中图法分类号:TP18; TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-14

论文引用格式:Li Zhe, Luo Jing, Liu Yu, Shi Weiwei, Gao Binzhi, Wang Xiaofan. Temporal Motion-Aware Dual-Branch Network for Video Tiny UAV Object Detection TMAD-Net[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260209.(李哲, 罗靖, 柳宇, 高彬智, 石伟伟, 王晓帆, 黑新宏. 面向视频微小无人机目标检测的时序运动感知双分支网络TMAD-Net[J/OL]. 中国图象图形学报, XXXX:1-14. DOI: 10.11834/jig.260209.)[DOI:10.11834/jig.260209]

面向视频微小无人机目标检测的时序运动感知双分支网络TMAD-Net

李哲[1,2],罗靖[1,2*],柳宇[1,2],高彬智[1,2],石伟伟[1,2],王晓帆[1,2],黑新宏[1,2]

1. 西安理工大学计算机科学与工程学院,西安　710048; 2. 陕西省网络计算与安全技术重点实验室,　西安　710048

摘　要:**目的**　针对实际视频监控场景中,无人机目标像素占比小、外观特征微弱,易被复杂动态背景掩盖,传统单帧目标检测算法无法有效利用帧间时序运动信息,导致微小目标漏检、误检频发的核心问题,开展视频微小无人机目标检测方法研究,为低空安防系统提供可靠的视觉检测方案。**方法**　提出时序运动感知双分支网络(Temporal Motion-Aware Dual-Branch Network, TMAD-Net)。具体地,空间语义提取分支提取多帧堆叠图像的空间语义特征,显式运动先验分支通过去噪帧差图像捕捉目标高频运动特征,并使用运动-空间自适应融合模块融合双分支特征,从运动与空间维度自适应增强目标信号、抑制背景噪声,后输入主干网络完成微小无人机目标检测。**结果**　在公开的ARD-MAV数据集以及真实场景采集的Phone与DJI两个无人机数据集上开展对比实验,结果表明,在ARD-MAV数据集中,mAP50从基线的0.264提升至0.588,mAP50-95从0.155提升至0.334;在Phone数据集中,mAP50从基线的0.802提升至0.887,mAP50-95从0.590提升至0.638;在DJI数据集中,mAP50从0.316提升至0.822,mAP50-95从0.083提升至0.266;推理速度达122FPS,满足实时检测需求。**结论**　所提时序运动感知双分支网络模型通过显式分离并深度融合多帧空间与运动特征,结合运动-空间自适应融合模块精准校准,有效弥补了传统单帧算法对微小目标特征表征能力不足的缺陷,突破了复杂背景下视频微小目标检测的性能瓶颈,显著提升了微小无人机目标检测的准确率与鲁棒性。

关键词:视频目标检测;无人机;微小目标;运动解耦;时序运动感知;空间语义提取

Temporal Motion-Aware Dual-Branch Network for Video Tiny UAV Object Detection TMAD-Net

Li Zhe[1,2], Luo Jing[1,2*], Liu Yu[1,2], Shi Weiwei[1,2], Gao Binzhi[1,2], Wang Xiaofan[1,2]

1. Hei Xinhong; 2. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China; 3. Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China

Abstract: **Objective** With the rapid proliferation of unmanned aerial vehicles (UAVs) in civil and commercial fields, the demand for robust low-altitude security, airspace management, and intelligent visual surveillance systems has grown exponentially. Precise and real-time visual detection of UAV targets in continuous video streams is the core link of these systems. However, in practical uncontrolled surveillance scenarios, UAVs are usually located at extreme distances from the

收稿日期:2026-04-15;**修回日期**:2026-06-29

*通信作者:罗靖　E-mail:luojing@xaut.edu.cn

基金项目:国家自然科学基金(U25A20451,62476217);陕西省自然科学基础研究计划(2025JC-YBQN-806)

Supported by: National Natural Science Foundation of China(U25A20451,62476217); Natural Science Basic Research Program of Shaanxi (Grant No. 2025JC-YBQN-806)

camera, occupying only a tiny fraction of pixels on the imaging plane, thus lacking clear morphological outlines, observable textures, and structural details. They are highly susceptible to being obscured by complex and dynamic backgrounds such as moving clouds, fluctuating lighting, and cluttered ground landscapes. Traditional single-frame object detection algorithms segment continuous video streams into isolated static images for frame-by-frame processing, fundamentally ignoring the crucial temporal motion information between frames. This leads to severe representation limitations when facing extremely small targets, with persistently high missed detection and false alarm rates. Existing temporal detection methods based on 3D convolution or Transformer can capture temporal features, but their huge parameter amount and computational overhead make them impractical for deployment on resource-constrained real-time security systems. This paper aims to break through the performance bottleneck of pure spatial detection frameworks and achieve high-precision and real-time detection of tiny UAV targets in complex backgrounds. **Method** This paper proposes a Temporal Motion-Aware Dual-Branch Network (TMAD-Net). A dual-branch decoupling design is adopted at the data input end to physically separate spatiotemporal multimodal information. Specifically, the spatial semantic extraction branch processes multi-frame stacked raw RGB images to extract rich spatial appearance, local texture, and global contextual semantic features of the target and its surrounding environment. The parallel explicit motion prior branch is dedicated to processing inter-frame difference images between adjacent frames that have undergone absolute value operation and morphological denoising filtering. This branch filters out static background interference, camera jitter, and sensor noise, and generates a high-purity motion saliency map that highlights the high-frequency temporal motion trajectory of the flying UAV. The two data streams independently pass through the shallow network to complete high-dimensional feature extraction, and are then cascaded and concatenated along the channel dimension to maximally preserve multimodal spatiotemporal information. Subsequently, a motion-spatial adaptive fusion module is introduced to perform adaptive weight calibration on the fused features from both motion and spatial dimensions, so as to enhance the feature response of weak target signals and suppress irrelevant background noise. Finally, 1×1 convolution is applied to accomplish cross-channel dimensionality reduction and feature mapping, and the calibrated spatiotemporal features are seamlessly fed into the subsequent backbone network to complete target detection. **Result** Extensive comparative tests and ablation experiments were conducted on the public ARD-MAV dataset and two private UAV datasets (Phone and DJI) collected in real-world scenarios, which are characterized by small target scales, varying illumination conditions, and low signal-to-noise ratios. Quantitative experimental results show that, compared with the standard baseline model, the proposed method achieves steady performance improvement on the ARD-MAV dataset: the mean average precision at 50% intersection over union (mAP50) increases from 0.264 to 0.588, and the mean average precision over the intersection over union range of 50% to 95% (mAP50-95) increases from 0.155 to 0.334; on the Phone dataset: the mAP50 increases from 0.802 to 0.887, and the mAP50-95 increases from 0.590 to 0.638. On the highly challenging DJI dataset, the proposed model achieves a remarkable performance improvement: the mAP50 surges from 0.316 of the baseline to 0.822, with a performance gain of over 160%, and the mAP50-95 jumps from 0.083 to 0.266. Compared with mainstream lightweight detection algorithms including YOLOv9s, YOLOv11s, and YOLOv12s, the proposed model achieves optimal detection accuracy on all three datasets. Even compared with the larger-scale RT-DETR-L model based on the Transformer architecture, the proposed model still achieves better overall performance in both detection accuracy and inference speed. Meanwhile, the proposed model has a parameter count of 11.33M, which is basically consistent with the 11.13M of the baseline model. It reaches an inference speed of 122 frames per second (FPS), maintaining excellent real-time inference capability while realizing a steady improvement in accuracy. Ablation experiments further verify the effectiveness of the dual-branch architecture, spatial semantic extraction branch, and motion-spatial adaptive fusion module respectively, proving the rigor of the design logic of the proposed architecture. **Conclusion** The TMAD-Net proposed in this paper successfully compensates for the inherent limitation of insufficient feature extraction capability of traditional single-frame detection algorithms when facing tiny targets. Through explicit separation and in-depth fusion of multi-frame RGB spatial appearance features and computationally optimized motion features, combined with the precise calibration capability of the motion-spatial adaptive fusion module, this method effectively isolates real targets from heavily cluttered and dynamic backgrounds, and effectively alleviates the long-standing performance bottleneck of video small target detection in complex real-world environments. Empirical results confirm that the proposed model significantly improves the

detection accuracy and algorithmic robustness of UAV target recognition under non-ideal conditions, and provides a feasible, effective and novel visual detection paradigm for real-time video surveillance, public safety monitoring, and advanced low-altitude defense systems.

