中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Zhao Minghua, Wang Nan, Lyu Jiahao, Hu Jing, Du Shuangli, Shi Cheng, Wang Lin, You Zhenzhen. Anomaly detection in aerial and ground videos using dual-stream network with memory enhancement[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI:10.11834/jig.260199.(赵明华, 王楠, 吕佳豪, 胡静, 都双丽, 石程, 王琳, 尤珍臻. 融合双流网络和记忆增强的空地视频异常检测[J/OL]. 中国图象图形学报,XXXX:1-15. DOI:10.11834/jig.260199.)[DOI:10.11834/jig.260199]

融合双流网络和记忆增强的空地视频异常检测

赵明华[1,2],王楠[1],吕佳豪[1*],胡静[1],都双丽[1],石程[1],王琳[1],尤珍臻[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**　视频异常检测作为视频监控系统的核心任务之一,在公共安全与智能监控等领域具有重要应用价值。同时在无人机航拍场景下,由于视角变化剧烈、目标尺度不稳定以及背景复杂多变,使得异常检测任务更加具有挑战性。现有方法大多仅侧重于静态外观特征建模,忽视了动态运动信息所蕴含的关键时序特征,同时普遍依赖生成模型进行重建或预测,容易出现"过度泛化"问题,从而削弱对异常事件的判别能力。针对上述问题,提出一种融合双流网络和记忆增强的空地视频异常检测方法。**方法**　首先,通过双编码器架构分别提取视频序列的外观特征和运动特征,将同一尺度的两类特征进行融合。其次,将高维融合特征送入具有更新策略的记忆增强模块中,学习多样化的正常特征。最后,采用跳跃连接机制把多尺度融合特征和记忆增强后的特征送入具有注意力的解码器中预测未来帧。**结果**　在UCSD Ped2、CUHK Avenue和ShanghaiTech三个地面基准数据集上,所提方法的AUC分别达到98.8%、89.1%和74.7%;同时,在Drone-Anomaly无人机航拍数据集的多个子场景中均取得优异性能,尤其在Railway Inspection和Farmland Inspection场景中分别达到94.76%和91.41%,优于多种对比方法。**结论**　本文方法通过协同建模外观与运动信息,并结合记忆增强机制,有效缓解了过度泛化问题,提升了模型对复杂动态场景中异常事件的判别能力,在地面监控与无人机航拍场景下均表现出良好的鲁棒性与泛化性能。

关键词:异常检测;卷积神经网络;记忆网络;融合算法;注意力机制

Anomaly detection in aerial and ground videos using dual-stream network with memory enhancement

Zhao Minghua[1,2], Wang Nan[1], Lyu Jiahao[1*], Hu Jing[1], Du Shuangli[1], Shi Cheng[1], Wang Lin[1], You Zhenzhen[1]

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** Video anomaly detection (VAD) aims to automatically identify events that deviate from learned normal patterns in video sequences and has become a critical component of intelligent surveillance systems. It plays an important role in public security, intelligent transportation, industrial inspection, and urban management. With the rapid

收稿日期:2026-04-14;**修回日期**:2026-06-16

*通信作者:吕佳豪,男,博士研究生,主要研究方向为视频异常检测。E-mail: jhlyu25@stu.xaut.edu.cn

基金项目:陕西省自然科学基金项目(2024JC-ZDXM-35, 2024JC-YBMS-573, 2024JC-YBMS-458);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: Nature Science Foundation of Shaanxi Province, China(2024JC-ZDXM-35, 2024JC-YBMS-573, 2024JC-YBMS-458); 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)

deployment of unmanned aerial vehicles (UAVs), large-scale aerial video data have become increasingly available, creating new opportunities for anomaly detection. However, compared with conventional fixed-camera surveillance videos, UAV videos exhibit substantial viewpoint variations, severe object-scale changes, dynamic backgrounds, and complex motion patterns, which significantly increase the difficulty of anomaly detection. Although recent deep-learning-based methods have achieved promising performance, most existing approaches focus primarily on appearance modeling while insufficiently exploiting motion information, despite the fact that abnormal events are often characterized by unusual motion patterns such as sudden acceleration, abrupt direction changes, illegal crossings, or abnormal trajectories. Furthermore, reconstruction-based and prediction-based methods commonly rely on generative models to learn normal patterns and identify anomalies through reconstruction or prediction errors. These models often suffer from the over-generalization problem, where abnormal samples can still be reconstructed or predicted with relatively low errors, thereby reducing the discriminative capability between normal and abnormal events. Therefore, developing a unified framework that can effectively model both appearance and motion information while enhancing the representation of diverse normal patterns remains an important challenge for anomaly detection in both ground-based and aerial surveillance scenarios. **Method** To address the above issues, a dual-stream memory-enhanced network for ground-to-air video anomaly detection is proposed. The framework adopts a dual-encoder single-decoder architecture. First, a pretrained FlowNet2 optical-flow estimation network is employed to extract motion information from consecutive video frames. To prevent overfitting and reduce training complexity, the FlowNet2 parameters remain fixed throughout the training process. The original RGB frames and corresponding optical-flow maps are then fed into two independent encoders, namely an appearance encoder and a motion encoder, to extract multi-scale appearance features and motion features, respectively. To establish a stronger correlation between appearance and motion representations, a variance-attention appearance-motion fusion module is designed. Specifically, the spatial variance of motion features is computed to characterize the intensity of motion changes. Regions exhibiting large motion variance, such as sudden acceleration, abrupt direction changes, or unusual object movements, are assigned higher attention weights. These weights are subsequently used to enhance appearance features, enabling the network to focus on potentially anomalous regions while suppressing irrelevant background information. Through this mechanism, appearance and motion features can be deeply fused at multiple scales, resulting in more discriminative feature representations. To alleviate the over-generalization problem and improve the learning of diverse normal patterns, a memory-enhanced module with a dynamic update strategy is further introduced. The memory module consists of multiple learnable memory items that store representative normal feature prototypes. During the memory-reading process, query features extracted from high-level fused representations retrieve the most relevant memory items according to cosine similarity. The retrieved memory features are then combined with the original query features to generate memory-enhanced representations. During memory updating, a regularization score derived from prediction errors is employed to distinguish normal frames from potential abnormal frames. Only features associated with normal frames are allowed to update memory items, thereby preventing abnormal patterns from contaminating the memory bank. This strategy enables the memory module to continuously learn diverse normal behaviors while preserving its discriminative capability. In the decoding stage, multi-scale fused features are delivered to the decoder through skip connections, allowing low-level spatial details and high-level semantic information to be jointly utilized for future-frame prediction. Furthermore, efficient channel attention modules are embedded after the first three upsampling layers to adaptively recalibrate channel-wise feature responses and enhance prediction quality with negligible computational overhead. The entire network is optimized using a joint loss function composed of prediction loss, optical-flow loss, feature compactness loss, and feature separation loss. These losses collaboratively encourage accurate future-frame prediction, motion consistency preservation, compact intra-class representations, and discriminative memory-item distributions. **Results** Extensive experiments were conducted on three benchmark ground-surveillance datasets, namely UCSD Ped2, CUHK Avenue, and ShanghaiTech, as well as the Drone-Anomaly UAV dataset. Experimental results demonstrate that the proposed method achieves state-of-the-art or highly competitive performance across different scenarios. On the UCSD Ped2 dataset, the proposed method achieved an AUC of 98.8%, outperforming representative methods such as MemAE (94.1%), MNAD (97.0%), and AMMC (96.6%). On the CUHK Avenue dataset, an AUC of 89.1% was obtained, slightly surpassing DEDDnet (89.0%) and exceeding most recent prediction-based

approaches. On the challenging ShanghaiTech dataset, which contains more than 270, 000 training frames, multiple scenes, and diverse anomaly categories, the proposed method achieved an AUC of 74. 7%, demonstrating strong robustness under complex environments. To further evaluate the generalization capability in aerial surveillance scenarios, experiments were conducted on the Drone-Anomaly dataset, which contains multiple UAV-captured inspection and traffic-monitoring scenes. The proposed method achieved 94. 76% AUC with an EER of 0. 09% on Railway Inspection and 91. 41% AUC with an EER of 0. 10% on Farmland Inspection, significantly outperforming Frame-Pred, MNAD, ANDT, MKD, and SSPCAB. In addition to detection accuracy, computational efficiency was evaluated. The proposed model contains only 20. 39M parameters and achieves an inference speed of 62. 66 FPS on an NVIDIA RTX4090 GPU, outperforming several existing methods in terms of both efficiency and accuracy. Even when the computational cost of FlowNet2 is included, the end-to-end processing speed remains approximately 39. 76 FPS, satisfying the requirements of most real-time surveillance applications. Ablation studies further validate the effectiveness of the proposed components. Additional loss-function ablations confirmed that the combination of prediction loss, optical-flow loss, compactness loss, and separation loss yields the best detection performance. **Conclusion** This paper presents a dual-stream memory-enhanced framework for video anomaly detection in both ground-based and UAV surveillance scenarios. By jointly modeling appearance and motion information through a variance-attention fusion mechanism and learning diverse normal patterns via a dynamically updated memory module, the proposed method effectively alleviates the over-generalization problem and improves anomaly discrimination in complex dynamic environments. Extensive experiments on multiple benchmark datasets demonstrate its robustness, generalization capability, computational efficiency, and real-time performance. The proposed framework provides a practical and effective solution for intelligent anomaly detection across diverse surveillance applications and offers a promising foundation for future research on unified ground-to-air video understanding systems.

