DOI：10.12382/bgxb.2025.0480

面向多光谱无人机遥感图像的双流融合跨模态目标检测方法

廖欢[1]，顾成杰[2*]，朱东郡[3]，孟义[2]

(1.安徽理工大学 安全科学与工程学院，安徽 淮南 232001; 2.安徽理工大学 计算机科学与工程学院，安徽 淮南 232001;

3.安徽理工大学 公共安全与应急管理学院，安徽 合肥 231131）

摘要：针对当前无人机多光谱遥感目标检测中存在的模态内特征弱化与跨模态交互低效问题，提出一种融合双流单模态增强与跨模态特征交互的目标检测方法。设计一种基于深度可分离卷积的小波变换模块，通过提取并增强特征图的高频细节信息，强化可见光分支对纹理细节的提取能力。提出一种跨尺度全局信息融合模块，利用多尺度特征整合和全局信息提取，增强红外分支对热物体的检测能力。构建一种多模态混合自注意力融合模块，通过自适应引导可见光与红外模态的特征交互，充分挖掘光谱间互补性。实验结果表明，该方法在可见光-红外多光谱数据集 DVTOD 和 DroneVehicle 上的 mAP@0.50 分别达到 86.7%和 86.0%，较基线模型分别提升 1.3%和 1.9%，有效提升了无人机多光谱目标检测性能。

关键词：无人机；多光谱遥感目标检测；跨模态；小波变换；混合自注意力

中图分类号：TP391.4 **文献标志码：**A **文章编号：**1000-1093(2026)08-250480-XXX

An Object Detection Method for Multispectral UAV Remote Sensing Images

LIAO Huan[1], GU Chengjie[2*], ZHU Dongjun[3], MEI Yi[2]

(1. School of Safety Science and Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China;2. School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China;3. School of Public Safety and Emergency Management, Anhui University of Science and Technology, Hefei 231131, Anhui, China)

Abstract: To address the issues of weakened intra-modal features and inefficient cross-modal interaction in current UAV multispectral remote sensing object detection, an object detection method that integrates dual-stream single-modal enhancement and cross-modal feature interaction is proposed. A wavelet transform module based on depthwise separable convolution is designed to extract and enhance the high-frequency detail information in feature maps, thus strengthening the texture detail extraction capability of the visible light branch. A cross-scale global information fusion module is proposed, which utilizes multi-scale feature integration and global information extraction to enhance the detection capability of the infrared branch for thermal objects. A multimodal hybrid self-attention fusion module is constructed to adaptively guide the feature interaction between visible light and infrared modalities, fully exploiting spectral complementarity. Experimental results show that the proposed method achieves mAP@0.5 of 86.7% and 86.0% on the visible-infrared multispectral datasets DVTOD and DroneVehicle, respectively, and improves them by 1.3% and 1.9%, respectively, compared to the baseline model, effectively enhancing the performance of UAV multispectral object detection.

Keywords: unmanned aerial vehicle; multispectral remote sensing object detection; cross modal; wavelet transform; hybrid self-attention

收稿日期：2025-06-10

基金项目：国家长三角科技创新共同体联合攻关项目（2023CSJGG1100）；国家自然科学基金重大科研仪器研制项目（52227901）；国家重点研发计划项目（2024YFC3013800）；安徽省高等学校科学研究项目（2023AH051197）

*通信作者邮箱：cjgu@aust.edu.cn

0 引言

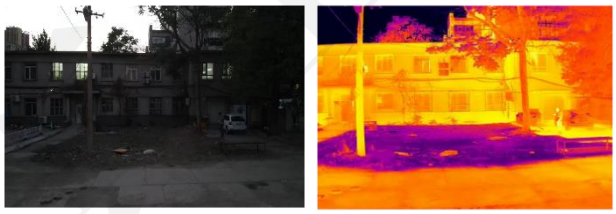
以人工智能技术为核心驱动力的低空经济蓬勃发展，正推动着城市治理、精准巡检与应急响应等领域实现革命性变革[1-5]。搭载各类传感器的无人机凭借高效率、低成本与高灵活性的特质，成为低空数据获取的核心载体，推动感知能力从平面向立体、从人工向智能转型[6]。无人机遥感图像目标检测技术作为智能感知的关键方法，在灾害救援现场勘查、军事侦察、基础设施巡检及公共安全监控等复杂场景中发挥着不可替代的作用[7]。图1展示了该技术在军事侦察与灾害救援场景中的典型应用，清晰呈现了其在复杂现场中的实践价值。然而，该技术的实际效能受限于所依赖的单一数据模态。例如，可见光图像虽然能呈现丰富的纹理与细节信息，成像质量却易受光照变化、云雾遮挡及夜间环境等因素严重干扰。与之互补，红外图像通过感知目标热辐射特征，具备恶劣天气及夜间工作能力，但空间分辨率较低且缺乏充足背景上下文信息，导致目标辨识困难。图2对比展示了可见光图像与红外图像的成像差异，直观呈现了两种单一模态的各自特点与局限性。单一模态的固有局限性，制约了无人机遥感检测在复杂环境下的鲁棒性与检测准确率[8-10]。

(a) 军事侦察　　(b) 灾害救援

(a) Military reconnaissance　　(b) Disaster relief

图 1 无人机应用领域

Fig.1 Application fields of UAVs



(a) 可见光图像　　(b) 红外图像

(a) Visible image　　(b) Infrared image

图 2 多光谱图像对

Fig.2 Multispectral images

近些年来，许多研究学者在多光谱图像目标检测领域取得了一定进展。Zhang等[11]提出了SuperYOLO方法，通过集成简单的压缩激励注意力机制与超分辨率技术有效提升了遥感图像中小目标的检测精度与效率。Chen等[12]设计的双重特征增强YOLO模（Dual-Feature-Enhancement YOLO, DEYOLO），则从通道与空间两个维度对多模态特征进行增强，旨在减少模态间的干扰。然而，上述方法主要依赖卷积神经网络（Convolutional Neural Network, CNN）或局部的注意力机制来构建特征，缺乏对全局上下文信息的有效建模，容易丢失长程语义关联。相比之下，跨模态融合Transformer[13]（Cross-Modality Fusion Transformer, CFT）首次在该领域引入了交叉自注意力机制，推动了基于Transformer架构的融合方法的发展。随后，迭代交叉注意力引导特征融合方法（Iterative Cross-Attention Guided Feature Fusion, ICAFusion[14]）通过设计参数共享、多次迭代的交叉注意力机制，进一步强化了对跨模态互补信息的挖掘。代小波等[15]引入了交叉注意力机制，并结合对齐模块以提升模态特征融合效果。广义多光谱检测 Transformer[16]（Generalized Multispectral Detection Transformer, GM-DETR）则通过交叉注意力及单模态增强，缓解了模态差异与特征不对齐问题。另外，跨模态对齐检测模型[17]（Cross-Modal Alignment Detector, CMADet）在交叉注意力机制的基础上引入了图像配准技术，从而提升了模型在空间对齐前提下的特征融合表现。但上述基于自注意力的方法均未考虑到在网络的深层，由于通道数的增加，红外模态的热辐射特征通常比可见光模态的纹理特征更为丰富和显著。此时，若直接以信息相对较弱的可见光特征作为查询，来检索红外特征作为键和值，则可能引入冗余甚至噪声，未能充分利用红外模态的主导信息。此外，这些方法大多侧重于模态间的交互，忽视了对模态内固有特征的增强。在跨模态交互之前，对每个独立模态进行增强，以提炼其最具判别性的特征且交互前的模态内增强对提升检测性能至关重要。

