

基于跨域双流网络的引导式特征融合小目标检测算法

陈志旺^{1,2†}, 孙艺萱^{1,2}, 彭勇³

- 燕山大学 智能控制系统与智能装备教育部工程研究中心, 河北 秦皇岛 066004;
- 燕山大学 河北省工业计算机控制工程重点实验室, 河北 秦皇岛 066004;
- 燕山大学 电气工程学院, 河北 秦皇岛 066004)

摘要: 针对无人机航拍图像中的目标存在尺寸小、分布密集、纹理细节模糊等问题, 提出基于跨域双流网络的引导式特征融合小目标检测算法. 首先, 在主干网络部分提出空间-频域协同作用的跨域双流检测架构, 所提出架构并行构建空间域和频域两条特征提取路径: 空间流侧重于捕获局部细节特征; 频域流设计边缘与频域增强模块, 通过频域变换和动态高斯掩码进行三频带划分, 并利用上下文感知门控机制自适应增强不同频率特征, 从而提升网络的全局上下文感知能力. 然后, 在跨域特征融合部分设计自适应空间-频域协同融合模块, 通过动态权重分配实现跨域特征的高效整合. 最后, 在颈部网络采用引导式三路融合模块, 以主路特征为引导自适应整合上采样、主路以及跨层特征间的语义和细节信息, 有效缓解多尺度特征间的语义差异. 在 VisDrone2019 和 TinyPerson 两个公开数据集上进行实验, 实验结果验证了所提出方法的有效性.

关键词: 目标检测; 跨域双流网络; 边缘与频域增强; 上下文感知门控机制; 跨域特征融合; 引导式三路融合
中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1220

引用格式: 陈志旺, 孙艺萱, 彭勇. 基于跨域双流网络的引导式特征融合小目标检测算法 [J]. 控制与决策.

Guided feature fusion algorithm for small object detection based on cross-domain dual-stream network

CHEN Zhi-wang^{1,2†}, SUN Yi-xuan^{1,2}, PENG Yong³

- Engineering Research Center of the Ministry of Education for Intelligent Control System and Intelligent Equipment, Yanshan University, Qinhuangdao 066004, China;
- Hebei Key Laboratory of Industrial Computer Control Engineering, Yanshan University, Qinhuangdao 066004, China;
- School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China)

Abstract: To address the challenges of small object detection in UAV aerial images, where targets are typically small in size, densely distributed, and lack clear texture details, this paper proposes a guided feature fusion algorithm based on a cross-domain dual-stream network. Specifically, a spatial-frequency collaborative dual-stream architecture is introduced in the backbone, in which spatial-domain and frequency-domain feature extraction pathways are constructed in parallel. The spatial stream focuses on capturing local detail features, while the frequency stream incorporates an edge and frequency enhancement module. This module performs three-band frequency decomposition via frequency transformation and dynamic Gaussian masking, and employs a context-aware gating mechanism to adaptively enhance features at different frequency bands, thereby improving the network's global context perception capability. Subsequently, an adaptive spatial-frequency collaborative fusion module is designed to efficiently integrate cross-domain features through dynamic weight allocation. Finally, a guided three-branch fusion module is adopted in the neck network, where the main-branch features serve as guidance to adaptively fuse semantic and detailed information from upsampling, main-branch, and cross-layer features, effectively alleviating semantic discrepancies across different scales. Experiments conducted on the VisDrone2019 and TinyPerson public datasets demonstrate the effectiveness of the proposed method.

Keywords: object detection; cross-domain dual-stream network; edge and frequency enhancement; context-aware gating mechanism; cross-domain feature fusion; guided three-branch fusion

收稿日期: 2025-11-25; 录用日期: 2026-04-05.

基金项目: 国家自然科学基金项目 (61573305); 河北省自然科学基金项目 (F2022203038, F2019203511); 河北省级重点实验室绩效补助经费项目 (22567612H).

责任编辑: 周平.

†通信作者. E-mail: czwaaron@ysu.edu.cn.

0 引言

随着无人机技术和目标检测算法的快速发展,基于无人机航拍图像的小目标检测已成为计算机视觉领域最具挑战性的问题之一.目标检测技术凭借其出色的场景理解和定位能力,已广泛应用于精准农业^[1]、灾害监测^[2]、交通管理、航空航天以及地质环境调查等应用场景,具有重要的研究和应用价值.然而,与传统地面拍摄图像相比,无人机航拍图像中目标尺寸较小、尺度差异显著且背景复杂^[3].此外,航拍影像具有视角高、覆盖范围广、场景复杂等特点,被检测对象往往存在尺度不均、姿态变化大或遮挡等情况,这些因素显著增加了小目标的检测难度,对检测模型设计和优化带来了挑战.

随着深度学习的快速发展,目标检测算法不断更新,主要可分为两阶段、单阶段两类检测算法.以R-CNN^[4]系列为代表的两阶段检测算法,通过区域候选和精细回归两步式结构来实现高精度目标识别;以YOLO系列^[5]为代表的单阶段检测算法,在统一框架下直接预测目标类别和边界框信息.尽管这些检测算法在常规目标检测任务中表现优异,但是,在小目标检测场景下仍然存在性能下降的问题.为此,研究者们提出了多种改进策略以增强模型的特征表达能力和小目标感知能力.如:Wu等^[6]提出的YOLO-LSM通过轻量化结构和小目标检测层设计提升了无人机小目标检测性能;Zhang等^[7]提出了FFCA-YOLO,通过特征增强与空间上下文感知模块,增强了小目标特征表达并抑制背景干扰;丁浩晗等^[8]提出的DI-YOLO引入动态特征选择和并行异构特征调制机制,实现了跨层特征的自适应融合与全局-局部信息的协同建模;Chen等^[9]提出的Leformer通过CNN-Transformer的双流特征提取方式,实现了全局语义与局部细节的协同建模.

然而,这些方法在跨域融合的协同性和多尺度特征融合方面仍然存在不足.一方面,当前主流方法主要依赖空间域进行特征提取,通过卷积操作对局部邻域加权计算来捕获边缘和纹理信息^[10].但是,受限于逐层传递机制,全局上下文信息需要经多层传递才能实现整合,导致全局语义表达能力受限.而频域特征提取具备更强的全局表达能力^[11],但是,频域建模通常会对不同频率成分进行增强或抑制,这会导致频谱中各频率响应强度发生变化,在反变换回空间域后会出现空间结构信息的弱化或重分布,从而影响目标的定位精度.此外,现有频域方法通常侧重于高频增强以突出边缘和纹理特征,未充分考虑不同场景下小目标的多样性和背景复杂性.因此,单

一域的特征提取方式难以兼顾局部细节与全局语义表达.而空间域-频域双主干结构可以很好地结合两者的优点,但是,现有跨域融合方式多采用直接拼接或加权,忽略了跨域特征的表达差异,易导致语义错配或特征贡献不均,难以充分发挥跨域信息的协同作用.另一方面,多尺度特征融合通常采用特征金字塔网络(FPN)为代表的多尺度结构,虽然能够将高层语义逐级传递至浅层特征图,但是,由于小目标像素面积较小,在多次下采样过程中不断被压缩,导致在深层特征图中仅保留少量特征,难以实现精确检测和定位.

针对上述问题,本文从跨域特征提取与多尺度协同融合的角度,提出基于跨域双流网络的引导式特征融合小目标检测算法,主要研究内容如下:

1) 针对单一域特征提取难以兼顾局部细节与全局语义的问题,本文在主干网络部分构建空间域与频域协同的跨域双流架构(CDF),以实现局部-全局特征的互补表达.同时,引入非对称的交互式双流融合策略(IDFC),以强化跨域特征间的信息交互,有效提升模型对小目标的感知能力.

2) 为充分发挥频域的全局建模优势,在频域流设计边缘与频域增强模块(EEFA).该模块通过动态高斯掩码和上下文感知门控机制,进行三频带划分并对不同频段特征进行自适应增强.此外,EEFA采用分层配置策略,在浅层和深层(C_2, C_5)仅保留频域增强分支,中间层(C_3, C_4)则启用边缘增强和空间注意力分支,以强化高频特征并抑制背景噪声.

3) 针对跨域特征融合难以充分协调空间域与频域特征表征差异的问题,设计自适应空间-频域协同融合模块(ASFCF).该模块通过注意力机制动态调节两域特征的融合权重,引导网络在融合过程中自适应调节空间流与频域流输出特征的贡献比例,实现跨域特征的互补增强和高效整合.

4) 为解决多尺度融合阶段不同层级特征语义表达与空间对齐不一致的问题,本文设计引导式三路融合模块(GTF).该模块以主路特征作为引导信息流,从空间对齐、通道选择和自适应加权3个方面实现多尺度特征的协同融合.此外,在检测头部分增加一个高分辨率的 P_2 检测层,进一步提升小目标的检测性能.