Key words: anomaly detection; convolutional neural networks; memory networks; fusion algorithms; attention mechanisms

论文引用格式：Zhao Minghua, Wang Nan, Lyu Jiahao, Hu Jing, Du Shuangli, Shi Cheng, Wang Lin, You Zhenzhen. Anomaly detection in aerial and ground videos using dual-stream network with memory enhancement[J/OL]. Journal of Image and Graphics, XXXX: 1-14. DOI: 10. 11834/jig. 260199. (赵明华，王楠，吕佳豪，胡静，都双丽，石程，王琳，尤珍臻. 融合双流网络和记忆增强的空地视频异常检测[J/OL]. 中国图象图形学报, XXXX: 1-14. DOI: 10. 11834/jig. 260199)[DOI: 10. 11834/jig. 260199]

0 引 言

视频异常检测(video anomaly detection, VAD)旨在从视频序列中自动检测出偏离预期模式的异常行为(Gong等, 2019), 在公共安全、智能交通等领域具有重要应用价值。随着监控摄像头的广泛部署以及无人机(unmanned aerial vehicle, UAV)技术的普及, 产生海量的空地视频数据, 依赖人工处理不仅耗时耗力, 也难以满足实际应用中对实时性与准确性的要求。与地面固定摄像头采集的视频相比, 无人机视频具有视角动态变化显著、目标尺度变化剧烈、背景复杂多变以及运动模式多样化等特点, 这使得异常检测任务更加具有挑战性。因此, 亟需构建一种能够同时适用于地面监控与航拍视频的统一异常检测框架。

尽管地面监控视频与航拍视频在视角、背景和运动模式上存在差异, 但二者在异常检测任务中具有高度共性: 第一, 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; 第二, 两类视频均同时包含外观信息与运动信息, 且运动信息(如突然加速、方向突变等)往往是判别异常行为的关键。为此, 大多数VAD采用无监督方法, 仅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。

随着深度学习的发展, 当前无监督VAD方法主要分为基于重建(Li等, 2013; Zhang等, 2022; Fang等, 2020)和基于预测(Le等, 2023; Li等, 2023; Guo等, 2023)两类。基于重建的方法通过比较输入视频与重建结果之间的差异实现异常检测。该类方法通常采用自编码器(autoencoder, AE)或生成对抗网络(generative adversarial network, GAN)作为基础框

架。在AE框架下，Hasan等人（2016）提出基于卷积自编码器的视频异常检测方法。Fang等人（2020）提出多编码器单解码器结构分别建模外观与运动信息。Wang等人（2022）引入可学习卷积注意模块增强时空建模能力。Wu等人（2020）提出结合时空融合特征的稀疏编码网络。Kommanduri等人（2023）利用双残差卷积自编码器提升特征提取能力。在GAN框架下，He等人（2024）采用对抗数据增强策略提升异常检测性能。Zaheer等人（2022）通过构造不同质量的重建样本增强判别能力。Zavrtanik等人（2021）提出端到端表面异常检测和定位方法，可在无需复杂后处理的情况下直接实现异常区域的定位。然而，由于重建模型往往倾向于学习输入数据的恒等映射，即使面对异常样本也能够获得较好的重建结果，容易产生过度泛化问题，从而降低模型对异常事件的敏感性。

相比之下，基于预测的方法通过利用历史视频帧预测未来帧，并依据预测误差判断异常，能够更好地利用视频时序信息，因此近年来受到广泛关注。根据输入数据的不同，该类方法又可分为单流预测和双流预测两类。单流预测方法使用单流生成模型来学习正常事件的时空模式。Le等人（2023）提出基于注意力机制的时空预测网络。Hao等人（2022）通过时空一致性约束增强预测能力。梁家菲等人（2023）结合Transformer与U-Net实现全局与局部信息融合。为了进一步利用运动信息，研究者提出了双流预测框架。Liu等人（2022）构建外观-运动联合自编码器学习正常事件的典型时空特征。Cai等人（2021）设计特征传递模块增强双流关联性。Ma等人（2025）利用交叉注意力和记忆模块实现外观与运动特征的深度融合。Yang等人（2023）通过分类流与预测流的决策级融合提升异常检测能力。

近年来，随着无人机异常检测研究的兴起，越来越多学者开始关注航拍场景下的异常行为分析。Tran等人（2024）提出了一种基于Transformer架构的未来帧预测网络。Hamdi等人（2021）提出端到端网络从原始图像生成光流并提取时空特征，并提取紧凑的时空特征以用于异常检测。Chriki等人（2021）对CNN特征和传统手工特征进行了系统比较。Pu等人（2022）提出基于卷积变分自编码器的未来帧预测方法，并构建Drone-Anomaly数据集。Ahmed等人（2025）进一步提出基于时空关系交叉变换器的异常检测框架。虽然上述方法在无人机场景中取得了一定进展，但由于视角变化剧烈、背景动态复杂以及目标尺度变化显著，现有方法在稳定建模正常模式及有效区分异常方面仍面临较大挑战，尤其是在外观与运动信息的协同建模方面仍有提升空间。

综上所述，现有视频异常检测方法虽然能够在一定程度上学习视频的时空信息并捕获动态变化特征，但仍存在两个主要问题：一是模型容易产生“过度泛化”现象（Lee等，2022），对正常特征分布的学习不够充分；二是外观特征与运动特征之间的关联建模不足，难以充分发挥二者的互补优势。

针对上述问题，提出了一种融合双流网络和记忆增强的空地视频异常检测方法。该方法通过双编码器分别提取外观与运动特征，并利用基于方差注意力的外观运动融合模块实现深度融合，从而增强对复杂运动模式和尺度变化的建模能力；同时引入记忆增强模块以学习多样化正常特征，缓解过度泛化问题；在解码阶段结合多尺度特征与通道注意力，提高预测精度。实验结果表明，该方法在多个公开地面监控数据集及无人机航拍数据集下均取得了较好的检测性能，验证了其在复杂动态环境中的有效性与鲁棒性。

1 本文方法

提出一种融合双流网络和记忆增强的空地视频异常检测方法，如图1所示。该方法由双编码器和基于注意力的单解码器架构、基于方差注意力的外观运动融合模块（variance-based attention appearance-motion fusion module，VA-AMFM）和记忆增强模块（memory-enhanced module，MEM）组成。光流信息通过FlowNet2模型（Ilg等，2017）进行提取，该网络能解决光流估计中的小位移场景下的性能问题，在保持高精度的同时通过模型压缩实现了轻量化。在训练过程中，FlowNet2模型保持冻结状态，不参与端到端的梯度更新。该网络仅用于从输入视频帧中提取光流作为运动特征，其参数在

