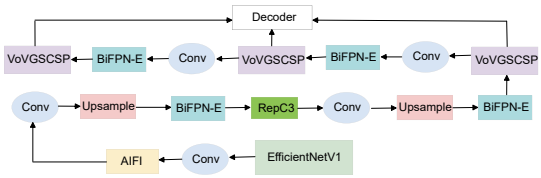
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 1 月, 第 53 卷, 第 1 期

轻量化 RT-DETR 与多尺度特征融合的火焰烟雾检测算法



孙光灵[1,2*]，王雪婷[1]

[1]安徽建筑大学电子与信息工程学院，安徽 合肥 230601；
[2]安徽建筑大学安徽省古建筑智能感知与高维建模国际联合研究中心，安徽 合肥 230601

摘要：针对火焰烟雾检测复杂性高，精度偏低，并且经常会产生漏检误检等问题，提出了轻量化 RT-DETR 与多尺度特征融合的火焰烟雾检测算法。该方法首先以 RT-DETR-r18 为基础架构，在主干网络中引入了一种轻量化网络结构 EfficientNetV1，降低了模型参数量；接着引入了 VoVGSCSP 模块，有效减少了特征冗余，并且增强了模型特征提取能力；最后结合双向特征融合金字塔网络 BiFPN 和 ELU 激活函数的优点，在 BiFPN 权重归一化过程中引入 ELU 激活函数，设计了 BiFPN-E 结构，提高了模型的跨尺度交互能力，有效提升了模型的检测精度。根据实验结果显示，改进的 RT-DETR 模型相较于原模型在公开数据集 NEWFire 和自建数据集 E-fire 上 mAP50 分别提高了 1.1% 和 1.8%，mAP50-95 分别提高了 1.7% 和 2.2%，参数量下降 32%，浮点计算量下降 65%。说明改进模型在提高检测精度的同时保持了轻量化要求，满足了在火焰烟雾检测时的精度和效率的实际需求。

关键词：EfficientNetV1；VoVGSCSP；激活函数；BiFPN

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250152 | CSTR: 32245.14.oee.2026.250152 | 中图分类号：TP391.4 | 文献标识码：A

孙光灵，王雪婷. 轻量化 RT-DETR 与多尺度特征融合的火焰烟雾检测算法 [J]. 光电工程，2026，**53**(1): 250152

Sun G L, Wang X T. Lightweight RT-DETR and multi-scale featurefusion algorithm for flame and smoke detection[J]. *Opto-Electron Eng*, 2026, **53**(1): 250152

1 引　言

随着现代化的推进和城市化的发展，各类公共场所、商业建筑以及居民住宅的火灾安全问题日益凸显。火灾的发生，不仅会直接危及生命和财产安全，其产生的烟雾还会对生态环境造成不可逆转的污染和破坏。传统的火灾探测方法主要依赖传感器[1]等物理设备，这些设备存在一定局限性，比如灵敏度不足、误报和漏报频繁、易受环境干扰等，难以实现早期预警，因此迫切需要更加高效的火灾检测算法。近年来，随着人工智能技术的快速发展，深度学习在计算机视觉领域取得了突破性进展，特别是在目标检测方面展现出卓越的性能优势，这为火灾检测研究提供了新的方法和研究思路。

基于深度学习目标检测领域，检测算法主要分为两种，一种是双阶段目标检测算法，另外一种是单阶段目标检测算法。双阶段目标检测算法需同时识别和定位两种不同类型的目标，并且还需要考虑两者之间的依赖关系，检测任务通常涉及多类别协同处理；而单阶段目标

收稿日期：2025-05-26；修回日期：2025-09-04；录用日期：2025-09-08；网络出版日期：2026-01-25

基金项目：2023 年安徽省住房城乡建设科学技术计划项目 (2023-YF058, 2023-YF113)；安徽省高等学校科学研究重点项目 (2023AH050164)；新时代育人质量工程项目 (2023zyxwjxalk112)；教育部协同育人项目 (22060651702475, 22060639218592)

*通信作者：孙光灵，sunguangling@163.com

版权所有©2026 中国科学院光电技术研究所

检测算法则专注于单一类别目标的快速识别与定位，通过直接在特征图上回归目标框坐标实现高效检测。双阶段目标检测算法主要有 CNN (convolutional neural networks)[2]、R-CNN (regions with convolutional neural networks) 和 Faster R-CNN (faster region-based convolutional neural networks)[3]。Majid 等[4] 提出了一种基于注意力机制的 CNN 模型，通过嵌入通道和空间注意力模块，动态增强火灾特征并抑制背景干扰，同时结合多尺度特征融合和双分支结构，有效提高了火灾检测的精度；Barmpoutis 等[5] 提出了一种结合 Faster R-CNN 目标检测框架和多维纹理特征分析的火灾检测方法，旨在提高复杂场景下的火灾检测精度。虽然此类方法对于火焰和烟雾的检测精度有了一定的提升，但是复杂的网络结构导致检测速度较慢，难以满足对实时性要求较高的火灾检测场景。单阶段目标检测算法比较典型的就是 YOLO (you only look once)[6] 系列算法。宋华伟等[7] 通过将 YOLOv5 中的路径聚合网络 (PAN) 替换为双向交叉尺度融合模块，并且在其检测头加入了协调注意力机制 (coordinate attention)，有效提升了小目标的捕捉能力；马耀名等[8] 提出一种面向复杂背景下火焰烟雾检测的 YOLOv8s 改进模型。设计了轻量级特征提取模块和多尺度上下文融合模块，有效提升了模型识别能力；最后使用 Inner-SIoU 损失函数，有效改善了边界框不匹配的问题。张小雪等[9] 通过在 YOLOv5s 网络中嵌入轻量化的级联稀疏查询模块，同时在主干部分融合高效通道注意力模块 (ECA, efficient channel attention)，并结合 Inner-SIoU 损失函数，以提升检测精度与鲁棒性。YOLO 算法虽然有着计算效率高、参数较低、易于部署等优点，但其轻量化导致特征提取能力较弱，并且 YOLO 主要基于卷积结构，导致全局建模能力有限。而 DETR (detection Transformer)[10] 是基于 Transformer[11] 架构的网络，其自注意力机制能够实现全局上下文建模，更擅长捕捉图像中远距离目标之间的关系，适用于复杂场景下的目标检测；同时 DETR 摒弃了传统的手工设计锚框与非极大值抑制 (non-maximum suppression, NMS)[12] 流程，后者在处理密集目标时易导致误抑制，而 DETR 通过端到端的匹配机制有效规避了该问题。虽然 DETR 在目标检测领域中有一定优势，但也因其 Transformer 模块参数多，推理速度较慢，难以在资源受限设备上进行部署。并且其缺失特征金字塔，对局部信息提取能力有限，导致对小目标检测效果不如预期。为解决上述问题，相关研究者们在此基础上相继推出了一些改进版本。Conditional DETR[13] 通过引入空间条件查询机制，将参考点坐标作为位置条件动态生成解码器查询，使解码器的交叉注意力能够动态聚焦于图像特征的局部相关区域，从而显著提升检测精度与训练收敛速度，而 RT-DETR (real-time detection Transformer)[14] 则引入了更高效的特征编码方式和 Query Selection 策略，相比其他改进的 DETR 模型而言，在保持高精度检测性能的同时更加轻量化，使其在实际应用中更具优势。