1 本文方法

所提出基于跨域双流架构的引导式特征融合小目标检测算法整体结构如图1所示.主干网络部分所改进的跨域双流架构(CDF)通过交互式双流融合

策略 (IDFC)、边缘与频域增强模块 (EEFA) 以及自适应空间-频域协同融合模块 (ASFCF), 有效解决了局部特性与全局语义表达不平衡的问题, 并针对性的增强了小目标的检测精度. 特征融合部分通过引

导式三路融合模块 (GTF) 来协调不同尺度特征间的语义差异, 强化小目标的细节捕捉能力. 所提出算法为无人机场景下的小目标检测提供了较优的特征提取和融合策略.

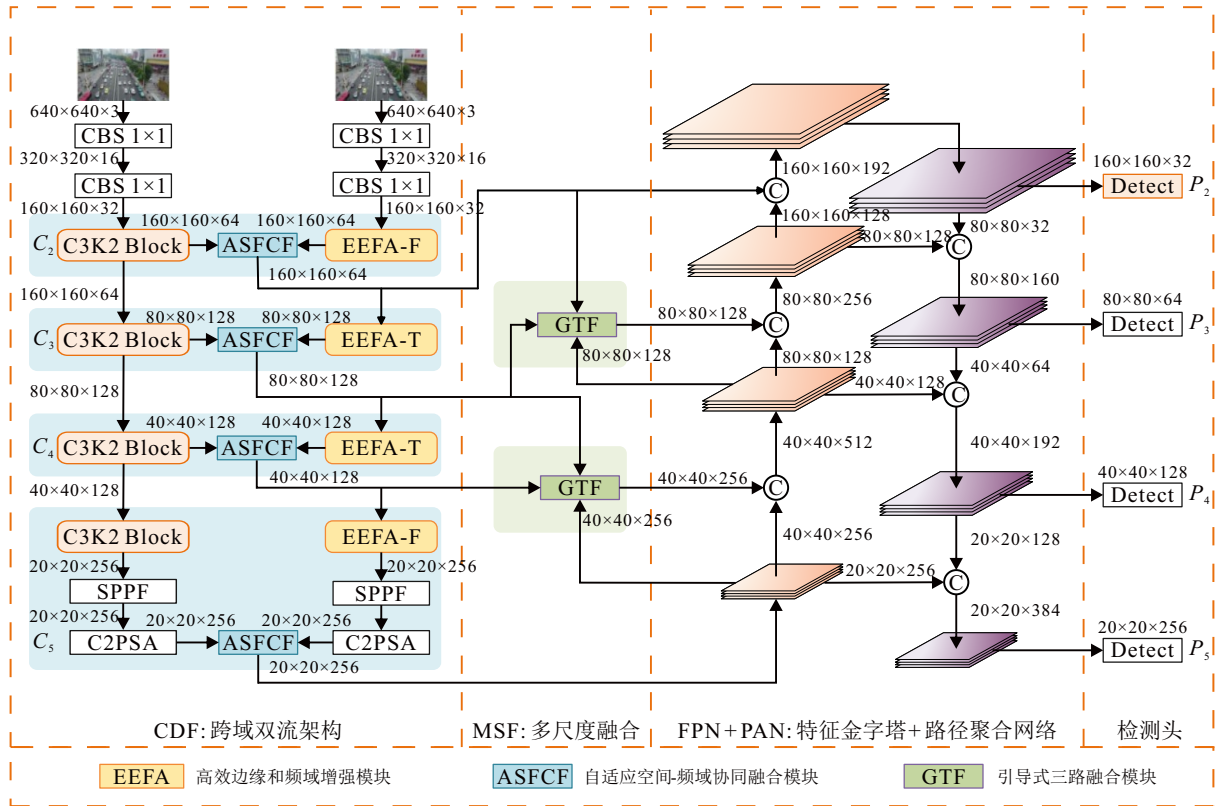


图1 模型网络结构

1.1 跨域双流架构

在小目标检测任务中, 大多数方法的主干网络依赖单一卷积神经网络 (CNN), 通过局部卷积核滑动进行加权计算来提取特征, 导致感受野受限. 尽管堆叠多层卷积可逐步扩大感受野, 但是, 这种扩展间接且有限, 难以有效捕获远距离上下文依赖. 因此, 空间域特征提取在捕获局部特征方面表现出色, 但是, 对全局语义的表征能力有限. 而频域特征提取通过频域变换获得全局频谱并分解为低频分量和高频分量, 再对不同频率成分进行增强或抑制, 可在一定程度上突出小目标特征. 但是, 频域建模通常依赖对频谱幅值的加权, 该过程会改变不同频率成分间的分布关系, 进而引起空间结构信息的弱化或重分布, 不利于目标的精确定位. 因此, 单一的空间域或频域特征提取方式难以兼顾局部细节与全局语义表达.

针对上述问题, 本文提出空间域-频域协同作用的跨域双流架构 (CDF), 如图 2 (c) 所示. 该架构由空间流与频域流两条互补特征提取路径组成, 旨在实现局部细节与全局语义的协同表达. 在空间流中, 沿用 CNN 主干结构, 通过 C3K2 等模块逐级下采样来

提取特征; 在频域流中, 采用边缘与频域增强模块 (EEFA), 首先, 通过快速傅里叶变换 (FFT) 将特征从空间域映射至频域, 获得全局频谱; 然后, 利用高斯掩码机制将频谱划分为低、中、高三段频带, 并通过上下文门控感知机制 (DGG) 自适应调节各频带的响应强度, 从而实现低频与高频特征的动态平衡, 提升网络在不同场景下的全局和局部表征能力.

在双流特征的传递机制上, CDF 引入了交互式双流融合策略 (IDFC). 不同于传统跨域双流网络在空间域与频域间独立传递、互不交互的特征提取方式, IDFC 在两条分支间构建了一种非对称的、逐层的传递结构. 具体而言, 空间流在每阶段输出的特征沿自身主干向下传递的同时, 通过融合模块与频域流融合, 并仅作为频域分支下一层的输入, 而不回传至空间流. 通过该非对称交互机制, 空间域的空间结构信息能够逐层参与频域流的特征提取过程, 从而在逐层特征传递过程中实现空间域与频域特征的信息交互和补偿.

为验证 CDF 融合策略的有效性, 本文对 3 种不同的融合策略进行对比, 如图 2 所示, 具体如下:

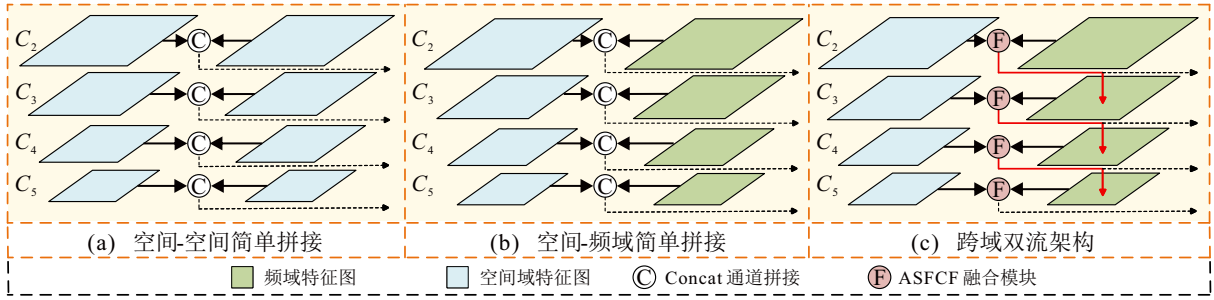


图2 双主干融合对比

1) 空间-空间特征的简单拼接. 该方法仅对同域特征进行线性叠加, 但是, 相同域特征的重复信息较多, 直接拼接不仅会导致通道维度成倍增加, 还易引入冗余特征干扰, 从而掩盖小目标的微弱响应.

2) 空间-频域特征的简单拼接. 该方式直接对两域特征进行线性叠加, 未考虑两域特征的表达差异和贡献比例问题, 易导致响应幅值较强或表达更连续的单一域特征主导网络的特征学习过程, 从而导致另一域的有效信息被削弱. 同时, 两域特征独立传递, 缺乏有效交互, 频域加权过程还可能会导致空间位置信息弱化.

3) 跨域双流架构 (CDF). 不同于两域简单堆叠的融合方式, CDF 采用交互式双流融合策略 (IDFC), 在各层建立了非对称的跨域特征交互机制. 空间流特征通过自适应频域-空间协同融合模块 (ASFCF) 与频域流特征进行融合, 自适应调节两域的融合比例, 并将融合后的特征传递至频域流的下一层, 从而缓解频域处理后空间结构信息弱化的问题. 通过这种逐层交互和自适应融合方式, CDF 同时整合全局语义与局部细节特征, 显著提升了模型在复杂背景下的微小目标检测性能, 具体实验对比结果分析见后文第 2.3.2 节.

1.2 边缘与频域增强模块

近年来, 许多基于频域的特征提取模块通过快速傅里叶变换 (FFT) 或小波变换将输入特征映射至频域, 并分解为低频和高频成分进行调制^[12]. 然而, 单一频段的强化难以适应复杂场景, 过度增强高频会放大噪声, 过度抑制低频则可能会破坏空间结构信息, 从而影响小目标的检测精度. 为此, 本文在频域流设计边缘与频域增强模块 (EEFA), 如图 3 所示.