ImageNet上预训练后固定，以避免因训练数据不足导致的过拟合问题。

具体而言，首先将视频帧和光流分别输入至对应的编码器，以提取多尺度的外观特征与运动特征；然后，在各尺度上通过VA-AMFM对外观与运动特

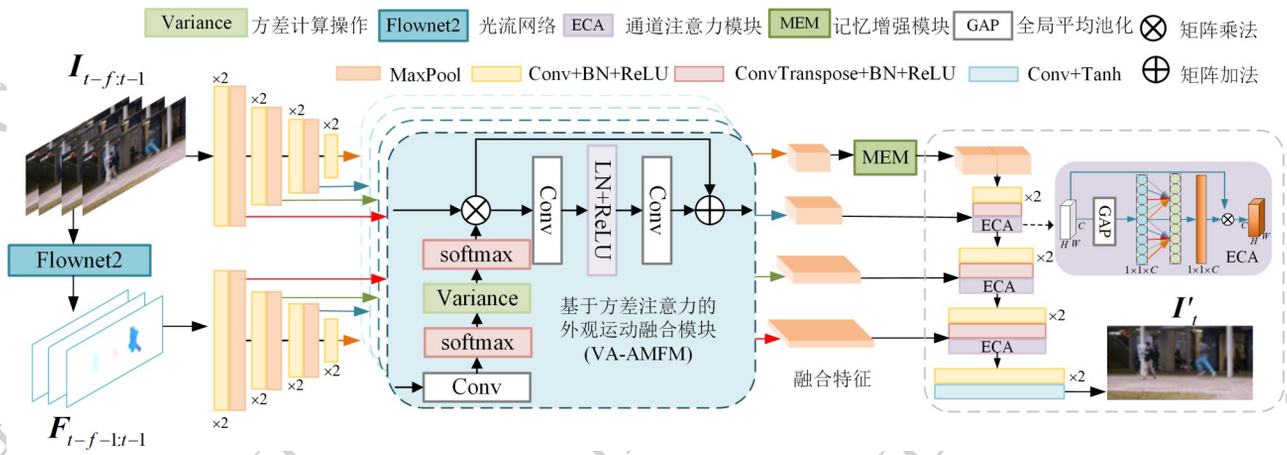


图1　网络总体架构

Fig. 1　The overall architecture of the network

征进行深度融合，获得多尺度融合特征；其次，将高维融合特征送入到MEM模块中，检索出最相关的特征再送入到具有通道注意力的解码器；最后，将多尺度融合特征跳跃连接到解码器，从解码器中预测出下一时刻的视频帧。所有模块将在下面的小节中详细介绍。

1.1　双编码器和基于注意力单解码器架构

为实现外观与运动信息的有效融合，采用双编码器单解码器结构，如图1所示。具体而言，基于U-Net(Ronneberger等，2015)框架，输入连续的f帧视频序列$\{\boldsymbol{I}_{t-f},\cdots,\boldsymbol{I}_{t-2},\boldsymbol{I}_{t-1}\}$和对应的$f$-1个光流序列$\{\boldsymbol{F}_{t-f-1},\cdots,\boldsymbol{F}_{t-2},\boldsymbol{F}_{t-1}\}$，通过外观编码器与运动编码器分别提取不同尺度的外观与运动特征。

在解码阶段，利用反卷积操作对高维融合特征进行上采样，同时将多尺度融合特征跳跃连接到解码器中，以充分利用不同层级的语义信息，从而生成预测帧$\boldsymbol{I'}_t$。为增强模型对关键通道信息的建模能力，在解码器的前三个上采样层之后均引入高效通道注意力(efficient channel attention，ECA)(Wang等，2020)模块，以实现跨通道特征的自适应加权，在提升检测性能的同时保持较低的模型复杂度。值得注意的是，解码器中连接的多个融合特征不是从同一个特征图上采样得到的，而是多个独立的特征，每个尺度的融合特征都充分融合了外观信息与运动信息。

1.2　基于方差注意力的外观运动融合模块

为充分挖掘视频序列中外观与运动信息之间的互补性，并建立两者之间的内在关联，提出基于方差注意力的外观运动融合模块VA-AMFM。该模块对双编码器提取到的同一尺度的两类特征进行基于方差注意力机制的融合，并通过跳跃连接将融合特征传递至解码器的对应阶段，从而更好的捕捉到运动信息的细小变化。具体的结构如图1所示。

在视频异常检测中，异常事件通常伴随着局部区域内运动幅度或运动模式的显著变化，例如突然加速、方向突变等。为突出这一特性，该模块引入方差作为度量运动特征在空间维度上变化强度的指标，使得运动变化剧烈的区域获得更高的响应权重。具体来说，首先通过1×1卷积对运动特征$\boldsymbol{f}_t$进行全局上下文建模，压缩并聚合通道信息，通过Softmax函数归一化，生成空间注意力分布：

$$\boldsymbol{o}_t = \mathrm{Softmax}(\mathrm{Conv}_{1\times1}(\boldsymbol{f}_t)) \tag{1}$$

随后，在空间维度上计算该特征的方差，用以刻画不同空间位置的响应离散程度：

$$\boldsymbol{\sigma}^2 = \|\boldsymbol{f}_t - 1/D\sum_{d=1}^{D}\boldsymbol{o}_t^d\|_2^2 \tag{2}$$

式中，b为批次大小，$D = W \times H$表示空间维度数。

最后，将方差映射为注意力权重：

$$\mathbf{ATT} = \mathit{S}\mathrm{oftmax}(\boldsymbol{\sigma}^2) \tag{3}$$

在获得运动方差注意力后，进一步利用该注意力对外观特征进行增强，从而实现外观与运动信息的深度融合。具体而言，使用1×1卷积和加法操作，对每个位置的特征进行外观全局上下文特征聚合，得到特征$\boldsymbol{q}_t$：

$$\boldsymbol{q}_t = \boldsymbol{a}_t + \mathrm{ReLU}(\mathrm{LN}(\mathrm{Conv}_{1\times1}(\boldsymbol{a}_t \odot \mathbf{ATT}))) \tag{4}$$

式中，$\boldsymbol{a}_t$表示从编码器中提取的外观特征；$\mathrm{Conv}_{1\times1}(\cdot)$表示1×1卷积。

1.3 记忆增强模块

无监督视频异常检测方法在训练时仅学习正常样本的特征，因此正常特征的学习尤为重要。为了加强正常特征的学习，抑制神经网络的过度泛化，引入了记忆增强模块MEM（Park等，2020），结构如图2所示。在融合特征上进行记忆增强，对正常的数据进行多样化学习。考虑到高维融合特征具有空间尺寸较小、语义信息更加丰富的特点，同时为了减少模型参数量与计算开销，模型仅在送入解码器之前的高维融合特征层引入记忆增强模块。

将高维融合特征$\boldsymbol{q}_t$在空间维度上划分为N个查询$\boldsymbol{q}_t^k \in \mathbf{R}^{1\times 1\times C}$，$k=1,2,\ldots,N$，$N=H\times W$。记忆存储模块由$M$个记忆项组成，分别表示为$\boldsymbol{m}_n$，$n=1,\ldots,M$，每个记忆项用于记录一种典型的正常特征。通过存储模块中的记忆项来对查询特征进行记忆更新。整个过程分为读取和更新两个部分。

读取阶段：具体结构如图2-(a)所示。首先计算每个查询$\boldsymbol{q}_t^k$和所有记忆项$\boldsymbol{m}_n$之间的余弦相似度来衡量两者之间的相似性，并通过Softmax函数计算出对应的权重$\boldsymbol{w}_t^{(k,n)}$：

$$\boldsymbol{w}_t^{(k,n)}=\frac{\exp((\boldsymbol{m}_n)^{\mathrm{T}}\boldsymbol{q}_t^k)}{\sum_{n=1}^{M}\exp((\boldsymbol{m}_n)^{\mathrm{T}}\boldsymbol{q}_t^k)} \tag{5}$$

式中，n为记忆项索引，k为查询索引。随后，对所有记忆项进行加权求和，得到与查询特征对应的聚合特征$\boldsymbol{m}_t^k$：

$$\boldsymbol{m}_t^k=\sum_{n=1}^{M}\boldsymbol{w}_t^{(k,n)}\boldsymbol{m}_n \tag{6}$$

$\boldsymbol{m}_t^k$可以理解为多个正常特征记忆项的组合表示，能够反映查询特征在不同正常行为模式下的响应情况。最后，为了不丢失原有的查询信息，将聚合项$\boldsymbol{m}_t^k$与查询项$\boldsymbol{q}_t^k$沿着通道维度进行拼接得到读取的特征$\boldsymbol{m}'^k_t$：

$$\boldsymbol{m}'^k_t=\mathrm{Concat}(\boldsymbol{q}_t^k,\boldsymbol{m}_t^k) \tag{7}$$

通过读取存储模块中所有记忆项，使模型学习不同的正常特征。

更新阶段：为了使记忆模块能够持续学习不同类型的正常行为，该模块引入了记忆更新策略，具体结构如图2-(b)所示。记忆更新遵循“最相似匹配更新”原则，即每个记忆项由所有与该记忆项最相似的查询特征共同更新。首先按照与读取阶段相同的方式，计算记忆项$\boldsymbol{m}_n$与所有查询特征$\boldsymbol{q}_t^k$之间的相似度，并在查询维度上进行Softmax归一化：

$$\boldsymbol{y}_t^{(k,n)}=\frac{\exp((\boldsymbol{m}_n)^{\mathrm{T}}\boldsymbol{q}_t^k)}{\sum_{\mathrm{k}=1}^{\mathrm{N}}\exp((\boldsymbol{m}_d)^{\mathrm{T}}\boldsymbol{q}_t^k)} \tag{8}$$

