

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-16

论文引用格式: Lyu Jiahao, Zhao Minghua, Hu Jing, Du Shuangli, Shi Cheng, Chen Yifei, Lv Zhiyong. Gated Spatial-Frequency Mamba for Unmanned Aerial Vehicle Video Anomaly Detection[J/OL]. Journal of Image and Graphics, XXXX: 1-16. DOI: 10.11834/jig.260211. (吕佳豪, 赵明华, 胡静, 都双丽, 石程, 陈逸飞, 吕志勇. 门控空间频率 Mamba 驱动的无人机视频异常检测[J/OL]. 中国图象图形学报, XXXX: 1-16. DOI: 10.11834/jig.260211.) [DOI:10.11834/jig.260211]

门控空间频率 Mamba 驱动的无人机视频异常检测

吕佳豪¹, 赵明华^{1,2*}, 胡静¹, 都双丽¹, 石程¹, 陈逸飞¹, 吕志勇¹

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

摘要: 目的 无人机凭借低成本、高分辨率的数据采集能力广泛应用于低空安防巡检,但其航拍视频常面临复杂背景干扰及异常行为捕捉困难的挑战。针对上述问题,提出一种门控空间频率 Mamba 驱动的无人机视频异常检测方法。**方法** 提出了基于门控空间频率 Mamba 的轻量化模型 GSF Mamba。该方法首先利用轻量化 ResNet 构建多尺度预训练编码器,并配合瓶颈融合层实现连续帧间的跨尺度时序特征聚合;随后,设计了空频双域并行 Mamba 架构,同步捕获空间全局演化轨迹与频率域细微扰动;最后,通过门控空间频率模块(GSF)实现双域特征的自适应融合。其中可训练参数量仅为 2.8M,显著降低了计算资源消耗。**结果** 实验结果表明,GSF Mamba 在两个无人机基准视频异常检测数据集上均取得了竞争性的 AUC 指标,在 Drone-Anomaly 数据集达到 72.8%,在 UIT-Adrone 数据集达到 60.7%。在保持高精度的同时,得益于轻量化的参数设计,模型在推理延迟方面表现优异,基于 ResNet18 模型的推理速度达到 102FPS,基于 ResNet34 模型的推理速度达到 89FPS,实现了检测效能与计算速度的最佳平衡。**结论** 实验结果表明,GSF Mamba 通过空频双域建模与自适应门控机制,有效增强了对复杂环境下异常行为的辨识力。其极低的参数量证明了该方法在无人机边缘侧进行实时、高效异常监测的实用价值与领先性。

关键词: 无人机;异常检测;低空目标;门控注意力;状态空间模型

Gated Spatial-Frequency Mamba for Unmanned Aerial Vehicle Video Anomaly Detection

Lyu Jiahao¹, Zhao Minghua^{1,2*}, Hu Jing¹, Du Shuangli¹, Shi Cheng¹, Chen Yifei¹, Lv Zhiyong¹

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

Abstract: Objective Unmanned Aerial Vehicles (UAVs) have rapidly emerged as pivotal and ubiquitous tools across diverse fields, including intelligent surveillance, urban traffic monitoring, crowd management, and rapid disaster response. This widespread adoption is largely credited to their exceptional spatial mobility, cost-efficiency, operational

收稿日期: 2026-04-15; 修回日期: 2026-08-12

* 通信作者: 赵明华, 女, 教授, 主要研究方向为计算机视觉和机器学习。E-mail: zhaominghua@xaut.edu.cn

基金项目: 国家自然科学基金(62571425); 陕西省自然科学基金项目(2024JC-ZDXM-35); 陕西省科协青年人才托举计划(20240146); 中国高校产学研创新基金-新一代信息技术创新项目(2024IT088); 全省微波空间智能云计算重点实验室开放课题资助(2025ZY01019); 西安理工大学博士创新基金资助项目(BC202621)

Supported by: National Natural Science Foundation of China (62571425); Nature Science Foundation of Shaanxi Province, China (2024JC-ZDXM-35); Shaanxi Provincial Association for Science and Technology Young Talent Lifting Program (20240146); China University Industry-Academia-Research Innovation Fund-Next Generation Information Technology Innovation Project (2024IT088); the Open Research Program of Laboratory for Microwave Spatial Intelligence and Cloud Platform under Grant No. 2025ZY01019; Doctoral Dissertation Innovation Fund of Xi'an University of Technology (BC202621)

flexibility, and superior comprehensive data acquisition capabilities from aerial vantage points. Despite these profound advantages, automated Video Anomaly Detection (VAD) in UAV-captured footage remains a formidable and unresolved challenge in the computer vision domain. The inherent complexity of this task is driven by a multitude of adverse aerial factors: highly dynamic and moving backgrounds, drastic and sudden perspective variations, camera jitter, varying flight altitudes, and the highly subtle, irregular nature of anomalous behaviors often obscured within high-density crowds or complex topographical environments. Furthermore, traditional deep learning models—particularly those relying heavily on standard Transformer architectures with global self-attention mechanisms—often struggle to strike an optimal balance between achieving high-precision detection and adhering to the stringent resource constraints characteristic of onboard edge devices. To systematically bridge this critical gap between high-performance sequence modeling and real-time edge deployment, this paper proposes GSFMamba, a highly sophisticated yet deliberately lightweight UAV video anomaly detection method fundamentally based on a novel Gated Spatial-Frequency Mamba architecture. **Method** The proposed methodology is engineered from the ground up to prioritize edge-optimized hierarchical feature representation without compromising modeling depth. The pipeline initiates by utilizing a lightweight Convolutional Neural Network encoder, specifically leveraging ResNet18 or ResNet34 backbones, to extract rich, multi-scale spatial features from the input video frames. These extracted features are subsequently processed by a custom-designed Bottleneck Fusion Layer, which functions to efficiently aggregate cross-scale temporal dependencies while aggressively compressing the channel dimensions to reduce the subsequent computational overhead. At the structural core of the GSFMamba method lies a meticulously crafted dual-branch parallel Mamba architecture. This core is explicitly designed to overcome the quadratic computational complexity inherent to traditional self-attention mechanisms, thereby maintaining robust and continuous long-range temporal modeling capabilities essential for video analysis, but with strictly linear complexity. The architecture bifurcates into two complementary processing streams. The first stream, the Spatial Mamba branch, is dedicated to capturing the global spatial evolution and structural correlations of the scene over time, leveraging the State Space Model (SSM) formulation to process long video sequences seamlessly. Concurrently, the second stream, the Frequency Mamba branch, innovatively integrates the Discrete Wavelet Transform (DWT) directly into the state space modeling process. This transform projects the spatio-temporal features into the frequency domain, decomposing the signal into distinct frequency sub-bands. This allows the network to effectively isolate and scrutinize high-frequency noise and subtle, abrupt motion disturbances that are highly indicative of microscopic anomalous activities, which might otherwise be smoothed over in the purely spatial domain. Finally, the features generated from these dual domains are adaptively synthesized and fused through a sophisticated Gated Spatial-Frequency (GSF) module. This module employs a data-driven, dynamic gating mechanism to evaluate the scene context in real-time, assigning learned attention weights to selectively emphasize the most informative diagnostic cues from either the spatial or frequency domain. The entire proposed GSFMamba is kept exceptionally compact, featuring a total trainable parameter count of only 2.8M, which significantly minimizes the memory footprint and ensures viable deployment on resource-constrained platforms. **Result** Extensive and rigorous experiments were conducted on two highly challenging, mainstream UAV video anomaly detection benchmarks—Drone-Anomaly and UIT-ADrone—to comprehensively evaluate the detection performance, generalization capability, and operational robustness of the proposed GSFMamba method. The quantitative results conclusively demonstrate that our method achieves highly competitive Area Under the Curve (AUC) scores, reaching 72.8% and 60.7% on the Drone-Anomaly and UIT-ADrone datasets, respectively. Compared to the baseline model ANDT, our proposed method achieves performance improvements of 4.2% and 0.2% on the respective datasets. While maintaining linear computational complexity, it realizes a significantly more efficient long-sequence modeling capability than standard Transformers, thereby exhibiting heightened sensitivity to anomalous events. When analyzing the performance across the five distinct sub-scenarios of the Drone-Anomaly dataset—namely Highway, Bike Roundabout, Vehicle Roundabout, Railway Inspection, and Solar Panel Inspection—our model consistently demonstrated superiority, achieving AUC values of 86.0%, 84.5%, 76.6%, 69.6%, and 80.8%, respectively, thereby outperforming the majority of existing state-of-the-art methodologies. Furthermore, on the UIT-ADrone dataset, recent weakly supervised methods such as MMVAD can reach a detection accuracy of 67.4%, ASTNet achieves 61.4%, and the UAV-based VAD method ASTT realizes 65.5% accuracy. However, these existing approaches heavily rely on either complex multi-modal information fusion or

