arxiv:2506.21135v1 中译本

YOLO-FDA：集成层次注意力和细节增强用于表面缺陷检测

Jiawei Hu[1][0009−0007−2476−3124]

College of Mechanical and Electrical Engineering, China Jiliang University,
Hangzhou, Zhejiang, China
2200110409@cjlu.edu.cn

摘要 工业场景中的表面缺陷检测既至关重要又具有技术挑战性，因为缺陷类型多样、形状和尺寸不规则、要求精细且材料纹理复杂。尽管基于 AI 的检测器在性能上有了显著提升，现有方法通常仍存在冗余特征、细节敏感度不足以及多尺度条件下的鲁棒性较弱等问题。为了解决这些问题，我们提出了一种新的 YOLO 基础检测框架——YOLO-FDA，该框架集成了精细细节增强和注意力引导的特征融合。具体来说，我们采用一种 BiFPN 样式的架构来强化 YOLOv5 主干网络中的双向多层特征聚合。为了更好地捕捉细小结构变化，我们引入了细节方向融合模块（DDFM），在次底层中加入方向性非对称卷积以丰富空间细节，并将次底层与低层级特征融合以增强语义一致性。此外，我们提出了两种新的基于注意力的融合策略——注意力加权连接（AC）和跨层注意力融合（CAF）——以改善上下文表示并减少特征噪声。在基准数据集上的广泛实验表明，YOLO-FDA 无论是在不同类型的缺陷还是尺度上，在准确性和鲁棒性方面都始终优于现有的最先进的方法。

Keywords: BiFPN · 细节方向融合模块 · 注意力加权连接 · 跨层注意力融合。

1 介绍

工业制造中的表面缺陷经常损害产品的结构完整性，并可能导致显著的经济损失。传统的手动检查方法劳动密集、容易出错且极易受到环境和人为因素的影响，难以保证一致的质量保障。随着深度学习的快速发展，智能缺陷检测方法因其高精度、可扩展性和鲁棒性而日益受到关注。特别是，由 AI 驱动的实时检测作为一种防止生产故障源头的有前景解决方案已经出现 [1]。

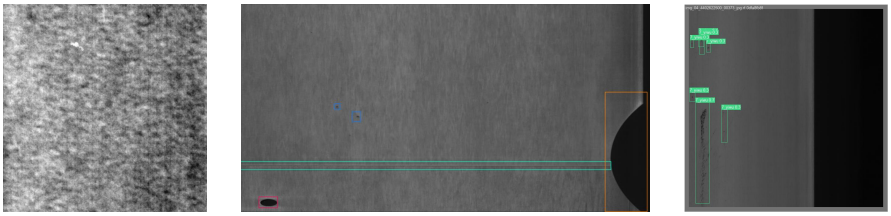

图 1: 来自 DAGM2007 和 GC10-DET 的表面缺陷示例。

然而，在复杂背景条件下识别小规模、不规则形状缺陷仍然存在挑战。这些困难主要源于当前检测框架中细节保留不足、多尺度特征融合不佳以及情境感知有限的问题 [2]。

早期的目标检测框架，如两阶段的 R-CNN 系列 [3–7] 展示了强大的定位能力，但由于依赖显式的区域建议，导致计算成本高和推理时间长。虽然后续的工作如 Guided Anchoring [8] 和 Cascade RPN [9] 优化了区域提议机制，但两阶段管道对于实时工业部署仍然效率低下。为了解决这些限制，单阶段检测器如 SSD [10] 和 YOLO 系列出现了更实用的替代方案，提供更快的推理速度和简化的训练过程。YOLOv3 [11] 引入了特征金字塔网络（FPN）以增强多尺度融合，而 YOLOv4 [12] 通过加入 PANet 进一步改进了融合架构。最近，YOLOv8 [13] 采用了无锚点的方法，并修改了 neck 结构以提高检测速度和泛化能力。这些进展已经通过定制的轻量级模块和增强的功能提取器扩展到了表面缺陷检测 [14–17] 。类似于虚拟试穿和以人类为中心的任务的趋势 [18, 19] ，在这些任务中设计了专门的结构来捕捉细粒度纹理和空间语义，工业缺陷检测也要求对局部细节和全局布局先验进行仔细建模。

尽管有了这些改进，当将基于 YOLO 的模型应用于表面缺陷检测时，仍然存在一些限制。首先，如图 1 中的第一幅图片所示，钢材、玻璃和织物等材料上的缺陷类型多样，以及纹理、光泽和反射率的微妙变化，通常会导致在复杂工业背景下出现高特征噪声和低辨别性 [2]。其次，许多缺陷非常小或拉长，在下采样过程中容易丢失，并且传统特征融合策略无法很好地表示这些缺陷 [20]。如图 1 中的第二幅图片的蓝色方框所示，GC10-DET 中称为夹杂 (In) 的缺陷类别通常以较小的形式出现，而绿色方框显示的 GC10-DET 中的焊缝缺陷类别通常具有更大的长宽比。第三，如图 1 的最后一幅图片所示，尽管近期的 YOLO 变体尝试整合复杂的融合模块，但它们经常引入信息冗余，这可能在不影响语义精度或空间一致性的情况下降低检测性能 [1]。