RT-DETR 作为一种端到端的实时目标检测框架，摒弃了传统检测模型中复杂的候选框生成与后处理流程，极大地优化了检测速度与系统响应能力，非常适用于火灾这类需要迅速反应的场景。但是由于局部信息捕获能力不足，细节提取能力有限，对小目标检测精度较低，容易出现漏检和误检问题，限制了其在高精度需求场景下的应用。因此本文提出改进的 RT-DETR 火焰烟雾检测模型。主要内容如下：

1) 在 RT-DETR 的主干网络中存在多个卷积和残差块的反复叠加，存在大量计算冗余，导致模型参数量偏大，因此本文重新设计了主干网络，引入了更轻量化的模型 EfficientNetV1[15]。

2) 由于原模型的 RepC3 结构仅通过卷积堆叠提取特征，缺乏显式的多尺度交互机制，在特征提取能力上存在一定的局限性，因此本文用 VoVGSCSP[16] 结构代替了原模型的 RepC3 结构，进一步优化该检测模型。

3) 引入双向特征金字塔网络 (BiFPN)[17]，并且在权重归一化时引入 ELU[18] 激活函数，增强跨尺度交互能力的同时提高梯度的稳定性，最终提升火焰烟雾的检测精度。

2 模型介绍

2.1 RT-DETR 模型

RT-DETR 主要针对 DETR 收敛速度慢、计算资源需求高等问题进行了优化，并且推出了多个不同版本的模型架构如 ResNet18[19]、ResNet50 等，以适应多样化的应用场景，大幅度提高了模型的处理能力和资源利用效率，本实验基于研究背景的要求和条件，选择 RT-DETR-r18 作为基准模型。

RT-DETR 模型如图 1 所示。RT-DETR 模型创新性地融合了 CNN 和 Transformer 架构的双重优势，主要由

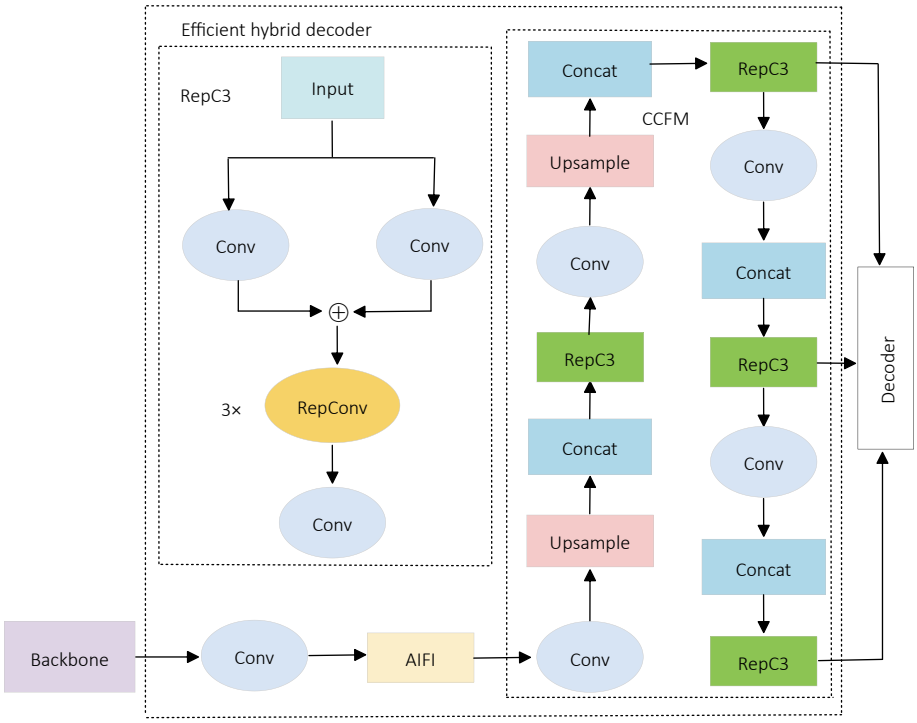


图 1 RT-DETR 模型

Fig. 1 RT-DETR model

主干网络、编码器与解码器组成。主干网络主要用于特征提取。RT-DETR 的编码器是一种高效的混合编码器，它通过解耦尺度内特征交互 (AIFI) 强化同层级特征的语义关联和跨尺度特征融合 (CCFM) 来实现不同分辨率特征的有效整合。这种设计能够在保持较高检测精度的同时，降低模型的计算复杂度。解码器则采用迭代优化机制，通过动态融合编码器输出的高级语义特征与空间位置信息，逐步优化检测结果。

2.2 改进的 RT-DETR 模型

改进的主干模型采用了轻量级的设计 EfficientNetV1 来替换传统骨干网络，并且在深层次用 VoVGSCSP 模块替换了 RepC3 结构，进一步增强模型的特征融合能力，同时降低了检测模型的复杂度，最后使用融入了 ELU 激活函数的双向特征融合金字塔结构 BiFPN，增强模型在复杂场景下的适应性，提高了检测的精度。改进后的 RT-DETR 如图 2 所示。

