 倍交叉验证自动划分训练集和验证集。其中 3 064 张图像作为训练集用于压缩注意力特征融合金字塔卷积神经网络模型的训练,766 张图像作为验证集用于神经网络模型的验证测试。数据集的每幅图像为灰度

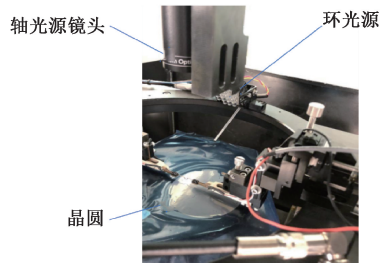
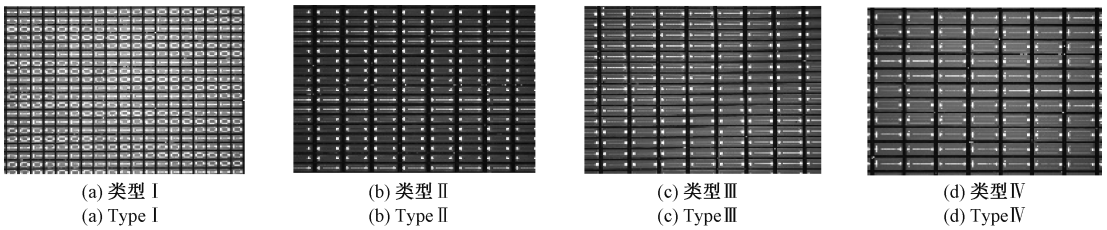


图 6　Mini/Micro-LED 光学图像采集系统

Fig. 6　Mini/Micro-LED optical image acquisition system



(a) 类型Ⅰ (a) Type Ⅰ　(b) 类型Ⅱ (b) Type Ⅱ　(c) 类型Ⅲ (c) Type Ⅲ　(d) 类型Ⅳ (d) Type Ⅳ

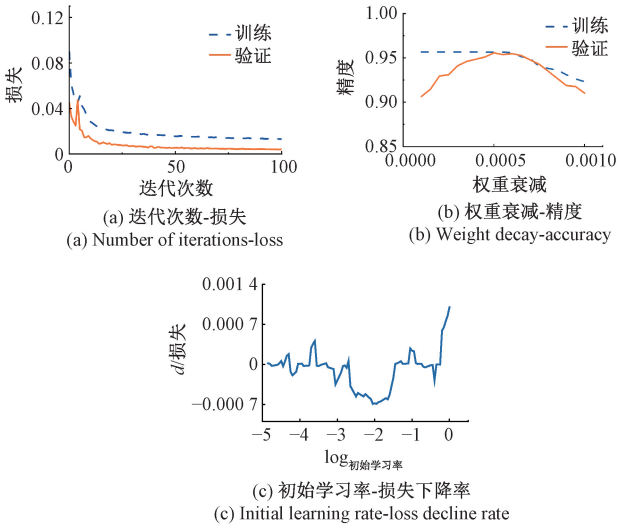
(e) 类型Ⅴ (e) Type Ⅴ　(f) 类型Ⅵ (f) Type Ⅵ　(g) 类型Ⅶ (g) Type Ⅶ

图 7　Mini/Micro-LED 数据集的 7 种芯片类型

Fig. 7　Seven chip types for the Mini/Micro-LED dataset

单通道图像,分辨率为 640×640。本次实验基于 Pytorch 神经网络框架和 Python 编程语言实现,深度学习框架为 pytorch1.13。实验在 NVDIA GeForce RTX3090 GPU 上完成。

实验训练采用随机梯度下降算法(stochastic gradient descent, SGD)优化训练过程中的损失,初始学习率为 0.01,训练迭代次数为 100 次,最大检测数量为 3 000,每批次的图像数量为 8,优化器权重衰减为 0.000 5,取其中平均精度均值(mean average precision,mAP)最高的一轮权重用于测试比较。本文采用 AP、mAP、召回率、检测速率作为评价指标。超参数的选择实验如图 8 所示。