对于第n个记忆项$\boldsymbol{m}_n$，为其建立一个查询索引集合$\boldsymbol{U}_t^n$，选取读取阶段与该记忆项目权重$\boldsymbol{w}_t^{(k,n)}$最大的查询索引k，将其加入对应的查询索引集合$\boldsymbol{U}_t^n$，该过程如图2-2(c)所示。随后，仅利用集合$\boldsymbol{U}_t^n$中的查询特征对记忆项进行更新，并对权重进行重新归一化：

$$\boldsymbol{y}'^{(k,n)}_t=\frac{\boldsymbol{y}_t^{(k,n)}}{\max_{k'\in U_t^n}\boldsymbol{y}_t^{(k,n)}} \tag{9}$$

最后，将索引集合中的查询与归一化之后的权重进行加权求和，实现对记忆项$\boldsymbol{m}_n$的更新。

在测试阶段，由于测试集中同时包含正常帧与异常帧，若继续更新记忆项，模型可能会错误地将异常特征存入记忆模块，导致异常检测性能下降。为此，引入一个正则评分函数F_t，用于判断当前帧是否参与记忆更新。

$$F_t=\sum_{x,y}\boldsymbol{W}_{xy}(\boldsymbol{I}'_t,\boldsymbol{I}_t)\|\boldsymbol{I}'^{xy}_t-\boldsymbol{I}_t^{xy}\|_2 \tag{10}$$

式中，x，y为空间坐标索引，$\boldsymbol{W}_{xy}(\boldsymbol{I}'_t,\boldsymbol{I}_t)$为：

$$\boldsymbol{W}_{xy}(\boldsymbol{I}'_t,\boldsymbol{I}_t)=\frac{1-\exp(-\|\boldsymbol{I}'^{xy}_t-\boldsymbol{I}_t^{xy}\|_2)}{\sum_{x,y}1-\exp(-\|\boldsymbol{I}'^{xy}_t-\boldsymbol{I}_t^{xy}\|_2)} \tag{11}$$

当得分F_t高于某个阈值γ时，该帧被认定为异常帧，该模块不更新记忆项，否则重新更新。

1.4 损失函数

网络的目标是从输入帧序列$\{\boldsymbol{I}_{t-f},\cdots,\boldsymbol{I}_{t-2},\boldsymbol{I}_{t-1}\}$中预测未来帧$\boldsymbol{I}_t$。使用预测损失$L_{\mathrm{predict}}$、光流损失$L_{\mathrm{flow}}$、特征紧凑度损失$L_{\mathrm{compact}}$和特征分离损失$L_{\mathrm{separate}}$作为损失函数来训练模型，由参数$\lambda_{\mathrm{c}}$和$\lambda_{\mathrm{s}}$平衡如下：

$$L=L_{\mathrm{predict}}+L_{\mathrm{flow}}+\lambda_{\mathrm{c}}L_{\mathrm{compact}}+\lambda_{\mathrm{s}}L_{\mathrm{separate}} \tag{12}$$

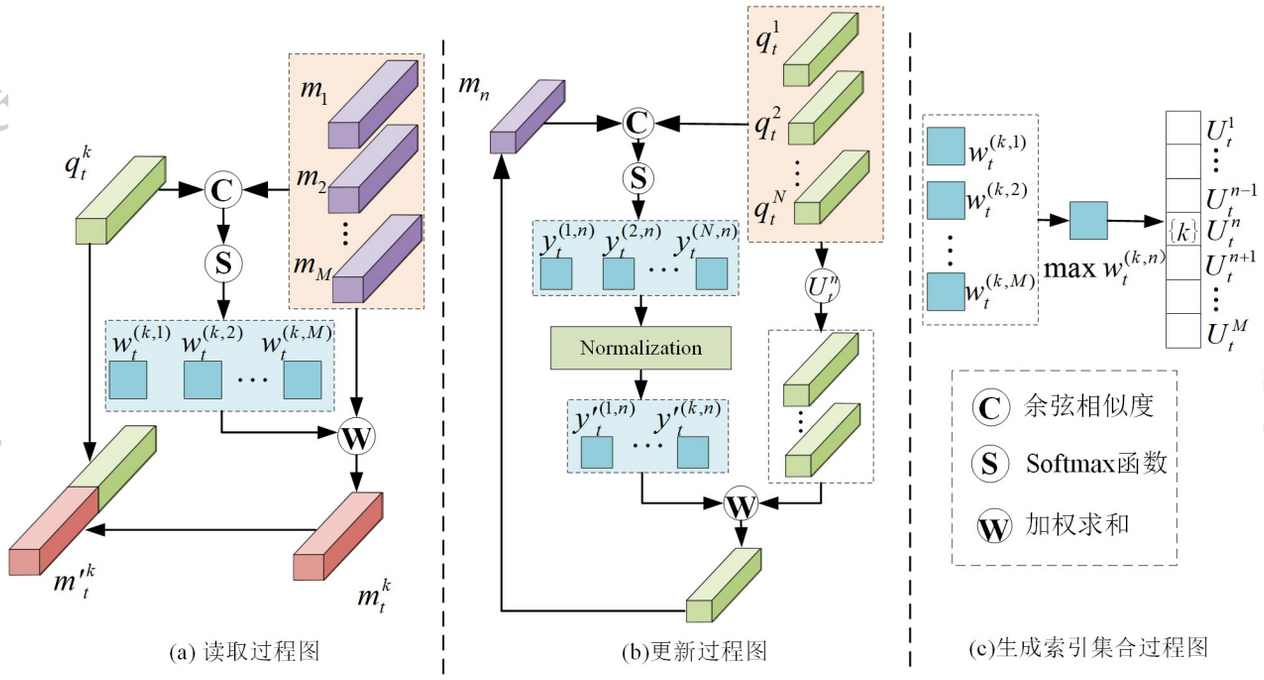
预测损失：为衡量的是预测帧$\boldsymbol{I}'_t$与对应真实帧$\boldsymbol{I}_t$之间的相似程度。具体来说，最小化这两者之间的L_2距离：

$$L_{\mathrm{predict}}=\|\boldsymbol{I}'_t-\boldsymbol{I}_t\|_2 \tag{13}$$

光流损失：为衡量的是从预测帧$\boldsymbol{I}'_t$中估计的运动场$\mathrm{Flow}(\boldsymbol{I}_{t-1},\boldsymbol{I}'_t)$与真实帧$\boldsymbol{I}_t$估计的运动场$\mathrm{Flow}(\boldsymbol{I}_{t-1},\boldsymbol{I}_t)$之间的相似程度。具体来说，最小化这两者之间的$\mathrm{L}_2$距离：

$$L_{\mathrm{flow}}=\|\mathrm{Flow}(\boldsymbol{I}_{t-1},\boldsymbol{I}'_t)-\mathrm{Flow}(\boldsymbol{I}_{t-1},\boldsymbol{I}_t)\|_2 \tag{14}$$

特征紧凑度损失：为促使查询接近存储模块中与之相似度最高的记忆项，减少类内差异。具体来



((a)read process diagram;(b)update process diagram;(c)process diagram of generating index sets)

图2　记忆增强模块结构图

Fig. 2　Structure diagram of the memory enhancement module

说，最小化查询与记忆项之间的L_2距离：

$$L_{\text{compact}} = \sum_{k=1}^{N} \| \boldsymbol{q}_t^k - \boldsymbol{m}_p \|_2 \tag{15}$$

式中，N为查询总个数，p是查询$\boldsymbol{q}_t^k$最近项的索引，定义为：

$$p = \underset{n \in M}{\operatorname{argmax}}\, \boldsymbol{w}_t^{(k,n)} \tag{16}$$

特征分离度损失：为了考虑到正常数据的多种特征，需要存储模块中的记忆项之间差别足够大，因此，使用特征分离损失，定义如下：

$$L_{\text{separate}} = \sum_{k=1}^{N} [\, \| \boldsymbol{q}_t^k - \boldsymbol{m}_p \|_2 - \| \boldsymbol{q}_t^k - \boldsymbol{m}_n \|_2 + \alpha \,]_+ \tag{17}$$

式中，N为查询总个数，将$\boldsymbol{q}_t^k$代表查询，$\boldsymbol{m}_p$和$\boldsymbol{m}_n$是它的最近项和次近项。用n表示查询$\boldsymbol{q}_t^k$的次最近项的索引：

$$n = \underset{m \in M, m \notin p}{\operatorname{argmax}}\, \boldsymbol{w}_t^{(k,m)} \tag{18}$$

