Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(1), 242-253
Published Online January 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.151024

Hans汉斯

MSFA：通过多尺度融合特征增强的检测框架

周凌峰，沈　航

西华大学汽车与交通学院，四川 成都

收稿日期：2025年11月26日；录用日期：2025年12月31日；发布日期：2026年1月13日

摘　要

复杂驾驶场景下的交通锥检测面临目标尺寸小、背景复杂、遮挡严重等挑战。传统检测方法计算冗余高、环境适应性差，难以满足自动驾驶系统的实时性和准确性要求。本研究提出 SBA-YOLOv11 轻量化检测算法，引入三个关键创新：1) 金字塔多尺度特征聚合(MSFA)模块，实现高效跨尺度特征融合；2) 空间双向注意力(CSBA)模块，采用自注意力和交叉注意力机制增强特征表达；3) 门控全维度卷积(GatedFDConv)模块，结合通道分离策略和门控机制优化特征交互。构建包含 8000 张图像的高质量交通锥数据集，涵盖多样化场景条件。实验结果显示，与基线 YOLOv11n 相比，SBA-YOLOv11 在精确度、召回率和 mAP@50 方面分别提升 3.7%、3.6%和 7.4%，达到 89.9%、88.8%和 94.9%。与最先进方法对比，整体平均精度提升 2.3%~27.5%。所提方法成功平衡了检测精度和计算效率，有效解决复杂驾驶场景下的交通锥检测挑战，满足自动驾驶应用需求。

关键词

交通锥检测，YOLOv11，多尺度特征融合，注意力机制，自动驾驶，实时目标检测

MSFA: A Detection Framework Enhanced by Multi-Scale Fusion Features

Lingfeng Zhou, Hang Shen

School of Automotive and Traffic Engineering, Xihua University, Chengdu Sichuan

Received: November 26, 2025; accepted: December 31, 2025; published: January 13, 2026

Abstract

Traffic cone detection in complex driving scenarios faces challenges such as small target size, complex background, and severe occlusion. Traditional detection methods have high computational redundancy and poor environmental adaptability, which make it difficult to meet the real-time and accuracy requirements of the auto drive system. This study proposes the SBA-YOLOv11 lightweight

文章引用: 周凌峰, 沈航. MSFA: 通过多尺度融合特征增强的检测框架[J]. 人工智能与机器人研究, 2026, 15(1): 242-253. DOI: 10.12677/airr.2026.151024

detection algorithm, which introduces three key innovations: 1) Pyramid Multi Scale Feature Aggregation (MSFA) module to achieve efficient cross scale feature fusion; 2) The Spatial Bidirectional Attention (CSBA) module utilizes self attention and cross attention mechanisms to enhance feature expression; 3) GatedFDConv module combines channel separation strategy and gating mechanism to optimize feature interaction. Build a high-quality traffic cone dataset containing 8000 images, covering diverse scene conditions. The experimental results show that compared with the baseline YOLOv11n, SBA-YOLOv11 has better accuracy, recall, and mAP@50. The aspects increased by 3.7%, 3.6%, and 7.4% respectively, reaching 89.9%, 88.8%, and 94.9%. Compared with the most advanced methods, the overall average accuracy has improved by 2.3%~27.5%. The proposed method successfully balances detection accuracy and computational efficiency, effectively solving the challenge of traffic cone detection in complex driving scenarios and meeting the requirements of autonomous driving applications.

Keywords

Traffic Cone Detection, YOLOv11, Multi-Scale Feature Fusion, Attention Mechanism, Autonomous Driving, Real Time Object Detection

1. 引言

在智能交通领域，交通锥作为重要的道路安全标志，其准确识别对于自动驾驶车辆的安全行驶至关重要。Chen 等人[1]提出了浅，中，深三阶段的协议聚焦模块，相对于传统固定核，协议在不显著增加计算量的前提下，使模型对多尺度目标的特征提取更具针对性，但在卷积核大小分配时，未实现动态自适应。Hu 等人[2]通过改进的网络架构保留原始的 YOLOv8 的 C2f 模块高效特征融合能力，同时采用了 GhostNet 的低成本特征生成机制，通过简单线性变换生成冗余特征，替代部分卷积操作减少模型计算量和参数量。Bai 等人[3]保留了 Faster R-CNN 骨干网络，通过深层卷积逐步提取目标的低层次纹理特征与高层次语义特征，避免了传统方法依赖单一手工特征的局限性，但固有的 RCNN 架构其在“微弱小目标特征提取”上存在短板，影响检测精度。Wang 等人[4]提出基于 RegNet 骨干网络的 Faster R-CNN 变体，通过减少采样过程中的信息损失，显著提高了航天器部件的检测效率。针对算法推理速度慢的问题，He 等人[5]通过优化迭代检测头和简化特征金字塔网络结构来加速 Sparse R-CNN，在降低计算量的情况下同时提升了小目标检测精度，但精简迭代次数减少虽能提速，但对高难度目标的优化不足，导致此类目标精度损失严重。Cai 等人[6]通过构建多尺度自适应注意力特征融合网络(MSAA-Module)，利用多尺度迭代注意力特征融合(MSIAFF)改进 C2F 模块，并引入自适应空间特征融合(ASFF)模块处理不同尺度特征图，解决了复杂交通场景下小目标检测难题，但在车载嵌入式设备部署上无法满足自动驾驶的低延迟需求。上述通用目标检测方法在自然场景中表现出良好的性能，但直接用于交通图像的目标检测任务时往往无法取得令人满意的效果。原因主要有两点：一方面，基于交通图像的地面目标检测面临着目标尺寸小、背景复杂多变、重叠、遮挡、尺度变化等更多问题；另一方面，由于需要在车辆设备中实现推理过程，高检测精度与模型轻量化往往难以兼顾。