(a) 迭代次数-损失
(a) Number of iterations-loss

(b) 权重衰减-精度
(b) Weight decay-accuracy

(c) 初始学习率-损失下降率
(c) Initial learning rate-loss decline rate

图 8 超参数选择实验

Fig. 8 Hyperparameter selection experiment

1)批量大小,本文模型较复杂,受显卡内存限制,选择 8 为批量大小。

2)优化器,本文选择 SGD 优化器,SGD 训练速度快,避免了批量梯度更新过程中的计算冗余问题。

3)最大检测数量,Mini/Micro-LED 图像芯片目标数量多,包含 100~2 000 个待测芯片,设置为 3 000。

4)迭代次数,由图 8(a)可见,网络的损失在 50 个迭代次数内迅速下降,然后趋于平稳,因此我们选取迭代次数为 100,从而达到更佳性能。

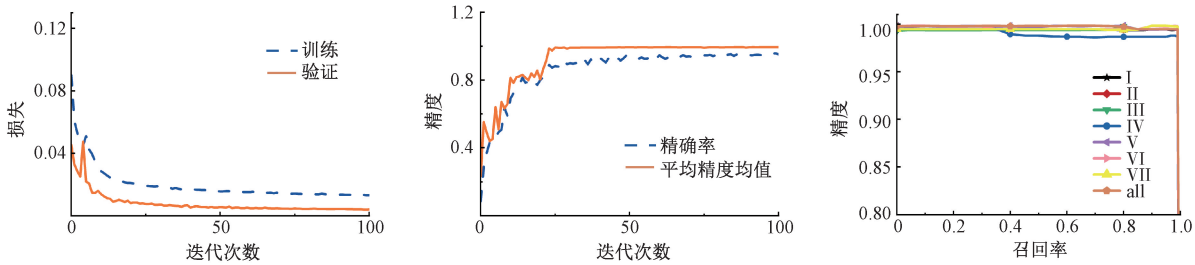
5)权重衰减,由图 8(b)可见,权重衰减取 0.000 5 时,模型不会出现欠拟合和过拟合现象,取得最高的检测精度,因此设置权重衰减为 0.000 5。

6)初始学习率,通过实验选择最优初始学习率,选择损失下降最快的学习率作为初始学习率。由图 8(c)可见,初始学习率为 0.01 时损失下降最快,因此选择 0.01 为初始学习率。

3.2 实验结果

1)CADSC-CNN 网络训练与检测性能

为了验证提出的 CADSC-CNN 模型的合理性,在 Mini/Micro-LED 数据集上进行了实验。在训练过程中记录损失、精度和精度-召回率曲线,如图 9 所示。结果表明,网络的损失在 50 个迭代次数内迅速下降,然后趋于平稳,精度在 30 个迭代次数内迅速饱和。本文模型没有出现过拟合和欠拟合现象,主要原因如下:模型的复杂度与 Mini/Micro-LED 数据集的复杂度相匹配;大量的 Mini/Micro-LED 训练图像能够覆盖数据的真实分布,减少模型学习到不具代表性的样本特征的机会;通过正则化技术约束模型参数,防止过拟合;模型在训练过程中通过实验选择了最优超参数;使用 4 倍交叉验证自动划分训练集和验证集,可以准确地评估模型在未见数据上的表现。