2.2.1 改进的主干网络

随着网络深度增加，模型会积累大量冗余信息，不仅显著增加了计算复杂度和参数量，还削弱了模型在复杂火灾场景下的泛化能力，因此本实验引入轻量化 EfficientNetV1 模型替代原 ResNet-18 主干网络，通过其

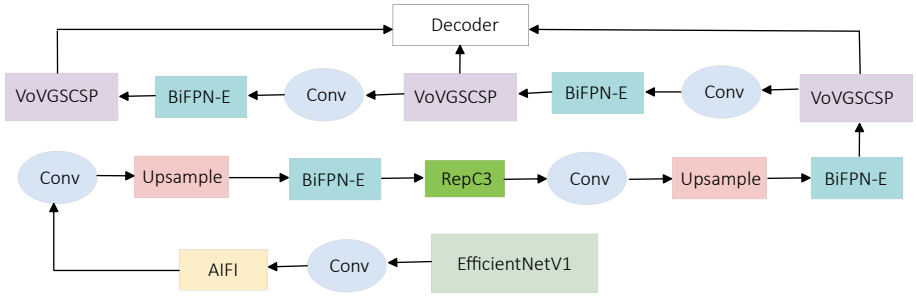


图 2 改进的 RT-DETR 模型

Fig. 2 Improved RT-DETR model

高效的复合缩放机制和深度可分离卷积，在保持特征提取能力的同时，有效降低模型参数量和浮点计算量，提升检测效率。

EfficientNet 是由 Google Research 团队提出的革命性神经网络架构，鉴于传统的神经网络架构通常采用单独调整深度、宽度或分辨率的方法进行缩放，单独增加一个维度可能会在提升网络性能的同时增加网络复杂性和计算量，而 EfficientNet 创新性地采用“复合缩放”(compound scaling) 方法，通过协同调节网络的这三个关键维度来实现计算效率与模型性能的平衡，同时还采用了 NAS (neural architecture search) 技术来自动调整这三者的比例，来建立最优的复合缩放系数。EfficientNetV1 神经网络结构如表 1 所示。

MBConv (Mobile inverted bottleneck convolution) 模块是 EfficientNetV1 的核心构建单元，该模块是来自 MobileNetV2[20] 中的倒残差模块。首先，标准卷积用于增加输入特征图通道数，然后是深度可分离卷积用于空间特征处理，同时也能够减少模型复杂度，最后是逐点卷积用于融合扩增后的特征。在此过程中，MBConv 采用了 Swish 激活函数，它结合了 ReLU[21] 的线性特性与 Sigmoid 的平滑性，具备两者的共同优点，从而提升了非线性表达能力。此外，为了增强通道维度上的特征表达能力，还在逐点卷积前引入了 SE 注意力机制。如图 3 所示，$K\times K$ 表示卷积核尺寸，s1/s2 表示步距。

2.2.2 重构 RepC3 结构

RepC3 结构通过重复堆叠卷积模块提升特征表达，但增加了计算复杂度，且缺乏有效的信息交互与多尺度融合，限制了检测性能。为此，本文引入 VoVGSCSP 模块，它采用高效卷积操作并强化多尺度特征融合，在更广泛的尺度范围内提升火灾检测的精度与鲁棒性。

GSConv 是构成 VoVGSCSP 模块的核心组件。它通过将卷积操作拆解成更小的部分 (组卷积)，从而减少了计算复杂度，同时保持了良好的性能，是一种轻量级卷积操作。首先，输入特征图会经过一次标准卷积进行下采样处理，接着应用深度卷积 (DWConv) [22]，在此过程中，每个输入通道都使用独立的卷积核进行处理；接下来，标准卷积与深度卷积的结果会被拼接在一起；最终通过 shuffle[23] 操作进行均匀混合处理。这样标准卷积的信息就能完全混入深度可分离卷积的输出，在不同的通道上均匀地交换局部的特征信息，从而让模型更好地学习到更多的多尺度信息。GSConv 模块结构如图 4 所示。

VoVGSCSP 是一种高效的轻量级模块，其核心是由 GSConv 和标准卷积组成的 GS bottleneck 基本单元，

表 1 EfficientNetV1 神经网络结构

Table 1 EfficientNetV1 neural network structure

Stage	Operate	Resolution	Channels	Layers
1	Conv 3×3	224×224	32	1
2	MBConv 1, *K* 3×3	112×112	16	1
3	MBConv 6, *K* 3×3	112×112	24	2
4	MBConv 6, *K* 3×3	56×56	40	2
5	MBConv 6, *K* 3×3	28×28	80	3
6	MBConv 6, *K* 3×3	14×14	112	3
7	MBConv 6, *K* 3×3	14×14	192	4
8	MBConv 6, *K* 3×3	7×7	320	1
9	Conv 1×1 & Pooling & FC	7×7	1280	1