1.5　异常检测

在测试阶段，使用(Zhong等，2022)中提出的评估策略对异常进行评分。计算预测帧和真实帧之间的峰值信噪比PSNR：

$$\mathrm{P}(\boldsymbol{I}_t, \boldsymbol{I}'_t) = 10\log_{10}(1/V) \tag{19}$$

$$V = \sum_{i=0}^{N} v_i \tag{20}$$

式中，P代表PSNR函数，v_i表示尺度i中的最大预测误差，N表示误差金字塔中包含的尺度总数。

由于在异常发生时模型的预测误差通常会显著增大，从而导致PSNR值降低，因此，PSNR可作为衡量异常程度的有效指标。进一步对PSNR进行归一化和取反，得到[0,1]区间内的异常分数：

$$\mathrm{S}(\boldsymbol{I}_t) = 1 - \frac{\mathrm{P}(\boldsymbol{I}_t, \boldsymbol{I}'_t) - \min_t(\mathrm{P}(\boldsymbol{I}_t, \boldsymbol{I}'_t))}{\max_t(\mathrm{P}(\boldsymbol{I}_t, \boldsymbol{I}'_t)) - \min_t(\mathrm{P}(\boldsymbol{I}_t, \boldsymbol{I}'_t))} \tag{21}$$

式中，$\boldsymbol{I}_t$是t时刻的真实帧，$\boldsymbol{I}'_t$是t时刻的预测帧，异常分数$\mathrm{S}(\boldsymbol{I}_t)$越高，说明异常发生的概率越大。

2　实　验

2.1　数据集

本文在UCSD Ped2(Li等，2013)、CUHK Avenue(Lu等，2013)、ShanghaiTech(Luo等，2017)这三个地面基准监控数据集和Drone-Anomaly(Jin等，2022)无人机航拍数据集上进行性能评估。表1统计了四个数据集的样本情况，包括视频数量、异常种类、来源及异常事件类型。表2进一步统计了Drone-Anomaly数据集的样本情况，共包含7个不同场景，由无人机航拍采集，总视频时长约5.6小时，共包含37段训练视频和46段测试视频。每个数据集的训

练集均只包含正常视频，测试集同时包含正常帧和异常帧。每帧的真实标注包含一个二进制标志，表示一帧是否包含异常事件。因此，标记0为正常帧，标记1为异常帧。

上述数据集覆盖了地面监控与无人机航拍两种典型视角，能够有效验证提出方法在不同场景下的检测与适应能力。

表1　四种数据集介绍

Table1　Describes the four datasets

数据集名称	视频数量	异常种类	来源	异常事件
Avenue	37	5	地面	投掷物品
Ped2	28	5	地面	骑自行车
ShanghaiTech	437	11	地面	追逐
Drone-Anomaly	92	10	空中	逆行车辆

2.2　实验设置

在地面监控和无人机数据集上，实验将每个视频帧的大小调整为256×256，并将视频帧像素归一化到[-1，1]。从训练集中获取5个连续的视频帧序列，将其中的前4帧作为输入，预测第5帧。为了与送入记忆增强模块的特征维度相匹配，将查询特征图的高度H、宽度W特征通道数C和记忆项数M分别设置为32、32、512和10。记忆项数M=10为经验设定值，该取值在常见记忆增强视频异常检测方法（Gong等，2019）中被广泛采用，能够较好地平衡记忆容量与计算开销。在所有数据集上均使用β=0.9的Adam优化器，批量大小为2。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）。在训练损失函数的时候，参数设置为λ_c=0.1，λ_s=0.1。在测试阶段更新记忆项时，设置的阈值γ=0.01，该阈值根据验证集上的经验观察确定，为手动调试得到的较优取值。所有模型实验均使用PyTorch进行训练和推理。

与先前的研究一致（Zhang等，2022；Fang等，2020），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。AUC值越大，表明算法的异常检测性能越好。

2.3　与其他异常检测方法比较

2.3.1　地面监控数据集实验结果对比

在地面监控场景下，将提出方法与多种异常检测方法在三种公开监控数据集上进行对比，结果如表3所示。可以看出，在UCSD Ped2、CUHK Avenue和ShanghaiTech数据集上均取得了较优性能，分别为98.8%、89.1%和74.7%。

在基于重建的方法中，MemAE等方法都是通过模型重建出输入帧，利用重建误差检测异常。但是与基于预测的方法相比，重建模型沿时间维度的推理能力较差以及卷积自编码器容易过拟合，现有的重建方法倾向于关注低级的像素级误差而不是高级的语义特征。本文重点对比预测方法，预测方法进一步分为单流预测与双流预测。

单流预测方法中，ASTNET、MGAN-CL、DEDDnet、Frame-Pred和STCEN方法都是单分支的，仅考虑外观信息，未能充分考虑运动信息。提出方法利用光流估计网络提取光流运动特征，将运动特征与外观特征进行融合，同时考虑了外观特征和运动特征。结果证明提出方法能够有效提升异常检测的性能。双流预测方法中，AMAE和MGAN-CL方法

表2　Drone-Anomaly数据集介绍

Table2　Describes the Drone-Anomaly datasets

场景	视频片段数量（训练/测试）	帧数（训练/测试）	异常事件
Highway	6 / 3	9045 / 2820	动物在道路上行走；汽车碰撞
Crossroads	10 / 5	15772 / 6244	逆行车辆；交通拥堵
Bike Roundabout	6 / 7	7950 / 18427	移动中的车辆
Vehicle Roundabout	4 / 2	5266 / 2643	行人横穿道路
Railway Inspection	3 / 1	1206 / 882	铁轨上的障碍物
Solar Panel Inspection	4 / 3	2848 / 2450	未知物体；面板缺陷
Farmland Inspection	4 / 1	9548 / 2387	不明车辆

虽然引入了运动分支，将外观特征与运动特征进行融合，但未能充分考虑正常特征的多样性。本文方法引入了记忆增强模块来加强正常特征的学习，扩大正常与异常的差异，获得了更好的检测性能。AMMC方法虽然也是双流融合，同时也记忆了正常特征，但由于异常事件往往涉及快速运动，提出方法引入方差注意力来突出视频中快速运动的物体，结果证明可以获得更精确的检测结果。

表3　地面监控数据集上与其他异常检测方法AUC(%)比较

Table3　Comparison of AUC (%) with other anomaly detection methods on ground-surveillance datasets

	方法	Ped2	Avenue	ShanghaiTech
重建	MemAE(Gong等,2019)	94.1	83.3	71.2
	MNAD(Park等,2020)	97.0	88.5	70.5
	MESDnet(Fang等,2021)	95.6	86.3	73.2
	Bi-READ (Kommanduri等,2023)	97.1	86.5	-
预测单流	Frame-Pred(Liu等,2018)	95.4	84.9	72.8
	ASTNET(Le等,2023)	97.4	86.7	73.6
	STCEN(Hao等,2022)	96.9	86.6	73.8
	HSTforU(Liu等,2022)	97.3	87.8	**75.3**
预测双流	AMMC(Cai等,2021)	96.6	86.6	73.7
	MGAN-CL(Li等,2023)	96.5	87.1	73.6
	DEDDnet(Zhong等,2022)	98.1	89.0	74.5
	AMAE(Liu等,2022)	97.4	88.2	73.6
	PDM-Net(Huang等,2024)	97.7	88.1	74.2
	GroupGAN(Sun等,2024)	96.6	85.5	73.1
	本文方法	**98.8**	**89.1**	74.7

注:加粗字体为每列最优结果,-表示原始文献中未计算该数据集下的AUC值。

ShanghaiTech数据集具有挑战性，因为它是一个大规模的数据集，包括超过270K的训练帧和42K的测试帧。由于其包含大量的数据，且正常和异常事件类型多样，因此其性能相对于其他数据集较低。同时，本文方法在该数据集上的结果略低于HSTforU方法，原因在于该数据集规模大、场景与异常类型多样且部分异常(如缓慢追逐)需要更长时间尺度的运动模式建模，而本文方法通过光流侧重局部运动刻画，HSTforU则采用层次化时空Transformer显式建模长时依赖，因而在多场景适应性上更具优势。

2.3.2　无人机航拍数据集结果对比

在无人机场景下，为进一步验证本文方法在复杂动态环境中的有效性，将其与多种异常检测方法在Drone-Anomaly数据集上的5个场景上进行对比，结果如表4所示。可以看出，本文方法在多个子场景中均取得了较优性能，尤其在Railway Inspection和Farmland Inspection场景中分别取得94.76%和91.41%的AUC，以及较低的EER(0.09%和0.10%)，优于多种对比方法，体现出较好的异常检测能力。