2. 方法

在本节中，为了构建一个具有强大多尺度能力的实时目标检测器，我们提出了 MSFA 模块和 CSBA

模块。我们的 MSFA 模块包含一个增强的分层多分支结构，拓展了结构特征空间，同时结合全局查询学习来提供跨阶段选择的控制参数，以减少有害的空间信息并增强多尺度表示。引入门控机制模块是为了在实时目标检测器中高效且有效地融入大内核卷积，以实现更好的多尺度能力。

2.1. MSFA 模块

本文所提改进模型的整体结构如图 1 所示。在骨干网络设计中，MSFA (Multi-Scale Feature Aggregation)模块采用分层渐进式特征聚合策略，通过级联具有不同感受野的卷积操作，实现多尺度特征的高效聚合与互补利用。该模块以“尺度逐步扩展”为核心逻辑：从较小感受野的局部特征提取出发，逐步扩大至大感受野的全局信息捕获；同时在各分支节点处设计特征保留机制，确保每个尺度层级的关键特征均能参与最终融合，并通过残差优化连接，最大限度维持特征传递过程中的信息完整性。尽管上述基础模块已显著提升模型的多尺度表征能力，但尚未充分挖掘不同核尺寸卷积的潜在价值尤其是在卷积神经网络(CNN)视觉识别任务中已被证实有效的大核卷积，长期因实时性约束被主流实时目标检测器忽视。将大核卷积融入实时检测器的核心障碍在于其高计算开销，且该问题在网络低阶段(处理高分辨率特征时)尤为突出。大核卷积应用于高分辨率特征图时会产生极高的计算成本，而将其迁移至低分辨率特征层，可在保障大感受野优势的同时，大幅降低整体计算复杂度，为实时性与性能的平衡提供可行路径。MFSA 模块设计与多尺度增强逻辑借鉴了 CSP 结构化模块的设计思想[7]，以具备通道分割能力的 MSFA 模块为核心组件，通过级联多感受野卷积操作，进一步强化模型对多尺度特征的深度融合与高效利用能力。

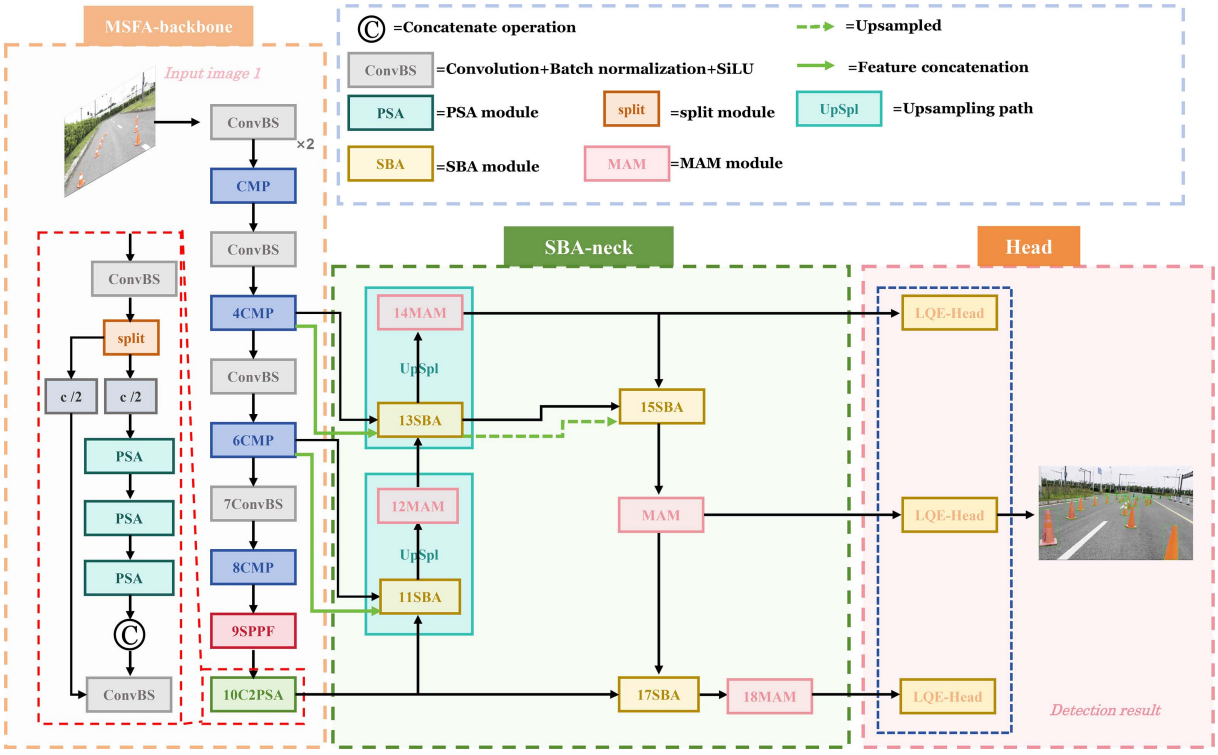


Figure 1. Overall MFSA structure
图 1. 整体 MFSA 结构

结构主要包含四个卷积层和特征融合部分。模块接受输入特征图 X：

$$X \in \mathbb{R}^{B\times C\times H\times W} \tag{1}$$

其中 B、C、H、W 分别表示批次大小、通道数、特征图高度和宽度。对于模块初始特征提取与第一次通道分割，首先对输入特征图进行 3 × 3 卷积操作，提取基础特征：

$$F_1 = \sigma\left(W_1 * X + b_1\right) \in \mathbb{R}^{B\times C\times H\times W} \tag{2}$$