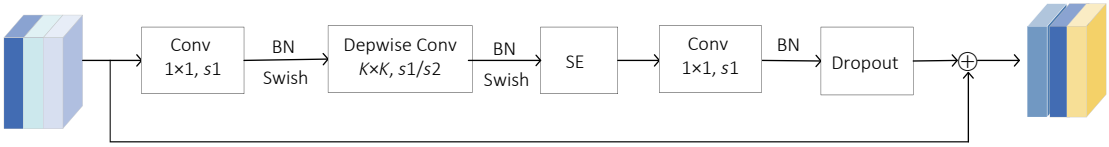


图 3 MBConv 模块结构

Fig. 3 Structure of MBConv module

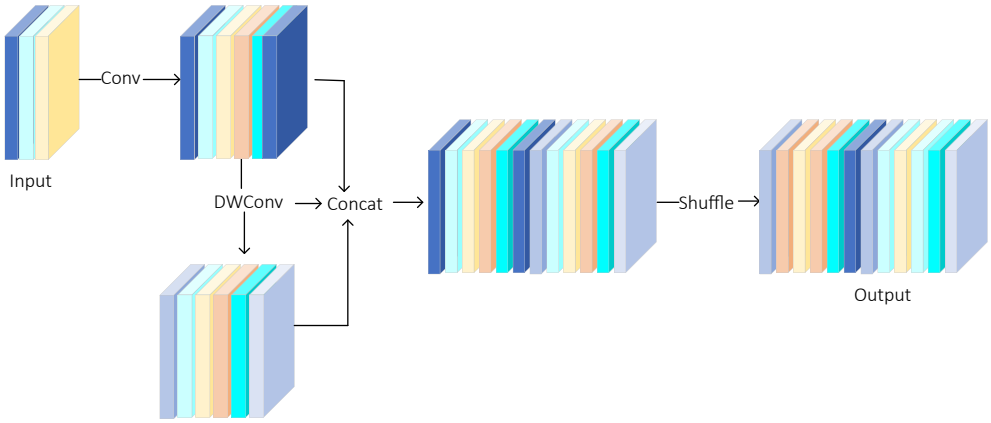


图 4 GSConv 模块结构

Fig. 4 Structure of GSConv module

采用一次性聚合策略实现跨阶段特征融合。该设计通过 GS bottleneck 模块的堆叠增强特征提取能力，在降低计算复杂度的同时保持了较高的检测精度，实现了轻量化和性能的良好平衡。GS bottleneck 模块结构和 VoVGSCSP 模块结构分别如图 5 和图 6 所示。

2.2.3 BiFPN-E 结构

在 RT-DETR 模型中，由于缺乏专门设计的多尺度特征融合模块，模型在处理小目标、细粒度信息时表现不足，同时原有的特征流动方式较为单一，浅层与深层特征之间的交互不充分，限制了特征表达能力的提升。因此，为了进一步增强局部细节感知和跨尺度信息交互能力，同时保持实时推理的优势，本文引入了更加高效灵活的双向特征融合金字塔网络 (BiFPN)。

PANet 模块结构如图 7 所示。BiFPN 模块结构如图 8 所示。BiFPN 是一种高效的多尺度特征融合网络，它在传统特征金字塔网络 (PANet)[24] 的基础上进行了优化。双向特征融合在 BiFPN 中指的是一种机制，它允许特征网络层中的信息在自顶向下和自底向上两个方向上流动和融合，使不同层级间的特征融合更高效，而且去除了冗余连接和计算节点，使得结构更加精简，计算量更少，火灾检测精度更高。BiFPN 通过对不同尺度的特征加权融合，确保在融合过程中，不同尺度的特征能根据其贡献程度被合理利用，这种方式能够减少信息丢失和冗余情况，从而增强了特征表达的有效性，特别是在火焰烟雾检测任务中表现最佳。

虽然 BiFPN 在特征融合时有不少优点，但在进行权重归一化时只是简单的加权平均归一化，这种方法通常难以捕捉到特征间相对复杂的非线性关系，导致融合的特征图不能充分利用某些尺度上的信息，特别是在比较复杂的背景下，会影响检测精度和鲁棒性。因此本实验在 BiFPN 进行权重归一化时结合了 ELU 激活函数，设计了 BiFPN-E 模块，对于不同尺度的特征，它能根据特征的响应强度进行灵活的调整，从而增强权重归一化的表达能力。其表达式为

$$\widehat{\omega}_i = \frac{ELU(\omega_i)}{\sum_{j=1}^{n} ELU(\omega_j)}, \tag{1}$$

式中：$\widehat{\omega}_i$表示归一化后的第 i 个特征图的权重；ω_i表示第 i 个融合分支的可学习权重，用于衡量输入特征的重要性；j 表示求和过程中的索引，用于遍历所有 n 个特征图；ELU(·) 表示数线性单元的激活函数，用于增强梯度传播的非线性处理。ELU 激活函数公式为

$$ELU(x) = \begin{cases} x, & x > 0 \\ \alpha(\exp(x) - 1), & x \leqslant 0 \end{cases}, \tag{2}$$

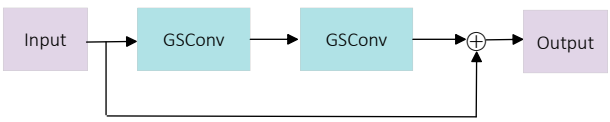


图 5 GS bottleneck 模块结构

Fig. 5 Structure of GS bottleneck module

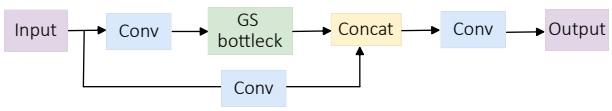


图 6 VoVGSCSP 模块结构图

Fig. 6 Structure of VoVGSCSP module

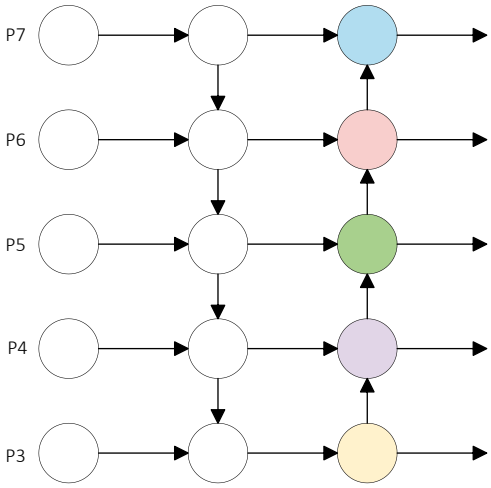


图 7 PANet 模块结构图

Fig. 7 Structure of PANet module

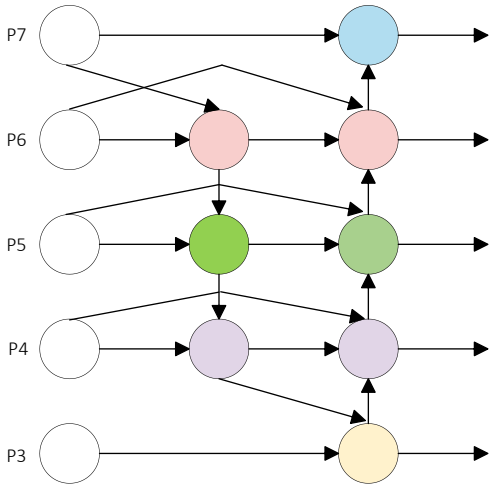


图 8 BiFPN 模块结构图

Fig. 8 Structure of BiFPN module

式中：α表示设置为 1 的超参数，用来控制负区域的斜率，可以发现 ELU 对所有负值输入均能产生非零梯度，有效避免了神经元死亡问题。作为非饱和函数，它也不会遇到梯度爆炸或者消失的问题。除此之外，该函数还保持了很好的非线性特性，不仅能够拟合更复杂的数据，还能增强模型的表达能力。

