文章编号:1671-4598(2025)11-0259-08 DOI:10.16526/j.cnki.11-4762/tp.2025.11.031 中图分类号:TP391.4 文献标识码:A

# 基于改进 MobileNetV2 的金属表面缺陷分类方法

姚 旭[1,2], 杨延宁[1,2], 白鸿冰[1,2]
(1. 延安大学 物理与电子信息学院, 陕西 延安 716000;
2. 陕西省能源大数据智能处理省市共建重点实验室, 陕西 延安 716000)

**摘要**: 金属表面缺陷检测是工业制造中质量控制的关键环节; 传统的人工检测方法由于成本高、效率低, 难以满足现代制造业对高精度与高效率的需求; 文章提出了一种基于 MobileNetV2 的改进网络模型, 用于提高金属表面缺陷检测的精度与效率; 在 MobileNetV2 网络基础上, 引入坐标注意力机制以增强特征学习能力, 采用深度可分离思想改进 Inception 模块, 在增强网络对多尺度特征的提取能力的同时保持模型参数量; 通过图像增强技术处理数据集, 以提升网络的鲁棒性; 实验在 NEU-DET 金属缺陷数据集上进行, 验证了模型的有效性; IC _ MobileNetV2 模型在验证集上取得了 92.8%的准确率, 与原始的 MobileNetV2、GoogleNet、DenseNet、ResNet34 和 ResNet50 相比, 准确率分别提高了 5.6%、2.8%、0.9%、1.7%和 1.7%; 实验结果表明, 该方法在金属表面缺陷分类任务中具有较好的应用潜力。

**关键词**: 缺陷图像检测; MobileNet 网络; 深度可分离卷积; 注意力机制; Inception 网络

# Metal Surface Defect Classification Method Based on Improved MobileNetV2

YAO Xu[1,2], YANG Yanning[1,2], BAI Hongbing[1,2]
(1. School of Physics and Electronic Information, Yan'an University, Yan'an 716000, China;
2. Shaanxi Provincial Key Laboratory of Energy Big Data Intelligent Processing, Yan'an 716000, China)

**Abstract**: Metal surface defect detection is a key link in quality control of industrial manufacturing. traditional manual inspection methods have the characteristics of high cost and low efficiency, making it difficult to meet the requirements of high precision and efficiency in modern manufacturing industry. an improved network model based on MobileNetV2 is proposed to improve the accuracy and efficiency of metal surface defect detection. Based on this, a coordinate attention (CA) mechanism is introduced to enhance feature learning ability and incorporates an improved Inception module with depthwise separable convolution (DSC), and to extract multi-scale features in the enhanced network while maintaining the model parameter Image augmentation techniques are applied to process the dataset to improve the robustness of the network. Experiments on the NEU-DET metal defect dataset verify the effectiveness of the model. The IC _ MobileNetV2 achieves an accuracy of 92.8%, which is 5.6%, 2.8%, 0.9%, 1.7%, and 1.7% higher than those of the MobileNetV2, GoogleNet, DenseNet, ResNet34, and ResNet50, respectively. Experimental results show that this method has a good practical significance in metal surface defect classification.

**Keywords**: defect image detection; MobileNet network; DSC; attention mechanism; Inception network

## 0 引言

金属表面缺陷分类是工业制造领域中的重要问题, 直接影响产品质量和生产效率。随着智能制造和工业自动化的发展, 快速、准确地检测并分类金属表面的缺陷, 成为提高产品合格率和降低生产成本的关键步骤。传统的金属表面缺陷检测方法多依赖于人工视觉检查和传统图像处理算法, 然而这些方法存在效率低下、准确性不足等问题, 尤其是在面对复杂的缺陷类型和背景噪声时, 性能明显受限。近年来, 深度学习技术, 特别是卷积神经网络 (CNN, convolutional neural network) 的应用, 为金属表面缺陷检测提供了新的解决方案[1]。

**收稿日期**:2024-10-21; **修回日期**:2024-12-04。
**基金项目**:国家自然科学基金(52365069);研究生教育创新计划项目(YCX2024095)。
**作者简介**:姚 旭(2001-),男,硕士研究生。
杨延宁(1969-),男,博士,教授。
**引用格式**:姚 旭,杨延宁,白鸿冰. 基于改进 MobileNetV2 的金属表面缺陷分类方法[J]. 计算机测量与控制,2025,33(11):259-266.