其中，W_1 为 3 × 3 卷积核权重，b_1 为偏置项，σ 为激活函数，$*$ 表示卷积操作。该步骤的目的是对原始特征进行初步处理，为后续的多尺度分解做准备。在模块中第一次特征分割将 F_1 沿通道维度均匀分割为两个子特征图：

$$F_1^{(1)}, F_1^{(2)} = \text{Split}\left(F_1, \text{dim} = 1\right) \tag{3}$$

其中：$F_1^{(1)} \in \mathbb{R}^{B\times\frac{C}{2}\times H\times W}$，$F_1^{(2)} \in \mathbb{R}^{B\times\frac{C}{2}\times H\times W}$

这种分割策略使得 $F_1^{(1)}$ 用于进一步的多尺度处理，而 $F_1^{(2)}$ 作为较小尺度的特征直接参与最终融合。

其中中尺度特征提取与第二次通道分割的部分，对 $F_1^{(1)}$ 应用 5 × 5 分组卷积，提取中等尺度的特征信息：

$$F_2 = \sigma\left(W_2 * groupF_1^{(1)} + b_2\right) \in \mathbb{R}^{B\times\frac{C}{2}\times H\times W} \tag{4}$$

其中，分组数 $g = \dfrac{C}{2}$，即每个输出通道对应一个输入通道。分组卷积的使用显著降低了参数数量和计算复杂度，同时保持了特征提取的有效性。类似地，将 F_2 进行二次分割：

$$F_2^{(1)}, F_2^{(2)} = \text{Split}\left(F_2, \text{dim} = 1\right) \tag{5}$$

其中：

$$F_2^{(1)} \in \mathbb{R}^{B\times\frac{C}{4}\times H\times W}, F_2^{(2)} \in \mathbb{R}^{B\times\frac{C}{4}\times H\times W} \tag{6}$$

考虑到 7 × 7 卷积核能够提供捕获更大感受野，能够有效捕获全局上下文信息，对 $F_2^{(1)}$ 应用 7 × 7 卷积，提取大尺度特征：

$$F_3 = \sigma\left(W_3 * groupF_2^{(1)} + b_3\right) \in \mathbb{R}^{B\times\frac{C}{4}\times H\times W} \tag{7}$$

其中分组数 $g = \dfrac{C}{4}$。

在多尺度特征融合部分，将三个不同尺度的特征图进行通道级联：

$$F_{cat} = \text{Concat}\left(F_3, F_2^{(2)}, F_1^{(2)}, dim = 1\right) \tag{8}$$

其中对通道数进行验证：$\dfrac{C}{4} + \dfrac{C}{4} + \dfrac{C}{2} = C$，确保融合后的特征图通道数与输入一致，特征整合之间通过 1 × 1 卷积进行整合，并添加残差连接：

$$F_{out} = \sigma\left(W_4 * F_{cat} + b_4\right) + X \tag{9}$$

引入残差连接不仅可以有助于梯度的反向传播，还能够保留输入特征中的重要信息，防止网络退化。

综合上述各个阶段，MFSA 模块的完整前向传播过程可以表示为：

$$\text{MSFA}\left(X\right) = \sigma\left(W_4 * \text{Concat}\left(F_3, F_2^{(2)}, F_1^{(2)}\right) + b_4\right) + X \tag{10}$$

其中：$F_2^{(1)}, F_2^{(2)} = \text{Split}\left(\sigma\left(W_1 * X + b_1\right)\right)$，$F_2^{(1)}, F_2^{(2)} = \text{Split}\left(\sigma\left(W_2 * groupF_1^{(1)} + b_2\right)\right)$，

$F_3 = \sigma\left(W_3 * groupF_2^{(1)} + b_3\right)$。

为了进一步提升特征提取能力并优化计算效率，本文将 MSFA 模块集成到 CSP (Cross Stage Partial)结构中，形成 CSP-MSFA 模块，该模块继承自 YOLOv11 的 C2f 结构设计思路。具体表达式如下：

$$\text{CSP-MSFA}(X) = \text{Conv}_{1\times1}\left(\text{Concat}\left(X_{main}, \prod_{i=1}^{n}\text{MSFA}_i\left(X_{branch}\right)\right)\right) \tag{11}$$

其中，X_{main} 和 X_{branch} 分别表示通过初始卷积得到的主分支和处理分支特征，n 为 MSFA 模块的重复次数。

在卷积核的选择上，3 × 3 可以捕获细粒度的局部特征，5 × 5 捕获中等尺度的特征，7 × 7 捕获更大范围的上下文信息和全局特征，间隔增长可以避免尺寸过于接近，跳跃过大的问题，并保证不同尺度间有足够的差异性，奇数卷积核具备明确的中心点，便于实现对称填充、特征对齐，且在经典网络结构中已被广泛验证有效，兼顾模型稳定性与落地可行性，此外这个组合在多个经典网络结构中被验证有效，另外奇数尺寸的卷积核具有明确的中心点，便于对称填充，特征对齐以及网络设计的一致性。

2.2. 颈部结构设计

如图 2 所示，网络颈部的设计目标是融合低层次特征图的细节敏感信息与高层次特征图的丰富语义信息这对解决类内差异大、类间相似性高的检测难题至关重要。为此，本文提出 CSBA-FPN 架构，通过门控注意力策略与双向特征交互机制重构传统 FPN 元架构，其具体结构如图 2 所示。该架构中，两个核心模块(SBA 与 MAM)发挥关键作用：SBA 模块通过注意力门控机制动态调节特征融合权重，实现多尺度特征的高效聚合，增强对不同尺度目标的检测能力，保障信息传播的可靠性；MAM 模块则聚焦跨阶段多层次特征的深度整合，通过强化特征间的双向交互，有效避免关键信息丢失。下文将对这两个模块的设计细节与工作机制进行深入阐述。