相比之下，各对比方法在不同子场景中的表现存在一定差异。其中，Spatio-Temporal Dissociation和MNAD在部分场景中表现不稳定，而Future Frame Prediction和MLEP方法整体检测精度较低，说明仅依赖单一建模方式难以适应无人机视频中复杂多变的场景。从具体场景来看，由于Railway Inspection和Farmland Inspection中异常类型较为单一，任务复杂度相对较低，本文方法在这些场景中取得了较高的AUC和较低的EER。在Vehicle Roundabout和Crossroads等场景中，尽管异常模式更加复杂，本文方法仍优于多数对比方法，表明所提出的外观-运动融合机制能够有效捕捉关键动态变化。同

时,记忆增强模块有助于建模多样化的正常模式,从而提升异常判别能力。值得注意的是,本文方法在Highway场景中AUC低于Spatio-Temporal Dissociation方法结果。主要原因可能是由于高速运动导致FlowNet2光流估计存在模糊,运动特征提取不充分。

表4　无人机航拍数据集上与其他异常检测方法在AUC(%)和EER(%)的比较

Table4　Comparison of AUC (%) and EER (%) with other anomaly detection methods on UAV aerial datasets

方法	Railway Inspection		Highway		Vehicle Roundabout		Crossroads		Farmland Inspection	
	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
Frame-Pred(Liu,2018)	60.44	0.40	62.75	0.36	61.41	0.45	44.84	0.54	–	–
Spatio-Temporal Dissociation(Chang,2022)	33.45	0.63	**76.00**	**0.27**	38.55	0.61	38.09	0.59	–	–
MNAD(Park等,2020)	27.70	0.68	67.99	0.45	43.62	0.57	57.07	0.45	78.60	–
MLEP(Liu,2019)	68.35	0.40	55.58	0.47	56.18	0.48	49.80	0.51	–	–
ANDT(Jin,2022)	59.40	–	68.70	–	61.30	–	30.15	–	79.50	–
MKD(Salehi,2021)	62.40	–	64.30	–	62.70	–	**63.50**	–	75.20	–
SSPCAB(Ristea,2022)	59.10	–	67.80	–	62.30	–	60.40	–	79.00	–
本文方法	**94.76**	**0.09**	50.91	0.61	**65.17**	**0.43**	58.73	**0.41**	**91.41**	**0.10**

注:加粗字体为每列最优结果,–表示原始文献中未计算该场景下的AUC或EER值。

实验结果表明,本文方法在无人机场景中同样具有较好的检测能力,尤其在复杂动态环境和多样化异常模式下,能够有效提升异常检测性能。

2.3.3　模型参数量与FPS

视频异常检测作为较高应用价值的任务,计算效率尤为重要。模型的参数量和FPS是评估方法实时性的重要指标,因此,将本方法与其他异常检测方法的模型参数量与FPS进行比较,如表5所示。

该对比不考虑FlowNet2网络计算光流的代价。MemAE方法只有外观单分支,模型复杂度较低,但该方法使用3D卷积提取时空特征,需要更多浮点运算和内存带宽,导致处理速度较慢;MGAN-CL方法使用早期的低质量生成器和伪异常模块来训练判别器,在一定程度上增加了训练时间;ASTNET方法使用深度和广度网络WiderResNet34提取高级特征,该网络计算量高且内存带宽压力大,导致模型参数量远远高于其他方法,速度低于其他方法。在检测阶段,本章方法的模型复杂度为20.39M,实时帧率达到了62.66FPS,高于其他方法,能够满足实时视频异常检测要求。若进一步考虑FlowNet2提取光流的时间,则本文方法端到端处理速度约为39.76FPS,仍能满足多数实时视频异常检测场景的需求。

表5　与其他异常检测方法模型参数量与FPS比较

Table5　Comparison with other methods in model parameters and FPS

方法	Parameter(M)	FPS
AMMC(Cai等,2021)	25.1	45
MemAE(Gong等,2019)	6.49	38
MGAN-CL(Li等,2023)	–	30
ASTNET(Le等,2023)	150.97	16
本文方法	20.39	62.66

注:FPS测试硬件平台:NVIDIA GeForce RTX4090 GPU,Intel(R) Core(TM) i7-13700K CPU。

表6展示了网络各部分的FLOPs与平均推理时间。可以看出,模型的主要计算开销集中在编码器与解码器,其中解码器占比最高。相比之下,融合模块的计算量较低,在不同尺度下均保持轻量化设计。记忆增强模块虽然FLOPs较小,但推理时间相对较高,主要来源于特征匹配操作。总体来看,所提方法在保证性能的同时,实现了良好的计算效率与复杂度平衡。

表6　网络不同部分FLOPS与平均推理时间

Table6　FLOPs and average inference time of different parts of the network

模块	FLOPS	平均推理时间(ms)
外观编码器	13.801G	1.382
运动编码器	13.575G	1.371
融合模块512	563.326K	0.108
融合模块256	1.081M	0.082
融合模块128	2.198M	0.094
融合模块64	4.588M	0.137
记忆增强模块	9.96M	0.232
解码器	43.67G	3.132

注:融合模块后面的数字代表输入特征的通道数。

2.4　可视化

2.4.1　误差可视化

为验证提出方法中运动分支的有效性，在Avenue数据集上将单流方法(仅外观分支)与提出方法(外观-运动双分支)进行了误差可视化分析，重点关注误差图中目标部分，每张图右下角的数字代表误差图的均方根误差，从而量化异常或正常事件的偏离程度。对于异常事件，数值越大越好；正常事件则数值越小越好。结果如图3所示。

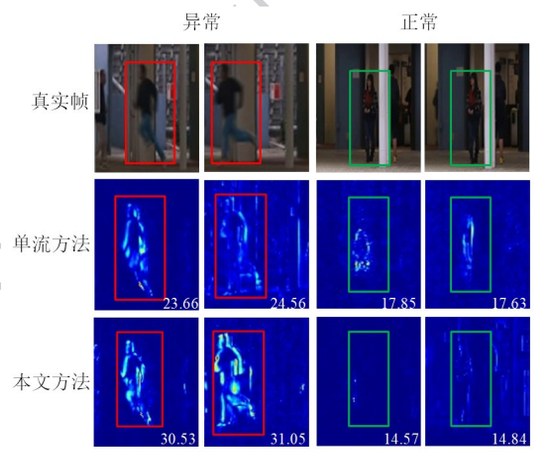


图3　单流方法与提出方法在Avenue数据集上误差可视化

Fig. 3　The single-stream method and the proposed method are used to visualize the error on the Avenue dataset

从图中可以看出，在正常场景下，单流方法仅依赖外观信息，对时序一致性建模不足，易受背景变化或局部扰动影响，导致部分区域预测误差较高；而提出方法通过引入运动分支并进行外观－运动协同建模，能够更准确刻画连续运动模式，使正常区域误差更低且分布更加平滑稳定。在异常场景中，单流方法对异常运动不敏感，误差响应分散，难以突出异常区域；相比之下，提出方法利用运动分支有效捕捉异常运动变化，在异常目标处产生更集中且显著的误差响应。综上，所提外观－运动双流结构在保证正常预测稳定性的同时，显著提升了对异常运动的区分能力，验证了运动分支在VAD任务中的有效性与必要性。

为进一步验证本章方法有效性，在UCSD Ped2、CUHK Avenue和ShanghaiTech三个数据集上的误差可视化结果如图4所示。对于每个数据集，第一行是真实的视频帧，第二行是模型预测出的预测帧，第三行是预测帧与真实帧之间的误差图。每个数据集可视化图的第一列为正常情况下的预测结果，预测误差较低。红色框标注出异常事件发生的区域。

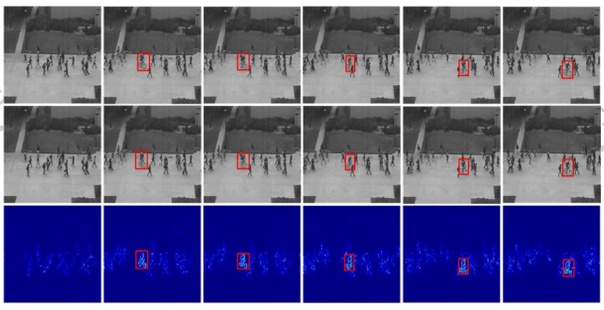
如图4-(a)所示，UCSD Ped2数据集是单场景下固定机位拍摄的，异常事件种类较少，模型能够准确的检测出异常。如图4-(b)所示，CUHK Avenue数据集的异常区域从左到右逐渐增大，面对这种多尺度的异常变化，模型依然能够准确地检测到每一帧中的异常区域，因此模型具有检测多尺度异常的能力。如图4-(c)所示，ShanghaiTech数据集有多个场景且光照条件都不同，异常事件类型较多，然而，即便在复杂环境下，模型仍然能够较为准确地检测出异常事件，表明模型在处理复杂场景、多种异常类型任务时具有很好的鲁棒性。