考虑到频域流不同层级特征在微小目标检测中的作用差异, EEFA 模块内部采用可选择机制, 通过参数配置对不同层采用差异化处理策略, 以实现多层特征的自适应增强. 在浅层和深层阶段 (C_2 、 C_5), 仅保留频域增强分支 (EEFA-F), 该分支由环形频带划分、上下文感知门控机制 (DGG)、残差特征融合 3

部分组成. 在中间层阶段 (C_3 、 C_4), EEFA 模块采用完整模式 (EEFA-T), 在 EEFA-F 的基础上叠加拉普拉斯卷积和空间注意力机制, 以增强对小目标的检测能力并抑制背景噪声. 中间层特征既保留了小目标的关键纹理信息, 又具有一定语义表达能力, 因此, 适合作为高频增强的主要层级.

首先, 输入特征 $X_{F_1} \in \mathbb{R}^{B \times C \times H \times W}$ 经快速傅里叶变换 (FFT) 映射至频域空间 F , 如下所示:

$$F = \text{fft}(X_{F_1}), \quad (1)$$

其中 $\text{fft}(\cdot)$ 为快速傅里叶变换. 然后, 通过动态高斯掩码机制在频域中构建低频、中频和高频 3 条环形滤波带, 其中心位置和带宽分别由可学习参数 μ_k 和 σ_k 控制.

具体而言, 对于频谱中任一个位置 (h, w) , 计算该点到频谱中心的归一化欧氏距离, 将其作为该位置的频率半径 r , 有

$$r = \frac{\sqrt{h^2 + w^2}}{\sqrt{2}}, \quad h, w \in [-1, 1]. \quad (2)$$

其中 $r \in [0, 1]$ 为该点低频到高频的相对位置, 将所有位置计算得到的 r 整理为一个径向距离矩阵 $R \in \mathbb{R}^{H \times W}$.

在此基础上, 定义 3 条环形高斯滤波带 M_k , 即

$$M_k(R) = \exp\left(-\frac{(R - \mu_k)^2}{2\sigma_k^2}\right), \quad k \in \{1, 2, 3\}. \quad (3)$$

其中: $M_k \in \mathbb{R}^{H \times W}$; $\exp(\cdot)$ 表示 e^x ; μ_k 和 σ_k 分别为第 k 个频带的中心半径和宽度, 且均为可学习参数, 在训练过程中通过反向传播自适应更新.

这种基于高斯函数的连续分频方式能够自适应调整三频段的覆盖范围, 相较于传统固定阈值的硬分频方式, 能够形成平滑的环形频带, 从而有效避免硬分频时可能产生的边界不连续或频谱伪影问题.

在频域增强过程中, 如何在不同场景下动态调节高低频响应是关键问题. 为此, EEFA 引入上下文感知门控机制 (DGG). 首先, 通过全局平均池化提取输入特征的上下文信息; 然后, 通过两层 1×1 卷积和非线性激活函数生成 3 个频段的注意力权重 g_1 、

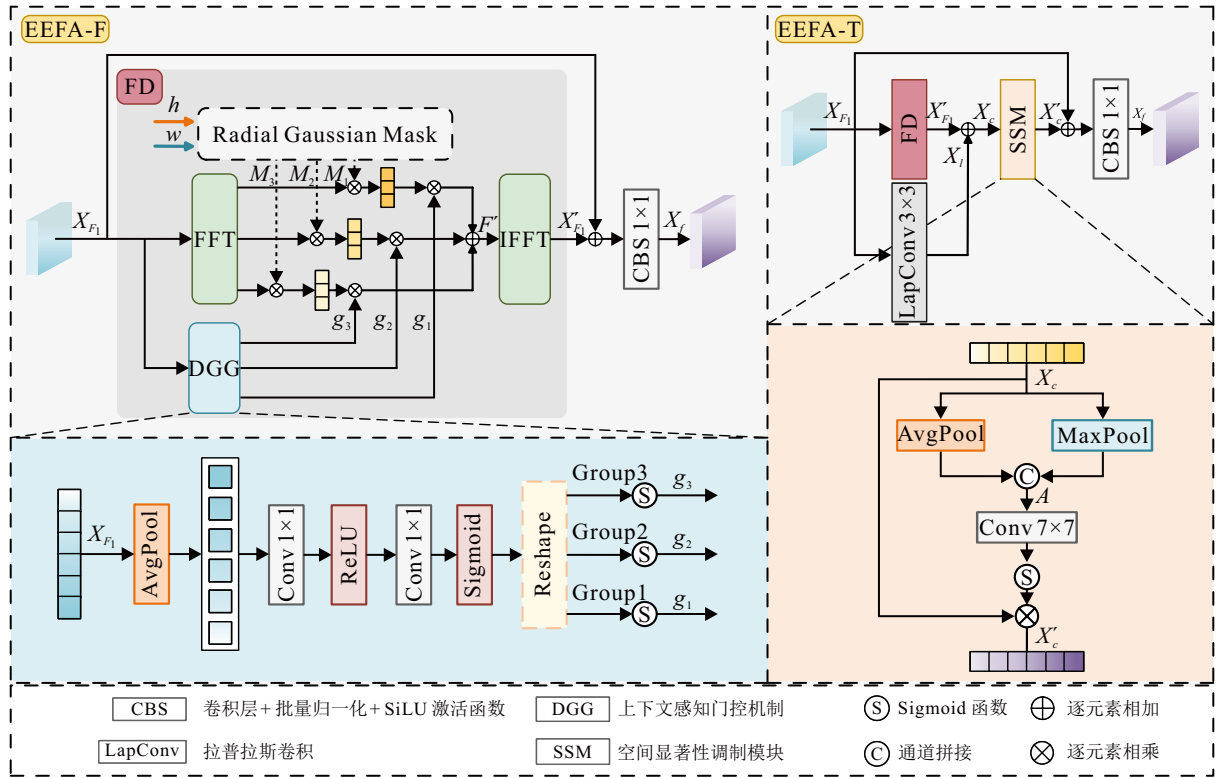


图3 边缘与频域增强模块

g_2 和 g_3 , 以表征各频段的重要性, 如下所示:

$$[g_1, g_2, g_3] = \sigma(\text{Conv}_{1 \times 1} \delta(\text{Conv}_{1 \times 1} \text{Avg}(X_{F_1}))). \quad (4)$$

其中: $\text{Avg}(\cdot)$ 表示全局平均池化, $\delta(\cdot)$ 为 ReLU 激活

函数, $\sigma(\cdot)$ 为 Sigmoid 函数。

该门控机制是可学习的, 网络通过反向传播自适应调整各频段权重, 从而实现对不同频率成分的动态调节, 其频域特征变化过程如图 4 所示。

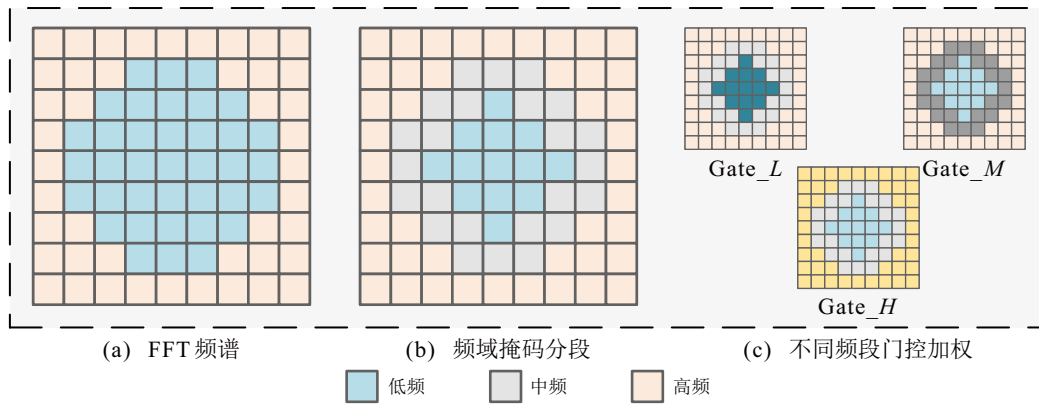


图4 频谱变化

为生成增强频谱特征, 采用加权求和方式对不同频率分量进行融合, 而未引入归一化操作, 以保留各频段间的幅值差异, 从而更有效地强化关键频率成分。3 段频带经加权组合形成增强频谱 F' , 有

$$F' = \sum_{k=1}^3 g_k (M_k \otimes F). \quad (5)$$

然后, 通过快速傅里叶逆变换 (iFFT) 将增强后的特征映射回空间域 X'_{F_1} , 有

$$X'_{F_1} = \text{ifft}(F'). \quad (6)$$

在中间层阶段, EEFA 通过固定权重的拉普拉斯卷积核 K_{lap} 进行边缘提取, 通过深度可分离卷积在每个通道上独立执行该操作, 获得边缘增强特征 X_l , 如下所示:

$$K_{\text{lap}} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \quad (7)$$

$$X_l = \text{LapConv}_{3 \times 3}(X_{F_1}), \quad (8)$$

其中 $\text{LapConv}(\cdot)$ 为 3×3 拉普拉斯卷积核。

由于直接增强高频可能会误强化背景噪声, 为

此引入空间显著性调制模块 (SSM) 以进一步抑制背景噪声并增强显著区域的响应, 输出的特征为 X'_c , 有

$$X_c = X'_{F_1} \oplus X_l, \quad (9)$$

$$A = \text{Concat}(\text{Avg}(X_c), \text{Max}(X_c)), \quad (10)$$

$$X'_c = X_c \otimes \sigma(\text{Conv}_{7 \times 7}(A)). \quad (11)$$

其中: $\text{Max}(\cdot)$ 表示全局最大池化, $\text{Concat}(\cdot)$ 表示通道拼接操作. 最后, 将 X_{F_1} 与原特征进行残差连接, 并通过卷积模块得到输出特征图为 X_f , 即

$$X_f = \text{CBS}_{1 \times 1}(X'_c + X_{F_1}). \quad (12)$$

其中: $X_f \in \mathbb{R}^{B \times C \times H \times W}$, $\text{CBS}_{1 \times 1}(\cdot)$ 表示 1×1 卷积操作、批量归一化和 SiLU 激活函数.