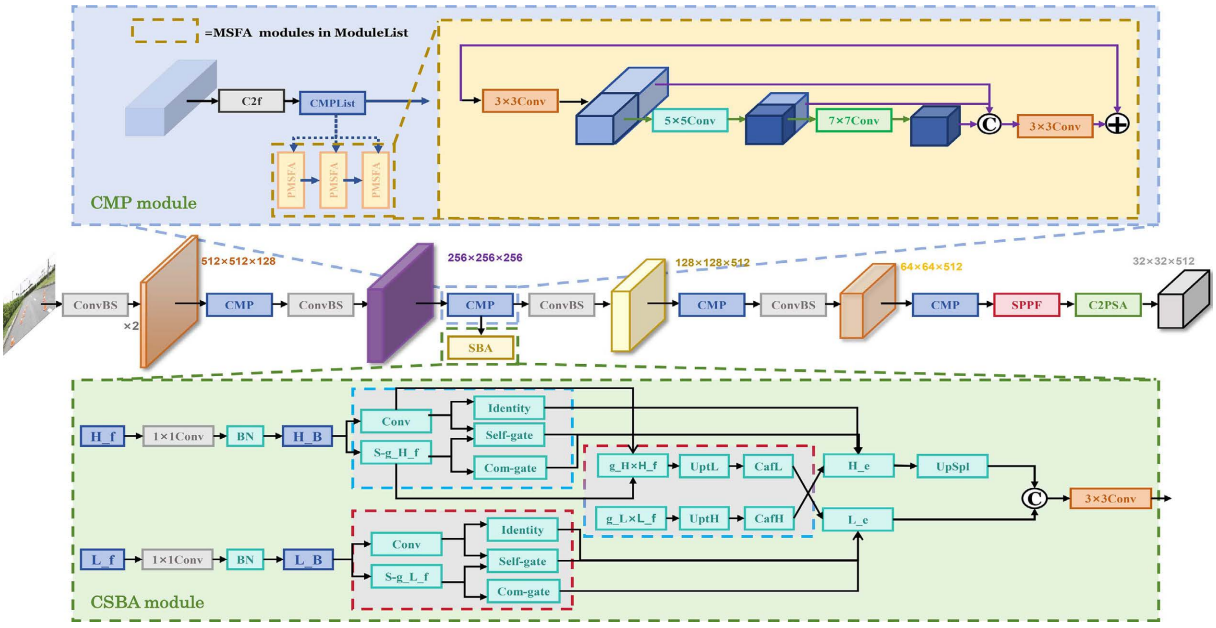


Figure 2. Module detail diagram
图 2. 模块细节图

2.2.1. CSBA 模块

该模块通过采用双向注意力机制与自适应对齐策略。针对多尺度特征的空间分辨率差异，聚焦特征

融合权重的动态优化，这是实现高低分辨率特征高效协同的核心。如图 2 所示，CSBA (Cross-Scale Bidirectional Attention Module)模块以双向门控注意力机制，通过自适应权重分配实现高低分辨率特征的协同优化，对融合权重进行门控机制动态调节，增强关键信息互补和多尺度信息特征聚合的能力。具体来说，模块输入为高分辨率特征图 *FH* 和低分辨率特征图 *FL*，并通过两个并行分支完成多尺度特征融合。为降低后续注意力计算的复杂度，首先对两类特征进行通道降维映射，左右分支(对应低、高分辨率特征)的降维结果可表示为：

$$L' = \mathrm{fc1}\left(F_L\right) \in \mathbb{R}^{d/2\times H_L\times W_L} \tag{12}$$

$$H' = \mathrm{fc2}\left(F_H\right) \in \mathbb{R}^{d/2\times H_H\times W_H} \tag{13}$$

其中 fc1 和 fc2 分别为低、高分辨率分支的降维映射函数，通过卷积权重 W_1 (fc1)与对应输入特征作用，将原始通道数 *CL* (低分辨率)、*CH* (高分辨率)统一降至 $d/2$，为注意力机制的高效计算奠定基础。

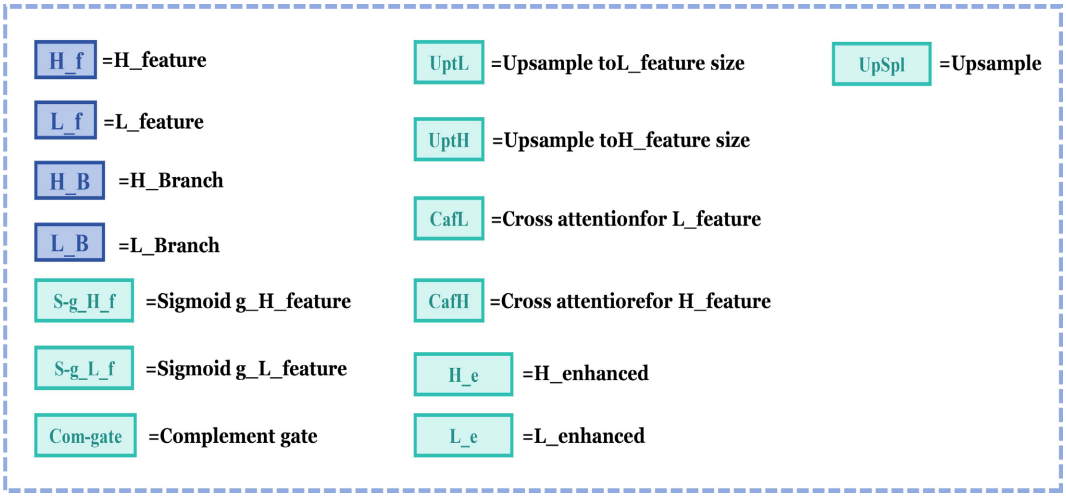


Figure 3. Module detail diagram
图 3. 模块细节图

在第二个分支中，核心模块在于双向门控注意力机制，包含两个关键组件：注意力门生成和双向特征交互。在门控信号生成分支中，左右分支的门控信号可以分别表示为：

$$g_H = \sigma\left(H'\right) \in \mathbb{R}^{d/2\times H_H\times W_H} \tag{14}$$