为了解决上述表面缺陷检测中的挑战，我们提出了一种名为 YOLO-FDA 的新框架，该框架增强了 YOLO 架构中的细节提取和特征融合。具体来说，我们将 YOLOv5 中原有的 PANet 颈部替换为更强大的双向特征金字塔网络结构，以增强双向多尺度信息流。为了更好地保留并强调小或不规则缺陷的细粒度特征，我们设计了细节方向融合模块（DDFM）来捕捉方向纹理，并明确将低层次的空间细节与高层次语义融合。此外，我们引入了两种新颖的关注机制：注意力加权连接（AC）通过保持所有输入通道并分配自适应权重确保特征丰富性，而跨层注意力融合（CAF）执行选择性的逐通道加权以抑制冗余或不相关的特征。这些模块共同提高了模型在杂乱背景中检测细微缺陷的能力。

本文的主要贡献总结如下：

- 我们用 BiFPN 结构替换了 YOLOv5 中的 PANet 模块，以增强双向多尺度特征融合。然后，我们引入了细节方向融合模块（DDFM）来加强细粒度特征表示并提高对局部缺陷的敏感性。
- 我们提出了两种基于注意力的融合策略：注意力加权连接（AC），在融合过程中保持特征丰富性，以及跨层注意力融合（CAF），执行通道感知加权以减少特征冗余并提高可解释性。
- 在 GC10-DET 和 DAGM2007 数据集上的实验评估表明，YOLO-FDA 在准确性和鲁棒性方面优于最先进的表面缺陷检测方法。

2 相关工作

2.1 基于 YOLO 的表面缺陷检测

YOLO（一次只看一遍）系列从 YOLOv1 到 [21]，实现了快速的端到端检测，再到现代变种如 YOLOv5 [22] 和 YOLOv8 [13]，这些变种采用了 CSPDarknet、PANet 和无锚点设计来提高准确性和效率。Yolo-World [23] 通过视觉语言预训练进一步扩展了 YOLO 在开放场景中的适用性。对于表面缺陷检测，SLF-YOLO [14] 利用通道门控和轻量级颈部结构以提高效率，而 QCF-YOLO [24] 通过使用 GhostConv 和 P2-head 来增强对微小缺陷的检测。然而，这些模型在面对复杂背景和不规则形状的缺陷时仍然存在问题，这是由于多尺度融合有限以及冗余特征聚合造成的。为了解决这些问题，我们建议引入不对称卷积以进行细节建模，并使用跨层注意力融合来抑制冗余——这借鉴了频率域交互和布局感知增强在航空和 SAR 检测中的应用 [25, 26]。

2.2 缺陷检测中的注意力机制

Transformer [27] 的引入引发了视觉任务中注意力机制的广泛采用。模块如挤压和激励（SE）[28] 和 CBAM [29] 已被广泛用于自适应地重新校准通道和空间特征。这些注意力设计也被应用于缺陷检测，例如在 YOLO-HMC [30] 中使用的 MCBAM，有效地增强了对小目标的检测。然而，传统的注意力模块仍然存在局限性。SE 仅在各个单独的通道内操作，使得跨尺度特征比较变得困难，而 CBAM 则顺序地应用空间和通道注意力，并没有进行联合优化。受到选择性频率建模 [31] 和精炼金字塔架构 [32] 的先前工作的启发，我们通过两个新颖的模块增强了注意力集成：AC 保留了丰富的特征而不会造成信息损失；CAF 在执行加权合并的同时保持原有的通道维度，从而减少了信息冗余。这些注意力策略旨在保持不同尺度间的语义一致性，这一方法与近期遥感研究中的频域见解一致 [25]。

3 提出的方法

3.1 概述

我们提出的 YOLO-FDA 框架如图 2 所示。基本的 YOLOv5 模型由三个部分组成：特征提取的主干、特征融合的颈部和生成边界框、类别概率及对象得分的头部。基于这一基础，我们提出了两项创新：（1）我们在基础上增加了 DDFM 模块以增强细节信息的融合；（2）在多个路径融合的位置使用 AC 和 CAF 模块，保持网络前端的信息丰富性，并减少后期的信息冗余和相互干扰。

3.2 细节方向融合模块

基本的 YOLOv5 采用 PANet 的特征融合方法，在 FPN 的基础上添加了一个自下而上的路径，使得信息可以双向传输并加强了信息融合。然而，在表面缺陷目标检测领域，对细节特征以及不同方向和大长宽比的目标的要求很高，PANet 甚至 BiFPN 的特征融合无法满足检测要求。因此，我们提出了 DDFM，并将其集成到特征融合过程中。

如图 2 中的紫色虚线框所示，我们提出了 DDFM，该方法首先需要将 YOLOv5 主干的第 4 层输出上采样，使其长度和宽度增加 2 倍，然后再与第二层的输出进行拼接，进一步增强基本 YOLOv5 主干第四层输出特征图的