1.3 自适应空间-频域协同融合模块

在跨域双流网络中, 如何高效融合空间域与频域特征是实现跨域特征互补表征的关键问题. 频域增强通过对频谱幅值进行加权改变了频谱能量分布, 在反变换回空间域后可能会引起空间结构信息的弱化或重分布, 使其与空间流所提取的空间结构特征存在差异. 空间域侧重局部结构信息, 而频域更强调全局语义, 两域在特征表达方面具有天然互补性. 然而, 简单的拼接或相加两域的输出特征, 易导致某一域的特征响应占主导, 从而淹没另一域特征, 削弱跨域互补的优势, 导致融合结果无法同时兼顾局部细节与全局语义. 为此, 本文设计自适应空间-频域协同融合模块 (ASF CF), 其结构如图 5 所示.

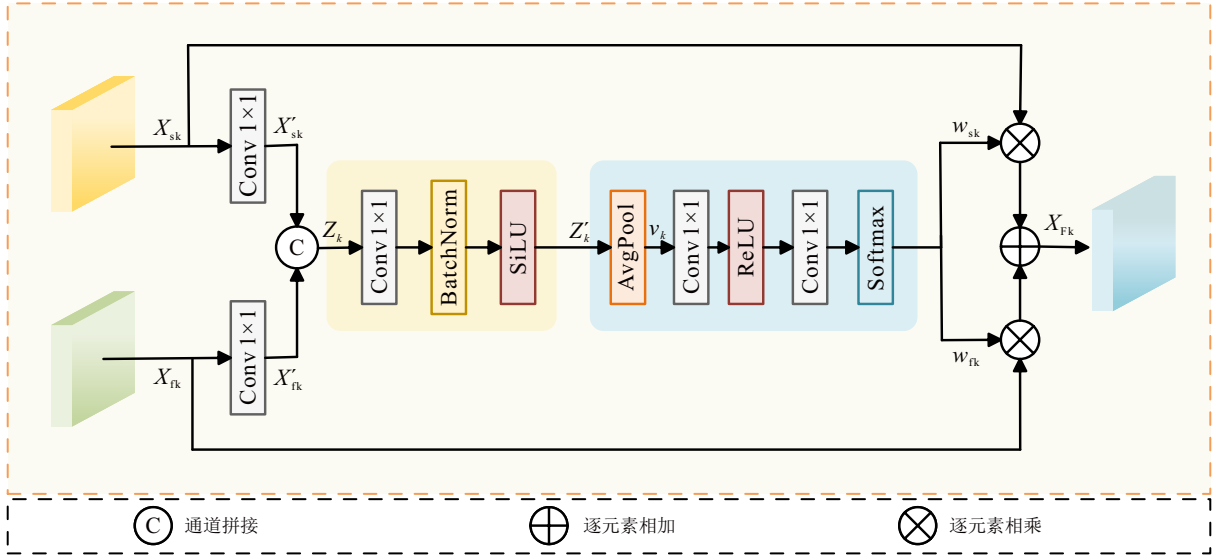


图5 自适应频域-空间协同融合模块

ASF CF 由通道对齐、特征融合以及注意力加权 3 个部分组成, 通过动态加权机制自适应调节空间域与频域特征的贡献比例, 从而充分发挥二者在局部细节与全局语义上的互补性, 提升小目标的表达能力和检测精度.

首先, ASF CF 采用 1×1 卷积对空间域特征 $X_{sk} \in \mathbb{R}^{B \times C \times H \times W}$ 与频域特征 $X_{fk} \in \mathbb{R}^{B \times C \times H \times W}$ 进行线性投影, 实现通道对齐, 保证跨域特征的可融合性, 有

$$X'_{sk} = \text{Conv}_{1 \times 1}(X_{sk}), \quad X'_{fk} = \text{Conv}_{1 \times 1}(X_{fk}), \quad (13)$$

其中 $k \in \{1, 2, 3, 4\}$. 对齐后的特征通过拼接操作在通道维度上进行信息整合, 如下所示:

$$Z_k = \text{Concat}(X'_{sk}, X'_{fk}). \quad (14)$$

然后, 利用轻量卷积块 (CBS) 进行非线性映射, 以促进跨域信息的交互融合, 如下所示:

$$Z'_k = \text{SiLU}(\text{BN}(\text{Conv}_{1 \times 1}(Z_k))). \quad (15)$$

其中: $\text{SiLU}(\cdot)$ 为 SiLU 激活函数, $\text{BN}(\cdot)$ 为批归一化.

在此基础上, 引入通道注意力机制生成跨域动态权重. 首先通过全局平均池化提取上下文特征, 如下所示:

$$v_k = \text{Avg}(Z'_k). \quad (16)$$

然后, 通过 MLP 结构预测生成频域和空间域的动态权重 w_{fk} 和 w_{sk} , 并通过 Softmax 归一化约束, 有

$$[w_{sk}, w_{fk}] = \text{Softmax}(\text{Conv}_{1 \times 1}(\delta(\text{Conv}_{1 \times 1}(v_k)))), \quad (17)$$

其中 $\delta(\cdot)$ 为 ReLU 激活函数. 最后, 通过自适应加权组合得到融合特征 X_{fk} , 如下所示:

$$X_{fk} = w_{sk} \cdot X_{sk} + w_{fk} \cdot X_{fk}. \quad (18)$$

该设计避免了单一域特征在融合中占主导的问题. 通过通道注意力生成动态权重, 自适应调节空间域与频域特征的贡献比例, 在 CDF 的融合过程中捕

获局部细节特征的同时补充了全局语义信息, 从而提升小目标在复杂场景下的检测性能. 此外, 得益于轻量化设计和通道共享机制, ASFCF 在增强表达能力的同时保持了较高计算效率.

1.4 引导式三路融合模块

在多尺度特征融合阶段, 特征金字塔网络 (FPN) 利用骨干网络产生的多尺度特征图, 将深层语义传递至浅层, 从而增强浅层特征的语义表达能力. 然

而, 在跨域双流主干结构中, 不同尺度特征在空间分辨率与分布特性上存在明显差异. 主路特征 S_k 表征当前层的核心语义信息; 深层上采样特征 U_k 具有更强的全局语义, 但是, 局部信息较弱; 跨层特征 P_k 则保留丰富的局部信息, 但是, 缺乏全局上下文. 直接采用拼接或逐元素相加进行融合易导致特征冲突. 为此, 本文设计引导式三路融合模块 (GTF), 如图 6 所示.