Key words: video object detection (VOD); unmanned aerial vehicle (UAV); tiny object; motion decoupling; temporal motion perception; spatial semantic extraction

0 引　言

随着航空电子技术发展，无人机(Unmanned Aerial Vehicle, UAV)在军事侦察、农业植保、物流运输、航拍测绘等军用与民用领域的应用呈现出高速的增长态势(周生辉 等，2026)。然而，微型无人机小巧灵活、获取便捷，不法分子常利用其开展越界侦察、偷拍窃密、违规空投，乃至恐怖袭击，安全风险持续走高。在机场净空区、核电站等关键敏感区域，“黑飞”无人机对公共安全和国家安全构成了严重威胁。因此，构建高效实时的低空安防与无人机反制系统，已成为当前学术界与工业界共同攻克的重大安全课题。

基于计算机视觉的视频目标检测抗干扰能力强、可输出方位与类别信息，是低空防御的核心技术。但实际低空监控场景中，微小无人机的精准检测仍面临严峻挑战，核心难点体现在两方面：

首先是微小尺度导致的弱特征属性。根据国际通用目标检测数据集MS COCO的定义，分辨率小于32×32像素的目标即被划分为小目标(Lin 等，2014)。在实际监控场景中，为追求更广阔的监控视野，摄像机通常采用广角镜头，当无人机飞行高度较高或距离摄像机较远时，其在成像平面上往往仅占据几个到十几个像素(Liang 和 Luo，2024)。尺度收缩使目标丢失纹理结构细节，经网络多层下采样后特征易被不可逆地湮没，纯空间维度优化存在性能瓶颈。

其次是复杂且动态变化的背景干扰。低空无人机的飞行背景通常包括动态飘动的云层、光照剧烈变化的天空、以及包含树木、建筑物、高压线塔等高频纹理的杂乱地面景观。对于缺乏显著外观特征的微小目标而言，其像素灰度或颜色分布易与复杂背景产生严重混叠，形成“视觉伪装”(Zhao 等，2021)。在强光干扰、逆光或雾霾等恶劣气象条件下，目标的对比度进一步下降，信噪比低，此时任何依赖纯空间纹理分析的检测算法都将面临失效的风险，漏检与误检率会急剧上升。

上述两方面挑战表明，在空间维度上对像素特征进行增强的方法存在本征局限，亟需引入空间之外的补充信息源。人类视觉系统之所以能够在杂乱的背景中迅速捕捉到远处的飞鸟或无人机，并非仅仅依赖目标的静态外观，而是高度依赖目标相对于静态或缓慢变化背景的独立运动(Wang 等，2021)。在面对伪静态背景下、几乎没有空间纹理的微小目标时，运动线索通常是区分目标与背景的可靠的依据。然而，以YOLO系列(Redmon 等，2016；Redmon 和 Farhadi，2018)为典型代表的单帧空间检测模型，将连续的视频割裂为孤立的静态图像帧进行逐帧处理，忽视了视频数据中蕴含的帧间时序运动线索，制约了微小目标检测精度(Wang 等，2022)。

针对视频目标检测，一些研究开始尝试引入3D卷积神经网络(3D CNN)(Chen 等，2024)、长短时记忆网络(Long Short-Term Memory Network, LSTM)(Hochreiter 和 Schmidhuber，1997；Shi 等，2015)或基于Transformer的时空注意力机制来隐式地学习时序特征(Vaswani 等，2017)。然而，在计算资源受限的实时低空安防场景中，3D卷积与Transformer模型因参数量庞大、计算开销过高，难以满足实时部署需求。因此，如何在充分利用帧间时序运动线索的同时保持模型的轻量化与实时性，是本文所要解决的核心问题。

针对上述痛点，本文提出了时序运动感知双分支网络TMAD-Net。该网络通过空间语义信息与时序运动信息的解耦式提取与自适应融合，充分捕捉连续视频帧间的目标运动线索，强化网络对微弱目标信号的感知能力。同时，网络在架构设计中全程兼顾轻量化与计算效率，在不显著增加模型参数量与计算开销的前提下，实现复杂背景下微小无人机目标的高精度实时检测。

本文的主要贡献总结如下：1)设计了空间语义提取分支以充分挖掘连续视频帧间的空间语义关

联,并且设计了显式运动先验分支以精准捕获无人机目标的高频运动特征;2)构建了运动-空间自适应融合模块,从运动与空间维度自适应校准双分支融合特征的权重,有效增强目标特征信号并抑制复杂背景噪声;3)提出时序运动感知双分支网络TMAD-Net,采用空间语义提取分支与显式运动先验分支并行的解耦式架构,解决了传统单帧检测算法割裂视频时序连续性、对微弱目标特征表征能力不足的问题。

1 相关工作

1.1 基于深度学习的小目标检测

目标检测作为计算机视觉的核心分支,在过去十年间经历了从手工特征(如SIFT、HOG)到深度学习特征的跨越式发展(Simonyan 和 Zisserman,2015)。以Faster R-CNN(Ren 等,2017)为代表的双阶段检测算法和以YOLO、SSD (Liu 等,2016)为代表的单阶段检测算法在常规尺度物体检测上已取得令人瞩目的成就。然而,小目标检测(Small Object Detection, SOD)始终是该领域公认的难点与痛点。

在深层卷积神经网络中,小目标在经过多层卷积与池化操作后,其在深层特征图上的感受野极速扩张,自身特征易被周围的背景噪声所同化,为了缓解这一问题,研究人员提出了多种优化策略。

第一类是多尺度特征融合机制:特征金字塔网络(Feature Pyramid Network, FPN)提出的深浅层特征融合是经典方案(Lin 等,2017),后续PANet(Liu 等,2018)、BiFPN(Tan 等,2020)等结构进一步强化了自底向上的信息流传递;吴军等人(2026)提出SOD-YOLO检测网络,通过特征适配与尺度优化针对性提升监控画面中小目标的表征能力。第二类是注意力机制:Woo等人(2018)提出的CBAM(Convolutional Block Attention Module)模块将通道注意力与空间注意力串联,使得网络能够自适应地聚焦于包含小目标的关键区域;冯琪涵等人(2025)提出融合前景细化与多维归纳偏置自注意力机制的方法,通过增强目标区域表征能力和抑制背景噪声,有效提升了小目标检测性能。第三类是上下文信息的利用:通过建模目标与邻域的空间关联、引入场景先验,借助全局上下文辅助判别,可缓解复杂场景下的误检漏检(Liang 和 Luo,2024)。第四类是基于生成对抗网络(Generative Adversarial Network, GAN)(Li 等,2017)的超分辨率重建技术:该类方法通过生成器将低分辨率特征映射至高维空间弥补细节,但算力开销大且训练不稳定,难以满足实时性要求。

尽管上述方法在一定程度上缓解了小目标特征消失的问题,但对于融入静态背景的远距离微小目标,纯空间维度的像素级增强仍难以避免大量漏检与误检,引入能够反映目标独立运动的时序线索是提升检测性能的重要方向。