the extensive stacking of model parameters to trade for high accuracy. Consequently, this heavy computational burden renders them ill-equipped to meet the stringent real-time requirements of low-altitude UAV patrol and inspection tasks in practical scenarios. In stark contrast, beyond merely top-line accuracy, detailed ablation studies unequivocally confirm the indispensability of our frequency-domain modeling approach, which provides a statistically significant boost in detecting subtle behavioral irregularities. Most notably, GSFMamba exhibits exceptional inference efficiency: with its minimal parameter count of 2.8M, it achieves an ultra-high-speed processing rate of 102 Frames Per Second (FPS) when equipped with the ResNet18 backbone, and 89 FPS with the ResNet34 backbone. This outstanding performance establishes an optimal trade-off between discriminative modeling power, hardware memory efficiency, and execution latency. **Conclusion** The comprehensive experimental evidence confirms that GSFMamba's innovative dual-domain modeling strategy and adaptive spatial-frequency gating mechanism provide a robust, highly scalable, and structurally elegant solution for high-efficiency, real-time anomaly monitoring deployed directly on UAV edge computing platforms. By effectively resolving the tension between structural simplicity and requisite modeling depth, our proposed linear-complexity architecture avoids the heavy hardware dependencies of contemporary models. Ultimately, this work establishes a formidable new baseline for resource-constrained aerial video surveillance, offering significant practical value and technical foundational support for the next generation of real-world autonomous security, smart city management, and proactive disaster mitigation applications.

Key words: unmanned aerial vehicle; anomaly detection; low-altitude object; gated attention; state space model

论文引用格式: Lyu Jiahao, Zhao Minghua, , Hu Jing, Du Shuangli, Shi Cheng, Chen Yifei, Lv Zhiyong. Gated Spatial-Frequency Mamba for Unmanned Aerial Vehicle Video Anomaly Detection [J/OL]. Journal of Image and Graphics. DOI:10.11834/jig.260211. (吕佳豪, 赵明华, 胡静, 都双丽, 石程, 陈逸飞, 吕志勇. 门控空间频率 Mamba 驱动的无人机视频异常检测 [J/OL]. 中国图象图形学报. DOI: 10.11834/jig.260211)

0 引言

随着大众级无人机 (unmanned aerial vehicle, UAV) 市场的广泛普及与相关技术的快速发展, 有关 UAV 的应用在交通管理 (李永福等, 2026) 和环境监测 (张积晗等, 2026) 等领域已经得到了广泛研究。特别是在交通领域中, 对于 UAV 所采集到的海量监控数据, 如何从中检测出偏离预期模式的异常行为, 保障低空经济形势下的智能化巡察与精准化检测的感知能力。因此, 基于无人机的视频异常检测 (video anomaly detection, VAD) 已展现出巨大的应用潜力。

由于异常事件发生概率较低, 且不同场景对异常的定义存在差异, 大多数 VAD 采用半监督方法 (Park 等, 2020), 亦称为单类分类 (one-class classification, OCC), 通过用正常数据训练模型, 让模型仅

学习正常事件特征, 在测试阶段将偏离该分布的事件定义为异常。例如在人行道上骑自行车、道路上发生事故、车辆在道路逆行等行为均可被定义为异常。这类异常事件所具有的稀疏性和突发性等属性决定了现有 VAD 难以通过标签化手段进行实时检测。尽管现有弱监督方法的检测性能优于基于 OCC 的方法, 但是依赖视频级或帧级标签仍然无法有效应对多变的场景与种类繁多的异常事件。

现有多数 VAD 方法均是基于地面固定视角的监控摄像头采集的视频数据展开研究。由于这类监控设备位置固定、视角不变, 其检测过程中仅需重点关注前景信息中存在的非理想运动流场 (如打架斗殴、异常轨迹等), 而静态背景始终保持稳定且可提前建模, 因此在一定程度上无需考虑静态背景对异常判别的干扰, 这类任务的核心难点集中于前景运动特征的提取与异常识别。例如, Yu 等人 (2023) 和 Zhang 等人 (2025) 均通过目标检测或前景裁剪等方式, 对视频帧进行预处理, 筛选出前景目标等关键信息, 但是这些方法都容易忽略了场景与行为的关联, 难以应对复杂多半的场景变化导致的事件属性突变, 即不同场景下的不同事件可能存在不同定义。

因此, 与大多数静态背景的监控视频相比, 无人机视频面临更多的挑战 (Tran 等, 2023), 包括动态视角下的背景变化、物体目标分辨率过小、以及不同飞行高度而导致的空间分辨率变化。为解决上述问题, 提出了门控空间频率 Mamba 驱动的无人机视频

异常检测方法 GSFMamba (gated spatial-frequency mamba)。首先,利用预训练的 ResNet 残差网络作为编码器,对连续视频帧进行多尺度特征提取与时序初步融合,构建起包含丰富空间细节的特征金字塔。其次,通过设计的瓶颈融合层对多层级特征进行跨尺度聚合与感受野扩张,实现语义信息的深度提炼。然后,在核心解码阶段,GSFMamba 设计了门控空间频率模块,将提取的特征解构为空间域与频率域两条并行支路,并利用 Mamba 结构优越的线性序列建模能力,同步捕获视频流中的长效空间演变与高频细节扰动。最后,通过门控机制自适应地融合双域特征,实现对无人机复杂动态背景下实现对未来帧的精准预测与异常事件判别。GSFMamba 不仅保持了高效的计算性能,更显著提升了模型对细微异常特征的辨识力。

本文主要贡献如下:

1) 针对无人机动态场景,提出了一种空间频率 Mamba 驱动的无人机视频异常检测方法,实现了局部卷积神经网络(convolutional neural network, CNN)特征与全局 Mamba 特征的高效耦合建模能力。

2) 设计了基于 ResNet18/34 的轻量化编码器,配合瓶颈融合层对连续帧的多尺度特征进行跨层级聚合与时序初步融合,有效应对无人机视角下的尺度变化与背景噪声。

3) 构建了基于门控空间频率 Mamba 模块,通过自适应地调控空间域变化与频率域扰动的融合比例,增强了模型对细微异常信号的捕获能力。

1 相关工作

现有视频异常检测任务依据监控视频的采集场景与视角,主要可划分为地面监控视频异常检测与无人机视频异常检测两大类。地面监控场景以人流密集、遮挡等为典型特征,无人机航拍场景则具有动态背景及目标尺度多变的特点。现有 VAD 主要通过重建和预测两类方法进行异常判别。基于重建的 VAD 方法侧重于空间信息重建,通过重建误差来衡量视频帧是否异常。而基于预测的 VAD 方法则更加关注时序信息,根据输入的视频帧序列来预测下一时刻的视频帧,通过预测误差来实现高效判别。此外,现阶段研究探索包括基于 CNN、Transformer 和

Mamba 的 VAD 模型,在追求检测精度的同时也需重点兼顾模型的实时性与轻量化需求。

1.1 基于 CNN 的视频异常检测

基于 CNN 的 VAD 方法能够有效提取图像特征并捕捉空间模式。具体而言,多数 VAD 方法均采用自动编码器(autoencoder, AE)和生成对抗网络(generative adversarial network, GAN)进行预测(Liu 等, 2018)或重建(梁家菲等, 2023)。针对视频中的特征原型(prototype), Gong 等人(2019)和 Park 等人(2020)率先提出记忆模型,通过记录正常模式特征,从而在测试阶段有条件的筛选和过滤异常特征。Fang 等人(2021)提出 MESDnet(multi-encoder single-decoder network),通过三个独立的编码器分别编码运动与外观特征,在解码阶段融合多个特征实现帧重建。Kommanduri 等人(2024)提出 DAST-Net(dense attention-aware spatiotemporal network),基于改进的密集注意力残差网络提取空间特征,并结合 ConvLSTM-AE 建模时序依赖,通过注意力增强残差连接聚焦关键区域。同时,相比于 AE 架构,部分研究(刘成明等, 2022)依赖 GAN 更强的特征表达能力取得了更好的性能。例如, Singh 等人(2024)设计双判别器分别优化重建帧与重建噪声质量,并在生成器中引入自注意力与通道注意力以捕捉精细特征,实现高质量视频重建与异常检测。Li 等人(2022)提出 VABD(variational abnormal behavior detection),融合概率 U-Net 与条件变分自编码器,引入基于光流运动损失保证帧运动连续性,并嵌入 WGAN(wasserstein GAN)来提升性能,实现对视频中异常帧的精准判别。针对模糊行为检测难题, Guo 等人(2023)提出 AMS-AE(appearance-motion synergy autoencoder),通过外观-运动融合协同强化规范模式表征,设计动态融合机制的记忆网络整合模态信息,从而扩大规范间隙以提升模糊行为检测能力。此外,进一步研究从单向预测跨越到双向预测,展现了优越的异常检测性能。Zhong 等人(2022)提出双向时空特征学习框架与多尺度异常评估策略,通过双编码器双解码器网络融合正反时序特征实现双向预测,并结合注意力特征融合模块与多尺度评估,提升复杂场景下不同尺寸异常目标的检测精度与鲁棒性。同样,为扩大无监督视频异常检测中正常与异常事件的判别边界, Lyu 等人(2026)提出双向跳帧预测方法,通过在训练阶段跳帧实现正反双向预测,而测