其中 σ 为 Sigmoid 激活函数。随后双向注意力机制对不同分辨率层之间的相互信息进行交换，交换后的特征图可以表示为：

$$L'' = d_{in}1\left(L'\right) \in \mathbb{R}^{d/2\times H_L\times W_L} \tag{15}$$

$$H'' = d_{in}2\left(H'\right) \tag{16}$$

其中 $d_{in}2$ 是一个卷积层，卷积核大小为 1×1。为保留原始特征信息并促进梯度反向传播，模块引入残差连接：通过自注意力增强特征相关性，同时利用跨尺度注意力融合互补信息，确保特征传递的完整性。针对多尺度特征的空间不匹配问题，模块采用自适应尺度对齐策略通过双线性插值上采样将特征图统一至相同空间分辨率，既保证特征空间的连续性，又提升网络的特征表示能力。基于上述设计，双向特征融合后的左右分支输出可表示为：

$$L_{\text{out}} = L'' + L'' \odot g_L + (1 - g_L) \odot \mathcal{U}\left(g_H \odot H'', (H_L, W_L)\right) \tag{17}$$

$$H_{\text{out}} = H^n + H^n \odot g_H + (1 - g_H) \odot \mathcal{U}\left(g_L \odot L^n, (H_H, W_H)\right) \tag{18}$$

最后特征聚合后生成的特征图和输出分别可以表示为：

$$H_{\text{Gmal}} = \mathcal{U}\left(H_{\text{cant}}, (H_L, W_L)\right) \tag{19}$$

$$\text{Output} = \text{Conv}_{3\times3}\left(\left[H_{\text{final}}; L_{\text{out}}\right]\right) \tag{20}$$

2.2.2. MAMF 模块

当前深度学习在目标检测任务中面临三大核心挑战：一是需有效捕获复杂的空间关系，二是受限于有限的计算资源需实现高效处理，三是需平衡模型性能与计算成本。为此，本文在 C3k2 模块的基础上，引入门控卷积块与频域卷积，提出 MAMF 模块。该模块通过频域处理与门控机制的协同融合，一方面借助选择性通道处理提升计算效率，另一方面通过频域变换以较低复杂度获取全局感受野，并结合嵌套结构完成多尺度特征提取。

MAMF 模块的核心运算流程如下：首先通过归一化组件在空间与通道维度分别施加层归一化，以此稳定训练动态过程，同时规范化空间与通道维度的特征分布；随后对特征进行通道扩展，并将扩展后的特征分割为三个组件，即门控特征(gate)、恒等特征(identity, i)与卷积特征(conv, c)；针对卷积特征组件，采用 FDConv (频域卷积)进行处理，充分利用快速傅里叶变换(FFT)的计算高效性；最后，门控特征经 GELU 激活函数处理后，与拼接后的恒等特征和卷积特征(i, c)进行逐元素相乘，再通过通道压缩恢复至原始维度，得到模块最终输出。

本模块的核心优势在于解决传统空域卷积的固有缺陷：传统大卷积核虽能覆盖大感受野，但计算复杂度极高，而 MAMF 通过频域处理可在降低复杂度的同时实现全局感受野覆盖。此外，模块通过 conv-ratio 参数实现自适应通道处理，能够精细调节计算效率与特征表达能力之间的平衡。为保障特征传递的完整性并促进梯度反向传播，模块在多个层次嵌入残差连接，不仅实现了各模块内部的跳跃连接，还构建了多个门控块独立的层次化残差路径。

3. 实验分析

3.1. 数据集

本研究实验中使用的数据集是自收集的锥桶数据建立的，并把数据集分成训练集和测试集两种，训练集主要是用来训练网络模型，测试集主要是用来测试验证算法的效果。本文的数据集都是使用双目相机在不同类型的环境下拍摄的锥桶图片，图片全都使用如图工具对数据集中的锥桶进行标注。数据集共有 8000 多张锥桶图片，测试集，验证集和训练集的比率为 2:1:7，模型的优化器的选择为 SGD 随机梯度下降法，其余的参数设置为默认值。

3.2. 实验细节和评价标准

实验详情：本研究中的实验使用主要配备 NVIDIA RTX 3090 GPU 和 i9-12600 KF CPU 的硬件进行。软件配置涉及 Ubuntu 操作系统上的 PyTorch 1.13.1 深度学习框架。输入图像大小为 640 × 640，总共迭代 300 次。批量大小为 16，初始学习率为 0.01，其余的参数设置为默认值。所有实验都是从头开始训练的，没有使用预先训练的模型。评估指标：为了评估在这项研究中提出的模型的有效性，模型的复杂性进行了评估的基础上的参数和内存占用的数量。在性能方面，采用精确度、召回率、mAP 50 和 mAP 作为评价指标。mAP 是指平均精密度，通过取在 IoU 阈值范围为 0.5 至 0.95 的步长为 0.05 的情况下计算的所有

平均精密度(AP)的平均值来计算，与 mAP95 相同。

3.3. 消融实验

为了评估 MSFA 模块、CSBA 模块和 MAMF 模块的有效性，本研究在自制的数据集上设计了 7 个消融实验。消融实验的结果如表 1 所示。考虑到模型轻量化设计的重点，表中除基线外，其余模型均采用新的检测层结构。实验 A、B 和 C 表明，三种改进方法都在一定程度上提高了模型的性能。其中，MSFA 模块(实验 A)带来的改善最为显著，mAP50 从基线的 91.0%提升至 93.6%，mAP95 从 69.7%提升至 72.7%，展现出最佳的单模块改进效果。CSBA 模块(实验 B)和 MAMF 模块(实验 C)也分别带来了一定的性能提升。通过组合不同模块的实验 D、E、F 可以看出，模块间存在协同作用，但并非简单的叠加效果。最终的完整模型(Ours)集成了所有三个改进模块，与基线相比，在精确度、召回率、mAP50 和 mAP95 方面分别提高了 3.7%、3.6%、3.9%和 7.4%，达到了 89.9%、88.8%、94.9%和 77.1%的优异性能。虽然模型尺寸从 5.2MB 增加到 10.1MB，但性能的显著提升证明了这种权衡的合理性。实验结果表明，本文提出的三个改进模块在提高小目标识别性能方面具有显著效果，验证了改进方法的有效性和可行性。