(a) 损失
(a) Loss

(b) 精度
(b) Accuracy

(c) 精度-召回率
(c) P-R curve

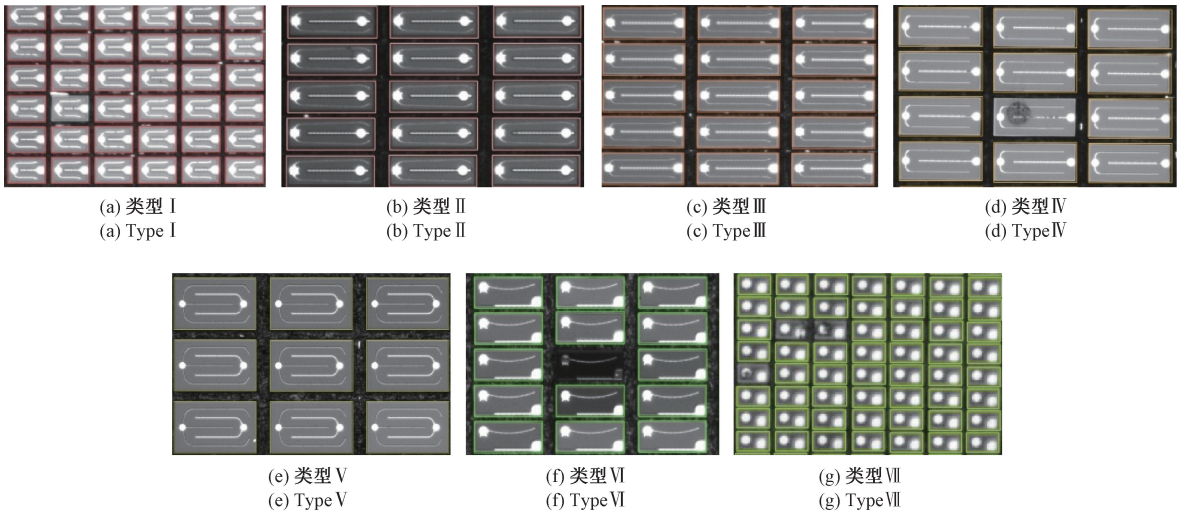
图 9 训练过程中记录的指标

Fig. 9 Indicators recorded during training

CADSC-CNN 在 Mini/Micro-LED 数据集上的检测准确率达到了 95.6%,其中尺寸最小密度最高的类型 VII 芯片的平均精度达到了 81.8%,同时,各类型的测试可视化结果如图 10 所示,本文提出的算法在各类型的 LED 芯片上都取得了较好的检测效果。

2)CADSC-CNN 与其他算法的对比

为了证明 CADSC-CNN 模型在 Mini/Micro-LED 数据集上的检测有效性,本文还选择了一些主流目标检测算法作为基准进行对比实验。一阶段的对比算法包括 SSD[22]、YOLOv5[23]、YOLOv8[24],二阶段的对比算法有



(a) 类型Ⅰ (a) Type Ⅰ　(b) 类型Ⅱ (b) Type Ⅱ　(c) 类型Ⅲ (c) Type Ⅲ　(d) 类型Ⅳ (d) Type Ⅳ

(e) 类型Ⅴ (e) Type Ⅴ　(f) 类型Ⅵ (f) Type Ⅵ　(g) 类型Ⅶ (g) Type Ⅶ

图 10　Mini/Micro-LED 测试集检测可视化结果

Fig. 10　Mini/Micro-LED test set detection visualization result

FasterR-CNN[25],基于自注意力的对比算法为 DETR[26]。训练过程中使用的超参数会影响模型的性能、速度和准确性。通过实验对比算法对超参数进行了微调。综合对比了 CADSC-CNN 与其他 5 种算法在 Mini/Micro-LED 数据集上的检测性能和检测速率,实验结果如表 3 所示。CADSC-CNN 算法在 7 种类型的 Mini/Micro-LED 检测中都达到了最高 AP,特别是尺寸最小、密度最高的类型 VII 达到了 81.8% 的 AP,显著高于其他 5 种算法。在所有的检测算法中,该方法的召回率最高,达到了 99.8%。检测速率 FPS 为 100.6,略低于一阶段检测算法 SSD、YOLOv5、YOLOv8,与二阶段检测算法 Faster R-CNN 和基于自注意力的对比算法 DETR 相比分别提高了 25.6% 和 34.7%。因此可得 CADSC-CNN 算法对不同类型的 Mini/Micro-LED 芯片都具有良好的检测效果,尤其对于小尺寸、高密度的芯片,该方法显著优于其他检测算法。