通过 CNN，系统可以从图像中自动提取复杂的特征，实现更高效的缺陷分类。

目前，研究者们已经在金属表面缺陷分类的任务中引入了多种改进技术，以提高分类的准确性和效率。文献［2］提出了一种基于 ResNet 的改进方法，通过在 ResNet 的基础上构建一个多尺度特征提取模块，能够捕捉到图像中不同层次的细节特征，尤其是在尺寸较小的缺陷识别上表现出色。通过这种方式，他们在金属表面缺陷数据集上的分类准确率显著提高[2]。文献［3］则采用了一种基于注意力机制的改进方法，通过使用通道注意力机制，使得网络能够自适应地调整对不同通道特征的关注度，从而提升对关键缺陷特征的感知能力[3]。这种方法不仅提高了网络在不同背景条件下的分类精度，而且增强了模型对图像中干扰因素的抑制效果，使得模型在处理噪声较多的工业场景时表现更为出色。文献［4］提出了一种基于多通道卷积神经网络的金属表面缺陷分类方法。该方法的核心是通过多通道输入，利用不同尺度的特征图来提升模型对细小缺陷的检测效果[4]。其模型通过多通道并行处理特征，能够有效提取来自不同尺度的缺陷信息，从而在识别小尺度、边缘模糊的缺陷时表现优异。文献［5］则通过引入自适应特征增强模块，进一步优化了网络的特征提取能力。主要通过对特征图进行自适应调整，增强网络对重要特征的表达能力，尤其是在处理复杂背景和相似纹理的缺陷时表现出色[5]。自适应特征增强模块能够根据输入图像的特征动态调整网络的参数，使得模型能够更加准确地捕捉到具有判别力的缺陷特征。通过这种方式，Zhou 等在多个实际应用场景中证明了该方法的有效性，提升了分类的准确性和鲁棒性。文献［6］通过采用稀疏卷积技术，提出了一种有效减少计算开销的缺陷分类模型。通过稀疏卷积减少了不必要的特征计算，显著降低了模型的计算复杂度和存储需求，同时保持了较高的分类精度。这种稀疏卷积技术通过选择性地跳过某些无关紧要的计算区域，能够在不影响模型性能的情况下减少冗余计算[6]。这些研究者通过引入多尺度特征融合、注意力机制、多通道处理、自适应特征增强和稀疏卷积等多种技术手段，显著提升了金属表面缺陷分类模型的性能。然而，这些方法在提升分类精度的同时，通常伴随着较大的计算开销，难以满足对实时性和高效性的需求。因此，在实际应用中，如何在精度和计算效率之间取得平衡仍然是一个需要进一步探索的问题。

为了应对这些挑战，MobileNetV2 作为一种轻量级神经网络结构，以其较少的参数量和较高的计算效率广泛应用于各种嵌入式设备和边缘计算场景中[7]。然而，MobileNetV2 在处理具有复杂背景和多样性特征的金属表面缺陷时，仍然面临特征提取不足和分类精度不高的问题。为了解决这一问题，在 MobileNetV2 网络的基础上进行了改进，提出了一种结合坐标注意力机制（CA，coordinate attention）和改进轻量化 Inception 模块的轻量级金属表面缺陷分类方法[8]。具体来说，改进包括以下三方面：（1）引入 CA 模块，通过自适应通道权重分配，增强网络对重要缺陷特征的关注度；（2）改进原始 Inception 模块，将其 3×3、5×5 标准卷积替换为深度可分离卷积，进一步减少模块的参数量及计算量。并且增强模型对不同尺度特征的融合能力，提升分类的准确率；（3）在保持模型轻量化的前提下，通过优化网络结构，降低了计算复杂度和参数量，使其更适用于资源受限的嵌入式设备和工业现场。

## 1 算法设计

### 1.1 MobileNetV2 网络

深度可分离卷积（DSC，depthwise separable convolution）[9]是 MobileNet 系列的核心技术，主要通过将标准 3×3 卷积分为 3×3 深度卷积（DWC，depth wise convolution）和 1×1 逐点卷积（PWC，point wise convolution）两步，减少了计算量和模型参数。

反向残差结构是 MobileNetV2 的创新点之一。它的设计灵感来自 ResNet 的残差结构[10]，但二者在实现方式上有所不同。ResNet 中的残差块是先对特征维度进行扩展，而 MobileNetV2 则相反，先对低维特征进行扩展，然后再通过深度卷积提取信息，最后再将特征维度压缩。首先通过一个 1×1 逐点卷积将低维特征扩展至高维空间，然后进行深度卷积，最后通过另一个 1×1 逐点卷积将特征维度还原至低维。残差连接则用于维持信息的连续性，仅在输入输出维度相同时启用。MobileNetV2 提出的另一个关键设计是线性瓶颈层。传统的卷积神经网络通常会使用非线性激活函数来提升模型的表达能力，但在低维空间中，这种非线性处理可能会导致信息丢失。因此，MobileNetV2 在反向残差块的最后一层去掉了 ReLU 激活函数，直接使用线性激活。这样可以在压缩特征维度的同时，尽可能保留特征信息，避免由于激活函数导致的信息丢失。反向残差块结构如图 1 所示。

MobileNetV2 网络的整体结构如表 1 所示，其中 Conv2d 是 2 维卷积操作，Bottleneck 为残差链接组成的瓶颈块，Avgpool 为全局池化操作，$t$ 为通道扩展因子、$c$ 为输出通道数、$n$ 为块重复次数、$s$ 为步长。

### 1.2 坐标注意力机制

在金属表面缺陷检测任务中，缺陷往往呈现为局部、不规则的形态，如裂缝、锈迹等，这些缺陷分布在金属表面的不同位置。因此，模型需要具备对表面纹理和缺陷形态的精确空间感知能力。传统的注意力机制，