试阶段用双向连续帧共预测中间帧,结合方差通道注意力与上下文空间注意力,强化特征表征以放大正常与异常的域内差异。

1.2 基于 Transformer 的视频异常检测

由于 CNN 仅通过有限的局部感受野来提取特征,难以获取全局信息。而 Transformer 模型能够捕捉全局依赖关系,从而能够捕捉视频或图像中的长期时间关联与跨空间关系。面向地面监控 VAD 任务, Qiu 等人(2024)基于 I3D 模块提取高维特征,设计双尺度深度聚类模块,压缩通道与空间特征以扩大正常与异常数据边界,并将聚类中心作为记忆特征向量优化模型性能。Le 等人(2025)利用多级 Transformer 作为编码器来生成多尺度特征图,然后将其传送到分层时空 Transformer,在解码阶段通过利用联合注意机制来捕获空间与时间信息。针对视角多变、运动复杂的无人机场景, Jin 等人(2022)首先将 Transformer 架构应用于无人机 VAD,提出的 ANDT(anomaly detection with transformers)将视频序列表示为时空小管嵌入,并利用 Transformer 的全局感受野进行未来帧预测。随后, Tran 等人(2024)针对连续视频采用提取单独视频帧的方式来学习空间特征和时序依赖性,最终将所有帧间特征通过交叉注意机制融合以预测未来帧。Le 等人(2025)同样基于 Transformer 网络,通过多阶段编码提取视频特征,结合分层时空 Transformer 模型实现多阶段编码,实现像素级精准预测与视频帧长程依赖的高效建模。针对无人机航拍场景的检测难点及现有方法适配不足问题, Biswas 等人(2025)提出视觉-语言多模态异常检测方案,通过自适应长短时视频分割结合对比学习,提升检测性能与场景适配度。

1.3 基于 Mamba 的视频异常检测

随着状态空间模型的出现,特别是 Mamba 模型(Gu 等, 2023)凭借优越的长序列依赖建模能力和轻量化线性复杂度计算性能。大量研究将 Mamba 应用到异常检测领域。例如, He 等人(2024)设计了基于 CNN 局部特征提取与 Mamba 全局特征优化的并行局部性增强状态空间模块,在工业异常检测数据集上取得了优越的性能。同样, 刘建明等人(2025)基于视觉 Mamba 和块特征分布,解决传统方法在工业领域中逻辑异常检测不敏感的问题。Lyu 等人(2025)和 Li 等人(2026)在地面监控 VAD 领域分别设计基于 Mamba 的重建-预测混合任务与基于外观

与差分信息的双流网络。Liu 等人(2025)提出 Mamba 驱动的多尺度时空学习方法 M²S²L,依托多粒度时空编码与特征分解机制强化行为建模与异常评估能力。而在弱监督 VAD 中, Shen 等人(2026)构建了文本重写驱动的 CLIP 模型(text rewriting CLIP, TrCLIP-VAD),并设计局部-全局多尺度 Mamba 模块挖掘全局局部时序关联,促进视觉与文本特征深度融合。Chen 等人(2026)提出跨模态注意力融合网络(cross-modal attention fusion network, CMAF-Net),通过双向注意力机制关联 RGB 与骨架特征,并结合双目标学习提升各类异常检测鲁棒性。上述方法都实现了与之前方法近似的精度前提下,都实现了较好的推理能力。然而,现有主流的异常检测方法仍然是通过基于大参数量模型,为了追求极致的检测精度,从而忽略了推理效率。特别是,针对无人机视频异常检测,应当首要考虑的是保障部署模型的所具有的实时性推理能力,其次才是具有优势的检测性能。

2 本文方法

本文提出了一种门控空间频率 Mamba 驱动的无人机视频异常检测 GSFMamba,通过输入连续 n 个视频帧 $I_{t-n:t-1}$ 来预测第 $n+1$ 帧 I_t ,如图 1 所示。由三个主要组件组成:基于 CNN 的预训练编码器,基于多尺度融合的瓶颈模块和基于 Mamba 的解码器。首先,输入连续帧,基于预训练编码器提取每个帧的不同尺度特征并特征相加。其次,将连续帧的三个尺度特征输入到瓶颈处进行高维特征融合。同时,将编码器的多个特征通过跳跃连接逐元素相加到对应的解码器。最后,将融合特征输入到基于门控空间频率 Mamba 的解码器中得到预测帧。

2.1 基于 CNN 的预训练编码器

本文设计两种预训练编码器,分别采取 ResNet18 和 ResNet34 两种模型作为基准预训练编码器。由于两者在编码阶段的通道数上保持一致,因此后续解码阶段亦保持一致。基于 CNN 的预训练编码器通过逐层下采样操作提取多尺度特征,其输出的空间分辨率呈现金字塔结构,分别为原始输入的 $1/4$ 、 $1/8$ 和 $1/16$ 。

针对输入的 n 个连续视频帧 $\{I_{t-n}, \dots, I_{t-1}\}$,编码器对每一帧图像独立执行特征提取。设第 i 个尺

度下的特征映射为 F^i , 则对于每一帧图像 I_k , 编码器生成一组多尺度特征集合 $\{F_k^1, F_k^2, F_k^3\}$ 。为了捕获时序上下文信息并降低冗余度, 提出方法通过对序列

中各帧对应的同尺度特征进行元素级融合, 得到融合后的特征向量 x_i , 如式(1)所示:

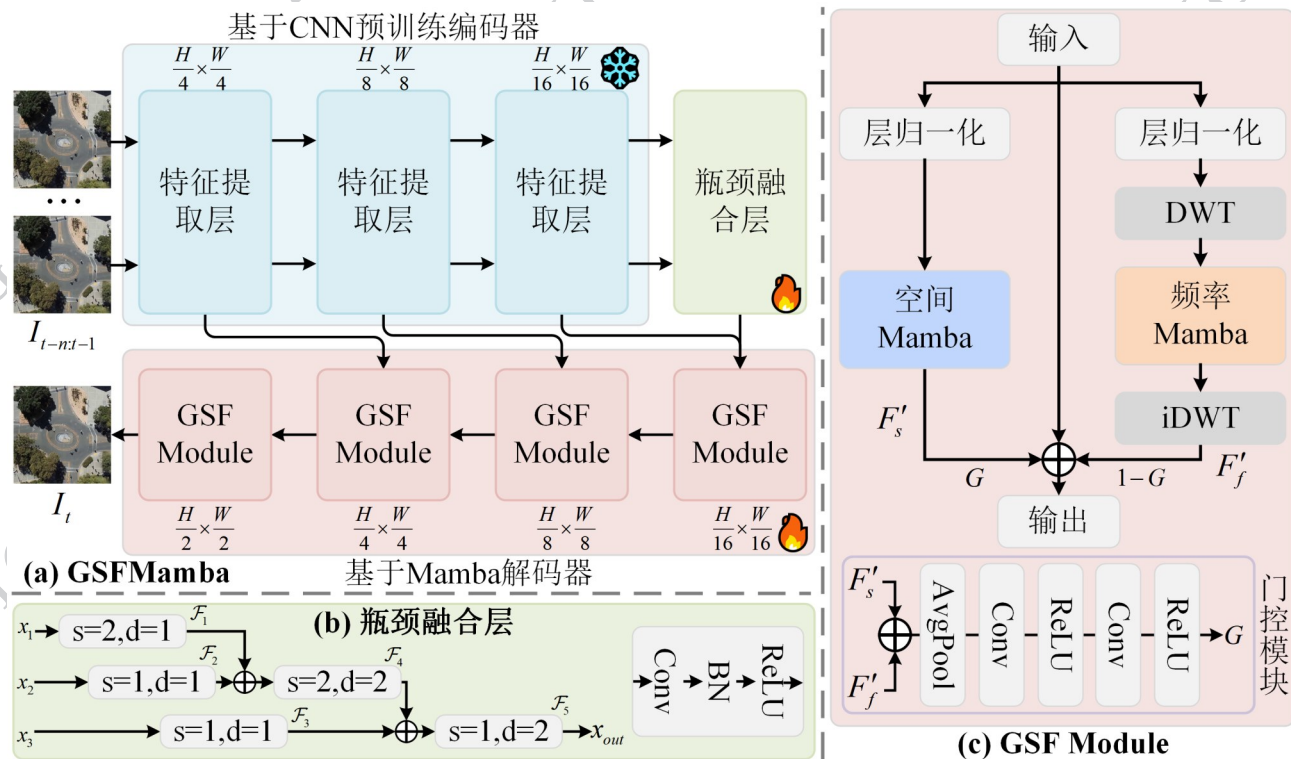


图1 提出的GSFMamba的总体架构。

Fig. 1 Overall architecture of the proposed GSFMamba. ((a)GSFMamba; (b)Bottleneck fusion; (c)GSF Module)