总体而言，提出方法能够准确学习正常事件的外观和运动特征，从而对正常区域进行准确预测，使误差保持在较低水平。相比之下，对异常区域的预测效果较差，误差相对较大。通过误差，可以区分正常和异常事件，准确检测出异常。

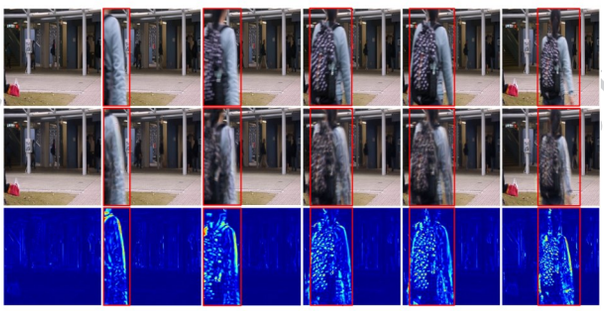
2.4.2　异常得分可视化

为直观展示提出方法在异常检测中的性能，图5展示了在三个地面监控数据集及无人机数据集三个场景中的异常评分曲线。异常分数通过多尺度计算获得，粉色区域表示异常帧。

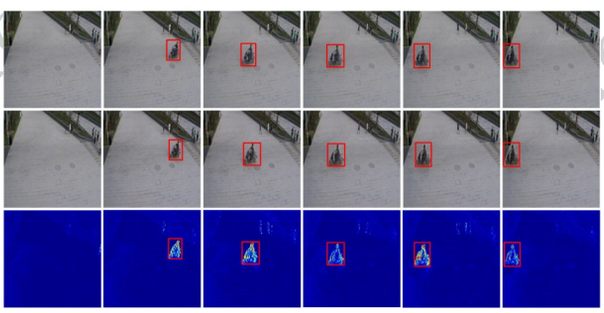
从得分结果可以看出，无论是在地面监控数据集还是无人机数据集上，当骑车闯入、跑动现象、横穿马路、车辆压线等异常出现时，异常得分显著增高，当异常结束后，异常得分回归正常，可以明显区分正常和异常。综合所有数据集结果，所提方法能



(a) USCD ped2



(b) Avenue



(c) Shanghai Tech

((a) UCSD Ped2;(b) CUHK Avenue;(c) ShanghaiTech)

图4　三个数据集上的预测误差可视化

Fig. 4　Visualization of prediction error on three datasets

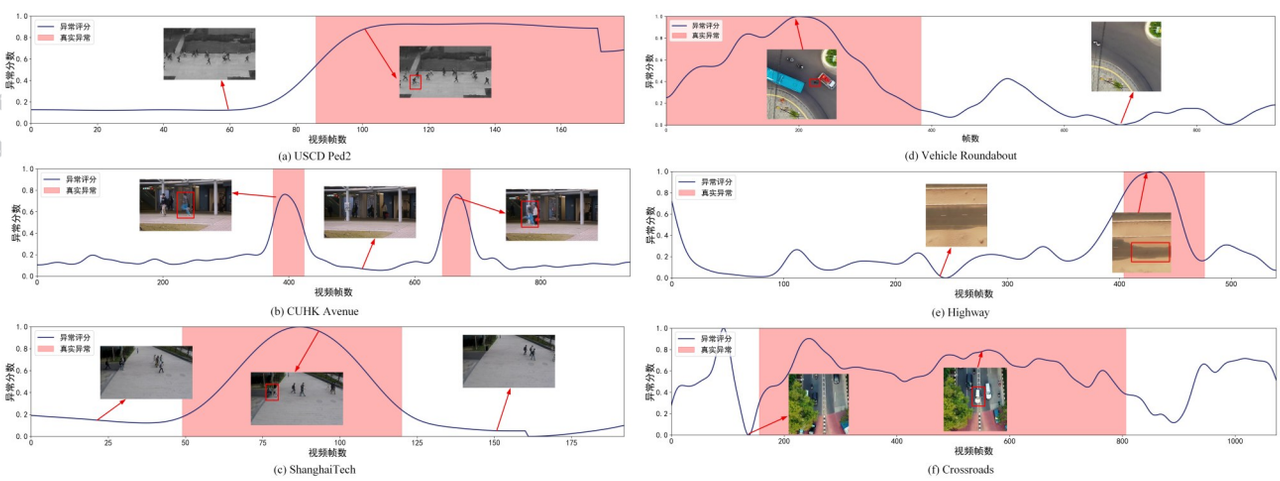
够稳定地检测不同场景中的异常事件,并清晰区分异常与正常行为。

2.5　消融实验

为了验证各模块的有效性,本文针对基于方差注意力的外观运动融合模块、记忆增强模块与通道注意模块这三个主要模块,在地面监控 UCSD Ped2 和 CHUK Avenue 数据集与无人机航拍数据集 Railway Inspection 场景上进行了详细的消融实验,结果如表7所示。

由于UCSD Ped2数据集的数据量少且异常特征更加明显,因此任意三个模块的组合在该数据集上得到的AUC值均比在CHUK Avenue上的结果高,但都低于提出方法的AUC值。模型3显示,当不使用融合模块时,模型的AUC最低,这说明仅依赖外观特征而忽略运动信息会显著制约模型对动态异常的表征能力,从而影响检测精度。模型1和模型2分别移除了MEM和ECA模块,其AUC结果均低于完整模型,说明每个模块对最终性能均有正向贡献。总体来说,消融实验证明了各个模块均能有效提升模型性能,证明提出方法的有效性与可行性。

为进一步验证不同损失函数对本文方法性能的影响,本节针对预测损失 L_{predict}、光流损失 L_{flow}、特征紧凑度损失 L_{compact} 和特征分离度损失 L_{separate} 这四个损失函数,在UCSD Ped2和CHUK Avenue数据集上展开消融实验,结果如表8所示。模型1显示,当仅使用预测损失 L_{predict} 时,模型的AUC最低,这说明仅依靠帧预测而忽略运动信息与特征约束会显著制约



(a) USCD Ped2　(d) Vehicle Roundabout

(b) CUHK Avenue　(e) Highway

(c) ShanghaiTech　(f) Crossroads

((a) UCSD Ped2;(b) CUHK Avenue;(c) ShanghaiTech;(d) Vehicle Roundabout;(e) Highway;(f) Crossroads))

图5　六个场景上的异常曲线可视化

Fig. 5　Visualization of score curves on six scenarios

表7　模块消融实验结果

Table7　Results of module ablation experiments

模型	VA-AMFM	MEM	ECA	Ped2	Avenue	Railway Inspection
1	√	√	–	98.44 (−0.36)	87.41 (−1.69)	89.15 (−5.16)
2	√	–	√	98.42 (−0.38)	87.55 (−1.55)	89.26 (−5.5)
3	–	√	√	98.39 (−0.41)	87.22 (−1.88)	85.48 (−9.28)
本文	√	√	√	**98.80**	**89.10**	**94.76**

注：以完整模型为Baseline（提升幅度记为“–”），每个数据下方括号里的数字表示提升幅度。

模型对动态异常的表征能力，从而影响检测精度。模型2和模型3分别移除了记忆模块相关的特征紧凑和分离损失（L_{compact}、L_{separate}）或光流损失（L_{flow}），其AUC结果均低于完整模型，说明每个损失函数对最终性能均有一定的提升。

表8　损失函数消融实验结果

Table8　Results of loss function ablation experiments

模型	L_{predict}	L_{flow}	L_{compact}	L_{separete}	Ped2	Avenue
1	√	–	–	–	97.94 (−0.86)	87.27 (−1.83)
2	√	√	–	–	98.20 (−0.60)	87.60 (−1.50)
3	√	–	√	√	98.13 (−0.67)	87.55 (−1.55)
本文	√	√	√	√	**98.80**	**89.10**

注：以完整模型为Baseline（提升幅度记为“–”），每个数据下方括号里的数字表示提升幅度。

3　局限性分析与未来工作

尽管本文方法在地面监控和无人机航拍多个数据集上取得了优异性能，但仍存在以下局限性：（1）对高速运动场景的检测能力不足：如Highway场景所示，当目标运动速度过快或存在严重运动模糊时，FlowNet2提取的光流质量下降，导致运动特征表征失效。（2）记忆模块容量受限：对于包含超过10种正常行为模式的复杂场景（如ShanghaiTech中的多个子场景），固定容量的记忆项难以充分覆盖所有模式多样性，导致部分正常行为被误判为异常。未来工作将围绕上述问题展开：（1）引入时序自适应光流估计或可变形卷积以增强高速运动建模；（2）探索动态可扩展记忆网络，根据场景复杂度自适应调整记忆项数量。

