基于红外视频与音频联合的无人机目标检测

常聚忠,姚其新,杨振东,黄瑶玲,胡生辉
(国网湖北省电力有限公司直流公司,湖北 宜昌, 443001)

摘 要 针对小型无人机检测中存在的目标弱小、背景复杂及易混淆等问题,提出了一种基于多模态信息融合的检测方法.为了提高目标的检测精度与鲁棒性,所提方法通过红外视频与音频分别进行目标的初步检测,然后通过决策级融合获得最终的检测结果.针对红外视频,采用"跟踪后检测"技术路线,设计动态显著性差异增强模块,通过多模态特征融合机制整合梯度-灰度特征与运动信息,结合局部熵引导的窗口缩放策略和时域运动验证,以提升弱小目标与背景的对比度.同时,引入时空轨迹编码与关联模块,利用LSTM网络进行短时轨迹特征提取和轨迹-量测匹配度计算,通过动态时空融合因子优化数据关联过程,克服了目标遮挡与轨迹断裂问题.在音频处理方面,通过提取梅尔频谱图并将其转换为对数梅尔频谱特征,采用CNN模型进行声学特征识别.实验结果表明:该文方法在精确率(89.5%)、召回率(85.7%)和平均精度(75.4%)等指标上均优于现有方法,为复杂环境下的无人机目标检测提供了有效可行的手段.

关键词 无人机目标检测;动态显著性增强;多模态融合;时空轨迹编码
中图分类号 TP391.4 **文献标志码** A **文章编号** 1672-4321(2026)05-0674-09
doi:10.20056/j.cnki.ZNMDZK.20260740

UAV target detection based on the integration of infrared video and audio

Chang Juzhong, Yao Qixin, Yang Zhendong, Huang Yaoling, Hu Shenghui
(State Grid Hubei Direct Current Company, Yichang 443001, Hubei China)

Abstract To addressing the challenges of detecting small drones, such as weak targets, complex backgrounds, and potential confusion, a detection method based on multimodal information fusion is proposed. To enhance the accuracy and robustness of target detection, the proposed method employs infrared video and audio for initial target detection, subsequently obtaining the final detection result through decision-level fusion. For infrared video, the "tracking-then-detecting" approach is adopted, incorporating a dynamic saliency difference enhancement module. The module integrates gradient-grayscale features and motion information through a multimodal feature fusion mechanism, combined with a window scaling strategy guided by local entropy and time-domain motion verification, thereby enhancing the contrast between weak and small targets and the background. Additionally, a space-time trajectory coding and correlation module is introduced, utilizing an LSTM network for short-term trajectory feature extraction and trajectory-measurement matching degree calculation. The dynamic space-time fusion factor optimizes the data association process, addressing issues such as target occlusion and trajectory discontinuity. Regarding audio processing, the Mel spectrum is extracted and converted into logarithmic Mel spectral features, with a CNN model employed for acoustic feature recognition. Experimental results demonstrate that the proposed method outperforms existing methods in terms of accuracy (89.5%), recall (85.7%), and average precision (75.4%), providing an effective and feasible approach for drone target detection in complex environments.

Keywords UAVs target detection; dynamic saliency enhancement; multimodal fusion; spatiotemporal trajectory encoding

收稿日期 2026-01-09
作者简介 常聚忠(1983-),男,高级工程师,研究方向:特高压直流输电运维管理,E-mail:15971674253@163.com
基金项目 国网湖北省电力有限公司科技资助项目(521521240005)

当前，无人机已成为涵盖监控[1]和农业[2]等多种实际应用领域的必备工具.无人机所具备的多功能性、便捷部署和高效运作特性，使其在民用和军用领域得到广泛应用.然而，随着无人机数量持续增加，未经授权或恶意的无人机活动对人民的生产生活构成重大威胁，尤其针对机场、军事基地、政府设施和关键基础设施等敏感区域.因此，为降低这些风险并保障公共安全，实时可靠的无人机检测与追踪研究已变得愈发迫切.

无人机检测与追踪面临的主要挑战在于其体积小巧、机动性强且隐蔽性高[3]，这些特性使其能够规避传统安防措施.具体而言，无人机与鸟类常表现相似的飞行模式和外形特征[4]，导致传统检测方法难以精准区分.因此，现有方法容易出现误报并降低检测精度.

由于无人机目标通常体积较小，对于高清的可见光视频而言，容易受到复杂场景的严重干扰，因此采用红外传感器来检测无人机的研究逐渐受到了广泛的关注.目前，红外视频的目标检测方法主要分为逐帧检测算法和多帧联合检测算法两类.

在逐帧目标检测方面，针对红外图像中目标特征不显著的问题，王伯霄等[5]为了获取更精准的目标潜在区域，提出了一种基于难点的注意力感知的区域建议模块以滤除多余背景.林珊玲等[6]提出了一种融合全局信息与空间域和通道域双域注意力机制实现了对图像中的小尺度目标检测精度的提升.刘兆龙等[7]为特征图分配三维权重，提高了特征提取能力.在复杂场景下，单帧目标检测算法的性能往往会显著下降，其根本原因在于单张图像的信息有限.当无人机被厚云层等遮挡物严重遮蔽时，目标的亮度对比度和细节特征均会遭到模糊化削弱，导致其可供识别的本质特征严重匮乏，这大幅增加了检测难度.

