基于深度空时域特征融合的高动态空中多形态目标检测方法 (特邀)

孙 鹏，于 跃，陈嘉欣，秦翰林*

(西安电子科技大学 光电工程学院，陕西 西安 710071)

摘 要：针对复杂背景下，依靠高超声速飞行器搭载的红外探测器对高动态空中目标的可靠探测和精确识别问题，提出了一种基于深度空时域特征融合的空中多形态目标检测方法。设计了加权双向循环特征金字塔结构提取多形态目标静态特征，并引入可切换空洞卷积，增大感受野的同时减少空域信息损失。对于时序运动特征的提取，为了抑制复杂背景噪声的同时将角点信息集中到运动区域中，通过特征点匹配法生成掩膜图，之后进行光流计算，根据计算结果设计稀疏光流特征图，利用 3D 卷积提取多个连续帧图像中包含的时序特征，生成三维时序运动特征图。最后，通过对图像静态特征与时序运动特征进行通道维度的拼接，实现深度空时域特征融合。大量的对比实验表明，文中方法可明显减少复杂背景下的虚假识别概率，具备高实时性的同时目标识别准确率达 89.87%，满足高动态下的红外目标智能检测识别需求。

关键词：目标检测； 特征融合； 多尺度金字塔； 稀疏光流； 3D 卷积

中图分类号：TP391.4 **文献标志码：**A **DOI：**10.3788/IRLA20220167

Highly dynamic aerial polymorphic target detection method based on deep spatial-temporal feature fusion (*Invited*)

Sun Peng, Yu Yue, Chen Jiaxin, Qin Hanlin*

(School of Optoelectronic Engineering, Xidian University, Xi'an 710071, China)

Abstract: Aiming at the problem of reliable detection and accurate recognition of high dynamic aerial targets by infrared detectors carried by hypersonic vehicles in complex background, an aerial polymorphic target detection method based on deep spatial-temporal feature fusion was proposed. A weighted bidirectional cyclic feature pyramid structure was designed to extract the static features of polymorphic target, and switchable atrous convolution was introduced to increase the receptive field and reduce spatial information loss. For the extraction of temporal motion features, in order to suppress the complex background noise and concentrate the corner information into the moving region, the feature point matching method was used to generate the mask image, then the optical flow was calculated, and the sparse optical flow feature map was designed according to calculation results. Finally, the temporal features contained in multiple continuous frame images were extracted by 3D convolution to generate a 3D temporal motion feature map. By concatting the image static features and temporal motion features in channel dimension, the deep spatial-temporal fusion could be realized. A large number of comparative experiments showed that this method can significantly reduce the false recognition probability in complex background, and the target detection accuracy reached 89.87% with high real-time performance, which can meet the needs of infrared targets intelligent detection and recognition under high dynamic conditions.

收稿日期:2022-03-10； **修订日期:**2022-04-07

基金项目:国家自然科学基金 (62174128)

作者简介:孙鹏, 男, 硕士生, 主要从事目标探测与人机融合方面的研究。

导师(通讯作者)简介:秦翰林, 男, 教授, 博士, 主要从事目标探测、人机融合、自主协同方面的研究。

Key words: object detection; feature fusion; multi-scale pyramid; sparse optical flow; 3D convolution

0 引 言

传统空中红外目标检测方法[1−4]主要利用滤波、噪声抑制、目标增强等方式提高背景与目标的对比度，进而通过阈值分割实现目标检测；深度学习方法[5−8]利用深层卷积网络提取输入红外图像的空间特征，预测空中目标的类别和位置。然而在高超声速飞行器制导控制领域，由于复杂的战场背景以及高动态条件，利用空间特征信息的目标检测方法无法对高速运动的空中目标进行有效识别和跟踪。对时空多维深度特征信息进行融合，利用空间网络处理静态信息，利用时间网络处理动态信息，可以大大提高空中高动态目标的识别准确率，是提升高超声速飞行器制导作战性能的一种有效方式。