因此，本文提出一种多光谱目标检测方法（Unmanned Aerial Vehicle Cross-Modal Network, UAVCM_Net），通过模态特征强化与跨模态互补机制，旨在提升复杂场景下的检测性能。本文的主要贡献包括：

1)构建基于深度可分离卷积的小波变换模块，增强可见光分支对高频纹理细节的表征能力。

2)设计跨尺度全局信息融合模块，结合不同阶段尺度特征提取与全局上下文建模，提升红外分支对热目标的检测性能。

3)引入跨模态混合自注意力机制，通过空间权重分配实现双模态特征的动态语义对齐与互补性融合。

1 总体多光谱模型设计

本文基于单模态YOLOv8基线模型，提出一种可见光与红外光谱双分支并行特征提取的多光谱增强检测网络，如图3所示。

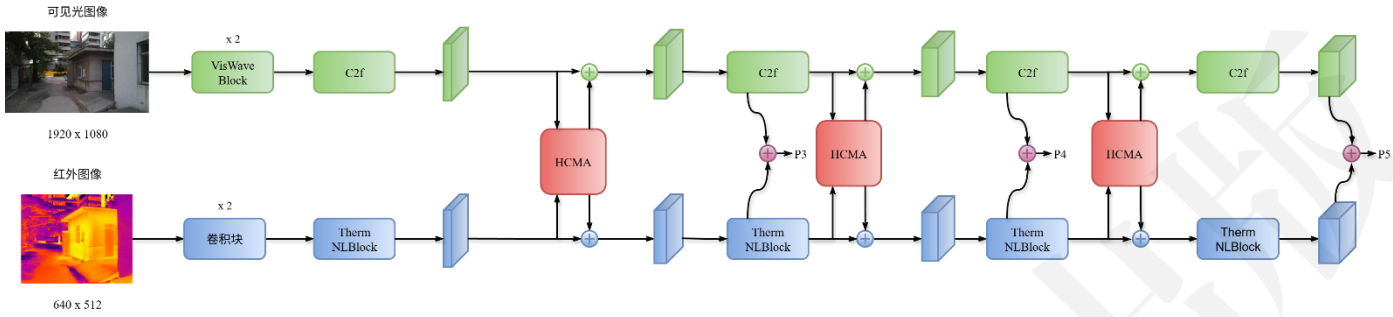


图 3 UAVCM_Net 网络结构

Fig.3 UAVCM_Net network architecture

UAVCM_Net采用可见光与红外双分支并行的异构架构。预处理阶段将双模态数据统一缩放至360×640×3分辨率，以保障尺度一致性。可见光分支网络前两层用深度可分离小波变换卷积(VisWaveBlock)模块替代原始卷积层以增强高频细节提取，红外分支主干的4个快速双卷积CSP瓶颈模块（Faster CSP Bottleneck with 2 convolutions，C2f）模块替换为(ThermNLBlock)单元，强化模态内的跨尺度上下文信息。针对小尺寸(P3)、中尺寸(P4)、大尺寸(P5)目标检测需求，在3个检测头层级嵌入混合跨模态自注意力机制(Hybrid Cross Modal Attention，HCMA)，通过空间相关性建模实现跨模态特征互补，并将优化特征反馈至下流各分支。其余模块继承YOLOv8基线架构。

1.1 基于深度可分离卷积的小波变换模块

在可见光分支的特征提取过程中，传统的卷积操作往往难以有效捕捉图像中的高频纹理细节，而这些细节正是区分复杂背景下目标与干扰物的关键特征。为了克服这一局限，引入小波变换技术。该技术通过多分辨率分析将信号分解为不同频率分量，能够精准分离并提取承载纹理信息的高频成分，从而有效弥补传统卷积在细节提取方面的短板。

小波变换[18]是一种基于多分辨率分析的信号处理技术，能够将信号分解为不同频率的分量，以分析其频域信息。小波变换卷积融合了小波变换的多尺度分析能力与卷积操作的特征提取优势，通过将输入信号分解为高频和低频分量，并分别提取其细节特征与主体特征进行融合，从而实现高效的多尺度特征建模。该方法基于二维Haar小波变换实现，如图4所示。图4中，*WT*为小波变换操作，X为输入特征图，$X_{LL}^{(i)}$、$X_{LL}^{(i)}$，$X_{LH}^{(i)}$、$X_{LH}^{(i)}$，$X_{HL}^{(i)}$、$X_{HL}^{(i)}$，$X_{HH}^{(i)}$、$X_{HH}^{(i)}$分别为第i层小波分解后的低频近似子带、垂直细节子带、水平细节子带、和对角细节子带特征图，其中i=1,2表示分解层级。

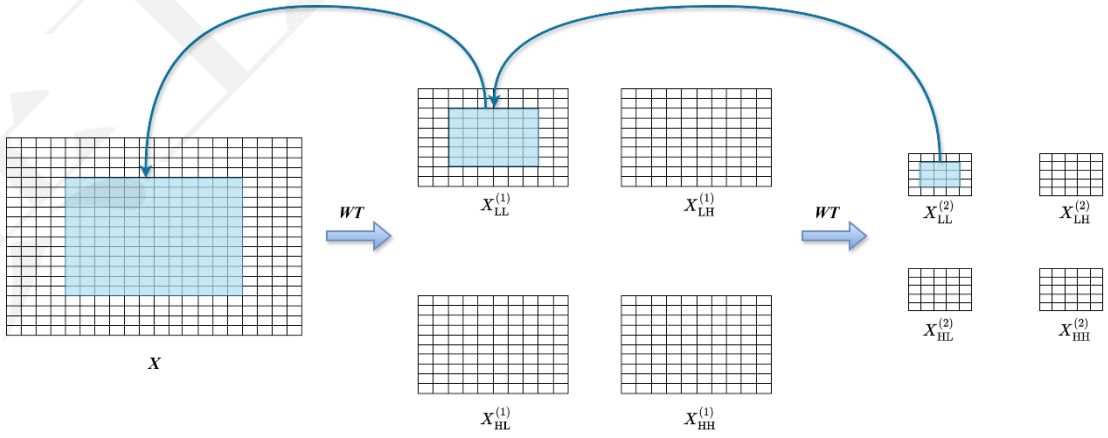


图 4 WTConv 工作原理

Fig.4 Operating principle of WTConv

WTConv工作的核心流程为首先，输入图像经由式（1）~式（4）定义的正交滤波器组进行分解，生成包含轮廓特征的低频子带以及包含细节信息的高频子带。

$$ff_{LL}=\frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} \tag{1}$$

$$ff_{\mathrm{LH}}=\frac{1}{2}\begin{pmatrix}1 & -1\\ 1 & -1\end{pmatrix} \tag{2}$$

$$ff_{\mathrm{HL}}=\frac{1}{2}\begin{pmatrix}1 & 1\\ -1 & -1\end{pmatrix} \tag{3}$$

$$ff_{\mathrm{HH}}=\frac{1}{2}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix} \tag{4}$$

式中：ff_{HH}、ff_{LH} 和 ff_{HL} 为高频子带；f_{LL} 为低频子带。

随后，各子带经由独立的卷积层进行特征增强，并通过融合得到复合特征F_{fused}，其过程如式(5)所示。

$$F_{\mathrm{fused}}=Concat\begin{pmatrix}Conv(X_{\mathrm{LL}}), & Conv(X_{\mathrm{LH}}),\\ Conv(X_{\mathrm{HL}}), & Conv(X_{\mathrm{HH}})\end{pmatrix} \tag{5}$$