Table 1. Results of ablation experiment
表 1. 消融实验结果

Models	MSFA	CSBA	MAMF	P (%)	R (%)	mAP50 (%)	mAP95 (%)	Size (mb)
Baseline				86.2	85.2	91.0	69.7	5.2
A	√			86.5	86.0	93.6	72.7	7.1
B		√		86.6	86.5	92.5	69.8	7.5
C			√	86.9	86.7	92.6	69.7	7.8
D	√		√	87.3	84.8	92.7	69.9	8.1
E		√	√	87.1	85.9	92.8	69.4	7.6
F	√	√		86.3	86.0	92.9	69.8	7.2
Ours	√	√	√	89.9	88.8	94.9	77.1	10.1

3.4. 对比实验

在 VisDrone 数据集上比较了现有的经典算法，验证了 MSFA-NET 在小目标检测中的性能。表 2 显示了经典算法的性能指标。与表中的其他算法相比，MSFA-NET 在 mAP 指标方面表现出色。虽然参数量为 27.7M 相对较大，但在复杂场景下的检测精度显著提升，完全满足实际应用要求。与在 VisDrone 数据集挑战中表现良好的 YOLOv8n [8]轻量级版本相比，MSFA-NET 的 AP 从 14.4%提升至 17.7%，AP50 从 25.9%提升至 31.8%，在小目标检测方面(APs)从 5.9%提升至 8.5%，展现出明显优势。综上所述，MSFA-NET 在小目标和复杂场景检测方面表现优异。图 3 显示了模型参数大小与其检测精度之间的对应关系。与 YOLOv11n 相比，MSFA-NET 达到了 17.7%的 AP，虽然参数量增加至 27.7M，但 AP 提升了 2.5%，AP50 提升了 6.0%，APs 提升了 2.7%。与最近的研究 DEIM-D-Fine-N [9]相比，两者在 AP 指标上基本持平，但我们的方法在 AP50 方面略有劣势，在小目标检测 APs 方面表现相当。这说明本文所提出的改进策略在保持整体检测精度的同时，在复杂场景下具有良好的鲁棒性。我们的实验是在自制的基准测试上进行的，使用的是目标检测的标准指标。所有模型均在自制的交通锥数据集检测(约 8000 张图像)上进行训练，并使用 test2019 (6k 张图像)进行评估。综上所述，PMA-YOLOv11 在小目标和低分辨率图像检测

方面表现优异。表 2 显示了模型在不同场景下的性能表现，与 YOLOv11n 相比，PMA-YOLOv11 在 AP 指标上提升了 2.5%，特别是在小目标检测(APs)方面提升了 2.7%，中等目标(APm)提升了 4.9%，大目标(APl)提升了 5.1%。这表明本文所提出的改进策略在各种尺度的目标检测中都具有显著的有效性。

Table 2. Comparative experimental results
表 2. 对比实验结果

Models	Input Shape	GFlops	Params	AP	AP50	APs	APm	APl
DEIM-D-Fine-N [9]	(640, 640)	7.123G	3.73M	0.177	0.322	0.090	0.262	0.376
YOLO8n	(640, 640)	8.1G	3.0M	0.144	0.259	0.059	0.225	0.339
YOLO10n	(640, 640)	6.5G	2.28M	0.142	0.261	0.063	0.224	0.292
YOLO10s	(640, 640)	21.4G	7.22M	0.179	0.323	0.086	0.278	0.361
YOLO11n	(640, 640)	6.3G	2.59M	0.142	0.258	0.058	0.225	0.316
YOLO11s	(640, 640)	21.3G	9.42M	0.176	0.313	0.080	0.272	0.364
YOLO12n	(640, 640)	6.3G	2.56M	0.142	0.259	0.057	0.224	0.346
YOLO12s	(640, 640)	21.2G	9.23M	0.176	0.312	0.081	0.274	0.356
(ours)	(640, 640)	40.1G	27.7M	0.177	0.318	0.085	0.274	0.367