$$x_i = \sum_{k=t-n}^{t-1} F_k^i, \quad i \in \{1, 2, 3\} \quad (1)$$

式中, 所得的 x_1, x_2, x_3 分别对应图 1(a) 中的三个特征提取层阶段输出, 并作为跳跃连接和后续瓶颈融合层的输入。

2.2 瓶颈融合层

为了对编码器输出的多尺度特征进行跨通道聚合与空间维度对齐, 提出瓶颈融合模块以消除不同层级特征间的语义差。如图 1(b) 所示, 输入来自编码器的融合特征 $\{x_1, x_2, x_3\}$ 。为了在扩大感受野的同时保持精细的空间结构, 设计了基于下采样卷积 (Downsampling Convolution) 与空洞卷积 (Dilated Convolution) 的瓶颈融合层。

在具体的处理流程中, 高分辨率特征通过步长 s 和扩张率 d 的变换, 与低分辨率特征进行跨尺度融合, 如式(2)所示:

$$x_{out} = \mathcal{F}_5(\mathcal{F}_3(x_3) + \mathcal{F}_4(\mathcal{F}_2(x_2) + \mathcal{F}_1(x_1))) \quad (2)$$

式中, $\mathcal{F}$ 代表不同参数配置的基础卷积块操作, 即

$\text{Conv} \rightarrow \text{BN} \rightarrow \text{ReLU}$ 。通过控制 $s = 2$ 实现下采样对齐, 利用 $d = 2$ 的空洞卷积获取更广的上下文信息。最终, 输出具有高阶语义信息的瓶颈特征 x_{out} , 随后被送入基于 Mamba 的解码器。

2.3 基于门控空间频率 Mamba 解码器

针对无人机航拍图像中目标尺度多变、背景干扰强以及边缘细节易丢失的问题, 文本提出基于状态空间模型的门控空间频率模块 (Gated Spatial-Frequency Module, GSF Module), 通过双并行路径分别在原始空间域和离散小波变换 (Discrete wavelet transform, DWT) 映射的频率域进行特征提取, 并利用门控机制实现特征的自适应融合。其中 Mamba 架构逻辑包含双路径特征提取与非线性映射过程, 如式(3)所示:

$$\begin{aligned} F^1 &= \text{LN}(\text{SS2D}(\sigma(\text{DWConv}(\text{Linear}(F)))))) \\ F^2 &= \sigma(\text{Linear}(F)) \\ F' &= \text{Linear}(F^1 \odot F^2) \end{aligned} \quad (3)$$

式中, $\sigma(\cdot)$ 为 SiLU 激活函数, $\text{LN}(\cdot)$ 为层归一化操作。

上述过程可以表示为 $X \leftarrow \mathcal{M}(F)$, $\mathcal{M}(\cdot)$ 为 Mamba 模块。

如图 2 所示, 输入空间域特征 F_i 或频率域特征 F_f 进入 Mamba 架构的两个平行分支进行处理, 生成最终的增强特征 F'_i 或 F'_f 。图 2(b) 展示了二维选择性扫描 (2D-Selective-Scan, SS2D) 模块的内部逻辑, 通过交叉扫描机制, 将输入的二维特征映射沿四个方向展平为一维序列。随后, 四个方向的序列并行进入选择性状态空间模型核心进行长程依赖建模, 其离散化过程及状态演化遵循图中所示的线性常微分方程。最后, 利用交叉合并操作将四个方向的增强特征重新恢复并融合成原始的二维空间特征。

空间长程建模分支。 为了捕获无人机航拍场景下动态区域语义关联, 引入作用于原始空间域的空间 Mamba 分支, 基于 SS2D 模块用于全局建模。首先, 将输入特征 F_{in} 经层归一化后得到空间域特征 F_i , 输入到空间 Mamba 模块 $\mathcal{M}_s$, 生成最终的增强空间特征 F'_i , 如式 (4) 所示:

$$F'_i = \mathcal{M}_s(F_i) \quad (4)$$

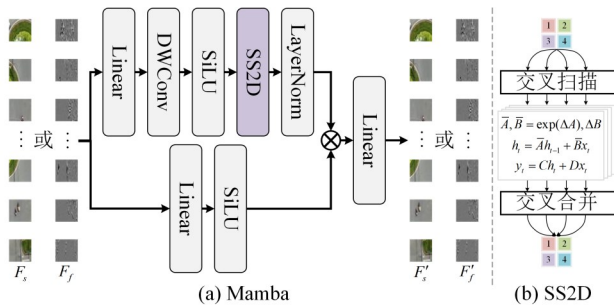


图 2 基于 SS2D 的空间/频率 Mamba 架构

Fig. 2 SS2D-based spatial and frequency mamba architecture.
(a) Mamba; (b) SS2D)

频率细节增强分支。 不同于空间 Mamba 分支提取空间域信息, 频率分支是为了关注航拍图像中的频率信息, 通过 DWT 将特征分解至不同频率, 实现不同频率域信息的高效特征提取。

频率 Mamba 模块 $\mathcal{M}_f$ 的扫描方式与空间 Mamba 保持一致, 即从不同的四个方向对频率信息选择性扫描, 通过互补的穿越路径建立一个全局感受场, 使每个像素能够高效地从多个方向收集所有其他像素的信息。最终通过逆向变换 DWT (Inverse DWT, iDWT) 生成增强频率特征 F'_f , 如式 (5) 所示:

$$\begin{aligned} F_f &= \text{DWT}(x_{in}), \\ \hat{F}_f &= \mathcal{M}_f(F_f), \\ F'_f &= \text{iDWT}(\hat{F}_f), \end{aligned} \quad (5)$$

式中, x_{in} 为经过层归一化的特征。值得注意的是, 经 DWT 处理后的频率 Mamba 由建立空间像素拓扑转变为构建频域长程关联。为克服微小异常易被 Mamba 空间平滑淹没的局限, 频率 Mamba 利用频域全局交叉扫描, 直接捕捉并放大异常对应的高频突变, 显著增强网络对细微扰动的感知。

门控融合模块。 由于航拍图不同区域的复杂程度各异, 设计门控机制以实现自适应融合。将空间分支特征 F'_i 与频率分支特征 F'_f 逐元素相加, 通过全局平均池化感知全局统计特性, 并经由双层卷积块与 ReLU 激活函数构建轻量化门控器, 生成像素级权重 G , 最终得到融合特征 F_{out} , 如式 (6) 所示:

$$F_{out} = F_{in} + G \cdot F'_i + (1 - G) \cdot F'_f \quad (6)$$

上述融合模块可以理解作为一种自适应信息路由器, 通过动态分配空间域和频率域的置信度, 平衡视觉信息与频率细节, 有效提升了网络对复杂异常的感知力。不同于 SENet 或 CBAM 模块, 提出的门控模块以极低计算开销动态评估空间拓扑与高频扰动的相对置信度, 实现双域自适应加权, 从而高度契合无人机边缘侧的轻量化需求。

2.4 损失函数

在训练阶段, 本研究构建了由强度损失与梯度损失组成的联合约束目标函数。通过多目标优化, 模型不仅能学习到像素级的颜色分布, 还能捕捉到物体运动引起的边缘变化, 如式 (7) 所示:

$$L = L_{pred} + L_{grad} \quad (7)$$

式中, 预测损失 L_{pred} 利用 L2 距离计算 $\hat{I}_t$ 和 I_t 之间的像素差异, 如式 (8) 所示。该项主要从全局强度的角度对模型进行约束, 旨在最小化预测帧与真实帧之间的绝对像素偏差, 确保生成图像在颜色和亮度分布上与基准保持一致。

$$L_{pred} = \left\| \hat{I}_t - I_t \right\|_2^2 \quad (8)$$

单纯的像素级损失容易产生平滑效应, 导致预测帧丢失关键的边缘细节。因此, 引入梯度损失 L_{grad} 用于计算 $\hat{I}_t$ 和 I_t 之间在水平和垂直方向上的梯度差:

$$L_{grad}(\hat{I}_t, I_t) = \sum_{ij} \left\| I_t^{ij} - I_t^{i-1j} - \left| \hat{I}_t^{ij} - \hat{I}_t^{i-1j} \right| \right\|_1 + \left\| I_t^{ij} - I_t^{ij-1} - \left| \hat{I}_t^{ij} - \hat{I}_t^{ij-1} \right| \right\|_1 \quad (9)$$