表 3　检测性能与对比算法的比较

Table 3　Comparison of detection performance with alternative algorithms

模型	mAP/%	AP_{I}/%	AP_{II}/%	AP_{III}/%	AP_{III}/%	AP_{V}/%	AP_{VI}/%	AP_{VII}/%	检测帧率/fps	召回率/%
YOLOv5	91.7	90.1	91.5	95.6	98.9	99.4	94.4	72.0	118.4	92.7
YOLOv8	92.0	90.6	92.8	95.1	99.0	99.5	93.1	73.7	109.0	89.9
SSD	90.4	88.7	90.6	94.8	97.5	98.8	92.6	69.9	116.2	88.2
Faster RCNN	90.1	88.3	90.5	94.9	97.2	98.6	92.5	68.7	80.1	86.4
DETR	94.3	93.4	94.8	97.9	99.2	99.5	96.8	78.6	74.7	96.3
本文	95.6	94.9	96.6	98.5	99.5	99.4	98.3	81.8	100.6	99.8

3)消融实验

为了验证压缩注意力细节-语义互补卷积神经网络中融入压缩多头自注意力 CSP 结构的有效性,在 Mini/Micro-LED 数据集上进行了消融实验,结果如表 4 所示。构建了两个基准模型,基准模型与本文所提出的模型在特征提取网络和检测网络阶段的结构和参数均相同,仅在特征提取网络阶段进行改动。具体改动如下:Base1 模型为纯卷积神经网络,其特征融合阶段不加入压缩多头自注意力 CSP 结构,从而验证了自注意力机制的全局特征提取能力对提高 Mini/Micro-LED 芯片缺陷检测精度的效果;Base2 模型的特征融合阶段加入基于多头自注意力 CSP 结构,不对自注意力机制进行压缩操作,因此验证了压缩多头自注意力相比于多头自注意力机制对芯片检测任务检测速率的提高。

由表 4 可以看出,与纯卷积编码网络 B1 相比,所构建的模型在 mAP 上提高了 6.2%,其中类型 VII 的 AP 提

表 4 本文算法的消融实验结果比较

Table 4 Comparison of ablation experiment results for the proposed algorithm

模型	mAP/%	AP_I/%	AP_{II}/%	AP_{III}/%	AP_{III}/%	AP_V/%	AP_{VI}/%	AP_{VII}/%	检测帧率/fps
Base1	90.8	90.0	90.7	94.1	98.2	98.9	93.6	69.8	126.4
Base2	95.9	95.8	96.9	98.8	99.5	99.5	98.7	82.1	87.3
本文	95.6	94.9	96.6	98.5	99.5	99.4	98.3	81.8	100.6

高幅度最大,达到了 13.9%。与加入自注意力机制的 Base2 模型相比,可以看出压缩多头自注意力模块使检测速率实现了 15.2%的增长。实验结果说明了在特征融合阶段融入基于自注意力机制的编码器结构有利于获取全局上下文信息,防止多次下采样后小目标的特征信息丢失,对小目标高密度的特征信息提取有重要意义;并且对自注意力机制进行压缩操作可以有效提高检测速率,减小模型参数,更适合工业缺陷检测。

4 结 论

本文提出了一种压缩注意力细节-语义互补卷积神经网络,用于 Mini/Micro-LED 芯片工业缺陷检测。针对 Mini/Micro-LED 芯片尺寸小密度高带来的挑战以及为了满足工业缺陷检测速率快、效率高的要求,分析了如何充分提取 Mini/Micro-LED 芯片的特征信息,以及如何减少模型计算复杂度。因此,本文提出了压缩注意力细节-语义互补卷积神经网络的模型框架,包含多尺度 Ghost 卷积特征提取网络、自注意力细节-语义互补特征融合网络和并行解耦头 3 部分。首先,该网络通过多尺度特征提高了对微小目标的局部特征提取能力;其次,通过细节-语义互补模块兼顾了高层特征的语义信息和低层特征的细节信息,并利用自注意力机制提取了全局信息;最后,对自注意力机制进行了压缩操作,通过卷积对输入向量进行降维,从而减少模型参数量,更适合工业检测。通过在 Mini/Micro-LED 数据集上验证表明,压缩注意力细节-语义互补卷积神经网络能够有效提高 Mini/Micro-LED 芯片缺陷检测精度和检测速率。

参考文献

[1] CHANG W, KIM J, KIM M, et al. Concurrent self-assembly of RGB microLEDs for next-generation displays[J]. Nature, 2023, 617: 287-291.