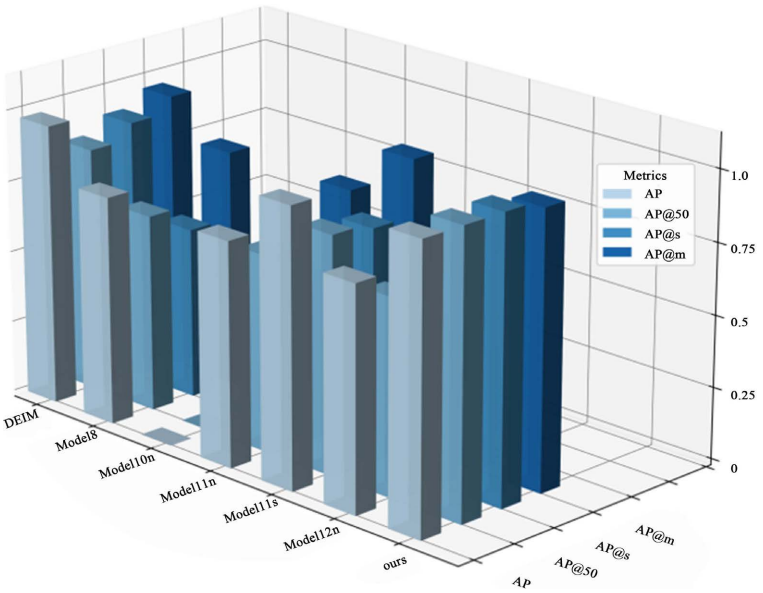
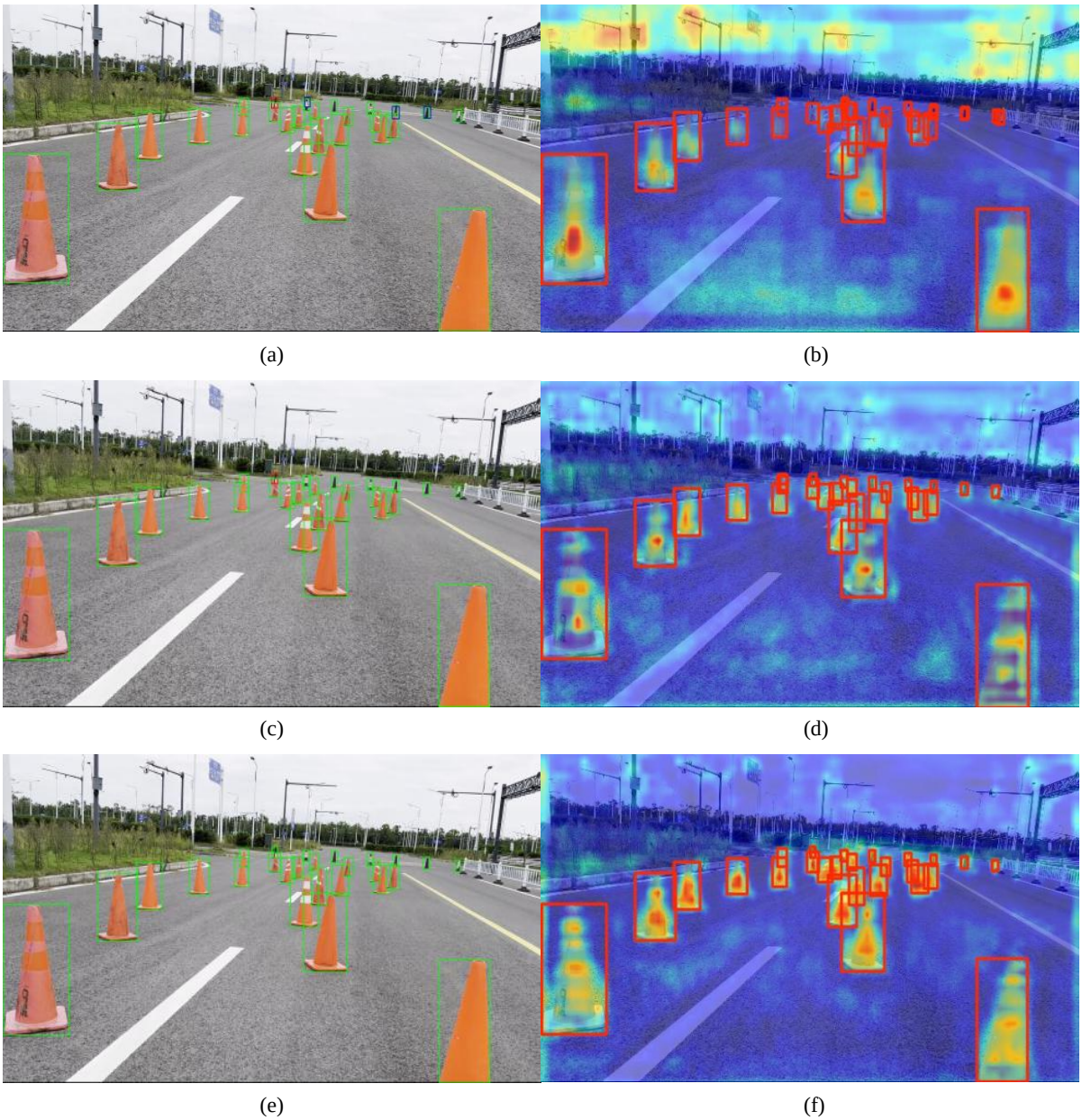


Figure 4. Performance comparison chart of object detection methods
图 4. 目标检测方法性能对比图

实验结果充分证明了本文方法在多项评估标准上的卓越性能。我们的方法获得了最高的整体平均精度(AP)，达到 0.181，相比现有方法取得了显著提升。具体而言，本方法比最接近的竞争对手 DEIM 提升了 2.3% (0.181 vs 0.177)，并且相比其他基线方法显示出大幅改进，与 Model10s、Model11s、Model8、Model11n 和 Model12n 相比，提升幅度在 2.8%到 27.5%之间。最显著的性能优势体现在 AP@50 指标上，我们的方法达到了 0.377 的得分，大幅超越了所有竞争方法。这相比第二名的 Model10s (0.323)提升了约 16.7%，充分展现了本方法在精确目标定位方面的卓越能力。AP@50 指标的大幅提升表明我们的方法在需要高定位精度的场景中表现优异，这对实际应用具有重要意义。

3.5. 实验检测结果及分析

为了在真实的场景中验证本文的模型，选取了交通锥数据集中检测难度较高的图像进行测试。结果分别见图 5。通过比较图 5(b)、图 5(d)和图 5(e)、图 5(f)中的突出显示区域，可以观察到 PMSFA-Net 对图像附近的小物体表现出更高的检测率边缘，有着出色的检测性能。Ours 在这张图像中检测到 21 个目标，而基线模型只检测到 19 个目标。总而言之，得益于针对性的改进，SFMSA-yolo11 对背景复杂、分布密集的交通锥图像表现出更好的检测能力，能够有效抑制图像背景噪声信息的干扰，并从中保留小目标特征信息。



(a) (b)

(c) (d)