3 实验与结果分析

3.1 实验环境介绍

训练过程中使用批处理大小为 8，图片尺寸为 640 pixel×640 pixel，学习率设置为 0.0001，优化器选择的是 AdamW，迭代次数为 200 轮。实验在 pytorch 框架下进行训练，实验环境配置如表 2 所示。

表 2 实验环境配置

Table 2 Experimental environment configurations

Configuration	Environmental parameter
CPU	Intel(R) Core(TM) i7-12700F 2.10 GHz
GPU	NVIDIA GeForce RTX 3090 24 G
Memory	32 G
Python	3.9.19
CUDA	11.8
Pytorch	2.0.0

3.2 数据集

本实验采用的是来自 Roboflow 网站的公开数据集，即 NEWFire 数据集。该数据集包含各种不同场景下的环境，该数据集被划分成 5923 张训练集、1681 张验证集和 839 张测试集 (用于评估模型性能)。同时，为了进一步验证模型的有效性，从网上搜集了相关火焰烟雾图像，去掉一些重复和不相关的图像，然后对没有标注的图片利用脚本工具进行标注，最终保留了 6000 张火焰烟雾数据集，最后数据集 (E-fire) 被划分成 4000 张训练集、1000 张验证集和 1000 张测试集。

3.3 评价指标

图像目标检测常用的评价指标主要有参数量 (parameters)、计算量 (GFLOPs)、精准率 (precision, *P*)、召回率 (recall, *R*)、平均精度 (average precision, *AP*)、平均精度均值 (mean average precision, *mAP*)。精准率和召回率的计算公式分别为

$$P = \frac{TP}{TP+FP}, \tag{3}$$

$$R = \frac{TP}{TP+FN}, \tag{4}$$

式中：*TP* (true positive) 代表真实为真样本，预测为真样本；*FP* (false positive) 代表真实为负样本，预测为真样本；*FN* (false negative) 代表真实为真样本，预测为负样本。

$$AP = \int_0^1 P(R)\mathrm{d}R. \tag{5}$$

使用不同的置信度阈值可以获得一系列对应的 Precision 和 Recall 坐标点，这些点连接形成的连续曲线即为 *P-R* 曲线。而 *AP* 就表示为 *P-R* 曲线与坐标轴围成

的面积，它衡量的是模型在不同召回率下的平均精度表现，*AP* 值越大，模型精确度越高。

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i\,, \tag{6}$$

式中：N 表示目标类别数量；AP_i 表示第 i 个类别的 *AP* 值。mAP50 代表将阈值 *IoU* 设为 0.50 时，模型的平均精度；mAP50-95 代表阈值 *IoU* 从 0.50 到 0.95 时，以步长为 0.05 计算得出的平均精度。

3.4 实验结果分析

3.4.1 RT-DETR-r18 模型在主干部分使用不同轻量化模型结果分析

为了初步减少模型参数，本实验在主干部分分别使用 EfficientViT[25]、GhostNetV3[26]、RepViT[27] 和本实验采用的 EfficientNetV1，其中 EfficientViT 和 RepViT 都是最小模型类别。其实验结果如表 3 所示，可以发现，EfficientViT 虽然参数和浮点数都很低，但检测精度下降不少；GhostNetV3 只有浮点数有所下降，其参数较于原模型还有一定上涨，并且精度也有所下降；RepViT 虽然相比原模型参数和浮点数都有下降，但相比 EfficientNetV1，效果还是不够突出。EfficientNetV1 很好地兼顾了精度和轻量化这两个方面。因此本研究选择 EfficientNetV1 作为特征提取的主干网络。

3.4.2 BiFPN 使用不同激活函数分析

在 RT-DETR-r18 模型中对 BiFPN 权重归一化时引入不同激活函数进行实验，根据表 4 数据可以发现，本实验方法效果最好。

3.4.3 消融实验

根据表 5 的消融实验结果，可以确定改进点的有效性。在引入 EfficientNetV1 后发现模型在 NEWFire 和 E-fire 数据集上 mAP50 分别提升了 0.5% 和 1.1%，mAP50-95 分别提升 1.3% 和 1.0%，参数减少了 6.5 M，浮点计算量下降了 33.3 G；在引入 BiFPN-E 后发现模型在 NEWFire 和 E-fire 数据集上 mAP50 分别提升了 0.5% 和 0.7%，mAP50-95 分别提升 1.0% 和 1.3%，参数减少了 0.2 M，浮点计算量下降了 0.4 G；在加入 VoVGSCSP 后可以发现相较于基准模型在 NEWFire 和 E-fire 数据集上 mAP50 分别提升 0.2% 和 0.5%，mAP50-95 分别提升 0.3% 和 0.8%，参数无明显变化，浮点计算量下降了 2.7 G；在只引入 VoVGSCSP 和 BiFPN-E 后，模型虽然在浮点计算量和参数量上都有所下降，但是精度指标 mAP50 和 mAp50-95 也都下降了，这是由于 VoVGSCSP 主要提取局部多尺度特征，而 BiFPN 则通过双向跨层级连接进行全局多尺度特征融合，说明两个模块对同一特征进行了重复处理，存在信息冗余的冲突；而且过度融合导致特征细节变得模糊和丢失，进而导致检测精度的下降。而在引入 EfficientNetV1 和 BiFPN-E 后，模型在 NEWFire 和 E-fire 数据集上 mAP50 分别提升了 0.6% 和 1.3%，mAP50-95 分别提升 1.3% 和 1.7%，参数减少了 6.6 M，浮点计算量下降了 34.5 G；在引入 EfficientNetV1