[2] LIN J Y, JIANG H X. Development of Micro-LED[J]. Applied Physics Letters, 2020, 116 (10): 100502.

[3] 姜也, 黄一凡, 熊美明, 等. PCBA 板载 DDR 芯片焊点缺陷检测研究[J]. 仪器仪表学报, 2023, 44(2): 129-137.
JIANG Y, HUANG Y F, XIONG M M, et al. Research on solder bump defect detection of DDR chip on PCBA[J]. Chinese Journal of Scientific Instrument, 2023, 44(2): 129-137.

[4] 赵朗月, 吴一全. 基于机器视觉的表面缺陷检测方法研究进展[J]. 仪器仪表学报, 2022, 43(1): 198-219.
ZHAO L Y, WU Y Q. Research progress of surface defect detection methods based on machine vision[J]. Chinese Journal of Scientific Instrument, 2022, 43(1): 198-219.

[5] HUANG H X, TANG X D, WEN F, et al. Small object detection method with shallow feature fusion network for chip surface defect detection[J]. Scientific Reports, 2022, 12(1): 3914-3914.

[6] EBAYYEH A A R M A, MOUSAVI A. A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry[J]. IEEE Access, 2020, 8: 183192-183271.

[7] 苏昊, 李文豪, 李俊龙, 等. 晶圆级 Micro-LED 芯片检测技术研究进展[J]. 液晶与显示, 2023, 38(5): 582-594.
SU H, LI W H, LI J L, et al. Recent progress of wafer level Micro-LED chip inspection technology[J]. Chinese Journal of Liquid Crystals and Displays, 2023, 38(5): 582-594.

[8] ZHONG F Q, HE S P, LI B. Blob analyzation-based template matching algorithm for LED chip localization[J]. The International Journal of Advanced Manufacturing Technology, 2017, 93(1/4): 55-63.

[9] CHEN M Y, CHEN J B, LI CH, et al. Defect detection of Micro-LED with low distinction based on deep learning[J]. Optics and Lasers in Engineering, 2024, 173(2): 107924.

[10] LIN H, LI B, WANG X G, et al. Automated defect inspection of LED chip using deep convolutional neural network[J]. Journal of Intelligent Manufacturing, 2019, 30(6): 2525-2534.

[11] SHU Y F, LI B, LIN H. Quality safety monitoring of LED chips using deep learning-based vision inspection methods[J]. Measurement, 2021, 168(1): 108123.

[12] CHEN S H, TSAI C CH. SMD LED chips defect detection using a YOLOv3-dense model[J]. Advanced Engineering Informatics, 2021, 47(1): 101255.

[13] PERWAIZ N, SHAHZAD M, FRAZ M M. Ubiquitous vision of transformers for person re-identification[J]. Machine Vision and Applications, 2023, 34(27): 1-14.

[14] SHIN H CH, ROTH H R, GAO M CH, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J]. IEEE Transactions on Medical Imaging, 2016,35(5): 1285-1298.

[15] GIL Y, PARK J H, BAEK J, et al. Quantization-aware pruning criterion for industrial applications[J]. IEEE Transactions on Industrial Electronics, 2022, 69(3): 3203-3213.

[16] CUI Y N, KNOLL A. PSNet: Towards efficient image restoration with self-attention[J]. IEEE Robotics and Automation Letters, 2023,8(9): 5735-5742.

[17] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More features from cheap operations[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1577-1586.