(e) (f)

Figure 5. Comparing the results of the comparison chart, the green box represents successful recognition, and the blue box represents false positives

图 5. 对比图结果，绿色框代表成功识别，蓝框为误检

4. 结论

针对复杂工况下交通锥背景复杂、目标尺度小等特点，提出了一种轻量级的小目标检测算法 SFMSA-YOLOv11。该算法基于 YOLOv11n 的改进。首先，在骨干网络中，采用了渐进式的特征聚合策略，通过级联不同感受野的卷积操作，实现多尺度特征的有效融合，通过分组卷积和通道分割降低计算开销，保持了不同层次特征的完整性和多样性。对于颈部网络，加入了 MAM-Conv 模块，减少了下采样过程中步幅卷积造成的特征损失，提高了骨干网络在低分辨率、小目标图像中的特征提取能力。使用空间双向注意力模块基于多尺度特征的互补性，通过自适应，双向且计算高效的解决方案，有效解决了计算机视觉任务中多尺度特征融合的问题，从而在更大的感受野上获得更显著的特征信息。而检测头则采用了更大胆的共享细节增强块，根据小目标的实际情况进行设计，在一定程度上提升了模型性能。在 VisDrone 小目标数据集上的结果表明，SFMSA-YOLOv 11 平衡了准确性和复杂性，具有出色的小目标检测特性。在我们未来的研究中，我们将进一步消除模型中的冗余成分，并采用知识蒸馏等技术，从更大的模型中实现更高的检测精度。这将有助于使我们的模型更轻，更准确。

为进一步提升复杂工况下交通锥检测的性能，本文提出了一种融合多尺度特征提取、通道注意力机制、大核卷积和门控机制的改进架构。该模型专门改进了 YOLOv11 的检测层结构、主干和检测头。总体而言，这项工作的贡献如下：

1) 提出了一种多尺度特征聚合模块 MSFA (Multi-Scale Feature Aggregation)及其 CSP 结构变体(CSP_MSFA)，通过分层级联和分组卷积策略实现跨尺度特征的高效融合与增强，显著提升了模型对多尺度目标的表征能力。

2) 设计了跨尺度双向注意力 CSBA (Cross-Scale Bidirectional Attention)模块，通过自注意力机制和交叉注意力机制协同作用，实现高分辨率特征与低分辨率特征间的双向增强，有效保留细节信息的同时提升语义表征能力

3) 提出了门控全维度卷积(GatedFDConv)模块及其 C3k2 结构变体，通过通道分离策略和门控机制，将高效的部分通道深度卷积与全维度特征交互相结合，在保持低计算成本的同时显著增强模型对复杂特征的捕获能力。

4) 构建了一个包含 8000 多张图像的高质量交通锥专用数据集，涵盖不同光照(如白天、夜晚、逆光)、天气(如下雨、雾霾)、场景(如城市道路、高速公路、施工区域)及姿态(如直立、倾倒、部分遮挡)的交通锥实例，并采用像素级语义分割与边界框双重标注方式，为交通锥检测与分割任务提供了丰富且具挑战性的基准数据支撑。

参考文献

[1] Chen, Y., Yuan, X., Wang, J., Wu, R., Li, X., Hou, Q., *et al.* (2025) YOLO-MS: Rethinking Multi-Scale Representation Learning for Real-Time Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **47**, 4240-4252. https://doi.org/10.1109/tpami.2025.3538473

[2] Jiang, H., Hu, F., Fu, X., Chen, C., Wang, C., Tian, L., *et al.* (2023) YOLOv8-Peas: A Lightweight Drought Tolerance Method for Peas Based on Seed Germination Vigor. *Frontiers in Plant Science*, **14**, Article 1257947. https://doi.org/10.3389/fpls.2023.1257947

[3] Bai, J., Zhang, H. and Li, Z. (2018) The Generalized Detection Method for the Dim Small Targets by Faster R-CNN Integrated with Gan. 2018 *IEEE* 3*rd International Conference on Communication and Information Systems* (*ICCIS*), Singapore, 28-30 December 2018, 1-5. https://doi.org/10.1109/icomis.2018.8644960

[4] Wang, Z., Cao, Y. and Li, J. (2023) A Detection Algorithm Based on Improved Faster R-CNN for Spacecraft Components. 2023 *IEEE International Conference on Image Processing and Computer Applications* (*ICIPCA*), Changchun, 11-13 August 2023, 1-5. https://doi.org/10.1109/icipca59209.2023.10257992

[5] He, Z., Ye, X. and Li, Y. (2023) Compact Sparse R-CNN: Speeding up Sparse R-CNN by Reducing Iterative Detection

Heads and Simplifying Feature Pyramid Network. *AIP Advances*, **13**, Article ID: 055205. https://doi.org/10.1063/5.0146453

[6] Cai, F., Qu, Z. and Yin, X. (2025) A Feature Fusion Network with Multiscale Adaptively Attentional for Object Detection in Complex Traffic Scenes. *IEEE Transactions on Intelligent Vehicles*, **10**, 4217-4230. https://doi.org/10.1109/tiv.2024.3476991

[7] Wang, C., Mark Liao, H., Wu, Y., Chen, P., Hsieh, J. and Yeh, I. (2020) CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 14-19 June 2020, 1571-1580. https://doi.org/10.1109/cvprw50498.2020.00203

[8] Xu, L., Zhao, Y., Zhai, Y., Huang, L. and Ruan, C. (2024) Small Object Detection in UAV Images Based on YOLOv8n. *International Journal of Computational Intelligence Systems*, **17**, Article No. 223. https://doi.org/10.1007/s44196-024-00632-3

[9] Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X. and Shen, X. (2025) DEIM: DETR with Improved Matching for Fast Convergence. 2025 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 15162-15171. https://doi.org/10.1109/cvpr52734.2025.01412