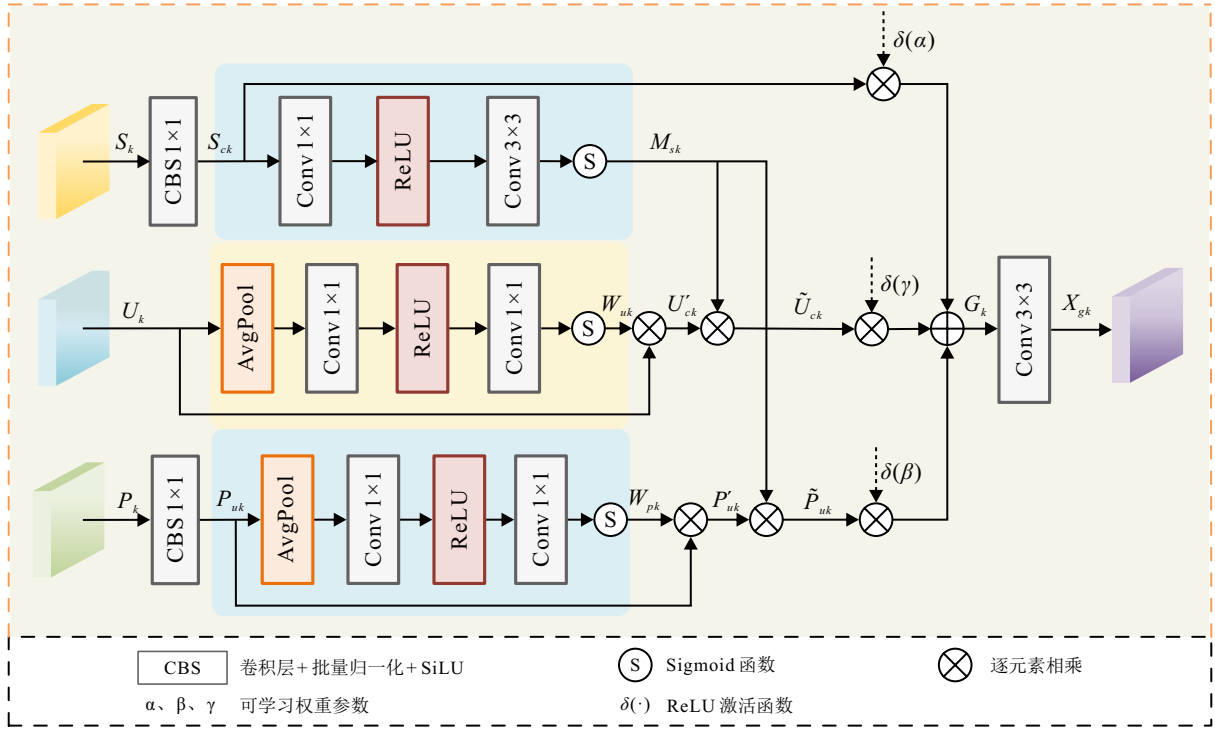


图6 引导式三路融合模块

GTF 以主路特征为引导信息流, 从通道选择、空间引导和动态加权融合 3 个层面实现多尺度特征的协同优化.

首先, 跨层特征 $P_k \in \mathbb{R}^{B \times \frac{C}{2} \times 2H \times 2W}$ 采用步长为 2 的 CBS 卷积单元进行下采样和通道映射, 以实现与主路特征的空间和通道对齐, 有

$$P_{uk} = \text{CBS}_{1 \times 1, s=2}(P_k). \quad (19)$$

其中: $P_{uk} \in \mathbb{R}^{B \times C \times H \times W}$, $k \in \{1, 2\}$.

对主路特征 $S_k \in \mathbb{R}^{B \times C \times H \times W}$ 则采用步长为 1 的 CBS 卷积单元进行通道嵌入和特征规范化, 如下所示:

$$S_{ck} = \text{CBS}_{1 \times 1, s=1}(S_k), \quad (20)$$

其中 $S_{ck} \in \mathbb{R}^{B \times C \times H \times W}$. 然后, 通过主路特征 S_k 生成空间引导掩码 M_{sk} , 用于辅助分支的选择性增强, 有

$$M_{sk} = \sigma(\text{Conv}_{3 \times 3}(\delta(\text{Conv}_{1 \times 1}(S_{ck}))))). \quad (21)$$

接着, 通过全局平均池化提取上采样特征 P_{uk} 和跨层特征 U_k , 并通过 MLP 生成通道权重 W_{pk} 和 W_{uk} , 从

而对各自通道进行重标定以抑制冗余噪声并突出有用通道, 如下所示:

$$W_{pk} = \sigma(\text{Conv}_{1 \times 1}(\delta(\text{Conv}_{1 \times 1}(\text{Avg}(P_{uk}))))), \quad (22)$$

$$W_{uk} = \sigma(\text{Conv}_{1 \times 1}(\delta(\text{Conv}_{1 \times 1}(\text{Avg}(U_k))))). \quad (23)$$

这里: $\text{Avg}(\cdot)$ 表示全局平均池化, $\delta(\cdot)$ 为 ReLU 激活函数, $\sigma(\cdot)$ 为 Sigmoid 函数. 接下来, 通过通道加权操作, 进行通道选择, 有

$$P'_{uk} = P_{uk} \otimes W_{pk}, \quad U'_k = U_k \otimes W_{uk}. \quad (24)$$

随后, 利用空间引导掩码 M_{sk} , 对已加权的辅助分支进行选择增强, 实现三路特征的协同增强, 如下所示:

$$\tilde{P}'_{uk} = P'_{uk} \otimes M_{sk}, \quad \tilde{U}_k = U'_k \otimes M_{sk}. \quad (25)$$

为实现三路特征间的自适应融合, GTF 设置 3 个可学习权重参数 α 、 β 和 γ (数值由网络在训练过程中学习得到, 经 ReLU 保证非负) 进行引导式动态加权融合. 该融合过程受到主路特征引导, 可根据不同

特征的重要性自适应调整各分支的贡献比例,即

$$w_{\text{sum}} = \delta(\alpha) + \delta(\beta) + \delta(\gamma) + \varepsilon, \quad (26)$$

$$G_k = \frac{\delta(\alpha) \cdot S_{ck} + \delta(\beta) \cdot \tilde{P}_{uk} + \delta(\gamma) \cdot \tilde{U}_k}{w_{\text{sum}}}, \quad (27)$$

其中 $\varepsilon = 10^{-6}$ 为防止除 0 的微小常数. 最后, 采用 3×3 卷积对 G_k 进行局部邻域增强与平滑处理, 如下所示:

$$X_{gk} = \text{Conv}_{3 \times 3}(G_k). \quad (28)$$

该模块通过主路引导的动态加权机制, 实现深层语义、浅层细节与当前层上下文信息的协同融合.

2 实验结果与分析

2.1 实验设置与评估指标

本文实验使用一张显存为 16 GB 的 NVIDIA RTX4080Ti 显卡, 显卡驱动版本 570.172.08, CUDA 版本 12.8. 实验中使用 python 版本 3.10, Pytorch 版本 2.4.1. 实验将输入图像尺寸调整为 640×640 像素, 学习率初始值为 0.01, batchsize 设置为 4, 最大训练轮数设置为 300.

为全面评估目标检测模型的性能, 实验采用在两个 IoU 阈值 (0.5 和 0.5 ~ 0.95) 下的平均精确率均值 mAP50% 和 mAP50% ~ 95%, 调和平均数 F1-score, 计算量 FLOPs、模型参数量 Params 以及每秒

处理帧数 FPS 等多个指标评估模型.

2.2 实验数据集

本研究在 VisDrone2019^[13] 和 TinyPerson^[14] 两个公开数据集上进行实验: VisDrone2019 数据集包含 10209 张无人机拍摄图像, 划分为训练集 6471 张, 验证集 548 张和测试集 3190 张, 共约 260 万个标注边界框, 目标被分为 10 个类别; TinyPerson 数据集将目标分为海上、陆地人类两类目标. 其中: 包含 1610 张图像, 且手动注释 72651 个目标边界框, TinyPerson 中的人相对较小, 图像尺寸主要为 1920×1080 , 人物尺寸通常小于 20 像素.

2.3 消融实验

2.3.1 不同模块的实验结果分析

为验证所提出基于跨域双流网络的引导式特征融合小目标检测算法在无人机航拍图像检测任务中的有效性, 本文基于 YOLOv11n 为基准模型, 在 VisDrone2019 和 TinyPerson 两个公开数据集上进行消融实验, 结果如表 1 和表 2 所示. 在 VisDrone2019 数据集上所提出方法的 mAP50% 和 mAP50% ~ 95% 相较于基准模型分别提升了 8.1% 和 5.1%, 在 TinyPerson 数据集上分别提升了 4.1% 和 1.37%, 该实验结果验证了所提出改进策略对小目标检测的有效性.

表1 不同模块在 VisDrone2019 数据集上的消融实验结果

Baseline	P2	IDFC	EEFA	ASFCF	GTF	P	R	F1	mAP50%	mAP50% ~ 95%	Params/M	FLOPs/G	FPS
✓						44.2	34.2	38.6	33.9	19.4	2.58	6.3	268.0
✓	✓					47.4	37.2	41.7	37.6	22.1	2.91	10.7	204.0
✓	✓				✓	49.6	39.2	43.8	40.0	23.7	4.64	23.5	179.5
✓	✓	✓				48.6	38.3	42.8	38.9	22.6	3.77	7.5	151.7
✓	✓	✓	✓			48.2	38.6	42.9	39.1	22.8	3.52	6.9	134.3
✓	✓	✓	✓	✓		50.4	39.6	44.4	40.1	23.5	3.63	7.0	121.4
✓	✓	✓	✓	✓	✓	50.8	41.2	45.5	42.0	24.5	4.58	9.8	114.3

表2 不同模块在 TinyPerson 数据集上的消融实验结果

Baseline	P2	IDFC	EEFA	ASFCF	GTF	P	R	F1	mAP50%	mAP50% ~ 95%	Params/M	FLOPs/G	FPS
✓						35.6	17.7	23.6	15.1	5.26	2.58	6.3	233.6
✓	✓					32.8	22.0	26.3	17.4	6.14	2.91	10.7	211.1
✓	✓				✓	35.6	22.7	27.6	18.9	6.35	4.64	23.5	167.3
✓	✓	✓				35.4	22.9	27.8	18.2	6.21	3.77	7.5	139.7
✓	✓	✓	✓			35.6	22.8	27.8	18.6	6.46	3.52	6.9	134.3
✓	✓	✓	✓	✓		35.2	23.3	28.0	19.0	6.61	3.63	7.0	114.3
✓	✓	✓	✓	✓	✓	35.9	23.4	28.3	19.2	6.63	4.58	9.7	104.5

此外, 为进一步分析 CDF 与 GTF 的效果, 对其引入前后的归一化混淆矩阵进行对比分析. 混淆矩阵通过统计各类别样本被正确识别以及误分类的比例, 直观反映模型在不同目标类别与背景间的区分

能力. 如图 7 所示, 随着 CDF 与 GTF 的逐步引入, 多个小目标类别 (如 people、bicycle 等) 在主对角线上的数值明显提高, 表明模型对真实类别的识别准确率得到了有效增强. 同时, 目标类别与 background

间的误检和漏检比例明显降低, 表明模型对目标与背景的区别能力进一步提升.