现有多帧联合检测算法，大致可以分为通道叠加、三维卷积计算和时序学习三种方式.姚胜波等[8]提出一种基于改进的单阶段红外小目标检测方法，采用了多分辨率并行骨干网络结构和通道分离注意力模块.Chen 等[9]提出了一种具有跨片段增强功能的切片时空网络，用于同时捕获片段内和片段间的时空运动特征，并提出了运动耦合颈部结构，以辅助特征提取器更好地从多帧图像中捕获和利用运动特征.Du 等[10]提出了有助于解决虚假收敛和样本误判问题的小交并比策略.针对目标信号微弱和背景复杂的难点，又提出了基于帧间能量积累增强机制的端到端时空特征提取与目标检测框架.Yan 等[11]提出了一种基于学习的多帧红外小目标检测方法.首先使用时序多尺度特征提取器从多时间尺度获取时空特征，随后通过空间多尺度特征精炼器在保持小目标位置信息的前提下增强时空特征的语义表达能力.这些多帧检测网络各有优点，但对于无人机目标检测任务中目标小、模型实时性强的实际要求，上述方法往往难以凑效.

近期无人机检测技术的进步主要集中在改进单阶段深度学习模型，以提高动态环境中的鲁棒性，优化多尺度检测，并增强小目标识别.Lenhard 等人[12]提出了 YOLO-Feder FusionNet，该模型将通用目标检测与伪装检测技术相结合，提高了在视觉挑战性背景下的无人机识别能力.基于小目标特征提取的困难，研究人员还引入了特征提取的改进.Zamri 等人[13]将卷积块注意力模块（CBAM）集成到 YOLOv8 的颈部，并增加一个高分辨率检测头，提高了对小型无人机的敏感度.Huang 等人[14]提出了 EDGS-YOLOv8，通过 Ghost 卷积降低了模型复杂度，同时整合了高效的多尺度注意力（EMA）和可变形卷积网络（DCNv2），在最小计算开销下确保了更高的准确性.最近，多传感器联合探测的方式在无人机目标检测中开始兴起，如有学者开始研究基于声学传感器的目标检测[15].一些论文[16-17]使用快速傅里叶变换（FFT）进行特征提取.近年来，梅尔频率倒谱系数（MFCC）是论文[18]中经常使用的特征，并取得了一定的效果.

综上，针对小型无人机目标检测任务，本文设计了一种结合红外视频和音频数据的联合检测方法.对于红外视频中无人机目标尺度小，与背景差异度低的问题，采用了“跟踪后检测”的路线，设计了动态显著性增强模块和时空轨迹编码与关联模块，通过挖掘时空域轨迹连续性，提高对动目标的准确跟踪检测.对于音频数据，由原始声音信号提取其梅尔频谱，再由 CNN 网络对其进行分类识别.最后，在决策阶段，采用决策级融合的策略，得到最终的检测识别结果.在实验部分，将本文方法与主流的无人机目标检测方法进行了对比，实验结果验证了本文方法的有效性.

1　方法

针对多传感（红外视频和音频）采集到的数据，本文设计了一种多模态联合的无人机目标检测方

法(其整体框架见图1).具体而言,对于红外视频数据,采用基于显著性差异的检测框架.对输入的红外视频数据,采用先跟踪后检测的方案,首先由动目标检测器将输入数据中的所有动目标检测出来,然后由特征提取器得到对于目标的特征描述,最后由分类器给出目标的具体类别.对于采集的音频数据,首先将时序信号转换为二维的梅尔频谱特征,然后采用基于CNN的特征提取器提取梅尔频谱图的特征,最后由分类器给出目标的具体类别.在分别获得红外视频和音频数据的识别结果后,采用基于不同源识别概率的逻辑回归模型进行决策级融合,最后给出具体的目标类别.