1.2 视频目标检测与时序特征提取

视频目标检测(Video Object Detection, VOD)通过挖掘时序相关性提升单帧性能,核心优势是利用运动线索打破视觉伪装,突破静态外观局限。

早期方法多基于后处理策略,例如Seq-NMS(Han 等,2016)通过将相邻帧的检测框进行图层链接与重评分,过滤掉时序上孤立的误检框。然而,这类方法高度依赖单帧检测器的输出,无法从根本上提升对微弱目标的特征提取能力。随后,以光流法为代表的特征级聚合方法成为主流。Zhu 等人(2017)提出的FGFA(Flow-Guided Feature Aggregation)利用光流网络计算相邻帧的运动场,并据此对相邻帧的特征图进行形变对齐与加权聚合。这种方法有效地提升了视频中运动模糊和遮挡目标的检测精度,但光流计算开销大,严重制约了算法的实时性。

近年来,随着硬件算力的提升,可在空间与时间维度同步滑动并显式提取时空联合特征的3D卷积网络,已被广泛应用于视频动作识别与检测领域。此外,基于Transformer的视频架构开始崭露头角,如TransVOD(Zhou 等,2022)利用跨帧注意力机制在全局时空域内建立像素级的长距离依赖。然而,这类前沿检测范式通常模型体量较大,难以直接部署于资源受限的监控端侧设备。

与之相比,帧差法处理简单、计算效率显著,且对高频运动边缘高度敏感(Lipton 等,1998)。将帧差图作为深度网络的显式特征输入,引导网络从中提取运动特征并自适应过滤噪声,是一种在保持轻量化的同时显式利用时序信息的极具潜力的方案。

1.3 多模态融合与双分支网络架构

在计算机视觉中,当单一模态(如RGB图像)的信息不足以支撑复杂场景下的高精度感知时,常引入第二模态并进行跨模态融合(Cross-Modal

Fusion)。典型的应用包括RGB-D(深度图)显著性检测以及RGB-T(热红外)目标检测。

按网络融合阶段划分,跨模态融合主要包括早期融合(Early Fusion)、晚期融合(Late Fusion)以及特征级融合(Feature-level Fusion)。早期融合在输入层拼接多模态通道特征,可实现信息初步交互,但会破坏模态独立性与语义特性,造成模态语义干扰、训练收敛困难。晚期融合采用各单模态网络分别推理,再通过加权融合检测结果,虽然规避了模态干扰,但忽视了底层特征互补关联,无法充分发挥多模态融合优势。

特征级融合中的双分支架构(Dual-Branch Architecture)能够有效平衡模态特异性与信息交互的需求,在保留各模态固有特性的同时,实现多模态信息的深度融合与互补(Bi 等,2023)。例如,Cong 等人(2023)在RGB-D检测中提出了一种双分支网络,通过空间注意力模块,利用深度图提供的几何轮廓信息去引导过滤杂乱的背景纹理。Xiao 等人(2024)进一步证明,在浅层独立提取异构特征,而在中层进行特征拼接与降维交互,既能避免早期融合的语义混淆,又能比晚期融合更好地利用多模态的协同优势。

受上述研究的启发,本文设计了空间语义提取分支与显式运动先验分支并行的双分支架构,在保留空间语义完整性的同时充分利用帧间运动线索,实现复杂背景下微小无人机目标的高精度实时检测。

2 方　法

2.1 总体网络架构

本文构建的时序运动感知双分支网络TMAD-Net,核心由空间语义提取分支、显式运动先验分支以及运动-空间自适应融合模块三部分组成。

在数据输入端,给定连续输入的视频帧序列,网络首先通过通道堆叠(channel stacking)构建空间语义提取分支张量;同时对相邻帧进行灰度化处理,计算绝对帧差构建显式运动先验分支张量。随后,双分支的高维特征在通道维度进行拼接,并通过运动-空间自适应融合模块动态校准运动与空间权重,经过1×1卷积降维后送入主干网络进行预测。如图1所示。

2.2 空间语义提取分支

为了在视频序列中捕获微小目标的上下文信息,本文在网络输入端进行多帧图像的通道堆叠。设有连续的 n 帧图像序列集合 $\boldsymbol{I}=\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_i,\cdots,\boldsymbol{I}_n\}$,其中 $\boldsymbol{I}_i$ 表示第 i 帧图像,$\boldsymbol{I}_i\in\mathbf{R}^{(H\times W\times 3)}$。本文通过在通道维度拼接 k 帧图像,构造空间语义提取分支的输入张量 $\boldsymbol{X}_{\text{rgb}}$,如式(1)。

$$\boldsymbol{X}_{\text{rgb}}=\text{Concat}(\boldsymbol{I}_i,\boldsymbol{I}_{i+1},\cdots,\boldsymbol{I}_{i+k-1})\in\mathbf{R}^{(H\times W\times 3k)} \tag{1}$$

式中,Concat表示沿通道维度的级联操作;k 为堆叠的帧数。$\boldsymbol{X}_{\text{rgb}}$ 输入到由两层Conv和一层C2f模块组成的空间特征提取模块中,得到高维空间外观特征 $\boldsymbol{F}_{\text{rgb}},\boldsymbol{F}_{\text{rgb}}\in\mathbf{R}^{(H'\times W'\times C)}$,至此完成空间语义提取分支的特征提取。

在检测任务中,对于多帧输入产生的联合特征,必须指定一个明确的"监督帧"(以哪一帧的真实标注即Ground Truth作为计算损失的基准)。为了探究时序感受野的最佳配置,本文在实验中系统对比了超参数 k 的不同值对模型精度的影响,并分别针对监督第一帧、中间帧和最后一帧三种场景开展了严谨的对照试验。最终选择了堆叠帧数为3且以最后一帧作为监督基准的策略,为双分支网络提供最优的基线特征。

2.3 显式运动先验分支

采用多帧堆叠的方式虽能提供丰富的时空上下文信息,但往往会致使网络难以从繁杂的静态背景信息中精准剥离出微弱的运动目标信号。针对该局限,本文设计了独立的显式运动先验分支,以实现运动目标信息与静态背景信息的精准分离,为后续网络的特征学习提供可靠支撑。

给定连续的三帧彩色图像 $\boldsymbol{I}_1,\boldsymbol{I}_2,\boldsymbol{I}_3$,首先将其转换为单通道灰度图像 $\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{G}_3$,以消除光照颜色变化带来的冗余干扰。随后,利用相邻帧计算绝对差分图像,如式(2)式(3)。

$$\boldsymbol{D}_1=|\boldsymbol{G}_1-\boldsymbol{G}_2| \tag{2}$$

$$\boldsymbol{D}_2=|\boldsymbol{G}_2-\boldsymbol{G}_3| \tag{3}$$

考虑到监控场景中常伴有传感器底噪、树叶微小晃动等高频噪声,直接使用绝对帧差图容易产生伪影并引发误检。因此,本文对生成的帧差图进行高斯平滑去噪处理与形态学处理(含膨胀、腐蚀操作)得到 $\boldsymbol{D}_1',\boldsymbol{D}_2'$。具体处理细节如下:高斯平滑采用5×5的高斯核,标准差设为1.5;形态学处理先执行