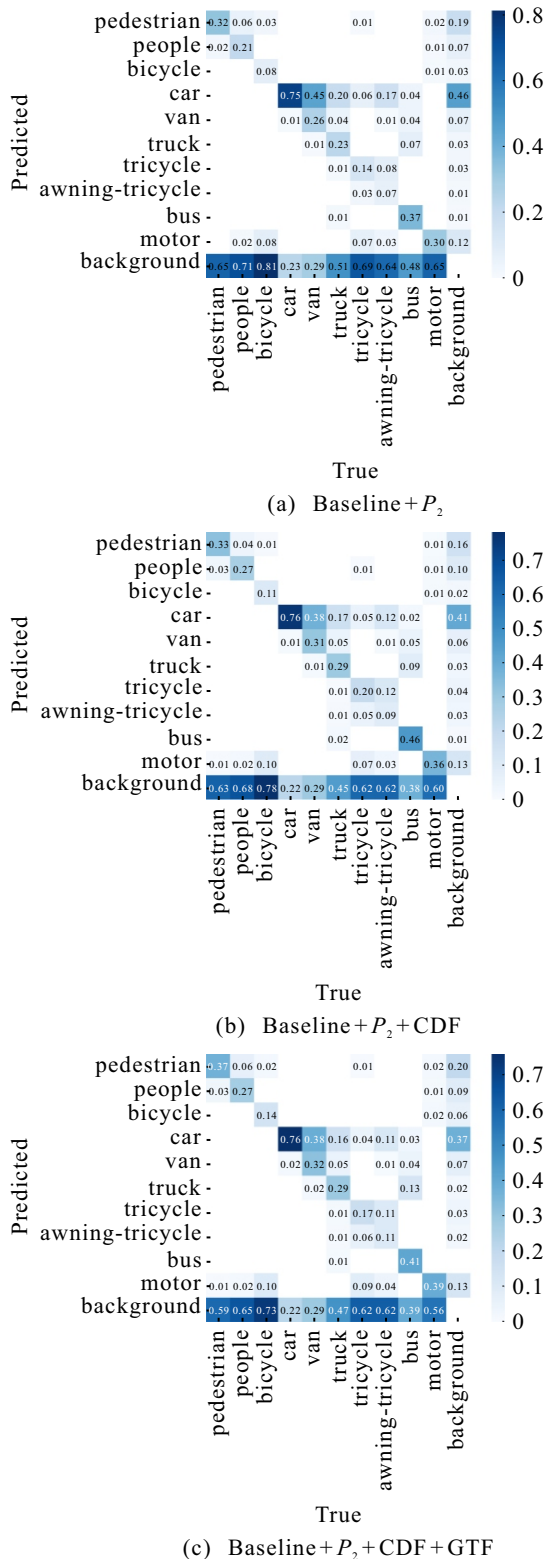


图7 CDF 与 GTF 模块集成前后的混淆矩阵对比

上述变化与特征响应的可视化结果保持一致, 如图 8 所示. 图 8(a) 的响应区域较为零散, 对于小目标的关注不稳定且存在较多背景激活; 图 8(b) 引入 CDF 后, 模型在更多小目标区域形成连续且清晰的响应, 并在目标边缘与轮廓处产生更强激活; 图 8(c)

进一步结合 GTF, 模型响应更加集中于真实目标区域, 背景噪声明显减少, 小目标区域的激活更加连续且聚焦.

2.3.2 不同双主干融合方式实验结果分析

为探究不同双流架构的融合方式对小目标检测性能的影响, 本文在 VisDrone2019 和 TinyPerson 数据集上分别设计并比较 4 种融合策略: 1) 空间-空间简单拼接; 2) 空间-空间交互拼接; 3) 空间-频域交互拼接; 4) 跨域双流架构 (CDF). 实验结果如表 3 所示, 跨域双流架构 (CDF) 在各项指标上均取得最优性能. 相比于空间-空间简单拼接的方式, 在 VisDrone2019 上 mAP50% 提升了 1.8%, 在 TinyPerson 上提升了 1.3%, 且该架构通过 IDFC 增强特征交互并避免通道膨胀, 使得参数量减少了 0.38 M.

2.3.3 不同主干的比较实验

为验证跨域双流架构 (CDF) 主干在小目标检测任务中的有效性, 本文进行多种主干替换实验, 将 CDF 与多种具有代表性的轻量化卷积主干以及主流 Transformer 主干进行系统对比. 各模型在 VisDrone2019 数据集上的实验结果如表 4 所示. 实验结果表明, 所提出 CDF 主干在所有关键指标上均取得最优性能, 其 mAP50% 为 40.1%, mAP50% ~ 95% 为 23.5%. 在保持较低参数量的同时实现了更优的检测精度.

2.3.4 EEFA 不同尺度采用 EEFA-T 的比较实验

为进一步探究边缘与频域增强模块 (EEFA) 中不同子结构的作用效果, 本文对跨域双流架构的频域分支进行实验. 分别在不同层级 ($C_2 \sim C_5$) 中引入 EEFA-T 模块, 其余层仅保留频域增强结构 EEFA-F, 实验结果如表 5 所示. 当 EEFA-T 模块作用于中间层 (C_3, C_4) 时, 模型在 mAP50% 和 mAP50% ~ 95% 上分别为 40.1% 和 23.5%, 性能达到最优.

2.4 对比实验

2.4.1 对比实验结果

如表 6 和表 7 所示, 所提出基于跨域双流架构的引导式特征融合检测算法在 VisDrone2019 和 TinyPerson 两个公开数据集上进行实验, 并与多种主流检测算法进行性能对比. 实验结果表明, 所提出方法在复杂航拍场景下表现出显著的检测优势. 在 VisDrone2019 数据集上, 两项关键的评价指标 mAP50% 和 mAP50% ~ 95% 分别达到了 42.0% 和 24.5%. 在 TinyPerson 数据集上, 分别达到了 19.2% 和 6.63%, 充分验证了所提出方法在精度和特征表

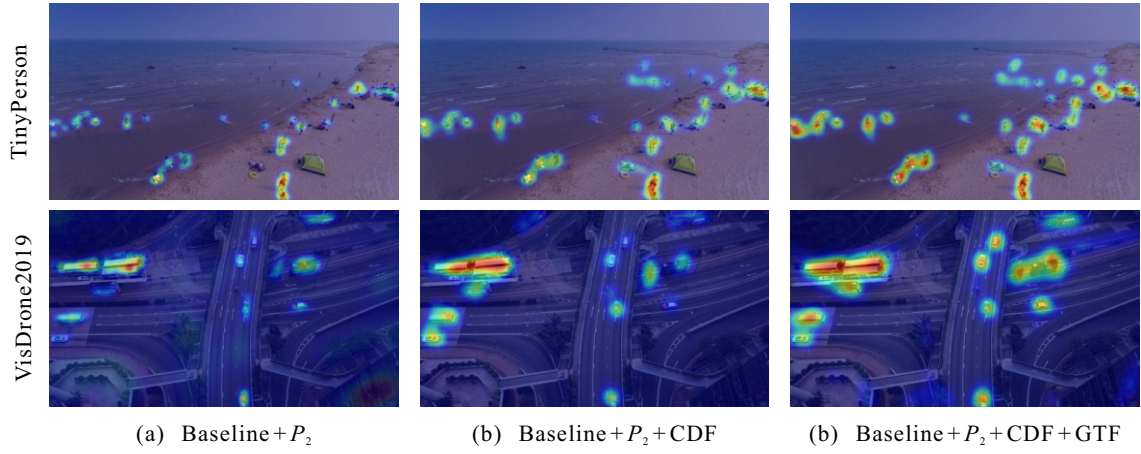


图8 CDF与GTF模块集成前后的热力图对比

表3 不同双流结构融合方式在 VisDrone2019 和 TinyPerson 数据集上的实验结果对比

Methods	double-backbone	IDFC	EEFA	ASFCF	VisDrone2019			TinyPerson			Params/M
					F1	mAP50%	mAP50%~95%	F1	mAP50%	mAP50%~95%	
1)	✓				42.4	38.3	22.9	26.3	18.1	6.19	4.15
2)	✓	✓			42.8	38.9	22.6	27.8	18.2	6.21	3.77
3)	✓	✓	✓		42.9	39.1	22.8	27.8	18.6	6.46	3.52
4)	✓	✓	✓	✓	44.4	40.1	23.5	28.0	19.0	6.61	3.63