式中：*Conv*表示标准普通卷积算子；*Concat*表示通道拼接操作。在此基础上，采用式(6)所示的单次卷积策略对低频子带进行多级分解，在保持多尺度特征表达能力的同时，显著降低计算复杂度。

$$X_{\mathrm{LL}}^{(i)},X_{\mathrm{LH}}^{(i)},X_{\mathrm{HL}}^{(i)},X_{\mathrm{HH}}^{(i)}=WT\left(X_{\mathrm{LL}}^{(i-1)}\right) \tag{6}$$

式中：*i*表示当前层级，*i*=1,2。最后利用式(7)所定义的转置卷积*ConvTransposed*对融合特征进行上采样，恢复至原始图像的空间分辨率。

$$X_{\mathrm{recon}}=ConvTransposed\begin{pmatrix}[X_{\mathrm{LL}},X_{\mathrm{LH}},X_{\mathrm{HL}},X_{\mathrm{HH}}],\\ F_{\mathrm{fused}}\end{pmatrix} \tag{7}$$

式中：X_{recon} 为经过转置卷积上采样操作后，空间分辨率恢复至与原始输入图像一致的输出特征；f_{LL}、f_{LH}、f_{HL}、f_{HH} 为小波分解后对应的低频近似子带特征、水平方向高频细节子带特征、垂直方向高频细节子带特征、对角方向高频细节子带特征；F_{fused} 为四个子带分别进行卷积特征提取后，沿通道维度拼接得到的融合特征。

与以往对小波分解后所有子带进行无差别处理的方法不同，针对可见光图像的高频细节提取需求，本文创新性地提出了 VisWaveBlock 模块，其结构如图 5 所示。

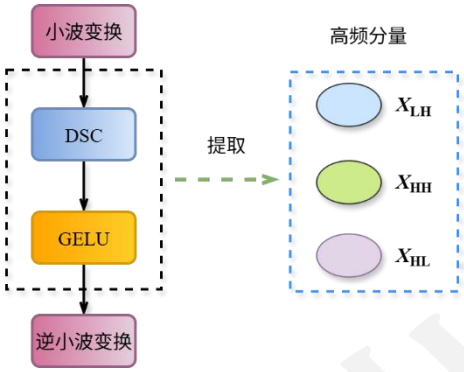


图 5　VisWaveBlock 网络结构

Fig.5 VisWaveBlock network architecture

该模块仅聚焦于承载主要纹理与边缘信息的 3 个高频子带（X_{LH}、X_{HL} 和 X_{HH}），通过直接处理这些纹理源头以规避对低频近似子带的冗余计算。在此基础上，模块引入深度可分离卷积（Depthwise Separable Convolution，DSC）替代传统卷积进行特征提取：DSC 通过对每个输入通道独立进行空间滤波，有效抑制了通道间的相互干扰，在减少参数量的同时，实现了更为纯净和专注的高频信息提取。为了进一步增强对复杂且细微纹理的表征能力，该模块采用高斯误差线性单元（Gaussian Error Linear Unit, GELU）激活函数替代传统的 ReLU。GELU 函数的平滑非线性特性有助于保留更多微弱的梯度响应，从而提升模型对细节特征的捕捉能力与训练过程的稳定性。最终，经过处理的高频特征通过逆小波变换重构至空间域，完成从频域到卷积域的信息融合。此外，基于 CNN 浅层特征富含丰富纹理与边缘先验的普遍认知，本文策略性地将 VisWaveBlock 模块仅嵌入可见光 Backbone 网络的前两层。这一设计使其在纹理信息最密集的层级上完成细节增强与提纯。该策略不仅显著提升了可见光单模态特征的细粒度表征质量，也为后续的跨模态融合任务提供了信息更完整、细节更突出的特征输入。

1.2　跨尺度全局信息融合模块

在红外分支的特征提取中，核心挑战在于：随着网络深度的增加，通道数量增多，特征信息愈发丰富，但仅增加网络深度会带来显著的计算负担，且无法实现全局上下文信息的有效增强。为了解决此问题，全局上下文网络(Global Context Network，GCNet[19])作为一种轻量级全局上下文建模技术被引入。该技术能够在较低的计算开销下实现长程依赖建模，进而建立热目标与全局背景的关联，以强化热特征的表征能力。

GCNet通过整合非局部网络的长程依赖建

模优势和压缩激励网络的通道注意力机制，提出全局上下文模块作为基础单元，实现了精准特征提取与计算效率的平衡。其执行流程主要由全局上下文建模与特征转换两部分构成，如图6的左半部分所示。在全局上下文建模阶段，输入特征通过双路径处理，主路径将特征展平为$C\times H\times W$维度，辅助路径则通过1×1卷积生成$HW\times1\times1$的注意力权重系数矩阵（其中，C为通道数，H为高度，W为宽度）。随后，辅助路径的权重系数经过Softmax归一化，与主路径的展平特征进行加权求和，生成全局上下文特征。其中，权重系数α_j的计算公式为

$$\alpha_j=\frac{exp(W_k x_j)}{\sum_{m=1}^{N_p}exp(W_k x_m)} \tag{8}$$

式中：W_k为线性变换矩阵，k为键特征变换标识，代表该矩阵用于生成注意力计算所需的键特征；x_j为输入特征图中第j个空间位置对应的特征向量，j为为特征图的空间位置索引；m为Softmax归一化项的求和遍历索引，用于遍历全部空间位置完成权重归一化计算；N_p为$H\times W$空间位置总数。全局上下文特征y的计算公式为

$$y=\sum_{j=1}^{N_p}\alpha_j x_j \tag{9}$$

在特征转换阶段引入通道压缩机制，全局上下文特征y通过1×1卷积进行通道维度压缩，压缩比为r，得到中间特征。随后，该特征经激活函数处理并再次通过1×1卷积恢复至原始通道维度，最终生成增强特征，其具体计算公式为

$$z_i=x_i+W_{v2}\cdot ReLU(LN(W_{v1}\cdot y)) \tag{10}$$

式中：z_i为经过上下文建模增强后的特征；W_{v1}和W_{v2}为可学习的1×1卷积核；LN表示层归一化。最终通过残差连接将增强后的全局特征与原始输入通过广播机制融合，既维持了特征维度的一致性，又实现了高效的全局上下文建模。