[18] 黄星华, 吴天舒, 杨龙玉, 等. 一种面向旋转机械的基于 Transformer 特征提取的域自适应故障诊断[J]. 仪器仪表学报, 2022, 43(11): 210-218.
HUANG X H, WU T SH, YANG L Y, et al. Domain adaptive fault diagnosis based on Transformer feature extraction for rotating machinery[J]. Chinese Journal of Scientific Instrument, 2022, 43(11): 210-218.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. ArXiv preprint arXiv: 1706.03762, 2017.

[20] BELAL M M, SUNDARAM D M. Global-local attention-based butterfly vision transformer for visualization-based malware classification[J]. IEEE Access, 2023, 11: 69337-69355.

[21] DENG D, JING L P, YU J, et al. Sparse self-attention LSTM for sentiment lexicon construction[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1777-1790.

[22] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]. Computer Vision-ECCV 2016, 2016.

[23] GUO G, ZHANG ZH Y. Road damage detection algorithm for improved YOLOv5[J]. Scientific Reports, 2022, 12(1): 15523.

[24] SAYDIRASULOVICH S N, MUKHIDDINOV M, DJURAEV O, et al. An improved wildfire smoke detection based on YOLOv8 and UAV images[J]. Sensors,2023, 23(20): 8374-8374.

[25] REN SH Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[26] CARION N, MASSA F, SYNNAEVE G, ex al. End-to-end object detection with transformers[J]. Computer Vision-ECCV 2020, 2020:213-229.

作者简介

田心如,2022 年于西安电子科技大学获得学士学位,现为西安电子科技大学硕士研究生,主要研究方向为缺陷检测与机器学习。

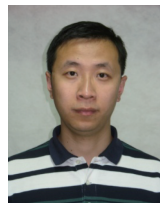
E-mail:xrtian@ stu. xidian. edu. cn

Tian Xinru received her B. Sc. degree from Xidian University in 2022. She is currently a M. Sc. candidate at Xidian University. Her main research interests include defect detection and machine learning.

褚洁,2022 年于西安电子科技大学获博士学位,现为西安电子科技大学博士后,主要研究方向为视觉检测与机器学习。

E-mail:jiechu@ stu. xidian. edu. cn

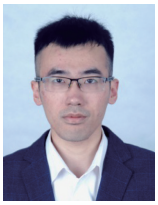
Chu Jie received her Ph. D. degree from Xidian University in 2022. She is currently a postdoctoral at Xidian University. Her main research interests include vision detection and machine learning.



蔡觉平(通信作者),分别在 1998 年和 2001 年于西安电子科技大学获得学士学位与硕士学位,2004 年于上海交通大学获得博士学位,现为西安电子科技大学教授、博士生导师,主要研究方向为信号处理,人工

智能芯片技术等。
E-mail：jpcai@ mail. xidian. edu. cn

Cai Jueping (Corresponding author) received his B. Sc. degre and M. Sc. degree both from Xidian University in 1998 and 2001, and Ph. D. degree from Shanghai Jiaotong University in 2004, respectively. He is currently a professor and Ph. D. supervisor at Xidian University. His main research interests include signal processing, artificial intelligence chip technology, etc.



温凯林,2024 年于西安电子科技大学获得博士学位,现为苏州鸿鹄骐骥电子科技有限公司总经理,主要研究方向为集成电路设计测试与图像处理。
E-mail：klwen@ stu. xidian. edu. cn

Wen Kailin received his Ph. D. degrees from Xidian University in 2024. Now he is general manager in Suzhou Honghu Qiji Electronic Technology Co., Ltd. His main research interests include integrated circuits design & test and image processing.

王宇翔,分别在 2021 年与 2024 年于西安电子科技大学获得学士学位和硕士学位,主要研究方向为图像识别与机器学习。
E-mail：wangyx@ foxmail. com

Wang Yuxiang received his B. Sc. degre and M. Sc. degree both from Xidian University in 2021 and 2024, respectively. His main research interests include image recognition and machine learning