表 3 主干使用不同轻量模型在 NEWFire 数据集上的实验结果

Table 3 Results of different lightweight backbones on NEWFire dataset

Models	mAP50	mAP50-95	Params/M	GFLOPs
Baseline	0.834	0.553	20.0	58.0
EfficientViT	0.820	0.540	10.9	28.2
GhostNetV3	0.830	0.550	25.8	41.5
RepViT	0.835	0.560	15.2	42.2
EfficientNetV1	0.839	0.566	13.5	24.7

表 4 BiFPN 结合不同激活函数在 NEWFire 数据集上的实验结果

Table 4 Experimental results of BiFPN combined with different activation functions on NEWFire dataset

Function	mAP50	mAP50-95	Params/M	GFLOPs
BiFPN	0.836	0.554	19.8	56.6
BiFPN+ReLU	0.835	0.560	19.8	56.6
BiFPN+ELU	0.839	0.563	19.8	56.6

和 VoVGSCSP 后，模型在 NEWFire 和 E-fire 数据集上 mAP50 分别提升了 0.4% 和 1.4%，mAP50-95 分别提升 0.4% 和 1.7%，参数减少了 6.0 M，浮点计算量下降了 36.4 G；而在引入这三个模块后的改进模型在 NeWFire 和 E-fire 数据集上 mAP50 分别提高了 1.1% 和 1.8%，mAP50-95 分别提升 1.7% 和 2.2%，参数下降了 6.3 M，浮点计算量下降了 37.7 G。结果说明改进模型在精度和轻量化方面均有不错表现。

混淆矩阵中每个数值代表真实类别为某行的样本，被预测为某列类别的具体数量。对角线上的数值表示正确分类的样本数，其高低直接反映对应类别的识别准确率；非对角线数值则揭示了类别间的混淆情况。改进前后的混淆矩阵对比如图 9 所示。通过对比改进前后的混淆矩阵可以发现，优化后的模型在性能上取得了显著提升。

3.4.4 对比实验

为了进一步验证模型的优越性，将本实验结果与 Faster R-CNN、RT-DETR-l、YOLOv5s、YOLOv8m、YOLOv8s、YOLOv9c 模型对比可以发现改进模型在精度和轻量化方面都有着一定的优越性。其结果如表 6 所示。

3.4.5 检测效果对比

改进前后检测结果对比如图 10 所示。为了直观验证改进模型的优越性，本文对比了改进前后模型在 NEWFire 测试集上的检测效果图，选取了在不同背景下的火焰烟雾检测情况。可以发现本实验的改进模型检测出的火焰和烟雾精度都比原始模型有所提高，表明了改进模型的优越性。

表 5　在 NEWFire 和 E-fire 数据集上的消融实验结果

Table 5　Ablation results on the NEWFire and E-fire datasets

EfficientNetV1	VoVGSCSP	BiFPN-E	NEWFire		E-fire		Params/M	GFLOPs
			mAP50	mAP50-95	mAP50	mAP50-95		
			0.834	0.553	0.773	0.509	20.0	58.0
√			0.839	0.566	0.784	0.519	13.5	24.7
		√	0.839	0.563	0.780	0.522	19.8	56.6
	√		0.836	0.556	0.778	0.517	20.0	55.3
	√	√	0.833	0.550	0.770	0.505	19.7	54.0
√		√	0.840	0.566	0.786	0.526	13.4	23.5
√	√		0.838	0.557	0.787	0.526	14.0	21.6
√	√	√	0.845	0.570	0.791	0.531	13.7	20.3

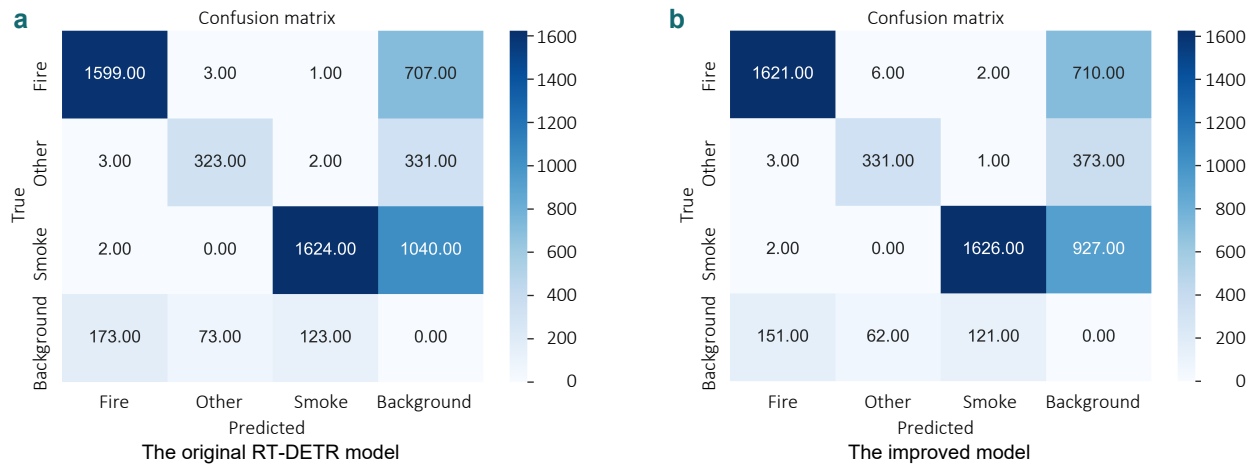


图 9　在 NEWFire 数据集上混淆矩阵对比结果

Fig. 9　Comparison of confusion matrices on the NEWFire dataset

表 6 在 NEWFire 和 E-fire 数据集上的对比实验结果

Table 6 Comparative results on NEWFire and E-fire datasets

Model	NEWFire		E-fire		Params/M	GFLOPs
	mAP50	mAP50-95	mAP50	mAP50-95		
Faster R-CNN	0.726	0.425	0.702	0.45	41.7	207.2
RT-DETR-l	0.840	0.542	0.785	0.512	32.7	107.4
YOLOv5s	0.830	0.545	0.770	0.504	9.1	23.8
YOLOv8m	0.834	0.559	0.776	0.514	25.8	78.7
YOLOv8s	0.832	0.554	0.778	0.518	11.1	28.4
YOLOv9c	0.830	0.552	0.776	0.512	25.3	102.3
ours	0.845	0.570	0.791	0.531	13.7	20.3