表4 不同主干的实验结果对比

Backbone	P	R	F1	mAP50%	mAP50%~95%	Params
FastNet ^[15]	45.3	35.3	39.3	34.9	19.8	5.48
Timm	48.1	36.0	40.6	36.7	21.4	14.68
EfficientVit ^[16]	42.8	31.1	35.3	31.2	17.6	5.28
ConvNeXtV2 ^[17]	44.3	35.2	38.7	34.9	20.0	6.97
RepViT ^[18]	47.2	37.1	41.0	36.9	21.5	8.02
SwinTrans ^[19]	49.8	40.6	44.3	40.0	23.1	31.41
CDF (ours)	50.4	39.6	44.4	40.1	23.5	3.63

表5 EEFA对不同尺度采用 EEFA-T 实验结果分析

Layers	P	R	F1	mAP50%	mAP50%~95%
C_2, C_3	48.9	39.1	43.5	39.8	23.1
C_2, C_3, C_4	49.4	38.9	43.5	39.6	23.3
C_2, C_3, C_4, C_5	48.2	39.0	43.1	39.9	23.5
C_3, C_4 (ours)	50.4	39.6	44.4	40.1	23.5

达方面的有效性。

此外,为进一步验证所提出模型在不同场景下的小目标检测能力,本文在 VisDrone2019 数据集上进行多场景可视化对比分析,如图9所示。在 VisDrone2019 数据集中,选取高空视角、低光照夜景以及复杂背景等典型无人机场景进行对比。结果表明,传统检测模型(YOLOv11、YOLOv12)在高空场景中易出现小目标漏检或边界框偏移;在低光照场景下由于图像对比度低、背景光源复杂,误检和漏检现象更加明显;在复杂交通背景中,则易产生框体重叠不准与部分小目标遗漏。相比之下,所提出模型在

上述场景中表现出更优的检测性能:在高空俯视图景中,模型能够准确定位远距离车辆和行人,检测框与目标位置一致性更强;在夜景条件下,能够在弱纹理区域保持良好的目标识别;在多目标密集场景中,能够有效区分邻近目标,减少框体重叠紊乱和误标注现象,呈现更清晰的检测结果。

2.4.2 模型复杂度分析

从结构复杂度角度分析,传统卷积网络依赖大量卷积核对局部邻域建模,参数规模随通道数和卷积核尺寸呈乘性增长,近似为 $O(k^2C^2L)$ 。其中: k 为卷积核尺寸, C 为通道数, L 为堆叠层数。简单拼接的空间-空间双主干结构通过两个完全相同的卷积主干并行提取特征,整体复杂度提升至 $O(2k^2C^2L)$,导致参数量显著上升(4.15 M)。

为此,本文在双主干结构在中引入 IDFC,通过跨层特征共享和引导传递机制减少辅助分支的重复卷积,使得有效卷积深度由 L 降至 L'_{aux} ($L'_{aux} < L$),整体复杂度降低为 $O(k^2C^2L + k^2C^2L'_{aux})$,参数量降至 3.77 M。然后,将辅分支中的 C3K2 替换为 EEFA,利用 FFT 进行全局频谱表示,其特征调制过程通过频带掩码和门控机制实现,仅引入少量轻量参数,整体复杂度约为 $O(C^2)$ 且不随空间建模深度累积增长,使得整体模型参数量进一步压缩至 3.52 M。

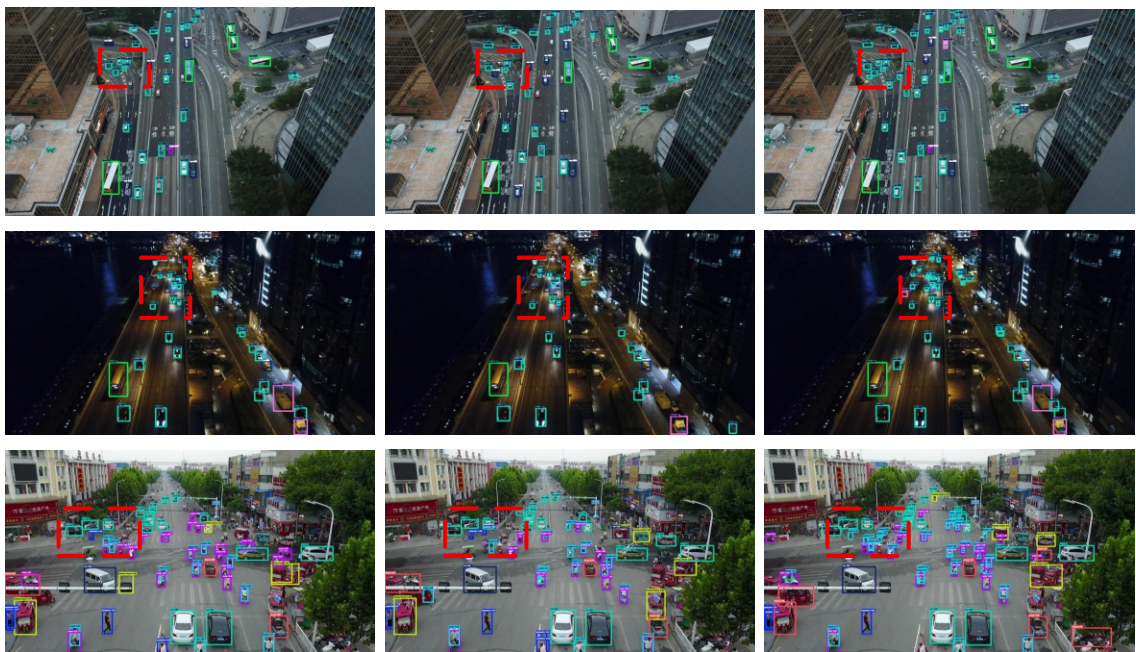
在此基础上,ASFCF 模块通过 1×1 卷积和轻量化注意力实现跨域特征的自适应融合,并在当前阶段恢复通道规模,从结构上抑制通道膨胀,参数仅增

表6 不同模型在 VisDrone2019 数据集上的性能对比

Methods	pedestrian	car	Precision	Recall	F1	mAP50%	mAP50% ~ 95%	Params/M	FLOPs/G
Faster-RCNN ^[4]	—	—	—	—	—	32.9	19.4	41.4	208.0
RetinaNet ^[20]	—	—	—	—	—	33.4	20.7	36.5	210.0
TOOD ^[21]	—	—	45.7	33.4	38.6	34.6	20.3	32.0	199.0
DETR ^[22]	—	—	43.9	32.7	37.5	35.1	18.3	41.7	96.5
OSD-YOLOv10 ^[23]	33.7	75.9	43.9	32.5	37.3	33.4	19.1	1.6	7.9
FFCA-YOLO ^[7]	36.9	73.6	47.8	34.5	40.1	35.1	19.4	7.0	15.8
CSFCANet ^[24]	—	—	46.7	36.6	41.0	35.6	20.5	9.1	35.8
YOLOv8-s ^[25]	43.8	80.2	49.6	39.8	44.2	40.1	23.9	11.2	28.6
YOLOv10-s ^[26]	42.3	79.8	49.8	38.6	43.5	39.4	23.8	7.2	21.4
YOLOv11-n	36.4	76.6	44.2	34.2	38.6	33.9	19.4	2.6	6.3
YOLOv11-s	43.9	80.4	50.3	39.0	43.9	39.8	24.0	9.4	21.3
YOLOv12-n ^[27]	36.1	76.8	44.9	33.9	38.6	33.7	19.3	2.6	6.3
ED-YOLO ^[28]	34.9	77.1	—	—	—	36.6	20.6	1.7	7.2
YOLO-LE ^[29]	44.2	63.5	—	—	—	39.9	22.5	4.0	8.5
Ours	47.8	82.8	50.8	41.2	45.5	42.0	24.5	4.6	9.8

表7 不同模型在 TinyPerson 数据集上的性能对比

Methods	earth person	sea person	Precision	Recall	F1	mAP50%	mAP50% ~ 95%	Params/M	FLOPs/G
YOLOv8-n ^[25]	15.2	13.9	34.6	16.7	22.5	14.6	5.03	3.01	8.1
YOLOv8-s ^[25]	16.4	16.2	35.8	17.5	23.5	16.3	5.76	11.13	28.4
YOLOv10-n ^[26]	14.4	12.7	31.8	16.5	21.7	13.6	4.75	2.27	6.5
YOLOv10-s ^[26]	16.2	15.0	34.1	17.6	23.2	15.6	5.52	7.22	21.4
YOLOv11-n	16.2	14.1	35.6	17.7	23.6	15.1	5.26	2.58	6.3
YOLOv11-s	16.8	15.5	35.8	17.9	23.8	16.1	5.60	9.41	21.3
YOLOv12-n ^[27]	14.7	13.0	34.0	16.5	22.2	13.8	4.71	2.56	6.3
YOLOv12-s ^[27]	17.5	16.0	36.9	18.7	24.8	16.7	5.81	9.23	21.2
FFCA-YOLO ^[7]	19.0	15.6	38.9	18.8	25.3	17.5	5.91	7.0	15.8
Ours	19.1	19.2	35.9	23.4	28.3	19.2	6.63	4.58	9.7



(a) YOLOv12

(b) YOLOv11

(c) ours

图9 不同模型在 VisDrone2019 数据集不同场景下的可视化比较

加至 3.63 M. 在多尺度融合阶段引入 GTF 模块, 用于协调当前层、上采样以及跨层特征. 该模块主要由轻量卷积和注意力构成, 其复杂度为 $O(C^2)$. 最终模型规模增至 4.58 M, 但是, 其增长主要来源于有限的轻量映射和特征调制, 整体复杂度仍然保持在可控范围内.