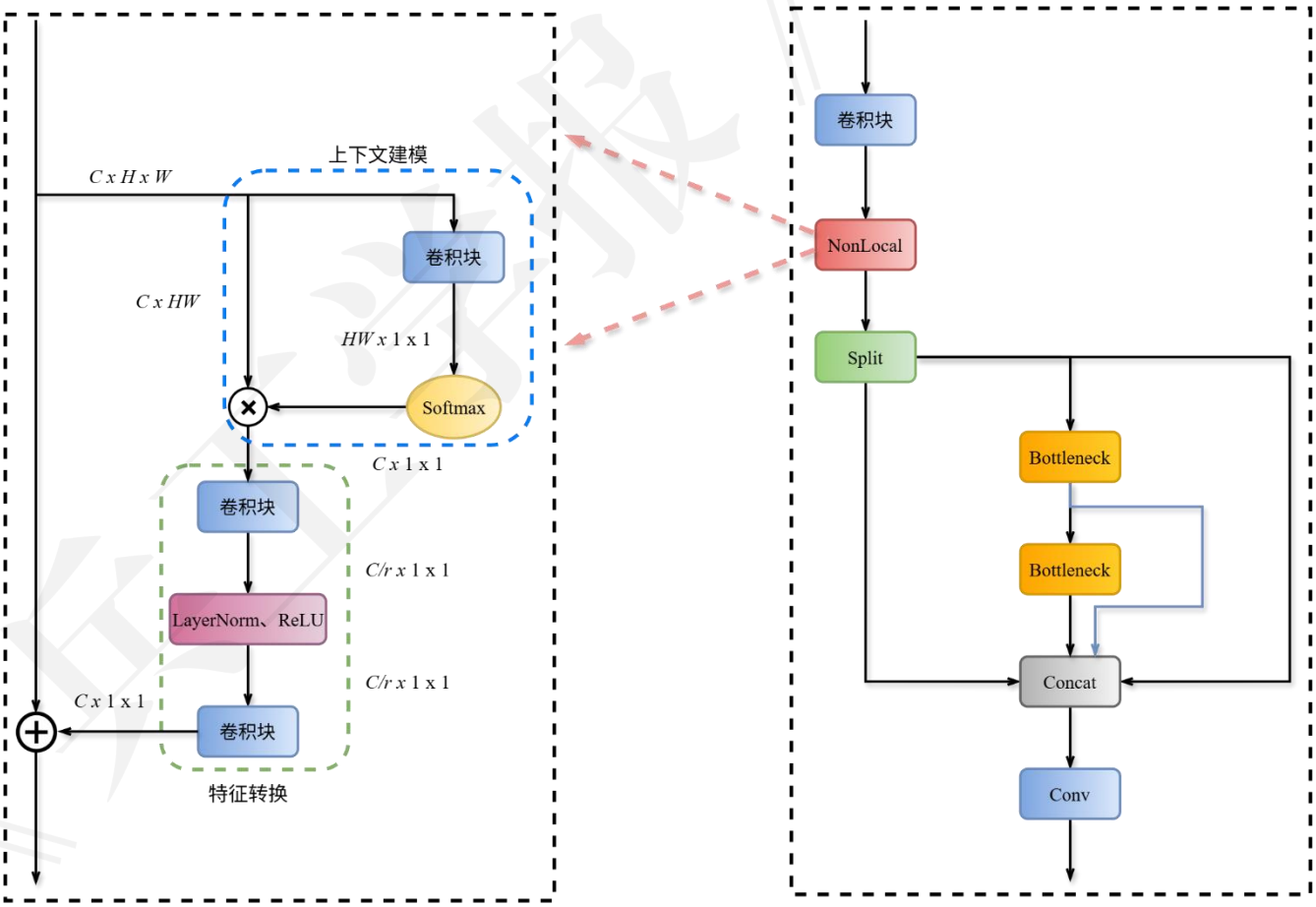


图 6 ThermNLBlock 网络结构

Fig.6 ThermNLBlock network architecture

在构建 ThermNLBlock 模块时，将 YOLOv8 的跨阶段特征融合结构 C2f 与轻量化的 GC 模块相集成，其结构如图 6 的右部分所示。该模块的设计旨在针对不同特征尺度阶段实现热特征增强，因此被依次部署于红外分支的主干网络中。具体而言，在四层原 C2f 处均部署了此模块，以确保覆盖并处理多尺

度的特征信息。在模块内部，C2f 结构承担着在当前阶段完成跨感受野特征融合的任务，而 GC 模块则专注于对当前阶段特征进行全局上下文建模，其通过通道注意力机制实现对特征的自适应加权。此种设计使得网络能够在从浅层到深层的不同尺度阶段，构建热目标与背景的差异化表征。在每个特征层级上，该结构均能实现热特征的全局增强，从而有效缓解红外图像浅层热特征表征能力弱的问题。最终，该方案实现了全尺度的热特征增强，为后续的跨模态融合任务提供了判别性更强的红外特征。

1.3 跨模态混合自注意力融合模块

Transformer 架构通过自注意力机制构建了数据元素间的动态关联网络，显著提升了模型对跨模态长程信息依赖的建模能力。在视觉任务中，该机制突破了传统卷积操作的感受野限制，通过对全局特征空间进行并行化语义关联，有效捕获了图像内容的整体表征规律。研究结果表明，这种特性在多模态特征融合检测中展现出显著优势，尤其在建立跨模态语义一致性方面具有理论先进性。然而，现有基于 Transformer 的多模态融合方法普遍采用串行特征级联策略，这种单向传递机制仅能实现浅层特征交互，导致模态间潜在的互补信息未被充分挖掘。更值得注意的是，简单的张量拼接操作会引入高维冗余特征，不仅使注意力权值矩阵的计算复杂度呈平方级增长，还易引发模态特征相互干扰，最终造成检测模型在复杂场景下的泛化性能衰减。因此，本文在传统自注意力机制的基础上，提出了一种多尺度跨模态混合自注意力操作 HybridCrossModalAttention，如图 7 所示。

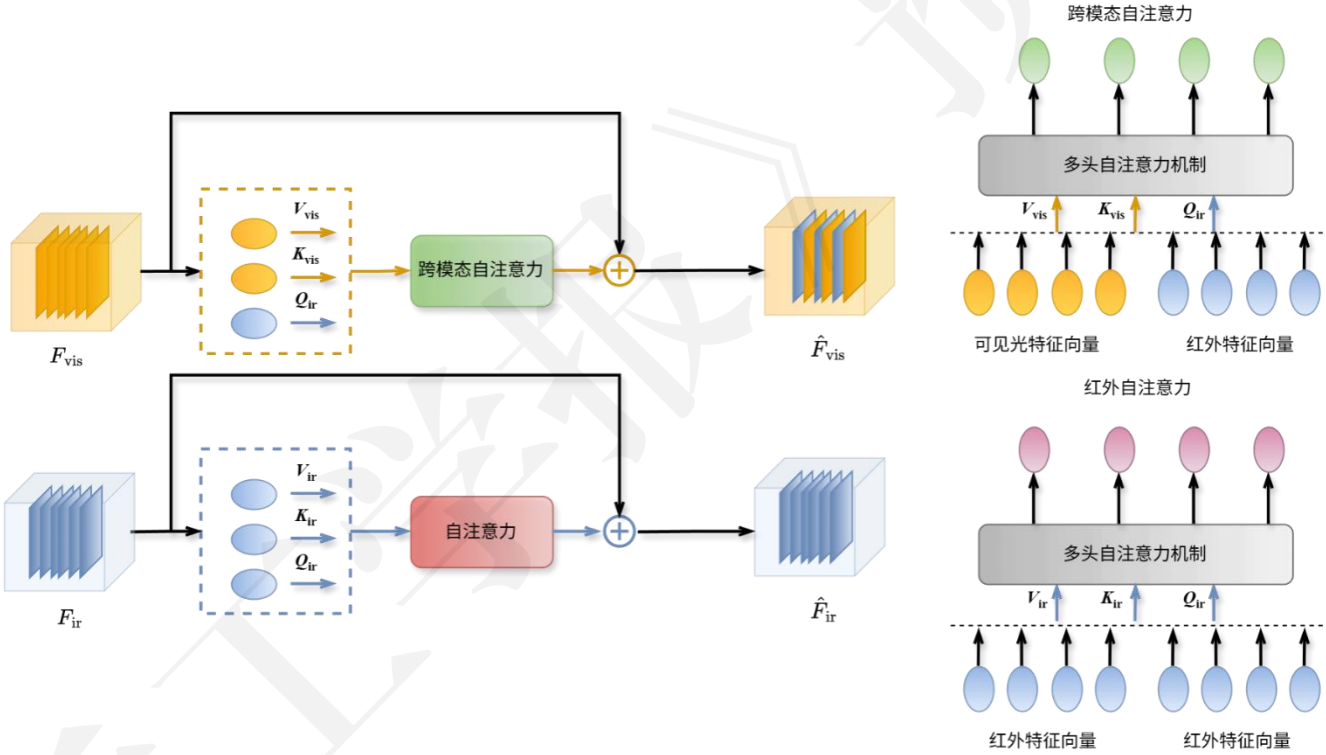


图 7 Hybrid cross modal attention 网络结构

Fig.7 Hybrid cross modal attention network architecture

该方法通过构建跨模态信息交互机制，在多尺度特征空间中实现模态间的信息融合。具体而言，通过设计跨尺度的注意力权重计算模块，使不同模态特征在 3 个空间尺度上进行信息交互，从而实现跨模态特征的深度融合。