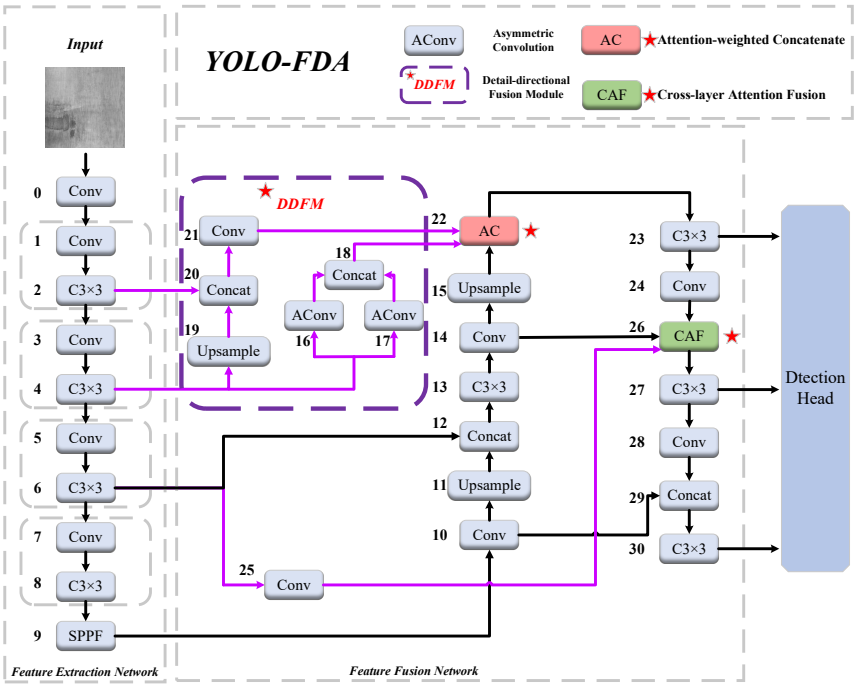

图 2: 提出的 YOLO-FDA 模型的整体架构。

详细信息，使模型能够捕捉到更多详细的特征。此外，主干的第四层输出特征图经过两次非对称卷积处理后再进行拼接。具体来说，在非对称卷积过程中，我们分别使用 1×7 和 7×1 核来提取水平和垂直特征。形式上，这两个操作可以表示为：

水平：

$$F_h(x, y, c') = \sum_{k=-3}^{3} \sum_{c=1}^{C} w^h_{c',c,k} \cdot F(x, y+k, c), \tag{1}$$

垂直：

$$F_v(x, y, c') = \sum_{k=-3}^{3} \sum_{c=1}^{C} w^v_{c',c,k} \cdot F(x+k, y, c). \tag{2}$$

这里，$F(x, y, c)$ 表示空间位置 (x, y) 和通道 c 处的输入特征图，而 F_h 和 F_v 分别表示水平和垂直过滤后的输出特征。不对称卷积权重分别表示为 $w^h_{c',c,k}$ 和 $w^v_{c',c,k}$，对应水平和垂直方向。

为了保持与原始输入相同的通道数量，我们将每个不对称卷积的输出通道维度设置为输入通道的一半。具体来说，如果输入特征图有 C 个通道，则 F_h 和 F_v 都将有 $C/2$ 个通道。这种方向感知表示增强了模型区分具有强烈方向性特征的缺陷类型的能力。

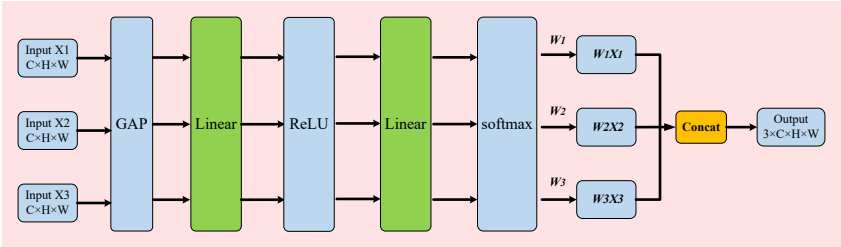

图 3: 所提出的注意力加权连接模块的示例。

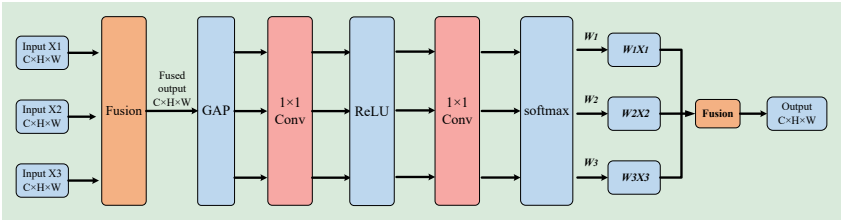

图 4: 所提出的跨层注意力融合模块的说明。

3.3 注意力融合模块

YOLOv5 的特征融合方法是普通的拼接。虽然实现成本低且计算量小，但在多路径融合过程中会无选择性地拼接所有特征，并且没有权重，导致信息冗余或冲突。因此，在 YOLO-FDA 的两个三路径融合节点中，我们考虑使用不同的注意力融合机制：**交流**和 CAF。