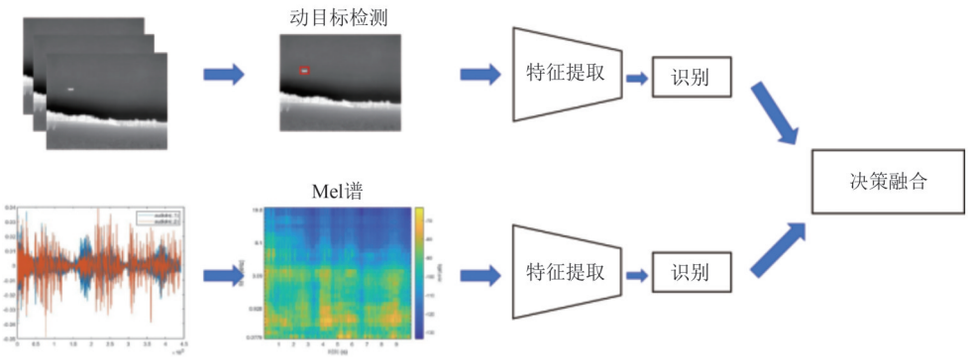


图1 小型无人机目标检测整体框架

Fig. 1 Comprehensive framework for target detection in small UAVs

1.1 红外视频无人机目标检测

基于红外视频的目标检测流程如图2所示.对于输入的红外视频,首先进行预处理与增强,以在抑制噪声同时并提升图像质量;其次,进入动态显著性差异增强阶段,此阶段并行计算局部对比度、深度特征和运动特征,通过自适应加权融合生成显著性图;再次,基于显著性图进行目标检测与特征提取,获取候选目标的位置和特征描述;然后,进行时空轨迹编码与关联,利用轨迹预测和关联概率计算实现目标与量测的数据关联,得到目标轨迹与状态;最后,在得到所有目标的运动轨迹后,利用成熟的CNN模型(本文中采用ResNet34)逐帧提取其深度特征,由此训练特征提取模型和分类器的权重参数.

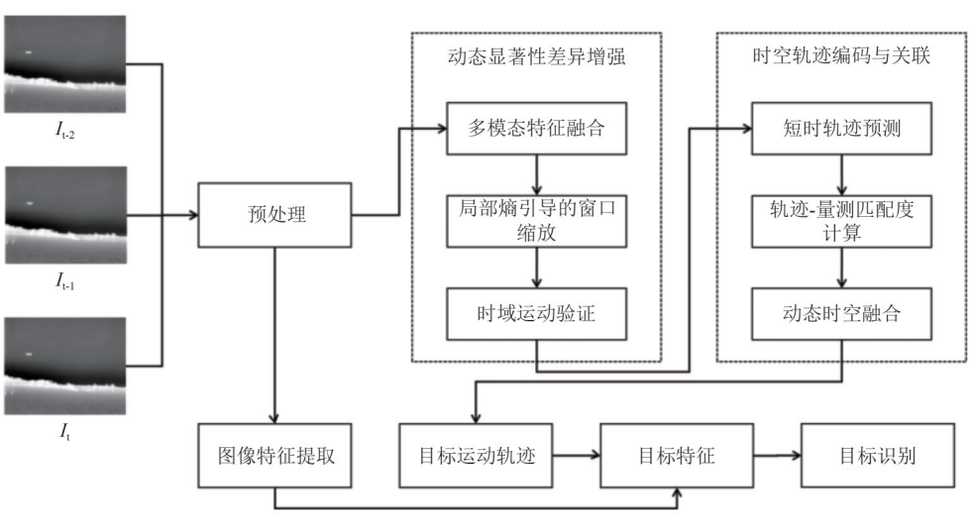


图2 红外视频目标检测流程

Fig. 2 Framework for infrared video target recognition

1.1.1 动态显著性增强模块

动态显著性差异增强模块是本方法的关键环节,旨在解决低对比度环境下弱小目标难以提取的问题.本模块通过多模态特征融合策略,同时利用目标的空域、时域和深度特征信息,提升目标检测的鲁棒性与准确性.

(1) 多模态特征融合机制

模块采用三级处理流程,首先进行梯度-灰度特征融合,利用梯度算子计算输入红外图像的梯度幅值,结合原始灰度信息形成初步融合特征.具体地,

对输入图像I_t,计算其x方向和y方向的梯度幅值,形成梯度图$G(x,y)$.随后,设计自适应加权融合机制,根据局部区域特征动态调整灰度与梯度的贡献权重.在纹理复杂区域,提高梯度特征的权重以增强边缘信息;在平滑区域,则侧重灰度对比度信息.这一自适应权重调整机制有效克服了传统固定权重融合的局限性,使算法能够适应多变的场景特性.

在局部对比度计算阶段,采用改进的多尺度块对比度度量方法,在多个尺度空间计算目标与背景的对比度差异.特别地,针对弱小目标的特点,这里设计了一种双窗口检测策略,内部窗口聚焦目标区域,外部窗口覆盖背景区域,通过计算两个窗口的统计特征差异来量化局部对比度.

为充分利用时间维度的信息,模块还集成了运动特征分析功能.对连续三帧图像(I_t,I_{t-1},I_{t-2})计算帧间差异,并结合目标的运动一致性验证,以保证仅有持续运动的区域才被判定为潜在目标.这种方式能够有效抑制静态背景干扰,从而突出运动目标.

(2) 局部熵引导窗口缩放

在局部对比度计算过程中,窗口尺寸的选择对检测性能有着重要的影响.过大的窗口会导致弱小目标信号被背景淹没,而过小的窗口则无法充分捕捉目标与背景的差异.针对这一问题,本模块引入了局部熵引导的窗口自适应调整机制.

窗口自适应调整机制的设计如下:

首先,对于图像中的每个候选区域,计算其局部信息熵:$E_{local}=-\sum p_i log_2 p_i$,其中$p_i$为灰度值$i$在局部窗口内的出现概率.信息熵量化了区域的纹理复杂程度:高熵值表示区域纹理复杂(如云层边缘、建筑轮廓),低熵值表示区域均匀(如天空、水面).