闭运算(先膨胀后腐蚀)以填补目标内部空洞,再执行开运算(先腐蚀后膨胀)以消除孤立噪声点,两次运算均采用3×3的矩形结构元素,迭代次数均为1。

处理后的单通道帧差图在通道维度拼接构造显式运动先验分支的输入张量$\boldsymbol{X}_{\text{diff}}$,如式(4)。

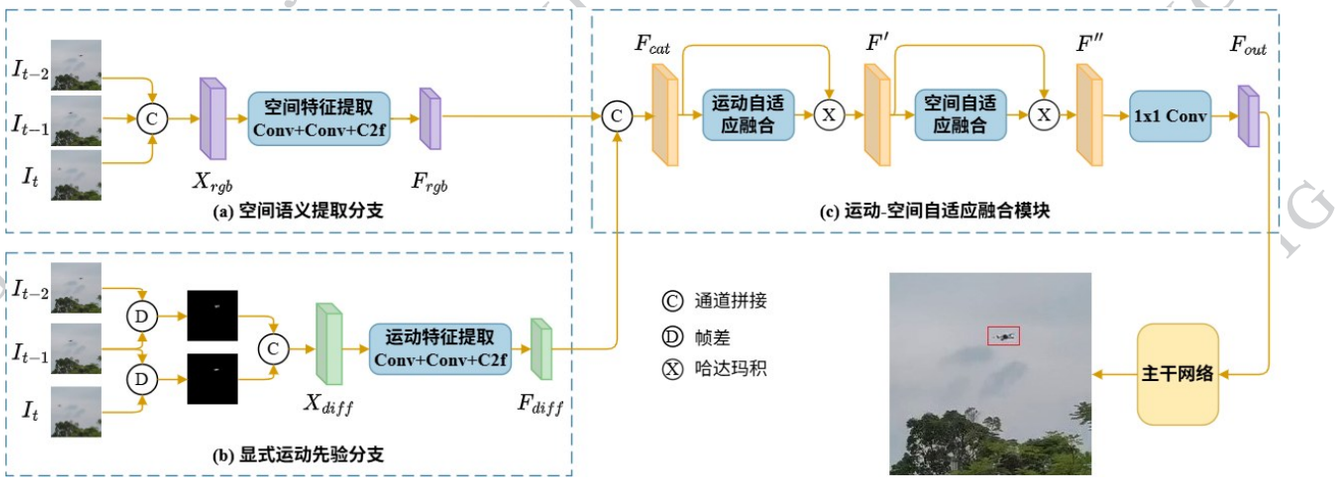


图1 时序运动感知双分支网络结构示意图

Fig. 1 Architecture of temporal motion-aware dual-branch network

$$\boldsymbol{X}_{\text{diff}} = \text{Concat}(\boldsymbol{D}_1', \boldsymbol{D}_2') \in \mathbf{R}^{(H \times W \times 2)} \tag{4}$$

$\boldsymbol{X}_{\text{diff}}$输入到由两层Conv和一层C2f模块组成的运动特征提取模块中,得到高频运动特征$\boldsymbol{F}_{\text{diff}}$,$\boldsymbol{F}_{\text{diff}} \in \mathbf{R}^{(H' \times W' \times C)}$,至此完成显式运动先验分支的特征提取。该分支信息能够有效过滤静态背景,迫使网络高度聚焦于目标的运动区域。

需要说明的是,空间语义提取与显式运动先验分支之所以采用相同的拓扑结构,一是保持架构简洁性与轻量化,避免为不同分支设计专用结构带来的额外计算开销与部署复杂度;二是消除结构差异带来的性能干扰,使两个分支的功能差异完全由输入模态与训练权重决定,更严谨地验证显式运动先验的独立贡献。但二者功能分化的核心机理在于输入模态本质差异与端到端训练驱动的特征分化。空间语义提取分支输入多帧堆叠的原始RGB图像,包含丰富静态空间信息;显示运动先验分支输入经预处理的运动显著性图,仅保留帧间动态变化区域,两者需要不同的特征提取策略,独立权重可实现更精准的特征提取。这种结构统一、输入异构、权重独立的设计,在保持架构简洁性的同时,实现了空间与运动特征的精准解耦,避免了模态间语义干扰。

2.4 运动-空间自适应融合模块

为实现多模态信息的深度交互,本文首先将空间语义提取分支得到的特征$\boldsymbol{F}_{\text{rgb}}$与显式运动先验分支得到的特征$\boldsymbol{F}_{\text{diff}}$拼接得到联合特征$\boldsymbol{F}_{\text{cat}}$,如式(5)。

$$\boldsymbol{F}_{\text{cat}} = \text{Concat}(\boldsymbol{F}_{\text{rgb}}, \boldsymbol{F}_{\text{diff}}) \in \mathbf{R}^{(H' \times W' \times 2C)} \tag{5}$$

随后,引入运动-空间自适应融合模块进行加权融合,该模块依次通过运动自适应融合模块和空间自适应融合模块。其中,运动自适应融合模块通过全局平均池化和全局最大池化聚合运动信息,送入共享的多层感知机计算运动权重;空间自适应融合模块则在通道维度上分别进行全局平均池化与全局最大池化,将结果拼接后通过卷积层生成空间权重。

全局平均池化聚合全局运动强度分布,抑制云层飘动、树叶晃动等大面积均匀背景噪声,生成全局一致的运动权重基线;全局最大池化提取局部运动峰值响应,精准捕捉微小无人机的高频运动边缘,避免微弱目标特征被全局平均平滑。二者结合可实现全局背景抑制与局部目标增强的精准校准,有效提升复杂动态背景下的检测性能。

设运动权重为$\boldsymbol{M}_{\text{m}}$,空间权重为$\boldsymbol{M}_{\text{s}}$,经过运动自适应融合模块加权后的特征图为$\boldsymbol{F}'$,经过空间自适应融合模块加权后的特征图为$\boldsymbol{F}''$,则其特征表示可由式(6)式(7)式(8)式(9)给出。

$$\boldsymbol{M}_{\text{m}}(\boldsymbol{F}_{\text{cat}}) = \sigma\begin{pmatrix}\text{MLP}(\text{AvgPool}(\boldsymbol{F}_{\text{cat}})) + \\ \text{MLP}(\text{MaxPool}(\boldsymbol{F}_{\text{cat}}))\end{pmatrix} \tag{6}$$

$$\boldsymbol{F}' = \boldsymbol{M}_{\text{m}}(\boldsymbol{F}_{\text{cat}}) \otimes \boldsymbol{F}_{\text{cat}} \tag{7}$$

$$\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{F}')=\sigma\left(f^{7\times7}\left(\begin{bmatrix}\mathrm{AvgPool}(\boldsymbol{F}');\\ \mathrm{MaxPool}(\boldsymbol{F}')\end{bmatrix}\right)\right) \tag{8}$$

$$\boldsymbol{F}''=\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{F}')\otimes\boldsymbol{F}' \tag{9}$$

式中，AvgPool与MaxPool分别代表全局平均池化与最大池化操作，MLP为多层感知机，$f^{7\times7}$为7×7尺寸的卷积运算，σ为Sigmoid激活函数，$\otimes$表示哈达玛积。最后，通过一个1×1卷积层进行跨通道降维映射，如式(10)。

$$\boldsymbol{F}_{\mathrm{out}}=\mathrm{Conv}_{1\times1}(\boldsymbol{F}'')\in\mathbf{R}^{(H'\times W'\times C)} \tag{10}$$

通过该模块，网络能够自适应地抑制静态背景带来的冗余特征，并增强运动区域的空间响应。降维后的融合特征$\boldsymbol{F}_{\mathrm{out}}$输入后续目标检测的主干网络，完成端到端的小目标检测。

3 实验结果与分析

3.1 数据集与评价指标

为验证TMAD-Net的有效性，在公开的ARD-MAV数据集以及两个真实监控场景下采集的私有无人机小目标数据集(Phone数据集与DJI数据集)上进行了对比与消融实验。