AC. 在早期融合阶段（第 22 层），保留详细且多样的信息是至关重要的。因此，我们提出一个 AC 模块，在连接之前对每个输入特征图进行加权，使网络能够强调更有信息量的来源，同时仍然保留所有输入通道。AC 模块的实现过程如图 3 所示。

给定一组 N 输入特征图 $\{F_i\}_{i=1}^{N}$，每个形状为 $B \times C \times H \times W$，我们首先计算它们的全局平均池化结果，为后续的全连接注意力网络做准备，可以表示为：

$$g_i = \mathrm{GAP}(F_i) \in \mathbb{R}^{B \times C}. \tag{3}$$

每个 g_i 都通过一个两层全连接注意力网络生成标量注意力分数：

$$s_i = \mathrm{FC}(g_i) = W_2 \cdot \mathrm{ReLU}(W_1 \cdot g_i). \tag{4}$$

其中，$W_1 \in \mathbb{R}^{C' \times C}$ 和 $W_2 \in \mathbb{R}^{1 \times C'}$ 是用于计算注意力分数 s_i 的两个全连接层的权重，而 $C' = C/r$ 是在缩减率为 r 的情况下减少的维度。

归一化的注意力权重通过 softmax 获得：

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)}. \tag{5}$$

最后，加权特征图被拼接在一起：

$$F_{\mathrm{AC}} = \mathrm{Concat}(\alpha_1 F_1, \alpha_2 F_2, \ldots, \alpha_N F_N). \tag{6}$$

该机制保留了所有特征图的完整通道信息，同时强调了它们的相对重要性。它在早期融合阶段特别有效，在这些阶段特征冗余较低但信息多样性较高。

CAF. 在更深层（第 26 层），不同的特征可能包含重叠或冗余信息。为防止过拟合并抑制对同一缺陷的重复检测，我们提出了一种 CAF 模块，该模块学习跨层加权求和而非拼接。CAF 模块的实现过程如图 4 所示。

给定输入特征图 $N\{F_i\}_{i=1}^{N}$，我们首先计算它们的平均值以获得一个代表性的特征图：

$$F_{\mathrm{avg}} = \frac{1}{N}\sum_{i=1}^{N} F_i. \tag{7}$$

接下来，我们应用全局平均池化将空间信息浓缩成通道描述符：

$$x = \mathrm{GAP}(F_{\mathrm{avg}}) \in \mathbb{R}^{B\times C\times 1\times 1}. \tag{8}$$

为了减少参数并提取紧凑的表示，一个 1×1 卷积将通道维度从 C 减少到 C'=C/r:

$$s^{(1)} = W_1 * x \in \mathbb{R}^{B\times C'\times 1\times 1}, \quad C' = \frac{C}{r}. \tag{9}$$

这里，$W_1 \in \mathbb{R}^{C'\times C\times 1\times 1}$ 表示该卷积层的学习权重。

然后我们应用 ReLU 激活函数以引入非线性和促进特征稀疏性：

$$z = \mathrm{ReLU}(s^{(1)}) \in \mathbb{R}^{B\times C'\times 1\times 1}. \tag{10}$$

简化后的特征映射到 N 个通道，生成每个输入图的原始注意力分数：

$$s^{(2)} = W_2 * z \in \mathbb{R}^{B\times N\times 1\times 1}. \tag{11}$$

这里，$W_2 \in \mathbb{R}^{N\times C'\times 1\times 1}$ 表示第二次卷积的学习权重。

为了获得总和为 1 的可比较权重，我们使用 Softmax 在 N 维度上对分数进行归一化：

$$\boldsymbol{\beta} = \mathrm{Softmax}(s^{(2)}) \in \mathbb{R}^{B\times N\times 1\times 1}. \tag{12}$$

最后，我们将归一化的权重 $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]$ 应用于融合原始特征图：

$$F_{\mathrm{CAF}} = \sum_{i=1}^{N} \beta_i F_i, \quad F_{\mathrm{CAF}} \in \mathbb{R}^{B\times C\times H\times W}. \tag{13}$$