图 10 检测结果对比

Fig. 10 Comparison of detection results

4 结 论

随着计算机科技的发展，基于深度学习的火焰烟雾检测已越来越成熟，本文将使用比较广泛的检测模型 RT-DETR 与轻量化模型 EfficientNeTV1 相结合，然后引入了 VoVGSCSP 模块，增强模型特征提取能力，再在双向特征金字塔网络 BiFPN 的基础之上融入 ELU 激活函数，进一步提高模型对火灾目标在不同尺度和复杂场景下的检测能力。火焰烟雾的检测相较于其他目标检测来说，对精度和轻量化方面都有更高的要求，后续可以在此基础上再朝着这两个方面继续优化，使得模型更加全面，提高其实际应用价值。

作者贡献

孙光灵：指导论文选题及论文的修改；王雪婷：研究资料和背景分析，实验实施与分析，论文写作。

利益冲突

所有作者声明无利益冲突

参考文献

1. Xie Y C, Yang L, Yan J. Design of smoke detection alarm based on MQ-2 sensor[J]. *Comput Meas Control*, 2021, **29**(8): 255−259.
谢永超, 杨利, 严俊. 基于 MQ-2 型传感器的烟雾探测报警器的设计[J]. 计算机测量与控制, 2021, **29**(8): 255−259.
2. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 2014: 580–587.
https://doi.org/10.1109/CVPR.2014.81.
3. Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2017, **39**(6): 1137−1149.
4. Majid S, Alenezi F, Masood S, et al. Attention based CNN model for fire detection and localization in real-world images[J]. *Expert Syst Appl*, 2022, **189**: 116114.
5. Barmpoutis P, Dimitropoulos K, Kaza K, et al. Fire detection from images using faster R-CNN and multidimensional texture analysis[C]//*2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, 2019: 8301–8305.
https://doi.org/10.1109/ICASSP.2019.8682647.
6. Sun G L, Wang X B. THI-YOLO: improved non-motorized drivers helmet detection of YOLOv8[J]. *Opto-Electron Eng*, 2024, **51**(12): 240244.
孙光灵, 王薪博. THI-YOLO: 改进 YOLOv8 的非机动车驾驶员头盔检测[J]. 光电工程, 2024, **51**(12): 240244.
7. Song H W, Qu X J, Yang X, et al. Flame and smoke detection based on improved YOLOv5[J]. *Comput Eng*, 2023, **49**(6): 250−256.
宋华伟, 屈晓娟, 杨欣, 等. 基于改进 YOLOv5 的火焰烟雾检测[J]. 计算机工程, 2023, **49**(6): 250−256.
8. Ma Y M, Zhang P F, Tan F S. Improved YOLOv8s model for smoke and flame detection in complex backgrounds[J]. *Comput Eng Appl*, 2025, **61**(3): 121−130.
马耀名, 张鹏飞, 谭福生. 面向复杂背景下烟雾火焰检测的改进 YOLOv8s 算法[J]. 计算机工程与应用, 2025, **61**(3): 121−130.
9. Zhang X X, Wang Y, Wu S Y, et al. An improved lightweight fire detection algorithm based on cascade sparse query[J]. *Opto-Electron Eng*, 2023, **50**(10): 230216.
张小雪, 王雨, 吴思远, 等. 基于级联稀疏查询机制的轻量化火灾检测算法[J]. 光电工程, 2023, **50**(10): 230216.
10. Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//*16th European Conference on Computer Vision*, Glasgow, 2020: 213–229.
https://doi.org/10.1007/978-3-030-58452-8_13.
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 2017: 6000–6010.
12. Bodla N, Singh B, Chellappa R, et al. Soft-NMS—improving object detection with one line of code[C]//*Proceedings of the 2017 IEEE International Conference on Computer Vision*, Venice, 2017: 5562–5570.
https://doi.org/10.1109/ICCV.2017.593.
13. Meng D P, Chen X K, Fan Z J, et al. Conditional DETR for fast training convergence[C]//*Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, Montreal, 2021: 3631–3640. https://doi.org/10.1109/ICCV48922.2021.00363.
14. Zhao Y A, Lv W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection[C]//*Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 2024: 16965–16974.
https://doi.org/10.1109/CVPR52733.2024.01605.
15. Tan M X, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks[C]//*Proceedings of the 36th International Conference on Machine Learning*, Long Beach, 2019: 6105–6114.
16. Li H L, Li J, Wei H B, et al. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles[Z]. arXiv: 2206.02424, 2022.
https://arxiv.org/abs/2206.02424v1.
17. Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 2020: 10778–10787.
https://doi.org/10.1109/CVPR42600.2020.01079.
18. Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs)[C]// *4th International Conference on Learning Representations*, San Juan, 2016.
19. He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016: 770–778. https://doi.org/10.1109/CVPR.2016.90.
20. Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018: 4510–4520.
https://doi.org/10.1109/CVPR.2018.00474.
21. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//*Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, 2011: 315–323.
22. Chollet F. Xception: deep learning with depthwise separable convolutions[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017: 1800–1807. https://doi.org/10.1109/CVPR.2017.195.
23. Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]// *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018: 6848–6856.
https://doi.org/10.1109/CVPR.2018.00716.
24. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for

object detection[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017: 936–944. https://doi.org/10.1109/CVPR.2017.106.

25. Liu X Y, Peng H W, Zheng N X, et al. EfficientViT: Memory efficient vision transformer with cascaded group attention[C]//*Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 2023: 14420–14430. https://doi.org/10.1109/CVPR52729.2023.01386.
26. Han K, Wang Y H, Tian Q, et al. GhostNet: More features from cheap operations[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 2020: 1577–1586. https://doi.org/10.1109/CVPR42600.2020.00165.
27. Wang A, Chen H, Lin Z J, et al. RepViT: revisiting mobile CNN from ViT perspective[Z]. arxiv: 2307.09283, 2023. https://doi.org/10.48550/arXiv.2307.09283.