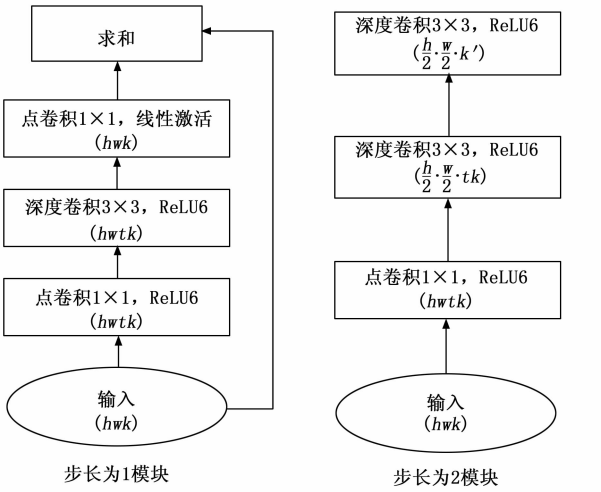



图 1　反向残差结构

表 1　MobileNetV2 网络结构图

| 输入 | 类型 | t | c | n | s |
|---|---|---|---|---|---|
| $224^2\times3$ | Conv2d | — | 32 | 1 | 2 |
| $112^2\times32$ | Bottleneck | 1 | 16 | 1 | 1 |
| $112^2\times16$ | Bottleneck | 6 | 24 | 2 | 2 |
| $56^2\times24$ | Bottleneck | 6 | 32 | 3 | 2 |
| $28^2\times32$ | Bottleneck | 6 | 64 | 4 | 2 |
| $14^2\times64$ | Bottleneck | 6 | 96 | 3 | 1 |
| $14^2\times96$ | Bottleneck | 6 | 160 | 3 | 2 |
| $7^2\times160$ | Bottleneck | 6 | 320 | 1 | 1 |
| $7^2\times320$ | Conv2d 1×1 | — | 1280 | 1 | 1 |
| $7^2\times1280$ | Avgpool 7×7 | — | — | 1 | — |
| 1×1×1280 | Conv2d 1×1 | — | class | — | — |

如 Squeeze-and-Excitation[11]，虽然能够有效地增强通道间的特征相关性，但在处理空间信息时表现较为有限，难以充分捕捉不规则缺陷的全局和局部特征。

坐标注意力机制（CA，coordinate attention）[12]通过分别对输入特征图的纵向和横向进行池化，生成两个一维的空间注意力映射。然后，将这些映射编码为通道间的依赖关系，依据这些依赖对输入特征进行加权。与 SE 不同，CA 不仅考虑了通道之间的关系，还通过保留空间坐标信息，使模型能够更专注于包含重要特征的区域。这样，网络能够重构出更具判别性的特征，从而提高分类器的准确性，增强模型在金属表面缺陷检测中的诊断性能。坐标注意力结构图如图 2 所示。

当坐标注意力模块输入一个高度为 $H$，宽度为 $W$，通道数为 $C$ 的图像时，坐标注意力模块首先会使用公式对特征图的 $H$ 和 $W$ 进行全局平均池化，分解为一对一维特征编码操作，以捕捉在这两个空间方向上的长程依赖：

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j) \tag{1}$$

接下来需要将这两个一维特征输入到共享的卷积层中，获得新的编码特征，并通过非线性激活函数调整，从而生成高效的空间注意力权重。此时，第 $C$ 个通道高度方向和 $H$ 宽度方向上的输出如式所示：

$$z_c^h(h)=\sigma\left(\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(h,i)\right) \tag{2}$$

$$z_c^w(w)=\sigma\left(\frac{1}{H}\sum_{0\leqslant i\leqslant H}x_c(j,w)\right) \tag{3}$$

通过沿宽度方向和高度方向进行编码操作可以使注意力模块不仅能够对每个通道的空间依赖关系进行建模，还能根据空间特征的变化，生成有效的注意力映射。最后利用生成的注意力映射和对原始输入特征进行加权操作。每个位置的特征不仅会根据通道间的依赖关系进行调整，还会根据空间位置的依赖进行加权。坐标注意力模块的最终输出如式所示：

$$\hat{X}(c,i,j)=X(c,i,j)\cdot z_c^w(w)\cdot z_c^h(h) \tag{4}$$

坐标注意力块不会改变输入图像的大小。在图像保持大小不变的情况下，注意力块建立 了特征卷积通道之间的联系，每个卷积通道空间内更受关注的区域有着更高权重，因而重 构后的图像能表现出更具判别性的特征。

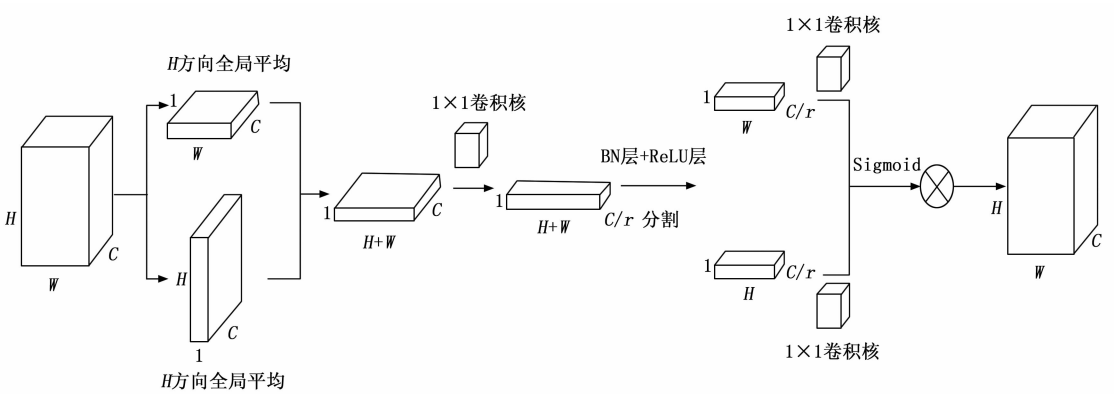



图 2　坐标注意力模块结构图

### 1.3 Inception 模块

Inception 模块最初由 Google 提出，其设计目标是通过并行的多尺度卷积操作丰富网络的特征表示能力，同时有效控制计算代价[13]。该模块通过包含不同大小卷积核（如 1×1、3×3、5×5）的多个分支以及一个池化分支，来捕捉输入图像的多尺度信息。Inception 模块的核心思想是使用 1×1 卷积来降低特征通道的维度，以减少计算量，同时提升网络的表达能力。在各卷积分支中，先通过 1×1 卷积进行降维，然后再使用较大的卷积核（如 3×3 和 5×5）提取深层次特征。通过这种设计，Inception 模块能够在不显著增加计算量的前提下，实现有效的多尺度特征提取。为了进一步降低参数量和计算成本，并提升网络的效率，本文对 Inception 模块进行了改进，借鉴了 MobileNet 系列中的深度可分离卷积（DSC，depthwise separable convolution）。DSC 能够有效分离空间卷积和通道卷积，使任意大小的标准卷积核具有较少的参数量和计算量。基于此，本文提出将 Inception 模块中的标准卷积替换为 DSC，以进一步提高计算效率[14]。本文将 Inception 模块中的 3×3 和 5×5 标准卷积分别替换为 3×3 深度卷积和 1×1 逐点卷积，以及 5×5 深度卷积和 1×1 逐点卷积。通过使用深度可分离卷积替代标准卷积，改进后的 Inception 模块在参数量和计算量上均显著降低，同时保持了对多尺度信息的良好捕捉能力。改进后的 Inception 模块如图 3 所示。