该策略有效抑制了冗余响应，并鼓励模型有选择性地关注最有用的逐层特征。由于输出具有与输入相同的维度，因此保持了结构一致性和计算效率。

4 实验与分析

为了验证提出的 YOLO-FDA 的优越性，将其与多个缺陷检测方法在两个数据集上进行了比较，即 GC10-DET 和 DAGM2007。

4.1 数据集

GC10-DET. GC10-DET [33] 数据集是由 2294 张灰度图像组成的集合，这些图像是从真实的工业环境中收集的，全部为各种钢材类型的表面缺陷。该数据集将 10 种不同类型的缺陷分类，即冲压（Pu）、焊接（Wl）、坑洞（Cg）、水斑（Ws）、油斑（Os）、丝状斑点（Ss）、夹杂物（In）、轧制坑（Rp）、褶皱（Crease）和腰折（Wf），为钢铁行业缺陷检测算法的开发和测试提供了广泛的资源。
DAGM2007 DAGM2007 [34, 35] 中展示了十类人工生成的缺陷。发布的数据集中的真实值是一个包含缺陷的椭圆区域。我们使用 DAGM2007 数据集将椭圆区域转换为用于缺陷检测的边界框。DAGM2007 包含 14,000 张无缺陷图像和 2,100 张有缺陷图像。这是一个弱监督数据集，并以灰度 8 位 PNG 格式保存。

4.2 评估指标

根据 [36] 中的标准，分别使用精确度、召回率和 mAP 来定量评估在 GC10-DET 和 DAGM2007 上的检测性能。

4.3 实现细节

训练详情。我们在 PyCharm 上使用 PyTorch 2.0.0 实现了我们的模型，并使用配备 24GB 内存的 NVIDIA 4090D 训练了网络。根据之前的研究所述 [36, 37]，我们分别将 GC10-DET 和 DAGM2007 数据集按照 8:1:1 的比例通过分层随机抽样划分为训练集、测试集和验证集。在训练阶段，我们将输入大小设置为 640 × 640。关于网络架构的细节，我们的模型基于 Yolov5，并使用 CSPDarknet [38] 和 SPPF [39] 作为主干。
参数设置。此外，该模型通过随机梯度下降（SGD）优化器进行优化，学习率为 0.01，权重衰减为 0.0005，批量大小为 8。epoch 数设置为 250。

表 1: 比较方法在 GC10-DET 数据集上的定量结果。

Model	Precision (%)	Recall (%)	mAP50 (%)
YOLOv3 [11]	68.9	52.2	61.8
YOLOv7_tiny [42]	59.0	63.8	62.6
YOLOv5n [43]	64.9	68.1	68.3
YOLOv8n [44]	61.9	66.5	68.3
YOLOv7 [42]	65.3	62.1	69.4
YOLOv9s [41]	68.2	67.7	70.2
YOLOv5s [43]	70.8	69.4	70.3
YOLOv5s-改进 [41]	71.9	68.4	71.0
领域特定语言-YOLO [40]	70.5	67.8	72.5
我们的	76.9	71.1	74.9

4.4 与最新方法的比较

GC10-DET **数据集上的比较**。我们将提出的方法与 GC10-DET 和 DAGM2007 上的代表性缺陷检测方法进行了比较。Table 1显示了我们在 GC10-DET 上与其他几种代表性缺陷检测方法的定量比较。YOLOv3、YOLOv7_tiny、YOLOv5n、YOLOV8n、YOLOv7 和 DSL-YOLO 的数据均来自文献 [40]，在 GPU 类型等方面与我们的论文略有不同，其训练次数为 200 轮。YOLOv9s、YOLOv5s-improved 的数据取自论文 [41]，同样在 GPU 类型等方面与本文略有差异，训练轮数为 500，批次大小为 32。YOLOv5s 的结果是在与本文描述相同的实验设置和环境下获得的。我们提出的 YOLO-FDA 模型在 GC10-DET 上实现了最高的 mAP50 74.9%，比基线提升了 4.6%，显著超越了许多其他 YOLO 系列模型，如 YOLOv7 和 YOLOv9，在 mAP50 指标上的得分分别为 69.4%和 70.2%。其次，YOLO-FDA 在精度指数上也表现出色，达到 76.9%，优于其他代表性 YOLO 模型。在召回率指标方面，我们的模型表现同样出色，实现了 71.7%的召回率，高于其他经典模型。

DAGM2007 **数据集上的比较**。DAGM2007 是一张人工生成的图像，包含十种不同的背景纹理图案，并且有更多的小目标。Table 2显示了我们提出的模型和一些代表性算法在 DAGM2007 上的训练结果。其中，SSD、Faster RCNN、RetinaNet 和 Cascade R-CNN 的数据均来自文献 [45]，使用的环境和参数与本文略有不同，例如训练周期的数量和学习率。YOLOv5n、YOLOv5s、HIC-YOLOv5、YOLOv7_tiny 和 YOLOv9 的数据均使用了与本文相同的环境和参数进行训练。从表格中可以看出，我们提出的 YOLO-FDA 模型在许多指标上表现良好。首先，值得注意的是，我们的模型在 mAP50-95 指标上达到了 67.9%，比 YOLOv5 提升了 2.7%，并且超过了表中的其他模型。在各类别的 mAP 比较方面，当测试 C1 时，我们模型的表现也极为出色，达