其次,基于局部信息熵,模块动态调整对比度计算的窗口尺寸:当局部熵高于阈值$\tau_{entropy}=1.8$时,将背景窗口尺寸从9×9缩小至5×5,以抑制复杂纹理干扰;否则保持较大窗口,确保充分捕捉目标与背景的对比度差异.

这一自适应窗口调整机制与多尺度分析相结合,在多个尺度空间重复进行局部对比度计算,最后通过双线性插值统一尺度后加权叠加,增强对尺寸变化目标的鲁棒性.通过这种设计,模块能够自适应不同背景复杂度,在保持高检测率的同时有效抑制虚警.

(3) 时域运动验证

为进一步提升检测可靠性,模块引入了时域运动验证机制.对连续三帧的显著性响应图进行时域一致性分析,仅保留那些在连续帧中持续出现的显著区域.

为此,首先对每个候选显著区域,计算其在连续帧中的显著性响应值,并计算其时间一致性度量:$S_{\text{temp}}(x,y)=\frac{1}{3}\sum S_{t-i}(x,y)$,其中$i=0,1,2$.

其次,设定自适应阈值$T_{\text{motion}}=\mu_S+2\sigma_S$(其中$\mu_S$和$\sigma_S$分别为$S_{\text{temp}}$的均值和标准差),只有那些显著性响应持续高于阈值的区域才被判定为真实目标:

$$S_{\text{final}}(x,y)=\begin{cases}S_{\text{temp}}(x,y) & ,if\ S_{\text{temp}}(x,y)>T_{\text{motion}}\\ 0 & ,\text{otherwise}\end{cases},\quad(1)$$

这种时域一致性验证能够有效过滤瞬态噪声和随机干扰,如飞鸟、云朵等的短暂缝隙等,有助于提升检测的可靠性.

在特征融合阶段,通过引入注意力机制,自动学习各特征通道的权重,实现局部对比度特征、深度特征和运动特征的自适应融合.最终输出综合显著性图,为后续目标检测与跟踪提供高质量的输入.

1.1.2　时空轨迹编码与关联模块

时空轨迹编码与关联模块旨在解决机动目标跟踪中的轨迹断裂与关联错误问题.传统跟踪方法多依赖于目标的位置和运动信息,难以有效处理目标遮挡、交叉等复杂场景.本模块通过引入轨迹编码机制与多特征关联策略,以利于提升复杂场景下目标跟踪的连续性与稳定性.

(1) 短时轨迹特征提取

模块采用分层处理架构,底层进行短时轨迹编码,通过轻量级LSTM网络学习目标的运动模式.具体而言,LSTM采用双层结构,每层128个单元,存储目标最近5帧的位置信息,构成短时轨迹序列,输入至LSTM网络进行特征编码.LSTM网络通过其门控机制,选择性记忆重要的运动状态变化,遗忘无关紧要的波动,从而提取出稳健的轨迹特征表示.这一轨迹编码过程能够捕捉目标的运动意图,为后续数据关联提供深层特征依据.训练使用Adam优化器(学习率0.001,衰减因子$\beta_1=0.9$,$\beta_2=0.999$),Dropout率为0.3以防止过拟合.

(2) 轨迹-量测匹配度计算

在轨迹-量测匹配度计算阶段,模块将轨迹特征与量测特征映射到同一度量空间,计算其相似度.具体采用余弦相似度度量,避免幅度差异对相似度计算的影响.量测点z_j通过MLP映射到与轨迹特征

相同维度的空间$\varphi(z_j)=\mathrm{ReLU}(W_z z_j+b_z)$,然后计算轨迹特征$h_t$与量测特征$\varphi(z_j)$之间的余弦相似度:

$$\mathrm{TM}_{jt}=\frac{h_t\varphi(z_j)}{\|h_t\|\|\varphi(z_j)\|},\tag{2}$$

同时,为应对目标形态变化,通过卷积神经网络提取目标的深度特征,计算表观相似性.这种多特征融合的相似度度量方法,能够综合运动与表观信息,提高数据关联的可靠性.特别地,对于红外弱小目标,由于缺乏丰富的纹理信息,利用目标的强度分布特征和局部形状特征,能获得更具判别力的特征表示.

(3) 动态时空融合因子

本模块设计了动态时空融合因子,将轨迹匹配度与传统的空间距离信息相结合.具体而言,重新定义目标-量测关联概率为:

$$\mathrm{STF}^{\mathrm{new}}(jt)=\lambda\frac{\mathrm{ED}(jt)}{\mathrm{MD}(jt)}+(1-\lambda)\mathrm{TM}_{jt},\tag{3}$$

其中,$\mathrm{ED}(jt)$表示量测j与目标t之间的欧氏距离,$\mathrm{MD}(jt)$表示马氏距离,λ为自适应权重系数.动态时空融合因子λ的优化过程基于目标机动性自适应调整.我们通过实验确定了λ的初始值为0.5,并设计了一种基于目标运动状态的调整机制:当目标运动速度变化率超过阈值时,λ以0.1的步长递减至0.3,以增强轨迹匹配度的权重;反之,在匀速运动状态下,λ恢复至0.5.该优化过程通过网格搜索在验证集上完成,以最大化跟踪精度.权重系数λ根据目标的机动性动态调整.当检测到目标处于高机动状态(如急转弯、快速变向)时,自动降低空间距离权重,相应提高轨迹匹配度的权重,以适应目标的机动行为.这种动态调整机制可以提升对机动目标跟踪的稳定性,缓解目标机动时易发生轨迹断裂的问题.