在 HCMA 混合自注意力机制中，跨模态特征投影实现流程如下：首先，根据式(11)和式(12)对可见光与红外模态的特征令牌进行线性变换，分别生成对应的查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$ 三元组，

$$\begin{aligned} \boldsymbol{Q}_{\text{vis}} &= \boldsymbol{X}_{\text{vis}} \cdot \boldsymbol{W}_{\boldsymbol{Q}}^{\text{vis}} \\ \boldsymbol{K}_{\text{vis}} &= \boldsymbol{X}_{\text{vis}} \cdot \boldsymbol{W}_{\boldsymbol{K}}^{\text{vis}} \\ \boldsymbol{V}_{\text{vis}} &= \boldsymbol{X}_{\text{vis}} \cdot \boldsymbol{W}_{\boldsymbol{V}}^{\text{vis}} \end{aligned} \tag{11}$$

式中：$\boldsymbol{Q}_{\text{vis}}$、$\boldsymbol{K}_{\text{vis}}$、$\boldsymbol{V}_{\text{vis}}$ 分别为可见光模态对应的查询向量、键向量、值向量，vis 表示可见光模态特征；$\boldsymbol{X}_{\text{vis}}$ 为可见光模态的输入特征令牌序列；$\boldsymbol{W}_{\boldsymbol{Q}}^{\text{vis}}$、$\boldsymbol{W}_{\boldsymbol{K}}^{\text{vis}}$、$\boldsymbol{W}_{\boldsymbol{V}}^{\text{vis}}$ 分别为可见光模态下查询、键、值对应的可学习线性变换权重矩阵。

随后，通过参数矩阵$\{W_{\boldsymbol{Q}}, W_{\boldsymbol{K}}, W_{\boldsymbol{V}}\}$构建模态间可学习的特征映射关系。该过程在保持特征空间一致性的前提下，为不同模态建立了参数共享的投影机制，有效增强了跨模态特征的拓扑对齐能力。

$$\boldsymbol{Q}_{\text{ir}} = \boldsymbol{X}_{\text{ir}} \cdot \boldsymbol{W}_{\boldsymbol{Q}}^{\text{ir}}$$

$$\boldsymbol{K}_{\mathrm{ir}}=\boldsymbol{X}_{\mathrm{ir}}\cdot\boldsymbol{W}_{\boldsymbol{K}}^{\mathrm{ir}} \tag{12}$$
$$\boldsymbol{V}_{\mathrm{ir}}=\boldsymbol{X}_{\mathrm{ir}}\cdot\boldsymbol{W}_{\boldsymbol{V}}^{\mathrm{ir}}$$

式中：$\boldsymbol{Q}_{\mathrm{ir}}$、$\boldsymbol{K}_{\mathrm{ir}}$、$\boldsymbol{V}_{\mathrm{ir}}$ 分别为红外模态对应的查询向量、键向量、值向量，ir 表示红外模态特征；$\boldsymbol{X}_{\mathrm{ir}}$ 为红外模态的输入特征令牌序列；$\boldsymbol{W}_{\boldsymbol{Q}}^{\mathrm{ir}}$、$\boldsymbol{W}_{\boldsymbol{K}}^{\mathrm{ir}}$、$\boldsymbol{W}_{\boldsymbol{V}}^{\mathrm{ir}}$ 分别为红外模态下查询、键、值对应的可学习线性变换权重矩阵。

针对 DVTOD 数据集可见光与红外模态未对齐且真实标签标注在红外图像的特点，受跨模态对齐检测(Cross-Modal Alignment Detector，CMA-Det)网络模型的启发，本文创新性地引入了混合跨模态注意力机制对可见光与红外模态语义特征难以对齐等问题进行处理。首先，在红外模态特征上应用自注意力机制，有效挖掘模态内部的高效特征。其次，在跨模态交互中，设计了基于红外查询 $\boldsymbol{Q}_{\mathrm{ir}}$ 与可见光键值对 $\boldsymbol{K}_{\mathrm{vis}}$、$\boldsymbol{V}_{\mathrm{vis}}$ 的跨模态自注意力操作。通过这种机制，红外查询能够从语义层面主动搜索可见光图像中语义匹配的区域，从而实现跨模态信息的互补与语义对齐。具体公式如式(13)和式(14)所示。

$$A_{\mathrm{ir}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{ir}}\cdot\boldsymbol{K}_{\mathrm{ir}}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{k}}}}\right)\cdot\boldsymbol{V}_{\mathrm{ir}} \tag{13}$$

$$A_{\mathrm{hybrid}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{ir}}\cdot\boldsymbol{K}_{\mathrm{vis}}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{k}}}}\right)\cdot\boldsymbol{V}_{\mathrm{vis}} \tag{14}$$

式中：A_{ir} 和 A_{hybrid} 分别表示通过全局红外自注意力和跨模态自注意力机制精炼后的特征；$d_{\boldsymbol{k}}$ 为键向量的维度，$\boldsymbol{k}$ 为键向量特征。其中，Softmax 函数通过对注意力分数进行归一化，将跨模态特征相似度映射为概率分布，从而实现对重要语义区域的动态增强。

为了进一步增强多模态特征，本文构建基于混合注意力机制的残差融合方法。如式(15)和式(16)所示。输入特征 F_{vis} 和 F_{ir} 通过跨层残差连接与自注意力、跨模态注意力分支融合，抑制深层网络的信息衰减，其输出特征 $\hat{F}_{\mathrm{vis}}$ 和 $\hat{F}_{\mathrm{ir}}$ 计算过程可表示为

$$\hat{F}_{\mathrm{ir}}=F_{\mathrm{ir}}+A_{\mathrm{ir}} \tag{15}$$

$$\hat{F}_{\mathrm{vis}}=F_{\mathrm{vis}}+A_{\mathrm{hybrid}} \tag{16}$$

最后，增强特征通过跨分支反馈循环机制重注入各模态通路，为后续层次化多模态融合提供语义一致性保障。

2　实验设置与参数评估

2.1　实验数据集

DVTOD 是由东北大学机械工程与自动化学院在 2024 年研发的无人机多光谱目标检测基准数据集。该数据集包含 1 606 对训练样本和 573 对测试样本，覆盖 16 种复杂属性和 54 个采集场景。数据通过大疆 MAVIC 2 Enterprise Advanced 无人机采集，其中可见光相机分辨率 1 920×1 080（48MP CMOS 传感器，f/2.8 光圈），热成像相机分辨率 640×512（8~14 μm 波长）。数据集囊括昼夜、四季场景，模拟真实世界的光照剧变（过曝/暗光）、极端天气、特殊材质遮挡等挑战性条件，如图 8 所示。