表 2: DAGM2007 数据集上比较方法的定量结果。

方法	mAP50-95 (%)	精度（%）	召回率（%）	每个类别的 mAP 值									
				C1 (%)	C2 (%)	C3 (%)	C4 (%)	C5 (%)	C6 (%)	C7 (%)	C8 (%)	C9 (%)	C10 (%)
YOLOv7_tiny [42]	60.5	88.3	95.2	67.0	71.0	65.4	74.8	67.9	32.3	46.9	40.8	67.1	62.6
SSD [10]	63.2	96.0	96.4	52.6	67.3	56.1	68.5	71.7	65.8	59.5	48.2	69.7	72.6
HIC-YOLOv5 [46]	64.2	87.7	94.3	73.5	72.9	61.6	64.2	75.0	50.5	56.3	45.4	70.6	71.7
更快的 RCNN [5]	64.6	75.2	98.6	60.5	64.4	56.5	69.2	69.6	70.6	58.8	51.9	72.2	72.0
YOLOv5n [22]	64.9	95.6	98.0	71.3	75.7	43.6	74.5	74.8	62.8	60.8	45.4	68.8	71.2
视网膜网络 [47]	65.2	–	–	62.5	66.9	56.0	70.2	72.8	71.7	60.9	53.7	69.8	71.9
YOLOv5s [22]	65.2	91.1	98.0	75.9	66.5	61.4	74.8	72.9	57.6	52.7	49.6	69.6	71.3
级联 R-CNN [7]	66.0	–	–	63.1	66.2	56.4	66.5	72.0	75.4	62.8	53.0	69.8	73.4
YOLOv9 [48]	67.1	95.6	93.8	61.6	78.7	60.5	71.0	78.7	63.3	64.8	47.6	71.2	71.7
我们的	67.9	94.7	98.4	84.5	76.7	52.9	72.7	77.1	74.4	48.3	52.6	70.5	69.6

到 84.5%，远远超过了许多经典模型。同时，在检测任务的 C2、C4、C5、C6、C8 和 C9 中，我们模型的表现同样处于领先地位，分别达到了 76.7%、72.7%、77.1%、74.4%、52.6% 和 70.5%。上述数据都反映了 YOLO-FDA 的优秀性能。

4.5 消融研究与分析

为了展示所提出的模型中每个创新点的作用和意义，我们设计了六项消融实验，并在 GC10-DET 上进行了实验。结果如图 5 所示。所有实验的实验环境和参数设置均与本文前面所述一致。第一次实验仅在 YOLOv5 上进行，用于与其他消融模型进行比较；第二次实验将 YOLOv5 的颈部部分从 PANet 替换为 BiFPN，以验证 BiFPN 架构的效果；第三次实验在 BiFPN 的基础上添加 DDFM，展示这一创新带来的性能变化；第四次实验在第三次实验基础上添加 CAF 模块，反映其效果；为了反映出 AC 和 CAF 模块在特征融合过程中不同的效果，我们设计了第五次实验，在第三次实验基础上添加 AC 模块代替 CAF；第六次实验结合本文提出的所有创新点，反映每个模块协作工作的效果。

以下是对所提出的 YOLO-FDA 从四个方面进行的全面分析，以探究其优越性的逻辑。

BiFPN **的作用**。来自实验 2，我们可以看到，在 YOLOv5 的特征融合部分使用 BiFPN 架构后，mAP50 和 mAP50-95 指标得到了提升，分别比基线高出 1.4%和 0.5%。这主要是因为 BiPFN 在同一尺度层级之间增加了额外的残差连接，使得水平方向上的特征传播更加充分。

DDFM **的影响**。从实验 3 中，我们可以看到在 BiFPN 基础上添加 DDFM 模块后，mAP50 和 mAP50-95 指标略有下降。通过分析检测结果，我们认为这是因为这种改进增强了对小目标细节及不同方向缺陷的敏感性，但在融合

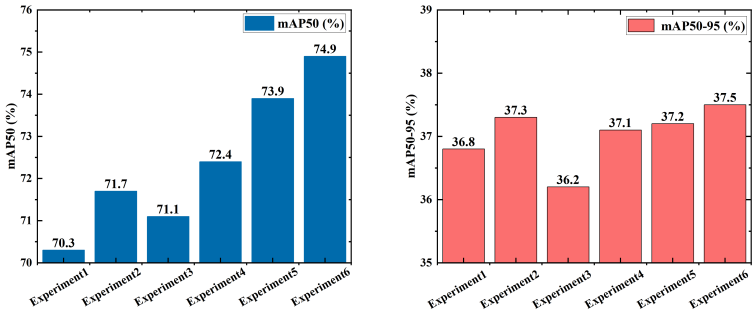

图 5: 不同消融模型在 GC10-DET 数据集上的定量结果。

时仍采用简单的拼接方法，会导致大量信息冗余，影响最终检测头的判断，并降低 mAP 指标。

AC 和 CAF 的影响。从实验 4 和 5 中，我们可以看到 AC 模块和 CAF 模块的不同特征融合方法会带来不同的效果。如果仅使用 CAF 模块，YOLOv5 的训练结果在 mAP50 和 mAP50-95 上分别可以提升 2.1%和 0.3%；而如果仅使用 AC，则分别可以提升 3.6%和 0.4%。我们相信这是因为虽然 AC 融合方法可以通过注意力机制学习不同特征图的权重，但最终仍然采用连接 (concatenate) 的方法，这可能会导致不适当的信息融合。仅使用 CAF 方法进行特征融合会在融合初期造成信息损失，导致在训练后期模型学习时特征饱和，并达到信息瓶颈。