1.2 音频信号识别

频谱图是一种分析声音波形的典型方法,其频率特性随时间变化,可通过在时频域排列声学数据的频谱获得,如图3所示.

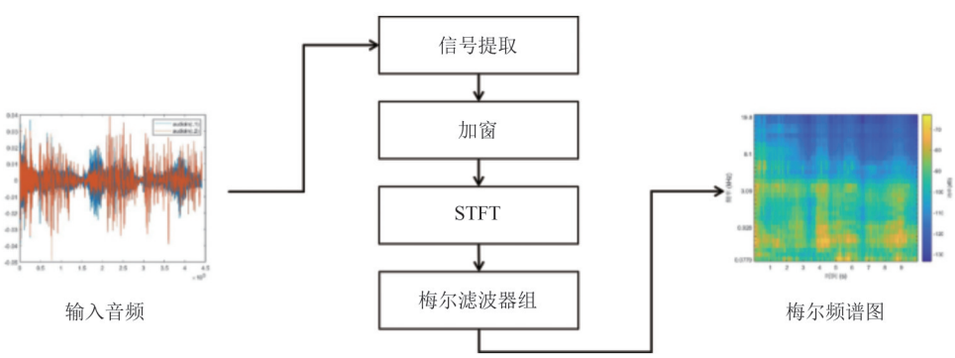


图3 音频频谱计算过程示意图

Fig. 3 Schematic diagram of the audio spectrum calculation process

首先将汉明窗应用于提取的音频信号,然后进行1440点STFT,分析窗口重叠长度为960个样本.当采样率为48 kHz时,梅尔滤波的频带数为32,帧数为89.最后,调整音频梅尔频谱图的大小以适应执行无人机检测或分类的CNN模型的输入尺寸.请注意,梅尔滤波器(或梅尔尺度三角滤波器)以高分辨率描述低频带,而以低分辨率表示高频带.因此,梅尔滤波器倾向于体现无人机在低频带的声学特性.

对数梅尔频谱图在语音识别和环境声识别中得到了广泛的应用.为了在保持音频信号频谱特征的同时大幅降低特征维度,可将对数梅尔频谱图的维度设定为128.提取梅尔频谱图后,将其转换为对数尺度以获得对数梅尔频谱图,最终得到的对数梅尔频谱图尺寸为128 × 128.对提取的特征进行z值标准化处理,计算并保存所有样本的均值和标准差至文件中,再从对数梅尔频谱图的每个数值中减去均值并除以标准差.

对于获取到的音频梅尔频谱图,可采用基于CNN的特征提取器提取梅尔频谱图的特征,音频信号识别采用CNN模型,其结构包括:输入层(128×128对数梅尔频谱图)、3个卷积块(每块含卷积层、ReLU激活和最大池化)、全连接层(512单元)及Softmax输出层.具体参数:卷积核尺寸3 × 3,步长1,池化窗口2 × 2;最后由分类器给出目标的具体类别.

2 结果与讨论

首先对该文所提出的动态显著性增强模块以

及时空轨迹编码与关联模块进行详细的消融实验，通过实验结果来验证各个模块带来的性能提升效果；其次，对比现有的优秀无人机小目标检测方法，通过客观指标来验证该文所提方法的优秀性能.

2.1　实验数据与评价

该文所用的数据[19]包含了多种模态数据（如表1所示），包括红外视频和音频信号.红外视频数据包含365段视频，每段视频的时长均相同，为10秒.红外视频的分辨率为320×256像素，30 Hz帧率.数据集中使用了三种不同的无人机，分别是为Inspire2、Mavic3和Phantom4.音频数据包含90段时长为10秒的音频文件，格式为wav.所有音频数据的采样频率均相同，为44100 Hz.数据集中的音频来自所采集的视频或单独由麦克风录制.背景类别包含在采集地点户外录制的一般背景声音，也包括安装有传感器的伺服电机移动云台平台的声音片段.图4和图5分别给出了该数据集中的部分视频帧和音频序列.