在推理速度方面, 本文首先在 NVIDIA RTX 4080 Ti (16 GB 显存) 上进行测试, 检测结果显示在 VisDrone2019 和 TinyPerson 数据集上的推理速度分别为 114.3 FPS 和 104.5 FPS. 此外, 本文还在算力较低的 NVIDIA RTX 3080 Ti (12 GB 显存)、NVIDIA RTX 1070 Ti (8 GB 显存) 和 NVIDIA RTX 3060 Ti (8 GB 显存) 上对模型进行了测试, 推理速度均高于 30 FPS^[30] 的实时处理性能标准, 进一步验证了所提出模型在不同硬件平台上的有效性.

3 结论

本文提出了一种基于跨域双流架构的引导式特征融合检测算法, 旨在提升无人机航拍场景下的小目标检测性能. 所提出算法构建了空间-频域双流网络结构, 充分发挥了空间域和频域特征提取的优势, 实现了跨域信息的高效融合和协同增强. 在此基础上, 本文设计了 EEFA 模块, 通过高频增强和分层配置策略, 有效提升了对小目标检测的表征能力; 同时, 提出了 ASFCF 模块, 利用动态加权机制实现了频域与空间域特征的自适应融合, 并在多尺度特征融合阶段引入了 GTF, 实现了不同尺度特征的语义增强和高效协同. 在 VisDrone2019 和 TinyPerson 等典型航拍数据集上的实验结果表明, 所提出方法在小目标检测精度方面取得了稳定且显著的提升, 验证了各模块设计的有效性. 未来工作将进一步优化跨域特征交互结构, 降低计算复杂度, 并探索该方法在旋转目标检测、跨模态感知等任务中的扩展潜力.

参考文献 (References)

- [1] Zhou H P, Yin W, Sun K L, et al. FO-YOLO for small object detection in drone aerial imagery[J]. *The Journal of Supercomputing*, 2025, 81(12): 1208.
- [2] 彭道刚, 邓玉澳, 王丹豪, 等. 面向复杂背景光伏电池红外图像的小目标缺陷检测研究[J]. *控制与决策*, 2026, 41(1): 186-200.
(Peng D G, Deng Y A, Wang D H, et al. Research on small-target defect detection for photovoltaic infrared images under complex backgrounds[J]. *Control and Decision*, 2026, 41(1): 186-200.)
- [3] 高卫峰, 易宇轩, 黄玲玲, 等. 一种高效的无人机航拍小目标检测算法[J]. *控制与决策*, 2025, 40(8): 2525-2533.
(Gao W F, Yi Y X, Huang L L, et al. An efficient algorithm for small object detection in unmanned aerial vehicle images[J]. *Control and Decision*, 2025, 40(8): 2525-2533.)
- [4] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
- [5] Zhou S L, Zhou H J, Qian L. A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images[J]. *Scientific Reports*, 2025, 15: 9255.
- [6] Wu C X, Cai C L, Xiao F, et al. YOLO-LSM: A lightweight UAV target detection algorithm based on shallow and multiscale information learning[J]. *Information*, 2025, 16(5): 393.
- [7] Zhang Y, Ye M, Zhu G Y, et al. FFCA-YOLO for small object detection in remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 1-15.
- [8] 丁浩晗, 贺万程, 万俊, 等. DI-YOLO: 一种面向无人机航拍图像的高效小目标检测框架[J]. *控制与决策*, 2025, 40(10): 3106-3116.
(Ding H H, He W C, Wan J, et al. DI-YOLO: An efficient small object detection framework for UAV aerial imagery[J]. *Control and Decision*, 2025, 40(10): 3106-3116.)
- [9] Chen B, Zou X C, Zhang Y, et al. LEFormer: A hybrid CNN-transformer architecture for accurate lake extraction from remote sensing imagery[C]. ICASSP 2024 — 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, 2024: 5710-5714.
- [10] Zhang H Y, Xiao P, Yao F F, et al. Fusion of multi-scale attention for aerial images small-target detection model based on PARE-YOLO[J]. *Scientific Reports*, 2025, 15(1): 4753.
- [11] Shi Z C, Hu J, Ren J, et al. HS-FPN: High frequency and spatial perception FPN for tiny object detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, 2025: 6896-6904.
- [12] Zhang S Z, Kong D X, Xing Y H, et al. Frequency-guided spatial adaptation for camouflaged object detection[J]. *IEEE Transactions on Multimedia*, 2025, 27: 72-83.
- [13] Du D W, Zhu P F, Wen L Y, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, 2019: 213-226.
- [14] Yu X H, Gong Y Q, Jiang N, et al. Scale match for tiny person detection[C]. IEEE Winter Conference on Applications of Computer Vision. Snowmass Village, 2020: 1257-1265.
- [15] Xiao P, Qin Z, Chen D J, et al. FastNet: A lightweight convolutional neural network for tumors fast identification in mobile-computer-assisted devices[J].

- IEEE Internet of Things Journal, 2023, 10(11): 9878-9891.
- [16] Liu X Y, Peng H W, Zheng N X, et al. EfficientViT: Memory efficient vision transformer with cascaded group attention[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 14420-14430.
- [17] Woo S, Debnath S, Hu R H, et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 16133-16142.
- [18] Wang A, Chen H, Lin Z J, et al. Rep ViT: Revisiting mobile CNN from ViT perspective[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 15909-15920.
- [19] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 10012-10022.
- [20] Wang Y Y, Wang C, Zhang H, et al. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery[J]. *Remote Sensing*, 2019, 11(5): 531.
- [21] Feng C J, Zhong Y J, Gao Y, et al. TOOD: Task-aligned one-stage object detection[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 3490-3499.
- [22] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: Deformable Transformers for end-to-end object detection[J/OL]. 2020, arXiv: 2010.04159.
- [23] Zhang Y, Chen X B, Sun S, et al. Vehicle detection in drone aerial views based on lightweight OSD-YOLOv10[J]. *Scientific Reports*, 2025, 15(1): 25155.
- [24] Li J L, Zheng C H, Chen P, et al. Small object detection in UAV imagery based on channel-spatial fusion cross attention[J]. *Signal, Image and Video Processing*, 2025, 19(4): 302.
- [25] Wang G, Chen Y F, An P, et al. UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios[J]. *Sensors*, 2023, 23(16): 7190.
- [26] Wang A, Chen H, Liu L H, et al. YOLOv10: Real-time end-to-end object detection[J]. *Advances in Neural Information Processing Systems*, 2024, 37: 107984-108011.
- [27] Tian Y J, Ye Q X, Doermann D. YOLOv12: Attention-centric real-time object detectors[J/OL]. 2025, arXiv: 2502.12524.
- [28] Li W Z, Xiao L S, Yao S, et al. ED-YOLO: An object detection algorithm for drone imagery focusing on edge information and small object features[J]. *Multimedia Systems*, 2025, 31(3): 195.
- [29] Chen Z, Zhang Y Y, Xing S H. YOLO-LE: A lightweight and efficient UAV aerial image target detection model[J]. *Computers, Materials & Continua*, 2025, 84(1): 1787-1803.
- [30] Liu S H, Zha J L, Sun J, et al. EdgeYOLO: An edge-real-time object detector[C]. Proceedings of the 42nd Chinese Control Conference. Tianjin, 2023: 7507-7512.

作者简介

陈志旺 (1978-), 男, 副教授, 博士, 主要研究方向为运动物体目标检测与跟踪、多旋翼飞行器导航及控制, E-mail: czwaaron@ysu.edu.cn;

孙艺萱 (2001-), 女, 硕士生, 主要研究方向为计算机视觉中的目标检测, E-mail: 1072814662@qq.com;

彭勇 (1963-), 男, 教授, 博士, 博士生导师, 主要研究方向为生物机器人控制和计算机视觉中的目标检测, E-mail: PY81@sina.com.