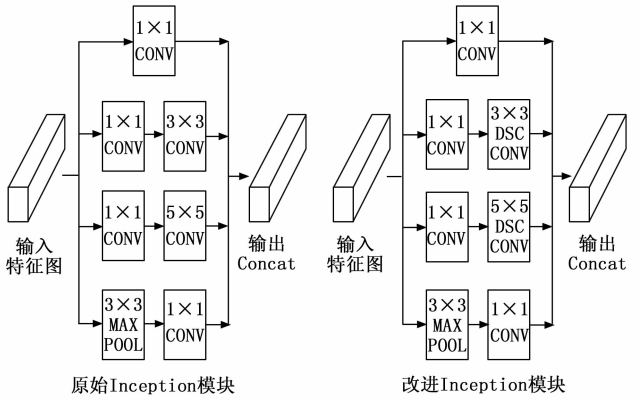



图 3 原始 Inception 模块与改进 Inception 模块

## 2 算法整体框架

由于特征图通过 MobileNetV2[15] 的反残差结构和深度可分离卷积在保证图像分类准确率的同时，提升了网络的图像训练速度。为了更有效地提取金属表面缺陷数据集上的多尺度缺陷特征，将 CA 注意力机制模块和 Inception _ DSC 模块添加到 MobileNetV2 的逆残差网络模块中，改进后的 MobileNetV2 网络结构如图 4 所示。

金属表面缺陷图片在数据集预处理阶段被统一调整为 224 像素×224 像素的 RGB 三通道格式，特征提取网络的第一个二维卷积作用是调整输入图像的通道数，降低输入图像大小，得到尺寸为 112 像素×112 像素、32 通道数的特征图 M1。M1 经过 16 个由逆残差连接组成的 Block 以及网络中插入的部分 CA 模块以及 Inception _ DSC 模块后，网络整体在参数量和计算量不显著增加的情况下，能够更有效的关注到空间中不同尺度的缺陷，提取更有效的缺陷特征。

在 MobileNetV2 网络中插入 CA 模块和 Inception _ DSC 模块时，需要考虑两者的作用特点及网络结构的特点，以确保这些模块能够有效地提升网络的提取能力。CA 模块主要是增强模型对于空间位置信息的捕捉能力，适用于需要细粒度信息的特征层。因此决定将 CA 模块插入到网络的低层或中层的特征提取阶段，在这些层中网络还保留了较高的分辨率和丰富的空间信

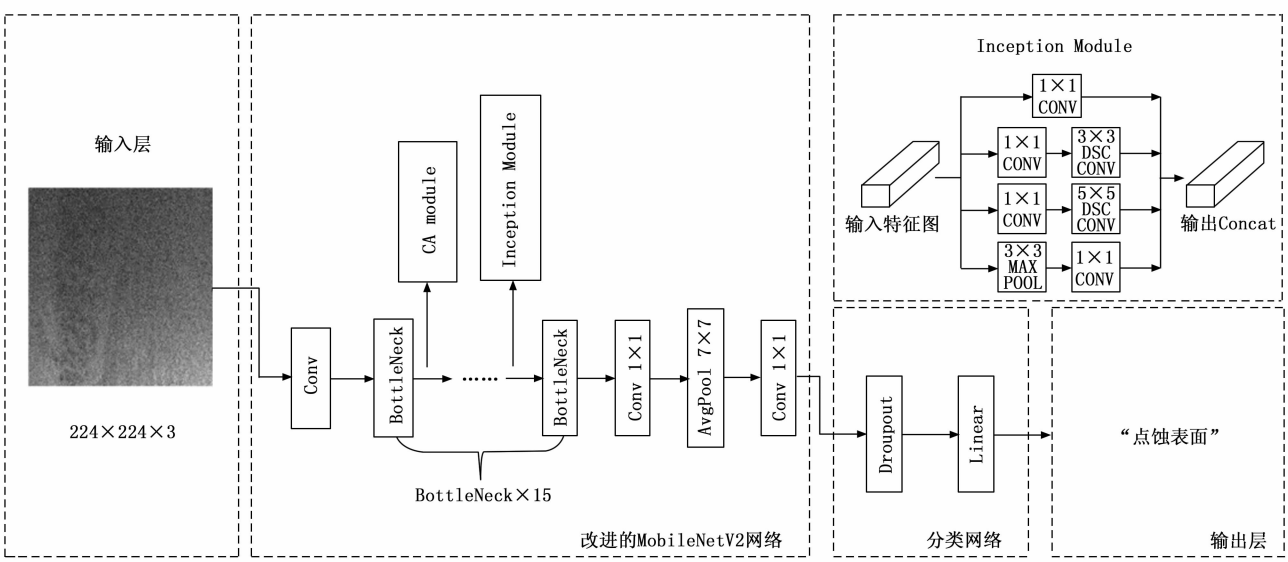



图 4 改进后的 MobileNetV2 网络结构图

息，加入 CA 模块可以帮助网络捕捉到更细粒度的缺陷和特征[16]。Inception _ DSC 模块主要用于增强网络的多尺度特征提取能力，决定将其插入在中层到高层的特征提取阶段，以便网络在更深的层次上进行多尺度特征的融合[17]。将 CA 模块和 Inception _ DSC 模块插入网络后得到新的 IC _ MobileNetV2 网络。IC _ MobileNetV2 网络结构如表 2 所示。