ARD-MAV数据集包含60段视频序列，总计106665帧图像，所有视频均由DJI Mavic2 Pro与M300型号无人机，在低空飞行环境下拍摄采集。每段视频仅包含单个微型飞行器目标，时长约1分钟，帧率30帧/秒，分辨率为1920×1080。数据集内目标尺寸跨度为6×3至136×75像素，平均尺寸仅占整张图像面积的0.02%。其中选取45段视频用于训练和验证，剩余15段视频用作测试。视频全部取自户外真实场景，涵盖多种现实难点：复杂背景、目标遮挡、相机剧烈抖动、高速运动以及微小尺寸。

Phone数据集使用华为P80+拍摄，共有14个视频序列，8397帧，分辨率为1280×720，目标尺寸跨度为5×3到41×22，平均尺寸仅占整张图像面积的0.02%。其中12个视频用于训练和验证，2个视频用于测试。DJI数据集使用DJI设备拍摄，共有14个视频序列，8387帧，分辨率为1920×1080，目标尺寸跨度为2×2到33×33，平均尺寸仅占整张图像面积的0.01%。其中12个视频用于训练和验证，2个视频用于测试。两个数据集中的每个视频均以30帧/秒的帧率进行记录。数据集中无人机目标通常距离摄像机较远，像素占比较低，在目标尺寸上相较于现有的公开数据集几乎做到了最小，背景包含了云层、树木、建筑等干扰，具有较高的挑战性。如表1和图2所示。

评估指标方面，本文采用目标检测领域通用的交并比(Intersection over Union，IoU)阈值为0.5时的平均精度均值(mAP50)和IoU阈值在0.5至0.95区间的多阈值平均精度均值(mAP50-95)作为核心评价指标。

3.2 实验环境与参数设置

本文实验基于PyTorch深度学习框架完成模型开发与部署，实验硬件采用NVIDIA GeForce RTX 4090 D显卡。训练阶段，在ARD-MAV数据集上网络输入图像尺寸均采用640分辨率，在Phone与DJI数据集上网络输入图像尺寸均为1280分辨率，采用Adam优化器进行梯度更新，初始学习率(lr0)设定为0.001，权重衰减系数(weight_decay)设定为0.0005，并启用自动混合精度(Automatic Mixed Precision，AMP)以加速训练。Mosaic与CutMix是为单帧静态检测设计的增强策略，本文依赖连续多帧的真实运动连续性，其跨图像拼接与区域裁剪替换会生成伪时序序列，严重干扰显式运动先验分支的运动特征建模，为确保多帧序列的物理一致性，训练过程中连同对比的模型也一同关闭了Mosaic和CutMix等强空间几何扭曲的数据增强策略。

3.3 主流算法的综合性能对比实验

本文在实验中采用YOLOv8s模型作为主干网络测试了TMAD-Net的性能。为全面评估TMAD-Net的检测性能与工程落地可行性，在公开的ARD-MAV数据集以及私有的Phone与DJI数据集上，将TMAD-Net与当前主流的轻量级单阶段检测算法YOLOv8s、YOLOv9s、YOLO11s、YOLO12s以及Transformer架构检测算法RT-DETR-L等展开对比实验。实验选取目标检测领域核心精度指标mAP50、mAP50-95作为检测效果评价依据，同时引入模型参数量(Parameters)与每秒推理帧数(FPS)(包含预处理，推理以及后处理时间)两个指标评估模型的计算复杂度与实时推理效率，综合验证所提算法在精度与速度层面的表现，实验结果如表2所示。

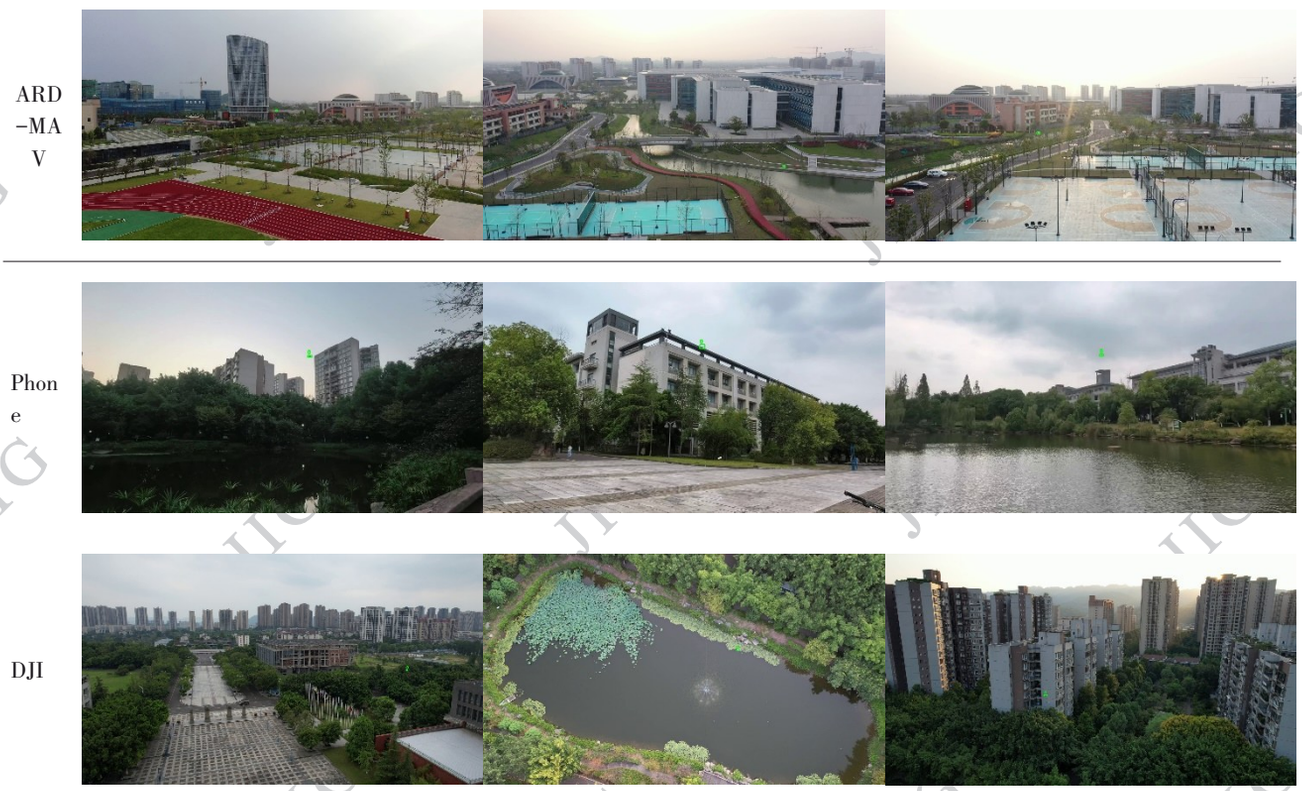
从表2的实验数据分析得出，所提的TMAD-Net在三个数据集的精度指标上均取得最优结果，充分验证了时序运动感知双分支架构在微小无人机目标

检测任务中的有效性与优越性，具体分析如下。

在精度层面：首先在公开数据集ARD-MAV上，TMAD-Net的mAP50和mAP50-95分别达到了0.588和0.334，相较于基线模型YOLOv8s分别提升了约122.7%和115.5%，超越了参数量近3倍的Transformer架构模型RT-DETR-L（mAP50=0.573），