表1　所用数据集的信息

Tab. 1　Information on the dataset used

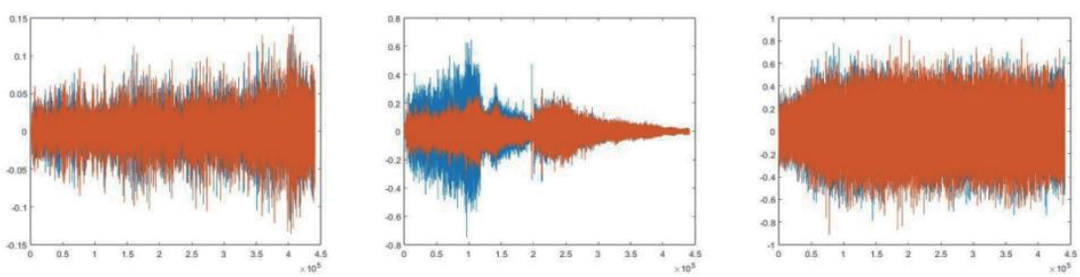
数据类型	视频分辨率	样本数量	具体类别
红外视频	320×256	365段	飞机、直升机、鸟、无人机
音频	/	90段	无人机、直升机、背景



(a) 无人机目标　(b) 飞机目标　(c) 飞鸟目标　(d) 直升机目标

图4　红外数据示意图

Fig. 4　Infrared data schematic diagram



(a) 无人机目标　(b) 直升机目标　(c) 背景

图5　音频数据示意图

Fig. 5　Audio data schematic diagram

为了便于图像特征提取模型的训练，可将视频数据逐帧分解.一段10秒的红外视频可抽取301帧图像，每一帧均有对目标的准确标记.经过数据的筛选，去除含强噪声等的脏数据后，数据集中的无人机目标数47257，非无人机目标数为62608.数据预处理包括：红外图像归一化至[0,1]，音频Z-score标准化；训练集与测试集按6∶4随机划分，确保类别平衡；预处理中应用了直方图均衡化（红外）和重采样（音频）以消除设备偏差.

按照目标检测评价的通用方式，本文的算法评估指标包括：精确率（Precision）、召回率（Recall）、F1值和平均精度（AP）.

2.2　消融实验

为了验证该文提出的先跟踪后检测的红外无人机目标检测方法中各模块的效果，本节进行了消融实验.消融实验中对比了增加不同模块的先跟踪后检测方法对后续检测结果的影响.表2中方法（1）代表不包含动态显著性差异增强模块和时空轨迹编码与关联模块的基础方法，方法（2）代表在（1）的基础上增加了动态显著性差异增强模块，方法（3）代表在（1）的基础上增加了时空轨迹编码与关联模块，方法（4）代表包含动态显著性差异增强模块和时空轨迹编码与关联模块的方法.

从消融实验结果可以看出，不同模块的引入对检测效果产生了显著的渐进式提升.基础方法（1）在不包含任何增强模块的情况下，精确率、召回率、F1值和平均精度分别为75.2%、68.1%、0.71和60.1%，这为后续改进提供了基准参考.当单独引入动态显

表2 消融实验结果

Tab. 2 Ablation experiment results

方法	Precision/%	Recall/%	F1	AP/%
(1)	75.2±0.13	68.1±0.15	0.71±0.0014	60.1±0.28
(2)	79.2±0.18	75.2±0.19	0.77±0.0016	62.4±0.26
(3)	78.7±0.21	77.2±0.15	0.78±0.0018	63.1±0.29
(4)	86.7±0.17	82.1±0.21	0.84±0.0019	70.5±0.34

著性差异增强模块形成方法(2)后,各项指标均有所提高,特别是召回率从68.1%提升至75.2%,表明该模块有效增强了系统对真实目标的捕捉能力,减少了漏检现象.同时加入时空轨迹编码与关联模块的方法(3)在召回率和F1值上进一步改善,显示出该模块在提升检测连续性方面的积极作用.

同时包含两个增强模块的方法(4)在所有评估指标上均达到了最佳性能:精确率86.7%、召回率82.1%、F1值0.84、平均精度70.5%.与基础方法相比,精确率提高了11.5%,召回率提升了14%,这充分证明两个模块之间存在明显的协同效应.动态显著性差异增强模块通过提升目标与背景的区分度改善了检测准确性,而时空轨迹编码与关联模块则通过利用目标的时间连续性增强了检测的稳定性.

消融实验结果同时也表明,动态时空融合因子λ的优化这种自适应机制能有效应对目标机动导致的轨迹断裂问题.在多目标交互密集场景下,该机制进一步展现出对遮挡恢复能力的提升作用.当目标被短暂遮挡后重新出现时,较低的λ值有助于强化外观特征匹配的贡献,从而提高重识别准确率.同时,为避免过度依赖运动模型导致漂移累积,系统引入了反馈校正模块,动态监控轨迹一致性并触发λ的局部调整.实验结果验证了其在复杂动态环境中的鲁棒性与适应性.

2.3 对比实验

为了对比该文方法和经典目标检测方法的效果,本节选取了多种目标检测方法,包括经典的双阶段有锚框目标方法Faster RCNN[20],单阶段有锚框目标方法RetinaNet[21]和YOLOv5[22],和无锚框方法CenterNet[23],RT-DETR[24].在对比实验中,分别采取了本文提出的先跟踪后检测的红外无人机目标检测方法(视觉模型)和基于红外和音频的联合检测模型(融合模型),对比实验的结果如表3所示.