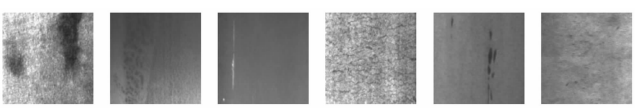
表 2　IC _ MobileNetV2 网络结构

| 输入 | 类型 | t | c | n | s |
|---|---|---|---|---|---|
| $224^2\times3$ | Conv2d | — | 32 | 1 | 2 |
| $112^2\times32$ | CA | 1 | 32 | 1 | 1 |
| $112^2\times32$ | Bottleneck | 1 | 16 | 1 | 1 |
| $112^2\times16$ | Bottleneck | 6 | 24 | 2 | 2 |
| $56^2\times24$ | Bottleneck | 6 | 32 | 3 | 2 |
| $28^2\times32$ | Inception | 1 | 32 | 1 | 1 |
| $14^2\times64$ | Bottleneck | 6 | 64 | 4 | 2 |
| $14^2\times96$ | Bottleneck | 6 | 96 | 3 | 1 |
| $14^2\times96$ | Inception | 1 | 96 | 1 | 1 |
| $14^2\times96$ | Bottleneck | 6 | 160 | 3 | 2 |
| $7^2\times160$ | Bottleneck | 6 | 320 | 1 | 1 |
| $7^2\times320$ | Conv2d 1×1 | — | 1280 | 1 | 1 |
| $7^2\times1280$ | Avgpool 7×7 | — | — | 1 | — |
| 1×1×1280 | Conv2d 1×1 | — | class | — | — |

## 3　实验结果与分析

### 3.1　数据集介绍

数据集采用由东北大学研究团队公开的 NEU-DET[18]（Northeastern University Detection of Surface Defects）金属表面缺陷数据集。数据集由金属表面图像构成，每张图像都包含一个或多个表面缺陷。每个缺陷都属于特定的类别，并带有相应的标签。该数据集共有 6 个类别 1 800 张图片，分别为轧制氧化（Rs）300 张，开裂（Cr）300 张，内含物（In）300 张，斑块（Pa）300 张，点蚀表面（Ps）300 张，划痕（Sc）300 张。每张图片分辨率均为 200 pixel×200 pixel。数据集各类别缺陷图像如图 5 所示。



(a)Ps　(b)Pa　(c)Sc　(d)Cr　(e)In　(f)Rs

图 5　金属表面缺陷数据集

### 3.2　小样本数据集 NEU-DET 的增强与划分

针对该数据集数据样本过少，可能存在过拟合风险的问题，对数据集进行扩充[19]。首先按照 8：2 的比例将数据集划分为训练集和测试集。之后通过模拟不同光照和环境条件下的图像变化、添加高斯噪声、以及通过生成对抗网络合成新的缺陷样本。此外，还对训练集进行了适度的几何变换来增强模型的鲁棒性和泛化能力。将原始训练集中每个类别的数据扩充到 700 张，共 4 200张图片。并将扩充后的数据集按照不同缺陷类别，放入不同文件夹中，方便后续训练。扩充后的训练集在一定程度上提高了模型的鲁棒性和泛化性[20]。经数据增强和划分后的数据集数量如表 3 所示。

表 3　金属表面缺陷数据集数量分布表

| 缺陷类型 | 轧制氧化 | 开裂 | 内含物 | 斑块 | 点蚀表面 | 划痕 |
|---|---|---|---|---|---|---|
| 扩充前 | 300 | 300 | 300 | 300 | 300 | 300 |
| 扩充后 | 700 | 700 | 700 | 700 | 700 | 700 |

### 3.3　实验环境及参数配置

本文实验的深度学习模型搭建在 Window11 操作系统下的 Pytorch 2.0.1 深度学习框架上，并且采用 python 3.9 版本语言编程实现。本实验硬件环境中，配备了高性能的 Intel Core i5-13490F 处理器和显存为 16 GB的 NVIDIA GeForce RTX 4060 Ti GPU，结合大容量的 16 GB 内存和 1 TB SSD 存储，能够为深度学习实验提供良好的支持。强大的并行计算能力与高速存储有助于提升模型训练的效率，缩短实验周期。

CNN 模型的参数配置对模型训练的结果至关重要。本文模型设置批处理大小（Batch _ size）为 64，为使模型充分收敛，设置训练集迭代次数（Epoch）为 100 轮。学习率使用 Adam 优化器，使其在学习过程中加速收敛。选取初始学习率为 0.0001。为防止过拟合问题，改善模型的鲁棒性，丢弃比率为 0.2。

### 3.4　Inception _ DSC 模块有效性验证

为了验证对 Inception 模块的改进效果，设计了对比实验，将原始 Inception 模块中的 3×3 和 5×5 卷积分别替换为深度可分离卷积。深度可分离卷积能够减少卷积层的参数量和计算量，以此来验证 Inception _ DSC 模块在减少参数量和计算量的前提下，能否帮助模型保持或提升分类性能。以 MobileNetV2 为基准，分别将原始 Inception 模块和 Inception _ DSC 模块插入到网络中，通过比较改进前后 Inception 模块在 MobileNetV2 骨干网络中的表现，评估其在实际任务中的有效性。本节实验评价指标为模型参数量、模型浮点计算量以及测试集最优准确率。对比结果如表 4 所示。