总体效果。实验 6 表明，在特征融合阶段混合使用 AC 和 CAF 模块可以带来比单独使用其中之一更好的结果。结合 DDFM，我们提出的 YOLO-FDA 模型在 GC10-DET 上可以达到 mAP50 为 74.9%和 mAP50-95 为 37.5%，分别比基线高出 4.6%和 0.7%。

4.6 可视化

如图 6 所示，由于增加了细节特征的融合路径并提高了对多尺度和多方向特征的敏感度，我们提出的 YOLO-FDA 模型可以更全面地检测具有较大纵横比或小目标缺陷。同时，与 YOLOv5 模型相比，我们的模型更有助于减少重复检测的现象。

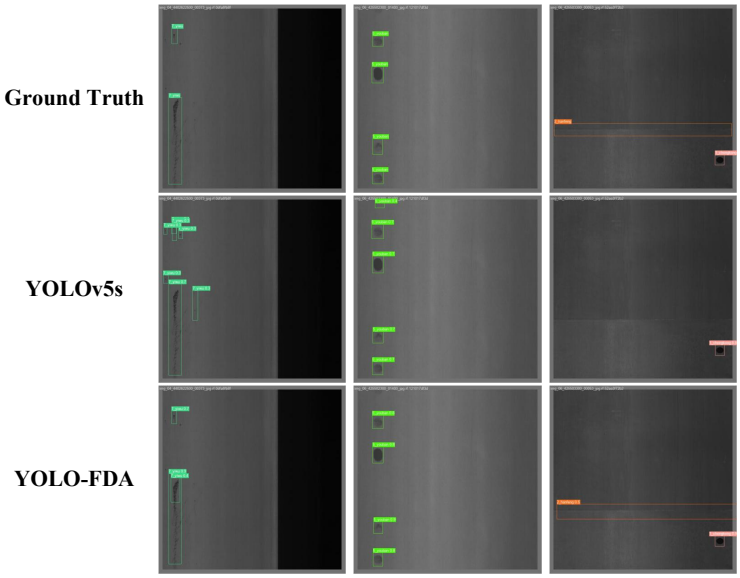

图 6: 可视化结果。将我们的方法与基线进行比较。青色框是一个称为夹杂的缺陷，绿色框是一个称为油斑的缺陷，橙色框是一个称为油斑的缺陷，粉色框是一个称为压印的缺陷。

5 结论

在本工作中，我们介绍了 YOLO-FDA，这是一个用于表面缺陷检测的细粒度和注意力增强型 YOLO 框架。通过整合双向特征融合、方向细节增强以及新型基于注意力的模块，YOLO-FDA 有效提升了多尺度鲁棒性和缺陷敏感性。实验结果表明，在各种类型的缺陷及不同尺度下，其性能优于现有方法，突显了该模型在工业检测场景中的有效性与通用性。

Bibliography

[1] F. Shen, C. Wang, J. Gao, Q. Guo, J. Dang, J. Tang, and T.-S. Chua, "Long-term talkingface generation via motion-prior conditional diffusion model," arXiv preprint arXiv:2502.09533, 2025.

[2] F. Shen, X. Du, Y. Gao, J. Yu, Y. Cao, X. Lei, and J. Tang, "Imagharmony: Controllable image editing with consistent object quantity and layout," arXiv preprint arXiv:2506.01949, 2025.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[4] R. Girshick, "Fast r-cnn," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448.

[5] R. Shaoqing, H. Kaiming, G. Ross, and S. Jian, "Faster r-cnn: Towards real-time object detection with region proposal networks," Proc. Adv. Neural Inf. Process. Syst., pp. 1–26, 2015.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[7] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.

[8] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2965–2974.

[9] T. Vu, H. Jang, T. X. Pham, and C. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," Advances in neural information processing systems, vol. 32, 2019.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 21–37.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[13] M. Hussain, "Yolov1 to v8: Unveiling each variant – a comprehensive review of yolo," IEEE Access, vol. 12, pp. 42 816–42 833, 2024.

[14] Y. Liu, Y. Liu, X. Guo, X. Ling, and Q. Geng, "Metal surface defect detection using SLF-YOLO enhanced YOLOv8 model," vol. 15, no. 1, p. 11105, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-94936-9

[15] Y. Du, H. Chen, Y. Fu, J. Zhu, and H. Zeng, "Aff-net: A strip steel surface defect detection network via adaptive focusing features," IEEE Transactions on Instrumentation and Measurement, 2024.