表3 对比实验结果

Tab. 2 Comparative experimental results

方法	Precision/%	Recall/%	F1	AP/%
Faster RCNN[20]	73.2±0.21	60.3±0.17	0.66±0.0018	58.7±0.23
RetinaNet[21]	70.1±0.15	62.4±0.26	0.66±0.0021	60.4±0.21
YOLOv5[22]	80.3±0.17	75.1±0.20	0.78±0.0018	67.2±0.20
CenterNet[23]	56.4±0.18	70.2±0.13	0.63±0.0017	52.4±0.19
RT-DETR[24]	82.4±0.16	78.7±0.19	0.81±0.0017	68.1±0.18
Ours(视觉模型)	86.7±0.17	82.1±0.22	0.84±0.0019	70.5±0.34
YOLOv5[22]+特征融合	82.4±0.23	76.8±0.21	0.80±0.0029	68.9±0.22
YOLOv5[22]+加权融合	84.7±0.19	79.5±0.17	0.82±0.0023	70.3±0.21
RT-DETR[24]+加权融合	86.1±0.18	82.4±0.15	0.84±0.0019	71.2±0.18
Ours(融合模型)	89.5±0.18	85.7±0.19	0.88±0.0018	75.4±0.23

根据表3的对比实验结果,不同检测方法的性能差异显著.整体来看,本文提出的视觉模型虽略逊于融合模型,但四项指标(86.7%精确率、82.1%召回率、0.80的F1值、70.5% AP)仍全面超越所有对比的视觉检测方法,这说明其独立架构设计已具备较强竞争力.在传统方法中,YOLOv5展现出相对均衡的性能,其精确率(80.3%)和召回率(75.1%)虽低于本文提出的方法,但F1值(0.77)和平均精度(67.2%)仍明显优于其他基准模型.Faster-RCNN和RetinaNet表现中等,前者召回率(60.3%)较低可能导致漏检风险,后者F1值(0.72)略优但整体仍有限制.值得注意的是CenterNet的异常表现:其召回率(70.2%)尚可,但精确率(56.4%)断崖式下跌,导致F1值(0.61)和平均精度(52.4%)均为最低,反映出该模型存在严重的误检问题,可能源于模型敏感度过高或训练数据偏差.RT_DETR的表现相对出色,但仍然在评价指标上略逊于本文方法.表3中对比了三种融合策略:决策级融合(本文方法)、特征级融合(早期融合红外与音频特征)以及平均融合(红外与音频结果的平均).实验显示,决策级融合在精

确率、召回率、F1 值和 AP 上均优于其他策略．这表明本文的融合策略优于简单的特征融合和加权融合策略．

图 6 给出了部分的检测效果，从中可以看出，无论是对于复杂的云层环境还是低对比度的情形，本文方法都可以完好的检测出目标，其他方法均存在不同程度的漏检现象．

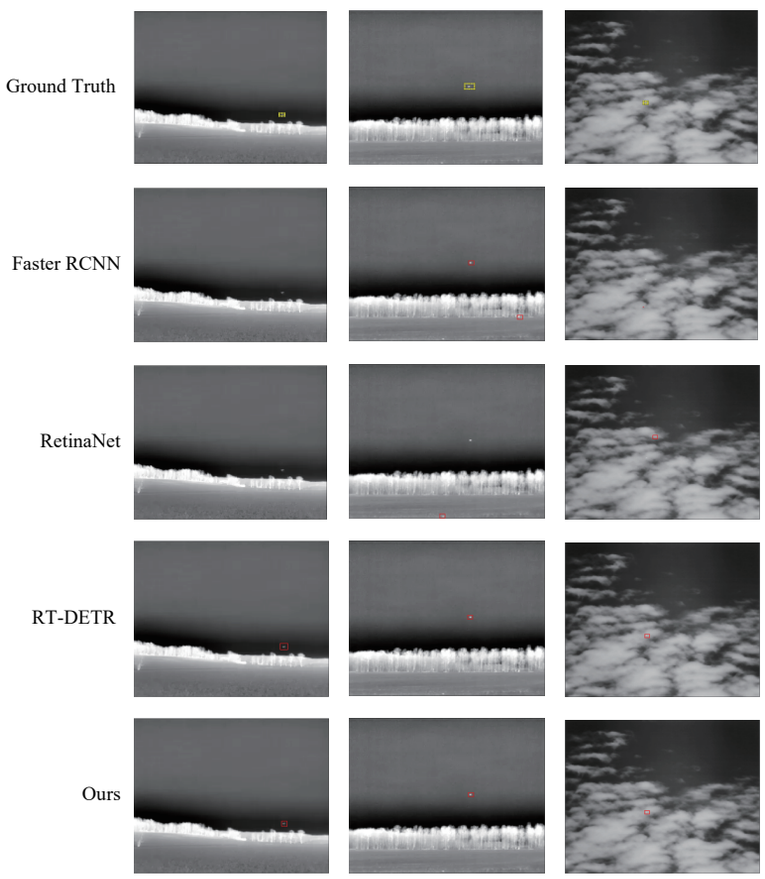


图 6　不同方法的检测效果对比图

Fig. 6　Comparison chart of detection performance across different methods

3　结论

本文针对复杂背景下的弱小目标检测问题，提出了一种基于红外视频与音频决策及融合的目标检测方法．针对红外视频，采用“跟踪后检测”技术路线，设计动态显著性差异增强模块，通过融合梯度-灰度特征与运动信息，并结合局部熵引导的窗口缩放和时域运动验证策略，改善弱小目标与背景的对比度．针对音频的分析，通过提取梅尔频谱图并将其转换为对数梅尔频谱特征，并采用 CNN 模型进行声学特征识别．最后基于上述两个分支的检测结果，通过决策级融合获得最终的目标检测结果．在开源数据集上的实验结果表明，本文方法对于低对比度和复杂背景的无人机目标检测具有良好的性能表现．