图2　数据集视频帧样例

Fig. 2　Sample video frames of datasets



取得了最优检测性能。TMAD-Net在Phone数据集中的mAP50和mAP50-95分别达到了0.887和0.638，相较于基线模型YOLOv8s分别提升约10.6%和8.1%，且显著优于同属轻量级阵营的YOLO系列算法；即便与参数量更大的RT-DETR-L相比，TMAD-Net的检测精度仍具有明显优势。在目标像素占比更低、背景干扰更剧烈的高挑战性DJI数据集中，TMAD-Net的性能提升尤为显著，mAP50从基线的0.316跃升至0.822，相对提升幅度超过160%，mAP50-95达到0.266。相比之下，同级别YOLO系列模型的特征提取能力明显不足，次优的YOLO11s的mAP50仅为0.469；而大参数量的RT-DETR-L在该场景下甚至出现严重的性能退化，mAP50仅为0.255。这有力地证明了，现有检测算法在应对微小目标与复杂动态背景时存在固有的局限性，而本文引入的时序运动感知机制成功破解了这一难题。

在计算复杂度层面：TMAD-Net的参数量与基线模型YOLOv8s基本持平，仅增加不到2%，完美保持了YOLOs系列特有的轻量化优势。与大参数量模型RT-DETR-L相比，TMAD-Net仅使用了其约1/3的参数量，却实现了更优的性能。这表明所设计的双分支融合架构较为高效，在引入时序运动模块时并没有以牺牲模型轻量化为代价，有效控制了模型的计算复杂度。

在实时推理层面：受限于时序双分支结构的引入，TMAD-Net的FPS相较于同系列YOLO模型有所下降，但122FPS的推理速度仍能满足实际视频

监控与低空防御系统的实时检测需求，实现了检测精度与实时性的良好平衡。

为进一步直观验证不同算法的实际检测效果，本文在三个数据集中进行了可视化对比，检测结果如图3所示。

综上所述，TMAD-Net在保持轻量化架构且未显

著增加计算开销的前提下，不仅全面超越了同量级的YOLO系列主流算法，更在检测精度与推理速度上对大参数量模型（如RT-DETR-L）实现了赶超。特别是在背景复杂、目标微小的严峻场景下，TMAD-Net有效突破了传统单帧检测的性能瓶颈，为微小无人机目标检测提供了一种兼具高精度、高实时性与低算力消耗的工程实用化解决方案。

表1　不同MAV数据集的比较

Table 1　Comparison of different MAV datasets

数据集	帧数	最大尺寸	最小尺寸	平均尺寸
NPS-Drones	70250	6.6e-04	8.2e-05	0.05%
FL-Drones	38948	1.4e-01	2.6e-04	0.07%
DUT Anti-UAV	10109	7e-01	1.9e-06	1.3%
Drones-vs-Bird	104760	2.5e-02	7.2e-06	0.1%
ARD-MAV	106665	3.5e-03	1.4e-05	0.02%
Phone	8397	9.7e-04	1.4e-05	0.02%
DJI	8387	5.3e-04	1.9e-06	0.01%

3.4　多帧堆叠与监督策略实验

为了探索在空间语义提取分支中，多帧通道堆叠的最优配置范式，本文在基线模型（YOLOv8s，单帧输入）的基础上，对超参数k的不同取值，以及监督策略（第一帧、中间帧和最后一帧）开展了详尽的实验。在DJI数据集与Phone数据集上的实验结果如表3所示。

对比基线模型与多帧堆叠的表现可知，纯空间维度的单帧检测在面对目标微小、背景复杂的DJI数据集时存在严重的特征表征失效（mAP50仅为0.316，mAP50-95低至0.083）。而引入多帧通道堆叠后，无论采用何种配置，模型的综合精度均获得了明显提升（mAP50普遍跃升至0.55~0.73之间）。这强有力地证明了，当目标自身空间几何特征极其微弱时，多帧提供的时序运动上下文是区分目标与复杂背景的重要线索。

实验数据表明，引入更长的时序窗口并未带来性能的持续正向增益，反而容易造成时序冗余。在Phone数据集中，k=5各策略的最高mAP50（0.862）未能超越k=3的峰值（0.872）；在更具挑战的DJI数据集中，k=5的各项指标更是出现明显回落（k=5监督最后一帧的mAP50跌至0.556）。其物理机理在于：对于快速不规则运动的微小无人机目标，过长的5帧窗口会引入过大的目标位移差和背景动态噪声，导致网络在通道聚合时发生空间特征对齐困难与运动模糊。而3帧恰好提供了一个短时、紧凑且高相关性的运动感受野，是捕获微弱动态特征的最佳窗口大小。

确定k=3的最优帧数后，以序列中哪一帧作为真值（Ground Truth）进行监督，对模型的回归能力产生了重要影响。在Phone数据集中，监督最后一帧取得了最优的检测性能，其mAP50达到了最高值0.872，相较与监督第一帧和中间帧分别提升了18.3%，10.2%；mAP50-95也达到了次优值0.613，

表2　不同检测方法的性能对比实验

Table 2　Performance comparison of different detection methods

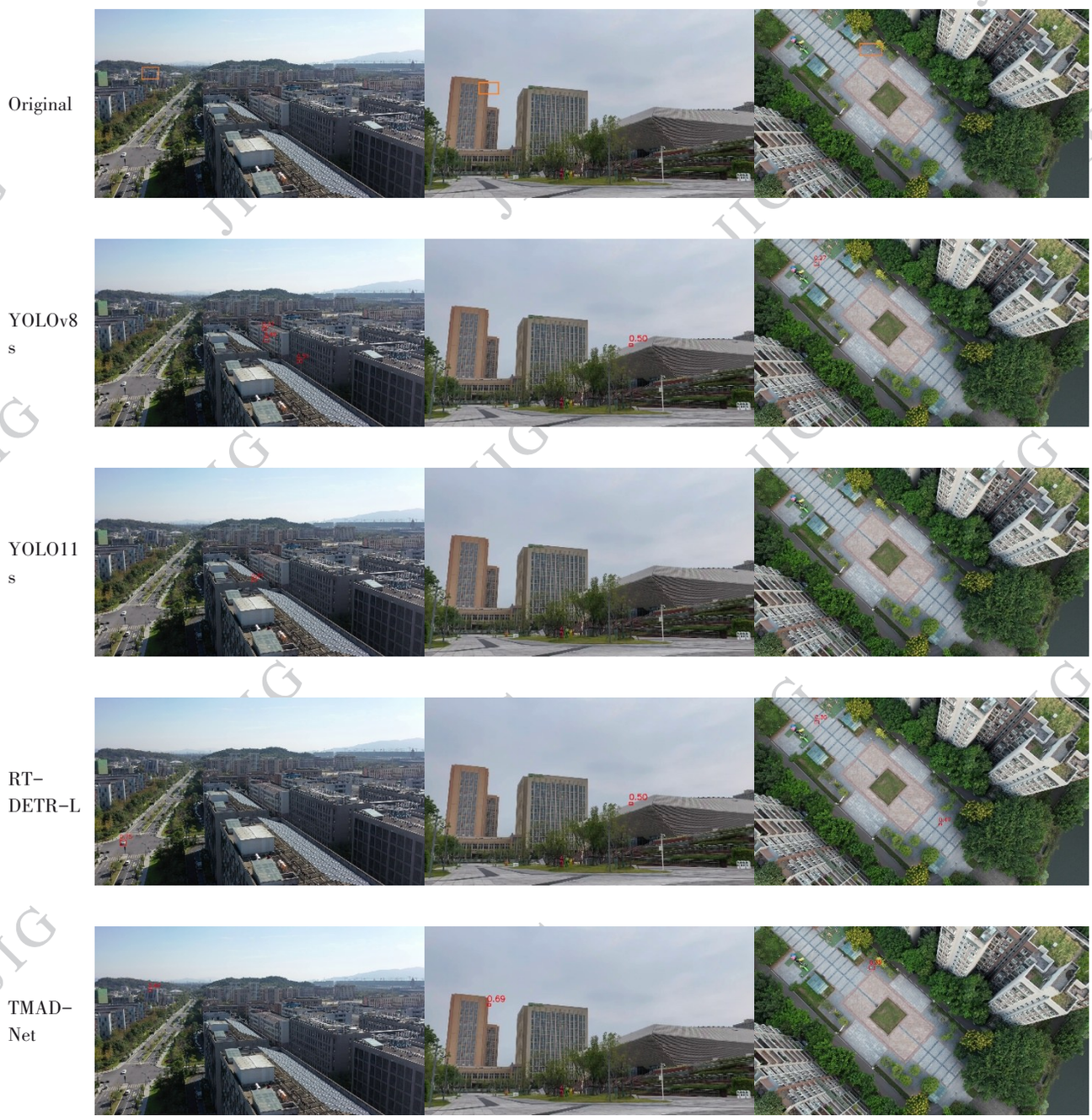
Method	ARD-MAV		Phone		DJI		Parameters	FPS
	mAP50	mAP50-95	mAP50	mAP50-95	mAP50	mAP50-95		
YOLOv8s	0.264	0.155	0.802	0.59	0.316	0.083	11.13M	233
YOLOv9s	0.229	0.135	0.819	0.557	0.327	0.058	7.17M	208
YOLO11s	0.284	0.159	0.761	0.553	0.469	0.131	9.41M	227
YOLO12s	0.290	0.167	0.709	0.469	0.306	0.071	9.23M	141
RT-DETR-L	0.573	0.310	0.840	0.553	0.255	0.059	31.99M	52
D-FINEs	0.308	0.156	--	--	--	--	3.73M	69
DEIM	0.259	0.133	--	--	--	--	38.54M	53
TMAD-Net	**0.588**	**0.334**	**0.887**	**0.638**	**0.822**	**0.266**	11.33M	122