表 4　Inception _ DSC 模块有效性验证

| 模型 | 参数量/M | 运算数/M | 准确率/% |
|---|---|---|---|
| MobileNetV2 | 2.23 | 332.95 | 87.2 |
| MobileNetV2＋Inception | 2.31 | 342.45 | 89.2 |
| MobileNetV2＋Inception_DSC | 2.24 | 336.20 | 90.3 |

基准 MobileNetV2 模型的参数量为 2.23 M，而在加入原始 Inception 模块后，参数量略有增加，达到了

2.31 M，增幅为 0.08 M。相较之下，当使用改进后的 Inception _ DSC 模块时，参数量的增加十分有限，仅增加了 0.01 M。虽然改进的 Inception _ DSC 模块加入了额外的结构，但它在参数量上的影响几乎可以忽略不计。在计算量方面，MobileNetV2 模型的浮点运算量（FLOPs）为 332.95 M。当加入原始 Inception 模块时，浮点运算量增加至 342.45 M，表明该模块引入了额外的计算开销，带来了约 9.5M 的 FLOPs 增加。这种计算量的增加意味着在模型的推理过程中需要更多的计算资源。然而，使用改进后的 Inception _ DSC 模块后，浮点运算量仅增加至 336.20 M，相比于原始 Inception 模块带来的计算开销，改进后的 Inception _ DSC 模块仅增加了约 3.25 M 的 FLOPs。这一结果表明，Inception _ DSC 模块的引入在计算效率方面更为优化，显著减少了额外的计算负担。在分类性能上，加入原始 Inception 模块后，模型的分类准确率从基准的 87.2%提升至 89.2%，准确率提高了 2 个百分点。这表明，Inception 模块在提升模型性能方面确实发挥了积极作用。更为显著的是，当使用改进的 Inception _ DSC 模块时，准确率进一步提升至 90.3%，相较于原始 Inception 模块又提高了 1.1 个百分点。这一结果说明，DSC 不仅在减少计算开销方面表现出色，还对模型的分类性能起到了显著的促进作用。

### 3.5 消融实验

为了深入验证 CA 模块和 Inception _ DSC 模块对 MobileNetV2 网络识别准确率的提升效果，本文设计了一系列消融实验。首先，以 MobileNetV2 网络作为基准模型，在此基础上，单独在网络中插入 CA 模块、Inception _ DSC 模块，以及同时插入这两个模块，构建多种组合模型。通过这些组合模型与基准模型的性能对比，本文系统地分析了 CA 模块和 Inception _ DSC 模块在单独作用及协同作用下对网络性能的提升情况。消融实验结果如表 5 所示。

表 5 IC _ MobileNetV2 消融实验

| 模型 | 参数量/M | 运算量/M | 准确率/% |
|---|---|---|---|
| MobileNetV2 | 2.23 | 332.95 | 87.2 |
| MobileNetV2＋CA | 2.23 | 333.81 | 91.4 |
| MobileNetV2＋Inception_DSC | 2.24 | 336.20 | 90.3 |
| IC_MobileNetV2 | 2.24 | 337.12 | 92.8 |

从表 5 的结果可以看出，所有模型的参数量差异非常小，均保持在 2.23 M 至 2.24 M 之间。这表明，无论是引入 CA 模块，还是应用 Inception _ DSC 模块，甚至是两者的组合，都未显著增加模型的参数规模，符合轻量化设计的预期。IC _ MobileNetV2 的参数量为 2.24 M，仅比基础模型 MobileNetV2 增加 0.01 M，表明在引入两个改进模块后，参数量的增加非常有限。在浮点运算量（FLOPs）方面，各模型也表现出相对轻量化的特征。基础的 MobileNetV2 的 FLOPs 为 332.95 M，加入 CA 模块后，FLOPs 仅略微增加至 333.81 M，增幅约为 0.26%，这说明 CA 模块的计算复杂度较低，对整体运算量的影响较小。引入 Inception _ DSC 模块后，FLOPs 增加至 336.20 M，表明该模块对计算复杂度的提升更为明显。最终的 IC _ MobileNetV2 的 FLOPs 为 337.12 M，比基础模型增加约 1.25%，尽管二者组合略微增加了运算量，但整体仍保持较高的计算效率。在测试集准确率方面，各模型的性能提升则更为显著。基础的 MobileNetV2 的准确率为 87.2%，加入 CA 模块后，准确率提升至 91.4%，增幅达 4.2 个百分点，表明 CA 模块在捕捉特征上下文信息方面有明显的增强作用。引入 Inception _ DSC 模块的 MobileNetV2 准确率达到 90.3%，相比基础模型提升了 3.1 个百分点，说明 Inception _ DSC 通过多尺度特征融合有效提升了模型的分类能力。IC _ MobileNetV2 结合了 CA 和 Inception _ DSC 模块后，准确率进一步提升至 92.8%，相较于基础模型提升了 5.6 个百分点。这一结果表明，CA 模块和 Inception _ DSC 模块在特征提取和多尺度信息处理上具有互补性，能够显著增强模型的泛化能力。消融实验结果表明，CA 模块与 Inception _ DSC 模块的结合在不显著增加参数量和计算量的前提下，能够有效提升模型的分类性能。

### 3.6 对比实验

为了验证 IC _ MobileNetV2 改进网络的有效性，将 IC _ MobileNetV2 网络与一些主流图像分类网络模型 MobileNetV2、AlexNet、GoogleNet、DenseNet、ResNet34 以及 ResNet50 在金属缺陷数据集上的训练结果进行对比。训练前将各模型超参数调整一致，迭代次数均为 100 轮，学习率为 0.000 1，模型均采用从头开始训练方法。实验结果如表 6 所示。