作者简介



【通信作者】孙光灵 (1976-)，男，硕士，教授，主要研究方向为机器学习、图像处理。

E-mail：sunguangling@163.com



王雪婷 (1999-)，女，硕士研究生，主要研究方向为图像处理、目标检测。

E-mail：18755443812@163.com

Lightweight RT-DETR and multi-scale featurefusion algorithm for flame and smoke detection

Sun Guangling[1,2*], Wang Xueting[1]

[1] School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei, Anhui 230601, China;
[2] Anhui Interational Joint Research Center for Intelligent Perception and High-dimensional Modeling of Ancient Buildings, Anhui Jianzhu University, Hefei, Anhui 230601, China

Abstract

Objective With the advancement of modernization and the development of urbanization, fire safety issues in various public venues, commercial buildings, and residential homes have become increasingly prominent. Fire incidents not only directly threaten lives and property but also generate smoke that often causes irreversible pollution and damage to the ecological environment. Traditional fire detection methods primarily rely on physical devices such as sensors, yet these approaches have limitations—including insufficient sensitivity, frequent false alarms and missed detections, susceptibility to environmental interference, and difficulty in achieving early warning. Consequently, there is an urgent need for more efficient fire detection algorithms. In recent years, rapid advancements in artificial intelligence have driven breakthroughs in deep learning within computer vision, particularly demonstrating exceptional performance in object detection. This has opened new avenues and research perspectives for fire detection studies.

Methods Addressing current limitations in flame and smoke detection technologies—such as insufficient accuracy in complex scenarios, susceptibility to false negatives and positives, and high computational costs—this research innovatively applies the cutting-edge RT-DETR model from computer vision to fire detection tasks. The aim is to thoroughly explore this model's capability to capture and recognize flame and smoke features in complex, dynamic environments, striving to develop a fire detection algorithm with high detection accuracy, low false alarm rates, and the ability to meet real-time early warning requirements. RT-DETR is an efficient real-time end-to-end detector based on the Transformer architecture. It organically combines the local feature extraction advantages of convolutional neural networks (CNNs) with the global modeling capabilities of Transformers through the introduction of a hybrid encoder. By employing an innovative IoU-aware query selection mechanism to precisely filter initial target queries, it achieves a good balance between detection accuracy and inference speed while significantly reducing computational latency. However, when handling complex, multi-scale, and morphologically variable targets like flames and smoke, this model still suffers from insufficient capture of local information and limited extraction of fine-grained features. This often leads to false negatives and false positives for small targets, limiting its application in scenarios demanding high

Foundation item: 2023 Anhui Province Housing and Urban-Rural Development Science and Technology Program Project (2023-YF058, 2023-YF113), Anhui Province Higher Education Institutions Scientific Research Key Project (2023AH050164), New Era Nurturing Quality Project (2023zyxwjxalk112), Ministry of Education Collaborative Nurturing Project (22060651702475, 22060639218592)

*Correspondence: Sun G L, E-mail: sunguangling@163.com

accuracy. To address this, this paper proposes a lightweight RT-DETR-based flame and smoke detection algorithm with multi-scale feature fusion to enhance the model's perception and discrimination capabilities for such targets. We first adopt RT-DETR-r18 as the baseline model, introducing the lightweight EfficientNetV1 architecture into the backbone network. Through evaluation and experimentation with mainstream lightweight networks like GhostNet, RepViT, and EfficientViT, results demonstrate that EfficientNetV1 achieves the optimal balance between accuracy and efficiency via its composite scaling strategy. It significantly reduces parameter count and computational burden while maintaining robust feature extraction capabilities, providing the optimal feature foundation for subsequent detection tasks. Second, the VoVGSCSP module is designed at the model's neck. This architecture integrates the efficient feature reuse of VoVNet with the gradient splitting concept of CSPNet, effectively enhancing the model's expression capabilities for key features like flame-smoke texture and color while suppressing feature redundancy. Finally, addressing the core challenge of extreme scale variations in flame-smoke targets, this paper introduces the more efficient multi-scale feature fusion mechanism bidirectional feature pyramid network (BiFPN)-E. During the weighted fusion process in the BiFPN, it innovatively employs the ELU activation function for weight normalization. The non-zero saturation property of the ELU function when handling negative inputs generates smoother and more discriminative fusion weights, thereby strengthening the interaction and complementarity between feature maps at different scales. This significantly improves the model's perception and discrimination accuracy in complex scenarios such as large-scale flames, faint smoke, and occluded objects.

Results and Discussions The final experimental results fully validate the comprehensive effectiveness of the proposed method. On the NEWFire dataset, the improved model achieved mAP50 and mAP50-95 scores of 84.5% and 57.0%, respectively, representing improvements of 1.1% and 1.7% over the baseline RT-DETR model. On the self-built E-fire dataset, the accuracy metrics mAP50 and mAP50-95 improved by 1.8% and 2.2%, respectively. More importantly, while achieving these accuracy gains, the model demonstrated significant lightweighting: model parameters were reduced by 32%, and floating-point operations were drastically reduced by 65%. In-depth ablation experiments further confirm that the BiFPN-E architecture is the most critical contributor to accuracy gains, while the EfficientNetV1 backbone serves as the core for achieving efficient model compression. Visual analysis also reveals that the improved model's feature heatmap responses are more concentrated on actual flame-smoke regions, demonstrating stronger suppression of fire-like distractors and complex backgrounds, higher recognition accuracy, and effectively reducing both false negatives and false positives.

Conclusions Through systematic enhancements to the RT-DETR model, this research achieves a synergistic breakthrough in accuracy, efficiency, and feature discriminative power for fire detection. The EfficientNetV1 backbone enables efficient model compression, while the introduced VoVGSCSP module enhances the model's ability to capture irregular fire shapes and dynamic changes. The BiFPN-E architecture significantly improves detection accuracy through optimized multi-scale feature fusion. The synergistic effect of these three components enables the model to maintain high accuracy while reducing parameters by 32% and computational load by 65%. This breakthrough overcomes computational resource constraints while enhancing performance, providing a practical solution for achieving high-precision real-time fire alerts on resource-constrained edge devices.

Keywords: EfficientNetV1; VoVGSCSP; activation function; BiFPN

Sun G L, Wang X T, . Lightweight RT-DETR and multi-scale featurefusion algorithm for flame and smoke detection[J]. *Opto-Electron Eng*, 2026, **53**(1): 250152; DOI: 10.12086/oee.2026.250152

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文