式中, i 和 j 表示当前视频帧的空间坐标。通过对相邻像素间的变化率进行建模, L_{grad} 强制要求模型关注图像的局部边缘结构。

2.5 异常检测

在测试阶段, 使用预测帧与真实帧间的像素误差来计算峰值信噪比 PSNR, 从而对正常与异常事件进行评估, 如式(10)所示:

$$\text{PSNR}(\hat{I}_t, I_t) = 10 \log_{10} \frac{1}{\left\| \hat{I}_t - I_t \right\|_2^2 / N} \quad (10)$$

式中, $\text{PSNR}(\cdot)$ 代表 PSNR 函数, $\hat{I}_t$ 和 I_t 分别表示预测帧与真实帧, N 表示视频帧中的像素数。

然后, 通过式(10)进行归一化获得 $[0, 1]$ 范围内的 PSNR 值, 并通过高斯滤波器平滑得到最终的异

常分数 $S(I_t)$ 。异常分数 $S(I_t)$ 越高, 表明异常发生的概率越大。

$$S(I_t) = \frac{\text{PSNR}(\hat{I}_t, I_t) - \min(\text{PSNR}(\hat{I}_t, I_t))}{\max(\text{PSNR}(\hat{I}_t, I_t)) - \min(\text{PSNR}(\hat{I}_t, I_t))} \quad (11)$$

式中, $\max(\cdot)$ 和 $\min(\cdot)$ 分别表示当前测试视频序列中所有帧 PSNR 值的最大值与最小值, t 表示当前测试视频序列中的帧索引号。

3 实验

3.1 数据集与评估指标

本文在无人机 VAD 基准数据集 Drone-Anomaly (Jin 等, 2022) 和 UIT-ADrone (Tran 等, 2023) 进行性能评估。表 1 统计了这些数据集的样本统计情况。对于每个训练集, 仅包含正常视频, 而测试集包含正常视频和异常视频。特别的是, 对于 Drone-Anomaly 数据集, 其包含 7 个独立的子场景。

表 1 无人机异常检测数据集统计数据

Table 1 Statistics of the UAV anomaly detection dataset

数据集	视频片段 (训练/测试)	视频帧数 (正常/异常)	异常类型
Drone-Anomaly	37 / 22	51,635 / 35,853	10
UIT-ADrone	41 / 51	59,186 / 147,005	10

与先前的研究一致 (Jin 等, 2022; Liu 等, 2026), 通过逐渐改变正常分数的阈值得到受试者工作特性 (receiver operating characteristic, ROC) 曲线, 然后计算 ROC 曲线下的面积 (area under curve, AUC) 作为评估指标。此外, 实验还额外计算了 Drone-Anomaly 数据集下的召回率 (Recall)、精度 (Precision)、F1 分数 (F1 score) 和总体准确度 (Overall Accuracy, OA)。上述指标的阈值通过训练集输入到训练模型中, 以获得所有训练样本的重建误差来确定, 具体为重建误差的平均值和标准偏差值的总和。

3.2 实验设置

所有实验基于单张 Nvidia GeForce RTX 4090 显卡, 使用 PyTorch 进行训练和推理。在预处理阶段, 将每个视频帧的大小调整为 224×224 , 并将像素归一化到 $[-1, 1]$ 。在 Drone-Anomaly 和 UIT-ADrone 两个数据集上使用 $\beta = 0.9$ 的 Adam 优化器, 批量大小为 4。将视频帧长度 n 设为 4, 通过获取 4 个连续的

视频帧序列, 从而预测第 5 帧。学习率初始设置为 $2e-4$, 并使用余弦退火方法进行衰减 (Loshchilov 等, 2016)。

3.3 与其他异常检测方法比较

在本节中, 将提出方法与地面监控 VAD 和基于无人机 VAD 的方法进行了全面对比。如表 2 所示, 提出方法在 Drone-Anomaly 和 UIT-ADrone 两个基准数据集上分别取得了 72.8% 和 60.7% 的 AUC 检测性能。相较于基线模型 ANDT, 本文方法分别提升了 4.2% 和 0.2%, 在保持线性复杂度的同时, 实现了比 Transformer 更高效的长序列建模能力, 从而对异常事件更敏感。在 Drone-Anomaly 数据集上超越了所有对比方法, 达到了最佳性能。而在 UIT-ADrone 数据集上, 尽管面临高度复杂的视角挑战, 弱监督方法 MMVAD 的检测进度可以高达 67.4%, 然而这些弱监督方法在训练阶段引入了额外的视频级或帧级异常标签, 这种监督信息的维度降维打击是造成本

文无监督框架在某些指标上落后的直接原因之一,因此两者在严格意义上不处于同一评估基准线。但提出方法相较于同类无监督方法仍展现出极强的竞争性。尽管部分无监督地面监控方法(如 ASTNet)略高于本文方法,但其同样依赖于堆叠模型参数来换取高精度,导致其难以应对实时低空巡检的实际需求。同样无人机 VAD 方法 ASTT 模型基于纯

Transformer 架构来预测未来帧,其模型具有 63.2M 的参数量,在复杂场景下更加有效。相比之下,在保持高性能输出的同时,提出方法通过极其精简的架构设计,将可训练参数量压缩至仅 2.8M,远低于目前主流的高性能模型,这充分证明了该算法在计算效率与检测精度之间取得了卓越的平衡,能够以极低的硬件开销实现高效的无人机异常检测。

表 2 与其他视频异常检测方法在 AUC(%)上比较
Table 2 Comparison with other VAD methods on AUC (%)

	方法	Drone-Anomaly	UIT-ADrone
地面监控 VAD	Frame-Pred (Liu 等, 2018)	57.9	53.6
	MNAD (Park 等, 2020)	52.5	57.1
	AMMCNet (Cai 等, 2022)	62.4	60.7
	Cluster-AE (Chang 等, 2022)	63.6	58.4
	ASTNet (Le 等, 2023)	63.2	61.4
	TEVAD* (Chen,等, 2023)	67.4	63.5
	MGFN* (Chen,等, 2023)	68.9	64.3
	MAAMNet (Wang 等, 2023)	61.5	60.1
	VADCLIP* (Wu 等, 2024)	68.1	66.3
	VADMamba (Lyu 等, 2025)	65.2	63.9
无人机 VAD	LGN-NET (Zhao 等, 2025)	64.3	62.7
	ANDT (Jin 等, 2022)	68.6	60.5
	ASTT (Tran 等, 2024)	67.8	65.5
	HSTforU (Le 等, 2025)	67.0	-
	MMVAD* (Biswas 等, 2025)	70.8	67.4
	本文方法 (ResNet18)	71.9	59.3
	本文方法 (ResNet34)	72.8	60.7

注:加粗字体为每列最优结果,*为弱监督方法。

表 3 对比了不同方法在模型效率与推理速度上的表现。现有 VAD 方法往往通过堆叠参数换取精度,如 Frame-Pred,AMMCNet 和 ASTNet 的计算量均超过 140 GFLOPs,难以适配无人机等应用场景。MNAD 虽达到 65FPS,但参数量仍显臃肿,不利于边缘侧部署。不同于无监督方法的帧级像素输入,弱监督方法 TEVAD 则是由预训练 I3D 网络提取视频级特征输入,导致 FLOPs 过大。相比上述方法,提出方法展现出了极高的计算效率,两类模型 ResNet18 和 ResNet34 均实现了性能与速度的优化,模型参数量仅为 5.6M 和 10.9M。两种轻量化模型的推理速

度最高分别可达 102FPS 和 89FPS,在参数量与计算开销也远低于传统模型,满足无人机低空场景下的实时检测需求。

为进一步评估本文方法的性能,表 4 展示了在 Drone-Anomaly 数据集五种不同场景下基于 ResNet34 预训练编码器的多种性能指标。

提出方法在 Drone-Anomaly 数据集的五个子场景下的多个指标均优于其他方法。相比 GANomaly、MNAD 等算法,提出方法在 AUC、Recall 及 OA 等核心指标上均实现大幅提升,特别是在 Highway 场景下,AUC 高达 86.0%,F1 分数突破 90%。实验结果