表 6 不同 CNN 网络性能对比

| 模型 | 参数量/M | 浮点计算量/G | 准确率/% |
|---|---|---|---|
| MobileNetV2 | 2.23 | 0.33 | 87.2 |
| GoogleNet | 5.97 | 1.56 | 90.0 |
| DenseNet | 9.85 | 3.10 | 91.9 |
| ResNet34 | 21.28 | 3.68 | 91.1 |
| ResNet50 | 23.52 | 4.14 | 91.1 |
| IC_MobileNetV2 | 2.24 | 0.33 | 92.8 |

从参数量和浮点计算量的角度来看，模型的复杂度与准确率之间并非完全线性相关。MobileNetV2 的参数量最少和浮点计算量最低，但其准确率为 87.2%，在所列模型中处于最低水平。而较复杂的模型，如 ResNet34 和 ResNet50，尽管参数量和计算量显著增加，分别达到 21.28 M/3.68 G 和 23.52 M/4.14 G，但其准确率为

91.1%，相比 DenseNet 并没有显著优势。这表明，模型复杂度的增加并不总是显著提升模型的分类性能。IC _ MobileNetV2 在仅略微增加参数量的情况下，其准确率显著提升至 92.8%，超过了所有其他模型。针对 MobileNetV2 进行的架构改进可以在保持轻量级的同时显著提升模型性能。这在资源受限的设备上尤为重要。ResNet34 和 ResNet50 作为典型的深度残差网络，尽管浮点计算量和参数量较高，但它们的准确率与 DenseNet 相近，且并未超越 IC _ MobileNetV2。在某些任务中，网络深度和复杂度的增加可能面临性能瓶颈。

在图 6 中，横轴表示不同 CNN 网络模型的浮点计算量，纵轴表示模型在测试集上的最佳准确率，圆的大小则代表模型的参数量。改进后的 IC _ MobileNetV2 模型在保持较低参数量和计算量的同时，实现了较高的准确率，其最佳准确率达到 92.8%，相比 MobileNetV2 提升了 5.6 个百分点。该模型的准确率显著高于其他对比模型，展示了在实际部署中具有潜在的应用价值。

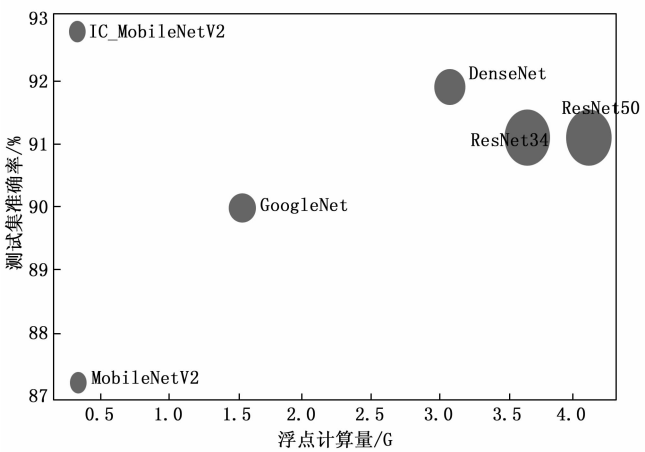



图 6　不同 CNN 网络性能比较

图 7 为 NEU-DET 测试集准确率曲线。MobileNetV2 和 GoogleNet 在早期的训练阶段收敛较快，但最终的验证准确率略低于其他模型，显示其在计算效率上的优势但可能牺牲了一定的精度。DenseNet 和 ResNet34 以及 ResNet50 达到了较高的最终验证准确率，尤其是 ResNet50 表现出最优的准确率，说明深层网络能够更好地捕捉数据中的复杂特征。IC _ MobileNetV2 在验证集上的表现接近 DenseNe 和 ResNet50，显示了改进后的轻量化模型在权衡计算复杂度与精度方面的优越性。

为了进一步验证 IC _ MobileNetV2 对于缺陷识别的有效性，同时选取来自德国模式识别协会（DAGM）于 2007 年发布的 DAGM 2007 数据集，该数据集包含 10 个不同类别的图像，每个类别包含 1 000 张“无缺陷”图像和 150 张“有缺陷图像，总计 11 500 张图

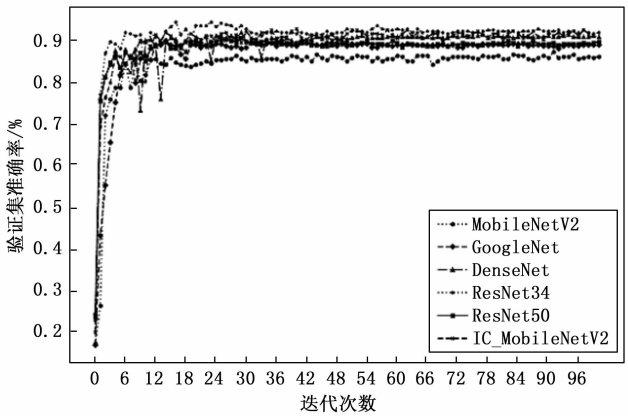



图 7　测试集准确率

像”。所有图像均为 8 位灰度 PNG 格式，模拟工业生产环境中的表面纹理和缺陷。IC _ MobileNetV2 在 DAGM 2007 数据集中准确率达到了 84.4%，相对于其他方法准确率也有一定提升。

表 7　DAGM 2007 数据集对比实验

| 网络模型 | 测试集准确率/% |
|---|---|
| MobileNet | 81.6 |
| DenseNet | 82.3 |
| GoogleNet | 82.4 |
| ResNet34 | 82.3 |
| ResNet50 | 81.7 |
| IC_MobileNet | 84.4 |