注：加粗字体为每列最优值。

略低于最高值0.617。在DJI数据集中，监督最后一帧同样取得了k=3配置下的最高mAP50(0.730)，

图3　不同目标检测算法的可视化检测结果对比

Fig. 3　Visual comparison of detection results among different models



相较与监督第一帧和中间帧分别提升了4.1%，19.9%；mAP50-95也达到了0.220，略低于最高值0.248。

综上所述，3帧堆叠并监督最后一帧在两个数据集中均展现出了最优的综合检测能力。因此，本文将其确立为最优基线配置，并在此基础上进一步引入显式运动先验分支与运动-空间自适应融合模块，以解决复杂背景下的特征解耦难题。

3.5　消融实验

如前文实验所示，在空间语义提取分支中，多帧

通道堆叠虽然能初步挖掘时序上下文信息，但仅仅依靠帧间通道拼接的方式，无法实现运动特征与静态背景的有效解耦，网络在复杂干扰场景下仍然面临微小运动目标特征提取困难的瓶颈。为此，本文在“3帧输入，监督尾帧”的最优基线配置基础上，引入独立的显式运动先验分支与运动-空间自适应融合模块，构建了时序运动感知双分支网络。为验证所提架构及各核心组件的性能贡献与必要性，本文在ARD-MAV，Phone与DJI三个数据集上，以YOLOv8s为基线模型开展消融实验，实验结果如表4所示。

表3　帧数堆叠与监督策略性能对比实验

Table 3　Performance comparison of different frame stacking and supervision strategies

堆叠帧数k	监督策略	Phone		DJI	
		Map50	mAP50-95	Map50	mAP50-95
Base(k=1)	当前帧	0.802	0.590	0.316	0.083
K=3	第一帧	0.737	0.540	0.701	**0.248**
K=3	中间帧	0.791	0.564	0.609	0.180
K=3	最后帧	**0.872**	0.613	**0.730**	0.220
K=5	第一帧	0.862	0.599	0.669	**0.248**
K=5	中间帧	0.798	**0.617**	0.577	0.162
K=5	最后帧	0.840	0.560	0.556	0.126

注：加粗字体为每列最优值。

在公开基准数据集ARD-MAV上，基线模型YOLOv8s受微小目标特征微弱与复杂动态背景干扰的制约，检测精度偏低，mAP50与mAP50-95仅分别为0.264和0.155。引入空间语义提取分支后，依托多帧堆叠强化时序上下文建模能力，两项指标分别提升至0.301和0.163，证实多帧时序输入能够有效弥补单帧微小目标语义表征不足的缺陷。在此基础上融入显式运动先验分支，通过独立建模帧间运动信息，从杂乱背景中有效分离目标运动轨迹，mAP50与mAP50-95大幅增至0.552和0.280。最后嵌入运动-空间自适应融合模块，自适应学习双分支多模态特征的权重分布，弥补简单拼接带来的语义对齐偏差与融合不充分问题，进一步将精度提升至0.588和0.334，充分验证了各核心组件逐级贡献、协同增益的设计逻辑。在DJI与Phone数据集上，各模块依旧保持一致的性能提升规律，双分支架构均可稳定带来精度增益，佐证了所提方法在不同干扰强度、不同目标尺度场景下的泛化有效性。

值得注意的是，显式运动先验分支在ARD-MAV与DJI数据集上增益显著，但在Phone数据集上提升有限，核心源于不同数据集背景动态程度与背景熵值互补性的差异。一方面，高动态背景会导致空间语义提取分支多帧融合出现严重运动模糊与特征混叠，显式运动分支可通过捕捉目标高频运动轨迹与背景低频运动区分；而Phone数据集低动态背景下，空间语义分支已能有效抑制噪声，运动分支去噪增益大幅削弱。另一方面，高熵值背景的复杂纹理易形成“视觉伪装”，运动特征作为独立线索可有效识别；而Phone数据集背景简单、遮挡较少，运动特征增益有限。

递进式消融实验完整验证了所提架构设计逻辑的严谨性与各模块的必要性：空间语义提取分支为微小目标检测提供了关键的时序上下文信息，解决了单帧场景下微小目标语义信息不足的核心瓶颈；双分支运动解耦架构通过显式运动建模，强化了运动先验引导，实现了运动特征与静态背景的有效解耦；而运动-空间自适应融合模块则实现了多模态特

表4　消融实验

Table 6　Ablation study results

空间语义提取分支	显式运动先验分支	运动-空间自适应融合模块	ARD-MAV		Phone		DJI	
			mAP50	mAP50-95	mAP50	mAP50-95	mAP50	mAP50-95
--	--	--	0.264	0.155	0.802	0.590	0.316	0.083
√			0.301	0.163	0.872	0.613	0.730	0.220
√	√		0.552	0.280	0.877	0.633	0.745	0.230
√	√	√	0.588	0.334	0.887	0.638	0.822	0.266

征的自适应融合，最大化挖掘了双分支架构的性能潜力。最终TMAD-Net在公开基准ARD-MAV数据集上实现了性能翻倍的显著提升，在DJI数据集上实现了约160%的相对性能增益，并且在Phone数据集上也保持了稳定的精度提升，充分验证了融合时序运动感知的双分支架构的有效性与工程价值。

4 讨 论

尽管TMAD-Net在常规视频微小无人机检测中取得了显著突破，但其核心依赖运动线索区分目标与背景的设计逻辑，使其在三类极端场景下存在明确性能边界：目标完全静止时，显式运动先验分支无有效响应，模型退化，性能回落，微小目标场景漏检严重，这是运动驱动算法的一个局限性；密集多目标交叉飞行时，不同目标的运动区域会发生混叠，导致运动先验无法精准分离个体边界，融合权重校准出现偏差，边界框回归误差增大甚至出现合并检测；剧烈相机全局抖动时，大面积背景运动噪声无法被简单滤波完全消除，会淹没目标微弱运动信号，引发大量虚假响应与误检。上述边界明确了本方法的适用范围，后续可通过实例级运动分割与全局运动补偿模块等进一步拓展场景适应性。