证明,基于 Mamba 的架构能更精准地捕获无人机视角下的复杂时空异常,在各种高动态、多变换的监控任务中展现出卓越的检测鲁棒性与泛化能力。

3.4 误差分数可视化

图3直观展示了提出方法在 Drone-Anomaly 数据集不同场景下的异常分数可视化,纵轴表示归一化后的异常分数,横轴表示每个场景的帧索引,蓝色曲线表示通过高斯滤波器平滑得到的异常分数,蓝色区域表示异常区域。其中异常分数曲线与标注的真实异常区间具有高度的一致性:当视频中出现动物闯入道路、违法占用自行车道或行人违法穿越道路等异常行为时,异常分数迅速攀升并形成明显的峰值,且对应的标红关键帧精准捕捉到了事件发生

的瞬间。特别是在 Highway 和 Vehicle Roundabout 场景中,得分曲线在正常与异常交界处波动极小,展现了模型极强的判别力。此外,面对背景复杂的 Solar Panel Inspection 场景,提出方法可以高效检测出太阳能板周围出现的异常物体,证明了该方法在无人机高空动态视角下具有出色的时空建模能力与鲁棒性,能够有效区分环境噪声与真实异常。

3.5 消融实验

为了验证模型中各核心组件的贡献,表5展示了基于 ResNet18 模型的消融实验。实验结果表明,仅包含空间 Mamba 或频率 Mamba 的基准架构虽能初步捕捉异常特征,但性能受限于单一维度的表征能力;而将两者融合后,模型性能在所有场景下均实

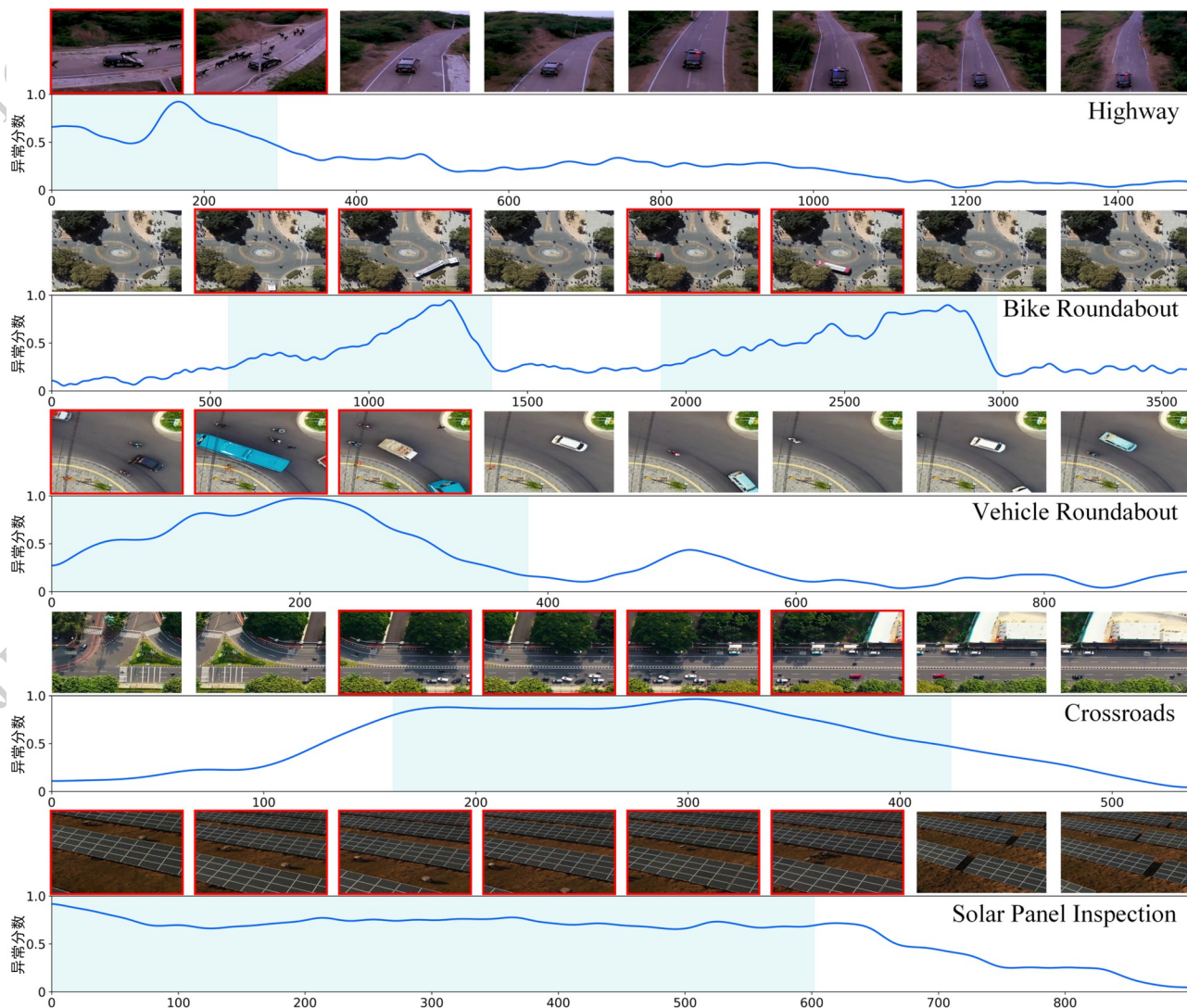


图3 Drone-Anomaly 数据集中五种子场景的误差分数可视化。上方标红的图像为具体的异常事件。

Fig. 3 We present the visualization of anomaly scores for five scenarios in Drone-Anomaly. The red image above represents the specific anomaly event. ((a)Highway;(b)Bike Roundabout;(c)Vehicle Roundabout;(d)Crossroads;(e)Solar Panel Inspection)

表3 不同方法的模型效率与推理速度比较

Table 3 Comparison of model efficiency and inference speed for other methods

方法	总参数量 (M)	FLOPs (G)	FPS
Frame-Pred(Liu 等, 2018)	24.2	148.1	25
MNAD(Park 等, 2020)	15.6	46.6	65
AMMCNet(Cai 等, 2022)	25.1	169.5	40
ASTNet(Le 等, 2023)	150.1	590.8	-
TEVAD*(Chen,等, 2023)	28.4	582.7	-
ASTT(Tran 等, 2024)	63.2	18.9	35
本文方法(ResNet18)	5.6	5.2	102
本文方法(ResNet34)	10.9	10.0	89

注:加粗字体为每列最优结果,*为弱监督方法。

现稳步提升,充分证明了挖掘空间分布与频率响应之间互补信息对复杂场景建模的重要性。进一步地,通过引入门控机制,模型不仅在各项指标上达到最优,更展示了该机制在自适应调节时频特征和抑制冗余噪声方面的卓越调度能力。此外,对比实验显示该门控机制在各项场景下的表现均优于 SENet 和 CBAM,进一步验证了其在复杂 UAV 环境下进行特征融合的针对性与优越性。从而验证了所提完整模型在 UAV 异常检测任务中的有效性与鲁棒性。

为了进一步分析不同预训练 CNN 编码器的特征提取能力和对无人机异常检测的性能影响。在 Drone-Anomaly 数据集的三个场景下进行了消融实验,如表 6 所示。ResNet50 凭借大参数量模型在多个场景中取得了最优的 AUC 表现,但由于其宽通道设计 $C_{50} = \{256, 512, 1024\}$ 。导致参与训练的参数量激增至 40.2M,这显著超出了边缘嵌入式设备对于实时推理的承载能力,不符合无人机轻量化部署的初衷。相反,MobileNetV3 虽具有极高的轻量化优势,但受限于较窄的通道配置 $C_m = \{24, 40, 112\}$,其特征表征能力难以有效处理高密度和复杂视角下的无人机图像,导致 AUC 降低。需要说明的是,Vehicle Roundabout 场景下,ResNet34 的结果低于 ResNet18,主要原因在于仅增加深度而未扩展通道的 ResNet34 在该场景下更易产生过拟合,降低对异常模式的泛化能力。相比之下,参数更紧凑的 ResNet18 具有更好的隐式正则化效果。而 ResNet50 通过显著增加网络宽度提升了多尺度语义表征能力

和训练参数容量,有效缓解了过拟合问题,因此性能重新提升。

相较于上述两种预训练模型,在综合权衡性能与效率的前提下,采用 ResNet18 和 ResNet34 两种模型来进行无人机视频异常检测,这俩种模型的通道数相同,仅为 $\{64, 128, 256\}$,在显著提升检测精度的同时,将训练参数量控制在仅 2.8M。这种配置在保留了足够的语义提取能力与满足边缘设备推理效率之间达到了理想的平衡。

4 结 论

本文提出了一种基于门控空间频率 Mamba 的无人机视频异常检测 GSFMMamba,旨在解决无人机低空场景下的实时性与高复杂度建模难题。通过构建并行的空间与频率 Mamba 块,在有效捕获长距离时空依赖的同时,显著降低了计算开销。实验结果表明,基于 ResNet18 主干的模型仅需 5.6M 参数量和 5.2GFLOPs,推理时可达 102FPS,在两种基准数据集上相较于主流方法检测性能更加出色或具有竞争性。证实提出方法能够精准捕捉复杂动态场景下的异常突发事件,亦实现检测精度与运行效率的良性兼容。

尽管提出方法在多数无人机监控任务中表现出色,但在部分极端光照变化或高度遮挡等极端复杂场景下的检测精度仍有提升空间。这反映出单一的结构化状态空间模型在应对非平稳环境噪声时,特征判别力尚存在一定的局限性。未来研究将在维持轻量化的前提下,探索引入因果推理与运动补偿机制以消除动态干扰,并结合多模态层次语义以提升微小目标的语义辨识力,以增强模型对复杂环境的抗干扰能力。

表4 提出方法与其他方法在 Drone-Anomaly 数据集下五个场景中的表现进行了更详细的比较,并评估了AUC、召回率、精确率、F1分数和总体准确率。

Table4 Compare the proposed method against other methods on five scenes of the Drone-Anomaly dataset , evaluating metrics such as AUC, Recall, Precision, F1 score, and overall accuracy (OA).