参　考　文　献

[1] Khan M, Ahmad J, El Saddik A, et al. Drone-HAT: Hybrid attention transformer for complex action recognition in drone surveillance videos[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA. IEEE, 2024: 4713-4722.

[2] Guebsi R, Mami S, Chokmani K. Drones in precision agriculture: A comprehensive review of applications, technologies, and challenges [J]. Drones, 2024, 8(11): 686.

[3] Coluccia A, Fascista A, Sommer L, et al. The drone-vs-bird detection grand challenge at ICASSP 2023: A review of methods and results[J]. IEEE Open Journal of Signal Processing, 2024, 5: 766-779.

[4] Coluccia A, Fascista A, Schumann A, et al. Drone *vs.*

bird detection: Deep learning algorithms and results from a grand challenge[J]. Sensors, 2021, 21(8): 2824.

[5] 王伯霄, 宋延嵩, 董小娜. 难点注意力感知红外小目标检测网络[J]. 中国光学(中英文), 2024, 17(3): 538-547.

[6] 林珊玲, 张雪, 陈燕, 等. 融合全局信息与双域注意力机制的光学遥感飞机目标检测[J]. 光学 精密工程, 2024, 32(20): 3085-3098.

[7] 刘兆龙, 曹伟, 高军伟. 基于改进YOLOv7的小目标焊点缺陷检测算法[J]. 液晶与显示, 2024, 39(10): 1332-1340.

[8] Yao S, Zhu Q, Zhang T, et al. Infrared image small-target detection based on improved FCOS and spatio-temporal features[J]. Electronics, 2022, 11(6): 933.

[9] Chen S, Ji L, Zhu J, et al. SSTNet: Sliced spatio-temporal network with cross-slice ConvLSTM for moving infrared dim-small target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5000912.

[10] Du J, Lu H, Zhang L, et al. A spatial-temporal feature-based detection framework for infrared dim small target[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 3000412.

[11] Yan P, Hou R, Duan X, et al. STDMANet: Spatio-temporal differential multiscale attention network for small moving infrared target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5602516.

[12] Lenhard T R, Weinmann A, Jäger S, et al. YOLO-feder fusionnet: A novel deep learning architecture for drone detection[C]//2024 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates. IEEE, 2024: 2299-2305.

[13] Najihah Muhamad Zamri F, Gunawan T S, Hajar Yusoff S, et al. Enhanced small drone detection using optimized YOLOv8 with attention mechanisms[J]. IEEE Access, 2024, 12: 90629-90643.

[14] Huang M, Mi W, Wang Y. EDGS-YOLOv8: An improved YOLOv8 lightweight UAV detection model[J]. Drones, 2024, 8(7): 337.

[15] Svanstrom F, Englund C, Alonso-Fernandez F. Real-time drone detection and tracking with visible, thermal and acoustic sensors [C]//2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy. IEEE, 2021: 7265-7272.

[16] Siriphun N, Kashihara S, Fall D, et al. Distinguishing drone types based on acoustic wave by IoT device[C]// 2018 22nd International Computer Science and Engineering Conference (ICSEC). Chiang Mai, Thailand. IEEE, 2018: 1-4.

[17] Park S, Shin S, Kim Y, et al. Combination of radar and audio sensors for identification of rotor-type Unmanned Aerial Vehicles (UAVs)[C]//2015 IEEE SENSORS. Busan, Korea. IEEE, 2015: 1-4.

[18] Anwar M Z, Kaleem Z, Jamalipour A. Machine learning inspired sound-based amateur drone detection for public safety applications[J]. IEEE Transactions on Vehicular Technology, 2019, 68(3): 2526-2534.

[19] Svanström F, Alonso-Fernandez F, Englund C. A dataset for multi-sensor drone detection[J]. Data in Brief, 2021, 39: 107521.

[20] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 1440-1448.

[21] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice. IEEE, 2017: 2999-3007.

[22] Jocher G, Stoken A, Borovec J, et al. Ultralytics/yolov5: v3.0-nn. SiLU () activations, weights & biases logging, PyTorch hub integration [EB/OL]. (2020-08-13)[2025-11-18]. https://zenodo.org/records/3983579.

[23] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//Computer Vision-ECCV 2020-16th European Conference. Glasgow: ECVA, 2020: 474-490.

[24] Zhao Y, Lyu W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16965-16974.

(责编&校对 雷建云)