(c) 树荫遮挡　　(d) 伞影

(c) Tree shade occlusion　　(d) Umbrella shade

图 8　DVTOD 数据集样本

Fig.8 Sample of the DVTOD dataset

另外，真实标签标注采用双模态互补策略，常规场景标注热成像模态，特殊场景标注可见光模态，总体上以红外为主，通过 LabelImg 工具生成 PASCAL VOC 格式的共享标注框。DVTOD 拥有更丰富的目标类别，例如行人、车和自行车，并首次系统性包含全天候、跨季节的多光谱未对齐图像对，为无人机多光谱目标检测方法在复杂场景下的模态互补性与鲁棒性研究提供高价值实验基准。

DroneVehicle[20]是一个专为可见光与红外双模态车辆检测设计的大规模基准数据集。该数据集总计包含 28 439 组严格配对的 RGB 与红外图像，全部由无人机平台采集其图像覆盖了从白天到夜晚的全天候时段、多样化的光照条件、不同的飞行高度与视角，以及城市道路、停车场、居民区等复杂场景。该数据集面临的主要挑战包括：无人机俯拍视角下的车辆目标尺度普遍较小以及可见光模态在深夜等低光照条件下成像质量显著下降；这些特性使得 DroneVehicle 成为评估各类互补数据融合模型鲁棒性与性能的关键平台。依据官方标准划分，本文采用 17 990 对图像进行训练，1 469 对用于验证。

2.2 实验环境与参数配置

本文实验平台的软件配置涵盖 Python 3.10.16 解释器与 PyTorch 深度学习框架版本 2.1.1，搭配 CUDA 加速库 11.8 实现 GPU 并行计算。硬件系统采用 2 张 NVIDIA RTX4090 24GB 系列显卡作为主要计算单元，配合 AMD EPYC 7643 48-Core 处理器完成辅助运算，

详细情况如表 1 所示。

表 1　实验软硬件环境

Table 1 Experimental software and hardware environment

实验环境	型号
编译环境	Python 3.10.16
深度学习框架	Pytorch 2.1.1
CUDA	11.8
GPU	NVIDIA RTX4090 24GB
CPU	AMD EPYC 7643 48-Core
操作系统	Ubuntu 20.04.6 LTS

在模型超参数配置策略方面，本实验采用 300 轮训练周期，设置 batch_size 为 16，使用随机梯度下降（Stochastic Gradient Descent，SGD）优化器（动量 0.937），使用原 YOLOv8s 的预训练权重，并结合 Mosaic4 数据增强方法。初始学习率设为 0.01，最终学习率降至 0.000 1。

在目标检测任务中，评估模型性能的关键指标包括精确度（Precision, P）、召回率（Recall, R）和平均精度均值（Mean Average Precision, mAP）。这些指标通过以下公式计算。

1）精确度 P。衡量模型正确预测的正样本占所有预测正样本的比例，计算公式为

$$P=\frac{TP}{TP+FP} \tag{17}$$

式中：TP表示正确预测的正样本数；FP表示错误预测为正的负样本数。

2）召回率 R。衡量模型正确预测的正样本数占所有实际正样本的比例计算公式为

$$R=\frac{TP}{TP+FN} \tag{18}$$

式中：FN表示错误预测为负的正样本数。

3）平均精度均值 mAP。作为综合评估指标，mAP 表示所有类别精度-召回率曲线下的平均面积比例。其计算公式为

$$AP_i=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 P(r) \tag{19}$$

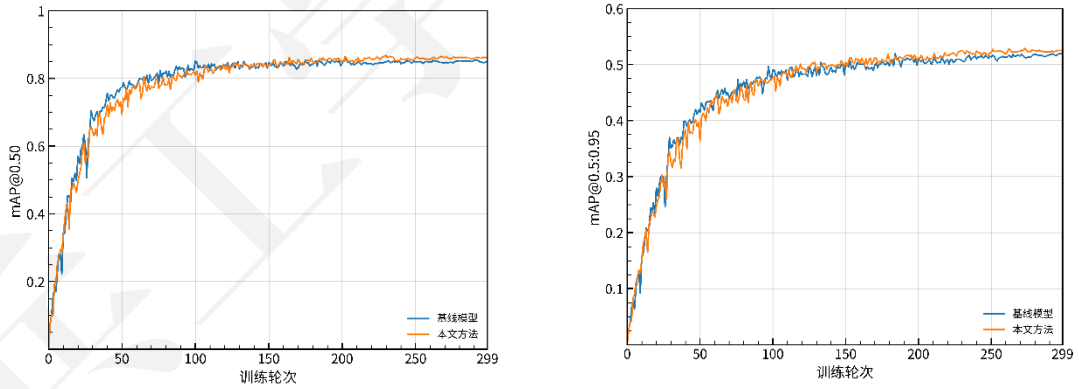
$$mAP=\frac{1}{n}\sum_{i=1}^{n}AP_i \tag{20}$$

式中：AP_i表示第i类的平均精度；n为目标类别总数量；P（r）为精度-召回函数。此外，mAP@0.50表示交并比(Intersection-over-Union，IoU)为0.5时的平均检测精度，而mAP@0.50:0.95则表示在IoU范围为0.50~0.95时的平均检测精度，后者对模型性能要求更高。

3　实验结果与可视化分析

3.1　实验训练过程分析

为了验证本文提出的 UAVCM_Net 方法性能，图 9 展示了其与基线模型 Two-Stream-YOLOv8 在 300 训练轮次中 mAP@0.50 与 mAP@0.50:0.95 的对比曲线。实验训练结果表明，UAVCM_Net 在两项指标上均展现出显著优势，并最终趋于收敛。



(a) mAP@0.50训练曲线
(a) Training curves of mAP@0.5

(b) mAP@0.50:0.95训练曲线
(b) Training curves of mAP@0.50:0.95

图 9 mAP 训练曲线对比

Fig.9 Comparison of mAP training curves

3.2　消融实验结果分析

为了验证本文所设计网络模块的有效性及其对目标检测性能的提升作用，在 DVTOD 多光谱数据集上进行系统的消融实验。实验以 Two-Stream-YOLOv8 为基础模型，通过逐步引入和组合不同的改进模块，系统地评估各模块对模型性能的贡献。实验中，本文选取了多个关键性能指标，例如 P、R、mAP、Param 以及不同 IoU 阈值下的识别精度，对模型的检测性能进行全面比较，通过这些实验，旨在验证各改进模块的有效性及其对模型整体性能的提升作用。此外，本文以双流基线模型为基础开展消融实验。其中，实验 A 为未引入任何改进模块的双流基

线模型，作为所有消融实验的性能基准；实验结果如 表 2 所示。

表 2 消融实验

Table 2 Ablation experiment

方法	VisWaveBlock	ThermNLBlock	HCMA	mAP@0.50/%	mAP@0.50:0.95/%	Param/10^6
A	×	×	×	85.4	52.0	16.2
B	✓	×	×	85.8	53.0	16.1
C	×	✓	×	85.7	51.9	16.3
D	×	×	✓	86.0	52.9	49.4
E	✓	✓	×	85.9	52.3	16.2
F	✓	✓	✓	86.7	52.5	49.4