5 总 结

针对视频中微小无人机目标外观特征微弱、易被复杂背景掩盖的检测难题，本文提出了一种时序运动感知双分支网络TMAD-Net。TMAD-Net通过空间语义提取分支与显式运动先验分支的协同作用，实现了空间语义与运动特征的有效解耦与互补，并结合运动-空间自适应融合模块解决了异构特征融合过程中的语义对齐难题。实验结果表明，该方法显著提升了检测性能。在公开基准ARD-MAV数据集上，核心指标mAP50从基线模型的0.264提升至0.588，实现了约123%的性能增益，充分证明了通过显式建模时序运动信息来增强微弱特征表征能力的必要性与有效性。同时，在私有的Phone与DJI数据集上也保持了稳定的精度提升，进一步验证了其泛化性，为复杂场景下的视频微小目标检测研究提供了严谨的理论依据与实证支撑。

参考文献(References)

Zhou S H, Rong C Z and Wang H L. 2026. A lightweight YOLOv8-based small object detection algorithm in UAV scenarios. Journal of Signal Processing, 42(1): 72-82 (周生辉, 荣传振, 王华力. 2026. 无人机场景下基于轻量化YOLOv8的小目标检测算法. 信号处理, 42(1): 72-82) [DOI: 10.12466/xhcl.2026.01.007]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. 2014. Microsoft COCO: Common objects in context // Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liang B C and Luo H. 2024. MEANet: an effective and lightweight solution for salient object detection in optical remote sensing images. Expert Systems with Applications, 238: #121778 [DOI: 10.1016/j.eswa.2023.121778]

Zhao X K, Li M L, Zhang G, Li N and Li J S. 2021. Object detection method based on saliency map fusion for UAV-borne thermal images. Acta Automatica Sinica, 47(9): 2120-2131 (赵兴科, 李明磊, 张弓, 黎宁, 李家松. 2021. 基于显著图融合的无人机载热红外图像目标检测方法. 自动化学报, 47(9): 2120-2131) [DOI: 10.16383/j.aas.c200021]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: Unified, real-time object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2018. YOLOv3: An incremental improvement [EB/OL]. (2018-04-08)[2026-03-19]. https://arxiv. org/abs/1804.02767. [DOI: 10.48550/arXiv. 1804.02767]

Wang W G, Lai Q X, Fu H Z, Shen J B, Ling H B and Yang R G. 2022. Salient object detection in the deep learning era: an in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 44(6): 3239-3259 [DOI: 10.1109/TPAMI.2021.3051099]

Wang W G, Shen J B, Lu X K, Hoi S C H and Ling H B. 2021. Paying attention to video object pattern understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 43(7): 2413-2428 [DOI: 10.1109/TPAMI.2020.2966453]

Chen Q, Zhang Z X, Lu Y Y, Fu K R and Zhao Q J. 2024. 3-D convolutional neural networks for RGB-D salient object detection and beyond. IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 35(3): 4309-4323 [DOI: 10.1109/TNNLS.2022.3202241]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco. 1997.9.8.

1735]

Shi X J, Chen Z R, Wang H, Yeung D Y, Wong W K and Woo W C. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting // Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS). Montreal, Canada: MIT Press: 802-810[DOI: 10.5555/2969239.2969329]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need // Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, USA: Curran Associates Inc.: 5998-6008 [DOI: 10.5555/3295222.3295343]

Lipton A J, Fujiyoshi H and Patil R S. 1998. Moving target classification and tracking from real-time video // Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV). Princeton, USA: IEEE: 8-14 [DOI: 10.1109/WACV.1998.730198]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition // Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR [DOI: 10.48550/arXiv.1409.1556]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. 2016. SSD: Single shot multibox detector // Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Lin T Y, Dollár P, Girshick R, He K, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 2117-2125 [DOI: 10.1109/CVPR.2017.106]

Liu S, Qi L, Qin H, Shi J and Jia J. 2018. Path aggregation network for instance segmentation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 8759-8768 [DOI: 10.1109/CVPR.2018.00913]

Tan M, Pang R and Le Q V. 2020. EfficientDet: Scalable and efficient object detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 10781-10790 [DOI: 10.1109/CVPR42600.2020.01079]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: Convolutional block attention module // Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Li J, Liang X, Wei Y, Xu T, Feng J and Yan S. 2017. Perceptual generative adversarial networks for small object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1222-1230 [DOI: 10.1109/CVPR.2017.211]

Han W, Khorrami P, Paine T L, et al. Seq-NMS for video object detection[EB/OL]. (2016-02-26)[2026-03-19]. https://arxiv. org/abs/1602.08465. [DOI: 10.48550/arXiv. 1602.08465]

Zhu X Z, Wang Y J, Dai J F, Yuan L and Wei Y C. 2017. Flow-guided feature aggregation for video object detection // Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 408-417 [DOI: 10.1109/ICCV.2017.52]

Zhou Q, Li X, He L, Sun J, Zhang J and Wang L. 2022. TransVOD: End-to-end video object detection with spatial-temporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 45(6): 7853-7869 [DOI: 10.1109/TPAMI.2022.3223955]

Bi H, Wu R, Liu Z, Zhu H, Zhang C and Xiang T Z. 2023. Cross-modal hierarchical interaction network for RGB-D salient object detection. Pattern Recognition, 136: 109194 [DOI: 10.1016/j.patcog.2022.109194]

Cong R M, Liu H Y, Zhang C, Zhang W, Zheng F, Song R, et al. 2023. Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection // Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 406-416 [DOI: 10.1145/3581783.3611982]

Xiao F, Pu Z D, Chen J Q and Gao X P. 2024. DGFNet: depth-guided cross-modality fusion network for RGB-D salient object detection. IEEE Transactions on Multimedia (TMM), 26: 2648-2658 [DOI: 10.1109/TMM.2023.3301280]

Feng Q H, Wang Z X, Sun C C and Shao Z W. 2025. Small object detection in drone images via foreground refinement and multidimensional inductive bias self-attention. Journal of Image and Graphics, 30(11): 3547-3563 (冯琪涵,王志晓,孙成成,邵志文. 2025. 融合前景细化和多维归纳偏置自注意力的无人机图像小目标检测。中国图象图形学报, 30(11):3547-3563 [DOI:10.11834/jig.250017]

Wu J, Cai G Z, Chu H X, Xu G, Zhao X M and Yin H. 2026. Small Object Detection Network for Wide-Field Surveillance Video SOD-YOLO. Journal of Image and Graphics: 1-18 (吴军,蔡广震,楚和轩,徐刚,赵雪梅,尹恒. 2026. 用于大视场监控视频的小目标检测网络 SOD-YOLO. 中国图象图形学报:1-18 [DOI:10.11834/jig.250491]

作者简介

李哲,男,硕士研究生,主要研究方向为目标检测。E-mail: lz_0520@163.com

罗靖,通讯作者,男,副教授,主要研究方向为深度学习、脑机接口。E-mail:luojing@xaut.edu.cn

柳宇，女，实验员，主要研究方向为机器学习。E-mail：liuyu@xaut.edu.cn
高彬智，男，本科生，主要研究方向为深度学习、计算机视觉。E-mail：3230913025@stu.xaut.edu.cn
石伟伟，男，副教授，主要研究方向为深度学习、计算机视觉。E-mail：wshi@xaut.edu.cn
王晓帆，男，副教授，主要研究方向为统计分析理论、机器学习。E-mail：wangxfok@xaut.edu.cn
黑新宏，男，教授，主要研究方向为安全关键计算机系统、人工智能。E-mail：heixinhong@xaut.edu.cn