4　结　论

本文提出了一种基于双流网络和记忆增强的空地视频异常检测方法。该网络具有双编码器和基于注意力单解码器架构，引入了基于方差注意力融合模块和具有更新策略的记忆增强模块，多尺度的融合了外观特征与运动特征，对正常数据特征进行多样化的学习，并将经过记忆增强的高维特征送入解码器，同时利用跳跃连接把多尺度融合特征连接到解码器的对应阶段，使得预测过程可以更好利用两者的特征信息。此外，利用通道注意力机制挖掘特征的通道关系，帮助模型更有效地学习。在地面监控和无人机数据集上的对比实验表明，能够有效提升模型在复杂动态环境中的异常检测能力，在多场景条件下表现出良好的鲁棒性与泛化性能，为实际应用中的视频异常检测提供了一种可行且高效的解决方案。

参考文献（References）

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. Washington, DC: AAAI Press: 938-946 [DOI: 10.1609/aaai.v35i2.16177]

Chang Y, Tu Z, Xie W, Luo B, Zhang S, Sui H and Yuan J. 2022. Video anomaly detection with spatio-temporal dissociation. Pattern Recognition, 122: 108213 [DOI: 10.1016/j.patcog.2021.108213]

Chriki A, Touati H, Snoussi H and Kamoun F. 2021. Deep learning and handcrafted features for one-class anomaly detection in UAV video. Multimedia Tools and Applications, 80(2): 2599-2620 [DOI: 10.1007/s11042-020-09774-w]

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, Korea (South): IEEE: 1705-1714 [DOI: 10.1109/ICCV.2019.00179]

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]

Hamdi S, Bouindour S, Snoussi H, Wang T and Abid M. 2021. End-to-end deep one-class learning for anomaly detection in UAV video stream. Journal of Imaging, 7(5): 90 [DOI: 10.3390/jimaging7050090]

Hao Y, Li J, Wang N, Wang X and Gao X. 2022. Spatiotemporal consistency-enhanced network for video anomaly detection. Pattern Recognition, 121: 108232 [DOI: 10.1016/j.patcog.2021.108232]

Hasan M, Choi J, Neumann J, Roy-Chowdhury A and Davis L. 2016. Learning temporal regularity in video sequences//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 733-742 [DOI:10.1109/CVPR.2016.86]

He P, Zhang F, Li G and Li H. 2024. Adversarial and focused training of abnormal videos for weakly-supervised anomaly detection. Pattern Recognition, 147: 110119 [DOI: 10.1016/j. patcog. 2023. 110119]

Huang C, Wen J, Liu C and Liu Y. 2024. Long Short-Term Dynamic Prototype Alignment Learning for Video Anomaly Detection// Proceedings of the International Joint Conference on Artificial Intelligence. Jeju, Korea (South): IJCAI: 866-874 [DOI: 10.24963/ijcai.2024/96]

Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A and Brox T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 2462-2470 [DOI: 10.1109/CVPR.2017.179]

Jin P, Mou L, Xia GS and Zhu XX. 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. 2023. Bi-READ: Bi-Residual AutoEncoder based feature enhancement for video anomaly detection. Journal of Visual Communication and Image Representation, 95: 103860 [DOI: 10.1016/j.jvcir.2023.103860]

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]

Lee J, Nam W J and Lee S W. 2022. Multi-contextual predictions with vision transformer for video anomaly detection//Proceedings of the International Conference on Pattern Recognition. Montreal, Canada: IEEE: 1012-1018 [DOI: 10.1109/ICPR56361.2022.9956507]

Li D, Nie X, Gong R, Lin X and Yu H. 2023. Multi-branch Gan-based abnormal events detection via context learning in surveillance videos. IEEE Transactions on Circuits and Systems for Video Technology, 34(5): 3439-3450 [DOI: 10.1109/TCSVT.2023.3325451]

Li W, Mahadevan V and Vasconcelos N. 2013. Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(1): 18-32 [DOI: 10.1109/TPAMI.2013.111]

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 W, Luo W, Li Z, Zhao P and Gao S. 2019. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China: IJCAI: 3023-3030 [DOI: 10.24963/ijcai.2019/419]

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, UT, USA: IEEE: 6536-6545 [DOI: 10.1109/CVPR.2018.00684]

Liu Y, Liu J, Lin J, Zhao M and Song L. 2022. Appearance-motion united auto-encoder framework for video anomaly detection. IEEE Transactions on Circuits and Systems II: Express Briefs, 69(5): 2498-2502 [DOI: 10.1109/TCSII.2022.3161049]

Loshchilov I and Hutter F. 2017. Stochastic gradient descent with warm restarts//Proceedings of the International Conference on Learning Representations. Toulon, France: ICLR: 1-16 [DOI: 10.48550/arXiv.1608.03983]

Lu C, Shi J and Jia J. 2013. Abnormal event detection at 150 fps in matlab//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 2720-2727 [DOI: 10.1109/ICCV.2013.338]

Luo W, Liu W and Gao S. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 341-349 [DOI: 10.1109/ICCV.2017.45]

Ma Q, Wang C and Zhou X. 2026. CF2M-Net: Cross-feature fusion and memory-constraint network for video anomaly detection. Information Sciences, 723: 122673 [DOI: 10.1016/j.ins.2025.122673]

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, WA, USA: IEEE: 14372-14381 [DOI: 10.1109/CVPR42600.2020.01438]

Ristea NC, Madan N, Ionescu RT, Nasrollahi K, Khan FS, Moeslund TB and Shah M. 2022. Self-supervised predictive convolutional attentive block for anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 13576-13586 [DOI: 10.1109/CVPR52688.2022.01321]

Ronneberger O, Fischer P and Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation//Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Munich, Germany: Springer: 234-241 [DOI: 10.48550/arXiv.1505.04597]

Salehi M, Sadjadi N, Baselizadeh S, Rohban MH and Rabiee HR. 2021. Multiresolution knowledge distillation for anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 14902-14912 [DOI: 10.1109/CVPR46437.2021.01466]

Sun Z, Wang P, Zheng W and Zhang M. 2024. Dual GroupGAN: An unsupervised four-competitor (2V2) approach for video anomaly detection. Pattern Recognition, 153: 110500 [DOI: 10.1016/j.patcog.2024.110500]

Tran TM, Bui DC, Nguyen TV 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]

Wang Q, Wu B, Zhu P, Li P, Zuo W and Hu Q. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 11534-11542 [DOI: 10.1109/CVPR42600.2020.01155]

Wang Y, Qin C, Bai Y, Xu Y, Ma X and Fu Y. 2022. Making reconstruction-based method great again for video anomaly detection//Proceedings of the IEEE International Conference on Data Mining. Orlando, FL, USA: IEEE: 1215-1220 [DOI: 10.1109/ICDM54844.2022.00157]

Wu P, Liu J, Li M, Sun Y and Shen F. 2020. Fast sparse coding networks for anomaly detection in videos. Pattern Recognition, 107(7): 107515 [DOI: 10.1016/j.patcog.2020.107515]

Yang Y, Fu Z and Naqvi S M. 2023. Abnormal event detection for video surveillance using an enhanced two-stream fusion method. Neurocomputing, 553: 126561 [DOI: 10.1016/j.neucom.2023.126561]

Zaheer M Z, Lee J H, Mahmood A, Astrid M and Lee S. 2022. Stabilizing adversarially learned one-class novelty detection using pseudo anomalies. IEEE Transactions on Image Processing, 31: 5963-5975 [DOI:10.1109/TIP.2022.3204217]

Zavrtanik V, Kristan M and Skočaj D. 2021. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 8330-8339 [DOI: 10.1109/ICCV48922.2021.00822]

Zhang X, Fang J, Yang B, Chen S and Li B. 2022. Hybrid attention and motion constraint for anomaly detection in crowded scenes. IEEE Transactions on Circuits and Systems for Video Technology, 33(5): 2259-2274 [DOI:10.1109/TCSVT.2022.3221622]

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]

作者简介

王楠，女，硕士研究生，主要研究方向为视频异常检测。E-mail: 2231220016@stu.xaut.edu.cn

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

王琳，女，讲师，主要研究方向为光学分子影像。E-mail: wanglin004@xaut.edu.cn

尤珍臻，女，讲师，主要研究方向为生物医学图像处理。E-mail: zhenzhen_you@xaut.edu.cn