注：×表示未加入对应模块，√表示加入对应模块。

1)实验 B：在实验 A 的基础上引入 VisWaveBlock 模块，该模块通过深度可分离小波变换机制，对可见光高频分量进行分解与重构。结果表明，mAP@0.50 从 85.4%提升至 85.8%，mAP@0.50:0.95 从 52.0%提高至 53.0%，分别实现了 0.4%和 1.0%的增益。这一改进验证了深度可分离小波变换对可见光高频细节特征的增强作用，其频率域分解特性有效提升了局部纹理信息的特征表示能力。

2）实验 C：在实验 A 的基础上引入了 ThermNLBlock 模块。该模块通过跨尺度全局信息融合机制，强化了红外热目标的特征表达，并提升了模型的多尺度特征整合能力。实验结果表明，mAP@0.50 由 85.4%提高至 85.7%。这一结果验证了 ThermNLBlock 模块在增强红外特征的跨尺度关联与语义表征方面的有效性。

3）实验 D：在实验 A 的基础上引入了 HCMA 模块，以构建可见光与红外模态间的信息交互机制。该模块在多尺度特征空间中设计了跨尺度注意力权重计算方式，促使双模态特征在 3 个空间尺度上进行深度交互，从而实现跨模态特征的充分融合。实验结果显示，mAP@0.50 从 85.4%提升至 86.0%，提高了 0.6%。这一提升验证了 HCMA 模块在跨模态特征优化融合中的有效性。

4）实验 E 和 F：实验 E 同时引入 VisWaveBlock 与 ThermNLBlock 模块，其 mAP@0.50 高于仅引入单一模块的实验 B 和实验 C。进一步，实验 F 融合了 VisWaveBlock、ThermNLBlock 及 HCMA 这 3 个模块，取得了所有实验中最高的 mAP@0.50。相比基线实验 A，其 mAP@0.50 分别提升了 1.3%与 0.5%，充分验证了模块协同的有效性。具体而言，VisWaveBlock 与 ThermNLBlock 分别对可见光与红外模态进行模态内特征增强，而 HCMA 则通过多尺度跨模态注意力实现模态间深度交互。三者构成互补协同机制，既体现了各模块的必要性，也证明了整体融合策略的合理性。此外，为了验证基于深度可分离的小波变换卷积模块在不同分支的有效性，本文在基线模型上分别对可见光与红外分支的前两层进行模块替换消融实验，实验结果如表 3 所示。

表 3 VisWaveBlock 消融实验

Table 3 VisWaveBlock ablation experiment

方法	可见光分支	红外分支	*P*/%	*R*/%	*mAP*@0.50/%	*mAP*@0.50:0.95/%
A	×	×	89.3	77.6	85.4	52.0
B	✓	×	86.7	80.9	85.8	53.0
C	×	✓	88.9	79.0	85.2	52.0

注：×表示未加入对应模块，√表示加入对应模块。

实验结果表明，基于深度可分离小波变换模块在可见光分支上在 mAP@0.50 和 mAP@0.50:0.95 指标上分别提升了 0.4%和 1.0%，其核心机制在于可见光图像的高频纹理细节可通过小波分解被有效提取与增强，从而优化目标表征能力；而将该模块嵌入红外分支时，因红外热成像特征以热辐射强度分布为主，其高频分量易受热扩散效应影响呈现稀疏性与噪声敏感性，导致小波分解难以有效建模关键特征，甚至引入干扰，从而导致 mAP@0.50 下降 0.2%，进一步验证了模态特性与模块设计的强相关性，即可见光分支适配高频增强策略，而红外分支需依赖热语义建模优化。

为了进一步验证 HCMA 模块中 Self Attention 与 Cross Modal Attention 在可见光与红外分支中的不同作用，以及考察其放置位置的合理性，本文开展了相关对比实验，具体结果如表 4 所示。

表 4 HCMA 消融实验

Table 4 HCMA ablation experiment

可见光分支	红外分支	*P*/%	*R*/%	*mAP*@0.50/%	*mAP*@0.50:0.95/%
跨模态自注意力	自注意力	88.6	79.6	**85.7**	**51.9**
自注意力	跨模态自注意力	88.7	79.7	85.3	51.2

注：加粗数字表示最优结果。

实验表明，在红外分支采用 Self Attention 并在可见光分支使用 Cross Modal Attention 时，模型在 mAP@0.05 与 mAP@0.50:0.95 上分别达到 85.7%与 51.9%，表现最佳。该配置相较于在可见光分支使用 Self Attention、红外分支使用 Cross Modal Attention 的方案，在 mAP@0.50 与 mAP@0.50:0.95 上分别提升了 0.4%与 0.7%。这是因为 HCMA 所处通道位置较深、分辨率逐渐降低，在特征层次较深、分辨率较低的阶段，红外特征相对显著，而可见光纹理信息则相对减弱。因此，在红外分支进行 Self Attention，并通过 Cross Modal Attention 引导其与可见光纹理信息对齐，能够更有效地融合双模特征。反之，若交换两种注意力机制的使用位置，性能反而下降。

3.3 对比实验结果分析

本节在 DVTOD 和 DroneVehicle 两个多模态遥感图像目标检测数据集上，将本文方法与多种先进的多模态目标检测方法进行了比较。

在 DVTOD 数据集上，不同先进方法的详细结果如表 5 所示。本文所提方法 UAVCM-Net 在 mAP@0.50 上取得了最优的检测性能，达到 86.7%。具体而言，相较于原 YOLOv8 基线模型，mAP 从 85.4%提升至 86.7%，提高了 1.3%。与基于 CNN 的方法例如 DEYOLO，或采用简单跨模态交叉自注意力机制的方法例如 ICAFusion、CMA-Det 相比，本文所提方法均表现出性能提升。这表明通过模态内增强、以红外为主导的自注意力机制和跨模态注意力机制，有效增强了模态内特征和跨模态交互特征，从而验证了本文所提方法的先进性。

表 5 不同主流方法在 DVTOD 数据集上的对比实验

Table 5 Comparative experiments of various mainstream methods on the DVTOD dataset

方法	模态	mAP@0.50/%
YOLOv5	可见光+ 红外	79.2
YOLOv8	可见光+ 红外	85.4
CFT	可见光+ 红外	82.7
CMA-Det	可见光+ 红外	85.0
DEYOLO	可见光+ 红外	85.5
ICAFusion	可见光+ 红外	86.3
UAVCM-Net	可见光+ 红外	**86.7**

在 DroneVehicle 数据集上，不同先进方法的详细结果如表 6 所示。本文方法 UAVCM-Net 同样取得了最优的检测性能，mAP@0.50 达到 86.0%。相较于基线模型 YOLOv8，性能提升了 1.7 个百分点。这一结果进一步验证了本文方法具有较强的鲁棒性。

表 6 不同主流方法在 DroneVehicle 数据集上的对比实验

Table 6 Comparative experiments of various mainstream methods on the DroneVehicle dataset

方法	模态	mAP@0.50/%
YOLOv5	可见光+ 红外	84.1
YOLOv8	可见光+ 红外	84.3
CFT	可见光+ 红外	84.3
CMA-Det	可见光+ 红外	82.0
DEYOLO	可见光+ 红外	85.6
ICAFusion	可见光+ 红外	84.9
UAVCM-Net	可见光+ 红外	**86.0**