### 3.7　宽度因子对于网络的影响

为了研究不同宽度因子对 MobileNetV2 网络性能的影响，设计了一组对比实验。宽度因子是 MobileNetV2 网络中的一个重要控制参数，通过调节宽度因子，可以缩放每一层卷积的通道数，从而影响模型的参数量和计算复杂度。选取了 0.25、0.5、0.75 和 1.0 四种宽度因子进行测试，并在相同的训练配置下对每种模型进行了训练和评估。采用分类准确率、模型参数量和浮点计算量作为评估指标。对比实验如表 8 所示。

表 8　不同宽度因子对网络性能的影响

| 模型 | 参数量/M | 浮点计算量/M | 准确率/% |
|---|---|---|---|
| IC_MobileNetV2(1.0) | 2.24 | 337.12 | 92.8 |
| IC_MobileNetV2(0.75) | 1.28 | 202.05 | 91.7 |
| IC_MobileNetV2(0.5) | 0.58 | 100.84 | 91.4 |
| IC_MobileNetV2(0.25) | 0.16 | 33.49 | 78.6 |

当宽度因子为 1.0 时，模型在 NEU－DET 数据集上的分类准确率达到 92.8%，这是性能最优的配置。而当宽度因子降至 0.75 和 0.5 时，准确率分别下降至 91.7%和 91.4%。虽然精度有所下降，但这些结果依然表明，在一定范围内降低宽度因子对模型的准确率影

响相对较小，且依旧保持了较高的性能。适度缩小模型宽度可以在性能与复杂度之间取得较好的平衡。然而，当宽度因子进一步减小至 0.25 时，准确率骤降至 78.6%，过度减少模型宽度会显著削弱其特征提取能力，导致分类性能的明显下降。

## 4 结束语

基于 MobileNet 系列的深度可分离卷积思想，提出了一种轻量化的 Inception _ DSC 模块，并将改进后的 Inception 模块与 CA 模块有效地嵌入 MobileNetV2 网络中，从而增强了网络对关键特征的提取能力。在保持模型轻量化的同时，显著提升了分类准确率。

通过在 NEU-DET 数据集上的对比实验，验证了 Inception _ DSC 模块的有效性；消融实验结果进一步表明，CA 模块和 Inception _ DSC 模块均对网络性能的提升具有重要贡献。最终，通过将改进后的 IC _ MobileNetV2 与主流模型 MobileNetV2、AlexNet、GoogleNet、DenseNet、ResNet34 和 ResNet50 进行对比，IC _ MobileNetV2 在准确率方面分别提升了 5.6、2.8、0.9、1.7 和 1.7 个百分点。实验结果表明，改进后的网络在模型复杂度和分类准确率之间达到了较好的平衡，展示了较强的实际应用潜力。

**参考文献：**


[1] BHATT P M, MALHAN R K, RAJENDRAN P, et al. Image-based surface defect detection using deep learning: A review [J]. Journal of Computing and Information Science in Engineering, 2021, 21 (4): 040801.

[2] SONG X, CHEN K, CAO Z. ResNet-based image classification of railway shelling defect [C] //Proceedings of the 39th Chinese Control Conference (CCC), IEEE, 2020: 6589 - 6593.

[3] XIE X, LI C, LIU Y, et al. An efficient channel attention-enhanced lightweight neural network model for metal surface defect detection [J]. Journal of Circuits, Systems and Computers, 2023, 32 (10): 2350178.

[4] CHEON S, LEE H, KIM C O, et al. Convolutional neural network for wafer surface defect classification and the detection of unknown defect class [J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32 (2): 163 - 170.

[5] LIU T, HE Z, LIN Z, et al. An adaptive image segmentation network for surface defect detection [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35 (6): 8510 - 8523.

[6] ZHOU S, CHEN Y, ZHANG D, et al. Classification of surface defects on steel sheet using convolutional neural networks [J]. Mater. Technol, 2017, 51 (1): 123 - 131.

[7] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510 - 4520.

[8] 胡 坤，吴国庆，胡祖辉，等. 基于改进的 VGG16 网络金属表面缺陷图像分类研究 [J]. 计算机应用与软件，2024，41 (6)：175 - 180.

[9] HOWARD A G. Mobilenets: Efficient convolutional neural networks for mobile vision applications [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2261 - 2270.

[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 - 778.

[11] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132 - 7141.

[12] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713 - 13722.

[13] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1 - 9.

[14] CHOLLET F. Xception: Deep learning with depthwise separable convolutions [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251 - 1258.

[15] XU Y, ZHANG K, WANG L. Metal surface defect detection using modified YOLO [J]. Algorithms, 2021, 14 (9): 257.

[16] TANG T, CUI Y, FENG R, et al. Vehicle target recognition in SAR images with complex scenes based on mixed attention mechanism [J]. Information, 2024, 15 (3): 159.

[17] RYBCZAK M, KOZAKIEWICZ K. Deep machine learning of mobileNet, efficient, and inception models [J]. Algorithms, 2024, 17 (3): 96.

[18] ZHANG J, KANG X, NI H, et al. Surface defect detection of steel strips based on classification priority YOLOv3-dense network [J]. Ironmaking & Steelmaking, 2021, 48 (5): 547 - 558.

[19] LV X, DUAN F, JIANG J, et al. Deep metallic surface defect detection: The new benchmark and detection network [J]. Sensors, 2020, 20 (6): 1562.

[20] 陆雅诺，陈炳才，陈德刚，等. 一种基于注意力模型的带钢表面缺陷识别算法 [J]. 激光与光电子学进展，2021，58 (14)：242 - 250.