[16] M. Yasir, L. Shanwei, X. Mingming, S. Hui, M. S. Hossain, A. T. I. Colak, D. Wang, W. Jianhua, and K. B. Dang, "Multi-scale ship target detection using sar images based on improved yolov5," Frontiers in Marine Science, vol. 9, p. 1086140, 2023.

[17] J. Lu, M. Zhu, K. Qin, and X. Ma, "Yolo-lfpd: A lightweight method for strip surface defect detection," Biomimetics, vol. 9, no. 10, p. 607, 2024.

[18] F. Shen, J. Yu, C. Wang, X. Jiang, X. Du, and J. Tang, "Imaggarment-1: Fine-grained garment generation for controllable fashion design," arXiv preprint arXiv:2504.13176, 2025.

[19] F. Shen, X. Jiang, X. He, H. Ye, C. Wang, X. Du, Z. Li, and J. Tang, "Imagdressing-v1: Customizable virtual dressing," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 7, 2025, pp. 6795–6804.

[20] F. Shen and J. Tang, "Imagpose: A unified conditional framework for pose-guided person generation," Advances in neural information processing systems, vol. 37, pp. 6246–6266, 2024.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[22] YOLOv5., in : https://github.com/ultralytics/yolov5.

[23] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16 901–16 911.

[24] L. Zhou, S. Yang, C. Wang, P. Huang, S. Wang, Y. Wang, and Q. Wang, "QCF-YOLO: A lightweight model of surface defect detection for quick-connect fittings," vol. 25, no. 1, pp. 1716–1731, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10752899/

[25] W. Weng, W. Lin, F. Lin, J. Ren, and F. Shen, "A novel cross frequency-domain interaction learning for aerial oriented object detection," in Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, 2023, pp. 292–305.

[26] C. Qiao, F. Shen, X. Wang, R. Wang, F. Cao, S. Zhao, and C. Li, "A novel multi-frequency coordinated module for sar ship detection," in 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2022, pp. 804–811.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[30] M. Yuan, Y. Zhou, X. Ren, H. Zhi, J. Zhang, and H. Chen, "Yolo-hmc: An improved method for pcb surface defect detection," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1–11, 2024.

[31] W. Weng, M. Wei, J. Ren, and F. Shen, "Enhancing aerial object detection with selective frequency interaction network," IEEE Transactions on Artificial Intelligence, vol. 1, no. 01, pp. 1–12, 2024.

[32] H. Li, R. Zhang, Y. Pan, J. Ren, and F. Shen, "Lr-fpn: Enhancing remote sensing object detection with location refined feature pyramid network," arXiv preprint arXiv:2404.01614, 2024.

[33] G.-D. dataset., in : https://github.com/lvxiaoming2019/GC10-DET-metallic-surface-defect-datasets.

[34] M. Jager, C. Knoll, and F. A. Hamprecht, "Weakly supervised learning of a classifier for unusual event detection," IEEE Transactions on Image Processing, vol. 17, no. 9, pp. 1700–1708, 2008.

[35] DAGM2007, in : https://hci.iwr.uni-heidelberg.de/content/weakly-supervised-learning-industrial-optical-inspection, 2023.

[36] Z. Huang, C. Zhang, L. Ge, Z. Chen, K. Lu, and C. Wu, "Joining spatial deformable convolution and a dense feature pyramid for surface defect detection," IEEE Transactions on Instrumentation and Measurement, 2024.

[37] X. Yu, W. Lyu, D. Zhou, C. Wang, and W. Xu, "Es-net: Efficient scale-aware network for tiny defect detection," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–14, 2022.

[38] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 390–391.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[40] Z. Wang, L. Zhao, H. Li, X. Xue, and H. Liu, "Research on a metal surface defect detection algorithm based on dsl-yolo," Sensors, vol. 24, no. 19, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/19/6268

[41] H. Li, M. Liu, Y. Yin, and W. Sun, "Steel surface defect detection based on multi-layer fusion networks," vol. 15, no. 1, p. 10371, 2025. [Online]. Available: https://www.nature.com/articles/s41598-024-74601-3

[42] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.

[43] Q. Song, S. Li, Q. Bai, J. Yang, X. Zhang, Z. Li, and Z. Duan, "Object detection method for grasping robot based on improved yolov5," Micromachines, vol. 12, no. 11, p. 1273, 2021.

[44] Z. Lv, Z. Zhao, K. Xia, G. Gu, K. Liu, and X. Chen, "Steel surface defect detection based on mobilevitv2 and yolov8," The Journal of Supercomputing, vol. 80, no. 13, pp. 18 919–18 941, 2024.

[45] Z. Xu, S. Lan, Z. Yang, J. Cao, Z. Wu, and Y. Cheng, "Msb r-cnn: A multi-stage balanced defect detection network," Electronics, vol. 10, no. 16, p. 1924, 2021.

[46] S. Tang, S. Zhang, and Y. Fang, "Hic-yolov5: Improved yolov5 for small object detection," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 6614–6619.

[47] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Chained cascade network for object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1938–1946.

[48] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in European conference on computer vision. Springer, 2024, pp. 1–21.