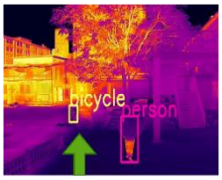
3.4 可视化分析

为了验证 UAVCM-Net 在复杂场景下的检测性能，本文选取多目标密集遮挡、小目标及复杂场景图像样本，与 YOLOv8 Visible、YOLOv8 Infrared 及 CMA-Det 进行对比实验，如表 7 所示。

表 7 可视化检测结果

Table 7 Visualization of detection results

场景	YOLOv8 可见光模态	YOLOv8 红外模态	CMA-Det	UAVCM-Net

远小目标场景				
自行车遮挡场景				
行人遮挡场景				
墙壁遮挡场景				

在远小目标检测中，UAVCM-Net 通过跨模态特征融合策略，解决了 CMA-Det 的微小目标识别失效问题（绿色箭头标注，下同）。针对单车与行人遮挡场景，模型通过单模态增强模块强化局部特征，准确识别遮挡目标，有效缓解基准模型的漏检缺陷。在墙壁遮挡场景中，跨模态互补信息有效区分背景干扰，实现遮挡目标精准定位。实验表明，UAVCM-Net 通过单模态增强与跨模态融合的协同优化机制，在远小目标检测、密集遮挡及复杂背景干扰场景中全面优于基准模型，验证了其实际应用鲁棒性。

4 结论

本文针对多光谱目标检测中模态特征判别性不足与跨模态互补性挖掘不充分的问题，提出了融合单模态增强与跨模态互补特征提取的 UAVCM-Net 方法，通过引入基于深度可分离卷积的小波变换模块、跨尺度全局信息融合模块以及跨模态混合自注意力融合模块，有效提升了可见光与红外模态的特征表征与交互能力。在 DVTOD 和 DroneVehicle 两个公开数据集上的实验结果表明，该方法在检测精度上均优于主流对比方法，并在复杂场景下展现出良好的鲁棒性，验证了所提模块能够有效增强特征判别性与模态间互补性。然而，跨模态融合机制仍存在进一步优化的空间，且所引入的模块在带来性能提升的同时也增加了计算复杂度；未来工作将围绕探索更高效的跨模态融合策略和像素级早期融合方法展开，以期在进一步提升特征互补性与判别能力的同时，实现精度与效率的更好平衡，从而满足无人机平台的实际应用需求。

参考文献（References）

[1] 苏向阳,汪洋,谭森起,等.PCA-YOLO11:全域复杂环境的轻量目标检测[J].兵工学报,2025,46(增刊 2):250584.
SU X Y, WANG Y, TAN S Q, et al. PCA-YOLO11: lightweight object detection in multi-altitude complex environments[J]. Acta Armamentarii, 2025,46(S2): 250584. (in Chinese)

[2] 赵紫臣,萧浩东,凌焕章.基于 YOLOv10n 的轻量化红外小目标检测模型[J].兵工学报,2025,46(11):250146.
ZHAO Z C, XIAO H D, LING H Z. Lightweight infrared small target detection model based on YOLOv10n[J]. Acta Armamentarii, 2025,46(11):250146. (in Chinese)

[3] 王泉,叶广飞,陈祺东.YOLO-SWR:无人机视角下轻量级交通车辆检测方法[J].计算机工程与应用, 2025,61(14): 112-122.
WANG Q, YE G F, CHEN Q D. Lightweight traffic vehicle detection algorithm from UAV perspective[J]. Computer Engineering and Applications, 2025, 61(14): 112-122. (in Chinese)

[4] 徐相华,周德靖,余超然,等.基于 DCA-YOLO 的受病虫害侵染树木农业无人机低空检测模型[J].农业机械学报,2025,56(10):479-491.
XU X H, ZHOU D J, YU C R, et al. Improved model for low altitude detection of trees infected by pests and doseases using agricultural dronesbased on DCA-YOLO[J]. Transactions of the Chinese Society for Agricultural Machinery， 2025, 56(10): 479-491. (in Chinese)

[5] 丁子天,喜文飞,钱堂慧,等.结合多特征的无人机雾天影像识别[J].红外技术,2025,47(7):833-841.
DING Z T, XI W F, QIAN T H, et al.Multiple feature fusion for unmanned aerial vehicle image recognition in foggy weather[J]. Infrared Technology, 2025, 47(7):833-841. (in Chinese)

[6] 孙备, 党昭洋，吴鹏，等. 多尺度互交叉注意力改进的单无人机对地伪装目标检测定位方法[J].仪器仪表学报, 2023, 44(6): 54-65.
SUN B, DANG Z Y, WU P, et al. Multi scale cross attention improved method of single unmanned aerial vehicle for ground camouflage target detection and localization[J].Chinese Journal of Scientific Instrument, 2023,44(6):54-65.(in Chinese)

[7] 付琨, 王佩瑾, 冯瑛超, 等.遥感跨模态智能解译:模型、数据与应用[J].中国科学:信息科学, 2023,53(8): 1529-1559.
FU K, WANG P J, FENG Y C, et al. Cross-modal remote sensing intelligent interpretation:method, data and application[J].Scientia
Sinica(Informationis),2023,53(8):1529-1559. (in Chinese)

[8] XIE Z X, SHAO F, CHEN G, et al. Cross-modality double bidirectional interaction and fusion network for RGB-T Salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4149-4163.

[9] WANG K P, TU Z Z, LI C L, et al. Learning adaptive fusion bank for multi-modal salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 7344-7358.

[10] LIU X Y, QI J H, CHEN C, et al. Relation-aware weight sharing in decoupling feature learning network for UAV RGB-infrared vehicle re-identification[J]. IEEE Transactions on Multimedia, 2024, 26: 9839-9853.

[11] ZHANG J Q, LEI J, XIE W Y, et al.SuperYOLO:super-resolution assisted object detection in multimodal remote sensing imagery:arXiv:2209.13351[R]. Ithaca,NY,US:Cornell University, 2022:2209.13351.

[12] CHEN Y S, WANG B R, GUO X Y, et al. Deyolo: dual-feature-enhancement yolo for cross-modality object detection[J].Lecture Notes in Computer Science, 2024,15317: 236-252.

[13] FANG Q Y, HAN D P, WANG Z K. Cross-modality fusion transformer for multispectral object detection:ArXiv:2111.00273[R].Ithaca,NY,US:Cornell University, 2021: 2111.00273.

[14] SHEN J, CHEN Y, LIU Y, et al. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection[J]. Pattern Recognition, 2024, 145:109913.

[15] 代小波,任思羽,何小海,等.基于特征增强和对齐融合的可见光–红外图像融合目标检测[J].工程科学学报，2025,47(12):2554-2565.
DAI X B, REN S Y, HE X H, et al.Visible–infrared fusion object detection based on feature enhancement and alignment fusion [J]. Chinese Journal of Engineering, 2025, 47(12):2554-2565. (in Chinese)

[16] XIAO Y M, MENG F M, WU Q Bet al.GM-DETR:generalized multispectral detection transformer with efficient fusion encoder for visible-infrared detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024:17157-17166.

[17] SONG K C, XUE X T, WEN H W, et al. Misaligned visible-thermal object detection: a drone-based benchmark and baseline[J] IEEE Transactions on Intelligent Vehicles, 2024,9(11): 7449-7460.

[18] FINDER S E, AMOYAL R, TREISTER E, et al. Wavelet convolutions for large receptive fields[C]//European Conference on Computer Vision. 2024: 363-380.

[19] CAO Y, XU J R, LIN S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop. 2019:1971-1980.

[20] SUN Y M, CAO B, ZHU P F, et al. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6700-6713.