场景	模型	GANomaly (Akçay等, 2018)	Skip- GAN (Akçay 等, 2019)	MemAE (Gong等, 2019)	MNAD (Park 等, 2020)	MKD (Salehi 等,2021)	SSPCAB (Ristea 等,2022)	ANDT (Jin等, 2022)	本文 方法
Highway	AUC(%)	62.7	64.8	67.2	66.9	64.3	67.8	68.7	86.0
	Recall(%)	65.1	63.7	67.3	65.9	62.8	67.5	68.4	88.6
	Precision(%)	62.9	66.7	68.2	66.5	65.3	69.7	66.9	84.4
	F1 score(%)	64.0	65.2	67.7	66.2	64.0	68.6	67.7	90.9
	OA(%)	61.5	64.6	66.1	65.7	63.9	66.3	65.9	83.8
Bike roundabout	AUC	71.7	77.7	79.5	77.4	74.8	76.8	82.2	84.5
	Recall	70.2	73.5	74.8	72.4	70.6	74.6	78.5	80.3
	Precision	77.5	74.3	73.4	75.2	75.1	76.0	79.0	82.5
	F1 score	73.7	73.9	74.1	73.8	72.8	75.3	78.8	78.1
	OA	69.3	67.7	75.2	69.8	73.2	70.4	76.7	79.2
Vehicle roundabout	AUC	55.1	58.5	64.1	61.9	62.7	62.3	61.3	76.6
	Recall	58.6	59.3	60.8	57.9	59.7	59.7	57.8	77.8
	Precision	55.7	62.8	59.0	61.6	63.7	63.8	64.1	71.8
	F1 score	57.1	61.0	59.9	59.7	61.6	61.7	60.8	72.8
	OA	54.0	57.1	59.2	59.4	58.7	60.4	58.0	74.4
Railway inspection	AUC	61.7	65.8	58.9	58.0	62.4	59.1	59.4	69.6
	Recall	55.7	60.7	58.0	61.3	59.7	62.0	60.7	65.0
	Precision	56.2	64.6	58.4	56.1	60.3	58.7	61.3	70.8
	F1 score	56.0	62.6	58.2	58.6	60.0	60.3	61.0	67.7
	OA	53.8	60.3	58.0	57.1	60.8	59.4	57.4	63.4
Solar panel inspection	AUC	64.6	65.7	65.8	64.7	63.5	65.0	64.2	80.8
	Recall	59.1	58.8	62.1	58.6	57.6	59.2	61.2	64.3
	Precision	63.7	57.5	57.6	58.0	54.7	60.9	66.0	79.2
	F1 score	61.3	58.1	59.8	58.3	56.1	60.0	63.5	74.7
	OA	57.3	60.2	57.7	59.6	56.5	58.7	60.8	70.5

注:加粗字体为每行最优结果。

表 5 模块消融实验结果
Table5 Results of module ablation experiments

模型	门控	空间	频率	Railway inspection	Highway	Vehicle roundabout
1	—	√	—	59.58	77.50	66.75
2	—	—	√	59.86	80.82	70.99
3	—	√	√	65.54	81.06	80.02
4	SENet	√	√	69.05	81.44	82.45
5	CBAM	√	√	68.92	82.77	82.34
6	√	√	√	69.19	83.91	82.87

注：“√”表示有，“—”表示无。

表 6 不同预训练编码器的消融实验结果
Table6 Results of ablation experiments with different pre-trained encoders

模型	参数量 (M) (冻结/训练)	Railway inspection	Highway	Vehicle roundabout
MobileNetV3	2.9 / 0.3	60.15	79.51	73.24
ResNet18	2.7 / 2.8	69.19	83.91	82.87
ResNet34	8.1 / 2.8	69.65	86.08	76.61
ResNet50	8.5 / 40.2	73.72	88.94	85.53

注：预训练模型仅加载并冻结前三层参数，其参数量约占完整模型的 1/2 ~ 1/3。

参考文献 (References)

- Akay S, Atapour-Abarghouei A and Breckon T P. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training//Asian conference on computer vision. Perth: Springer: 622-637. [DOI: 10.1007/978-3-030-20893-6_39]
- Akay S, Atapour-Abarghouei A and Breckon T P. 2019. Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection//2019 International Joint Conference on Neural Networks. Budapest: IEEE: 1-8. [DOI: 10.1109/IJCNN.2019.8851808]
- Biswas D and Tesic J. MMVAD: 2025. A vision - language model for cross-domain video anomaly detection with contrastive learning and scale-adaptive frame segmentation. Expert Systems with Applications, 285: 127857. [DOI: 10.1016/j.eswa.2025.127857]
- Cai R, Zhang H, Liu W, Gao S and Hao Z. 2022. Appearance-motion memory consistency network for video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual: AAAI Press: 35(2): 938-946 [DOI: 10.1609/aaai.v35i2.16177]
- Chang Y, Tu Z, Xie W and Yuan J. 2020. Clustering driven deep auto-encoder for video anomaly detection//European Conference on Computer Vision. Glasgow: Springer: 329-345. [DOI: 10.1007/978-3-030-58555-6_20]
- Chen B, Cao L, Liu W, Liu J, Zhao B and Liu Y. Cross-modal attention fusion of RGB and skeleton for multimodal-driven video anomaly detection. Pattern Recognition, 2026: 113815. [DOI: 10.1016/j.patcog.2026.113815]
- Chen W, Ma K T, Yew Z J, Hur M and Khoo D A. 2023. Tevad: Improved video anomaly detection with captions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 5549-5559. [DOI: 10.1109/CVPRW59228.2023.0057]
- Chen Y, Liu Z, Zhang B, Fok W, Qi X and Wu Y. 2023. Mgnf: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection//Proceedings of the AAAI Conference On Artificial Intelligence. Washington: AAAI Press: 37(1): 387-395. [DOI: 10.1609/aaai.v37i1.25112]
- Fang Z, Zhou J T, Xiao Y, Li Y and Yang F. 2021. Multi-Encoder Towards Effective Anomaly Detection in Videos. IEEE transactions on Multimedia, 23: 4106-4116 [DOI: 10.1109/TMM. 2020. 3037538]
- Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S and Hengel AV. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 1705-1714 [DOI: 10.1109/ICCV.2019. 00179]
- Gu A and Dao T. 2023. Mamba: Linear-time sequence modeling with selective state spaces[EB/OL]. [2026-04-15]
<https://arxiv.org/pdf/2312.00752>
- Guo C, Wang H, Xia Y and Feng G. 2023. Learning Appearance Motion Synergy via Memory-Guided Event Prediction for Video Anomaly Detection. IEEE Transactions on Circuits and Systems for Video Technology, 34 (3) : 1519-1531 [DOI: 10.1109/TCSVT. 2023.3297114]
- He H, Bai Y, Zhang J, He H Y, He Q D and Li X T, et al. 2024. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection//Advances in Neural Information Processing Systems. Vancouver: NIPS: 37: 71162-71187. [DOI: 10.52202/079017-2274]
- Jin P, Mou L, Xia G S and Zhu X X. 2022. Anomaly detection in aerial videos with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-13. [DOI: 10.1109/TGRS.2022.3198130]
- Kommanduri R and Ghorai M. DAST-Net: 2024. Dense visual attention augmented spatio-temporal network for unsupervised video anomaly detection. Neurocomputing, 579: 127444. [DOI: 10.1016/j.neucom.2024.127444]
- Le V T, Jin H and Kim Y G. 2025. HSTforU: anomaly detection in aerial and ground-based videos with hierarchical spatio-temporal transformer for U-net. Applied Intelligence, 55(4): 261. [DOI: 10.1007/s10489-024-06042-4]
- Le V T and Kim Y G. 2023. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence, 53(3) : 3240-3254 [DOI: 10.1007/s10489-022-03613-1]
- Li J, Huang Q, Du Y, Zhen X, Chen S and Shao L. 2021. Variational abnormal behavior detection with motion consistency. IEEE Transactions on Image Processing, 31: 275-286 [DOI: 10.1109/TIP.2021. 3130545]
- Li Y F, Huang X, Guo C Y, Wang Y R, Wu S M and Jian J B. 2026. UAV-UGV Cooperative Perception for Low-altitude Transportation: A Review. ACTA AUTOMATICA SINICA.52(02): 210-229 (李永福, 黄鑫, 郭常员, 王怡然, 吴三妹, 简金埠. 面向低空交通运输的无人机-无人车协同感知技术综述. 自动化学报, 2026, 52(02): 210-229). [DOI: 10.16383/j.aas.c250386.]
- Li Z, Zhao M Y, Yang X, Liu Y, Sheng J M and Zeng X H, et al. 2026. Stnmamba: Mamba-based spatial-temporal normality learning for video anomaly detection. IEEE Transactions on Multimedia. 1-12 [DOI: 10.1109/TMM.2026.3664969]
- Liang J F, Li T, Yang J Q, Li Y N, Fang Z W and Yang F. 2023. Video Anomaly Detection Combining Self-Attention and Auto-Encoder. Journal of Image and Graphics, 28 (4) : 1029-1040 (梁家菲, 李婷, 杨佳琪, 李亚南, 方智文, 杨丰. 2023. 融合自注意力和自编码器的视频异常检测. 中国图象图形学报, 28 (4) : 1029-1040) [DOI: 10.11834/jig.211147]
- Liu C M, Xue R, Shi L, Li Y H and Gao Y F. 2022. The gating self-attention mechanism and GAN integrated video anomaly detection.

- Journal of Image and Graphics, 27 (11): 3210-3221 (刘成明, 薛然, 石磊, 李英豪, 高宇飞. 2022. 融合门控自注意力机制的生成对抗网络视频异常检测. 中国图象图形学报, 27 (11): 3210-3221) [DOI: 10.11834/jig.210520]
- Liu J M and Zhuang W K. 2025. Industrial anomaly detection by combining visual Mamba and patch feature distribution. Journal of Image and Graphics, 30(10): 3215-3229 (刘建明, 庄维宽. 2025. 结合视觉Mamba和块特征分布的工业异常检测. 中国图象图形学报, 30(10): 3215-3229) [DOI: 10.11834/jig.240594]
- Liu W, Luo W, Lian D and Gao S. 2018. Future frame prediction for anomaly detection - a new baseline//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 6536-6545 [DOI: 10.1109/CVPR.2018.00684]
- Liu Y, Chen B, Zhu X, Liu J, Sun P, Zhou W. M2S2L: Mamba-based Multi-Scale Spatial-temporal Learning for Video Anomaly Detection [EB/OL]. [2026-04-15]
<https://arxiv.org/pdf/2511.05564>
- Loshchilov I and Hutter F. Sgdr: Stochastic gradient descent with warm restarts [EB/OL]. [2026-04-15]
<https://arxiv.org/pdf/1608.03983>
- Lyu J, Zhao M, Hu J, Huang X W, Chen Y F and Du S L. 2025. Vad-mamba: Exploring state space models for fast video anomaly detection//2025 IEEE International Conference on Multimedia and Expo. Nantes: IEEE: 1-6. [DOI: 10.1109/ICME59968.2025.11209020]
- Lyu J, Zhao M, Hu J, Xi R T, Huang X W and Du S L, et al. 2026. Bidirectional skip-frame prediction for video anomaly detection with intra-domain disparity-driven attention. Pattern Recognition, 170: 112010. [DOI: 10.1016/j.patcog.2025.112010]
- Park H, Noh J and Ham B. 2020. Learning memory-guided normality for anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 2020: 14372-14381 [DOI: 10.1109/CVPR42600.2020.01438]
- Qiu S, Ye J, Zhao J, He L, Liu L Y and E B C, et al. 2024. Video anomaly detection guided by clustering learning. Pattern Recognition. 153: 110550. [DOI: 10.1016/j.patcog.2024.110550]
- Ristea N C, Madan N, Ionescu R T, Nasrollahi K, Khan F S and Moeslund T B, et al. 2022. Self-supervised predictive convolutional attentive block for anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 13576-13586. [DOI: 10.1109/CVPR52688.2022.01321]
- Salehi M, Sadjadi N, Baselizadeh S, Rohban M H and Rabiee H R. 2021. Multiresolution knowledge distillation for anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 14902-14912. [DOI: 10.1109/CVPR46437.2021.01466]
- Shen S, Guo Z, Li Y, Wang L and Guo Z. TrCLIP-VAD: Weak supervised video anomaly detection by improving CLIP training with text rewriting. Neural Networks, 2026: 108951. [DOI: 10.1016/j.neunet.2026.108951]
- Singh R, Sethi A, Saini K, Saurav S, Tiwari A and Singh S. 2024. Attention-guided generator with dual discriminator GAN for real-time video anomaly detection. Engineering Applications of Artificial Intelligence, 131: 107830. [DOI: 10.1016/j.engappai.2023.107830]
- Tran T M, Bui D C, Nguyen T V and Nguyen K. 2024. Transformer-based spatio-temporal unsupervised traffic anomaly detection in aerial videos. IEEE Transactions on Circuits and Systems for Video Technology. 34 (9) : 8292-8309. [DOI: 10.1109/TCSVT. 2024. 3376399]
- Tran T M, Vu T N, Nguyen T V and Nguyen K. 2023. UIT-ADrone: A novel drone dataset for traffic anomaly detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 16: 5590-5601. [DOI: 10.1109/JSTARS. 2023.3285905]
- Wang L, Tian J, Zhou S, Shi H Y and Hua G. 2023. Memory-augmented appearance-motion network for video anomaly detection. Pattern Recognition. 138: 109335. [DOI: 10.1016/j.patcog.2023.109335]
- Wu P, Zhou X, Pang G, Zhou L R, Yan Q S, and Wang P, et al. 2024. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press: 38 (6) : 6074-6082. [DOI: 10.1609/aaai.v38i6.28423]
- Yu G, Wang S, Cai Z, Liu X W, Zhu E and Yin J P. 2023. Video anomaly detection via visual cloze tests. IEEE Transactions on Information Forensics and Security. 18: 4955-4969. [DOI: 10.1109/TIFS.2023.3300094]
- Zhang H, Cao C, Lv Q, Min L T and Zhang Y N. 2025. Autoregressive denoising score matching is a good video anomaly detector//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu: IEEE: 12057-12067. [DOI: 10.1109/ICCV51701.2025.01121]
- Zhang J H, Hong D Q, Huang Y J, Hong W J, Chen X and Chen B M. 2026. Methodology of UAV-based digital twins for smart environments. SCIENTIA SINICA Technologica, 2026, 56 (3) : 507-517. (张积哈, 洪端钦, 黄逸君, 洪文兴, 陈曦, 陈本美. 面向智慧环境的无人机数字孪生方法研究[J/OL]. 中国科学: 技术科学, 56, 1-11 [DOI: 10.1360/SST-2026-0047])
- Zhao M, Zeng X, Liu Y, Liu J and Pang C. Rethinking prediction-based video anomaly detection from local - global normality perspective. Expert Systems with Applications, 2025, 262: 125581. [DOI: 10.1016/j.eswa.2024.125581]
- Zhong Y, Chen X, Hu Y, Tang P and Ren F. 2022. Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. IEEE Transactions on Circuits and Systems for Video Technology, 32 (12) : 8285-8296 [DOI: 10.1109/TCSVT. 2022.3190539]
- Zhou H, Cai J, Ye Y, Feng Y H, Gao C X and Yu J Q, et al. 2025.

Video anomaly detection with motion and appearance guided patch diffusion model//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia: AAAI Press: 39 (10) : 10761-10769. [DOI:10.1609/aaai.v39i10.33169]

作者简介

胡静,女,副教授,主要研究方向为高光谱遥感影像处理。E-mail: jinghu@xaut.edu.cn

都双丽,女,副教授,主要研究方向为立体匹配和三维重建。E-mail: dusl@xaut.edu.cn

石程,女,副教授,主要研究方向为高分辨率遥感图像处理。E-mail: chengc_s@163.com

陈逸飞,男,硕士研究生,主要研究方向为视频异常检测。E-mail: yfchen@stu.xaut.edu.cn

吕志勇,男,教授,主要研究方向为遥感变化检测。E-mail: Lvzhiyong_fly@hotmail.com