为了将时域信息与空域信息融合，基于双流法、3D 卷积结构 (3D convolution, conv3D) 和长短时记忆 (Long Short Term Memory, LSTM) 网络的视频特征提取架构陆续被提出。Simonyan K 等[9]首次提出了双流法，对 RGB 图像的空间流 (Spatial Flow) 和采用稠密光流图的时域运动流 (Temporal Motion Flow) 两路输入流分别处理，再进行结果融合，实现行为识别任务。Zhang 等[10]针对双流网络对时序依赖较大的问题，提出一种基于改进双流时空网络的人体行为识别算法，增强网络的特征表达能力，以提高对时序主导行为的识别能力。Donahue J 等[11]结合 2D 卷积和 LSTM 结构，首先使用 2D 卷积对图像进行特征提取，获得一个视觉特征的序列，然后将特征序列直接输入到 LSTM 中，进一步挖掘上下文信息。Ji S 等[12]将 3D 卷积用于视频分析领域，但仅在浅层使用 3D 卷积，且使用了手工方式提取图像的灰度、梯度等信息，无法对视频进行实时处理。I3D[13](Inflated 3D ConvNet) 将 3D 卷积引入双流框架，利用稠密光流对图像中所有点逐一匹配，形成光流场。该方法具有良好的视频目标检测准确率，但计算量过大。

文中通过采用关键点匹配的方式，以稀疏光流替代稠密光流，利用 3D 卷积提取视频的动态特征；同时，设计了引入可切换空洞卷积[14](Switchable Atrous Convolution, SAC) 的双向循环特征金字塔结构[15]提取图像静态空间特征，与时域动态特征融合，实现对高动态空中多形态目标的实时高准确率检测。

1 空时域特征融合网络总体设计

文中基于行为识别任务中经典的双流法思路设计空时域特征融合网络，实现了复杂背景下高动态红外空中多形态目标的智能检测识别。网络主体可分为两“流”，即图像静态特征流和时序光流图，总体网络结构如图 1 所示。

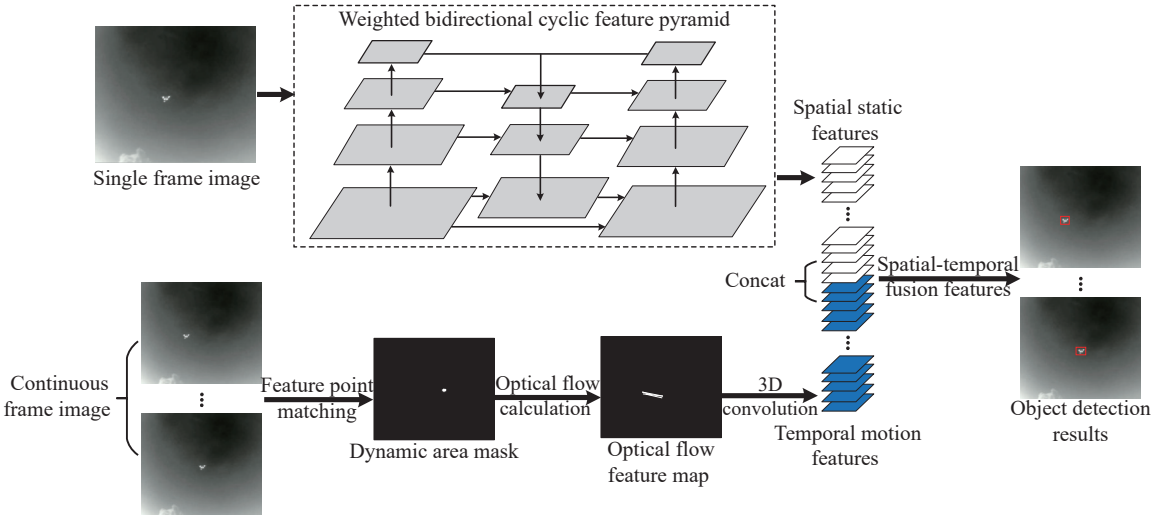

图 1 深度空时域特征融合检测网络

Fig.1 Deep spatial-temporal feature fusion detection network

其中，红外图像的静态多形态特征以加权双向循环特征金字塔为主体框架进行提取，同时引入可切换空洞卷积替代普通卷积，增大感受野的同时减少空域信息损失。在时域特征流方面，首先通过特征点匹配法生成运动区域的掩膜，集中目标特征，然后利用 LK(Lucas-Kanade) 金字塔结构[15] 进行光流计算，并根据计算结果设计二维 LK 稀疏光流特征图，最后利用 3D 卷积提取多个连续帧图像中包含的时序特征，生成三维时序运动特征图。文中对静态空间特征和三维时序运动特征分别进行单通道卷积操作，在通道维度进行拼接，从而实现特征图通道之间的信息融合，得到时空融合特征图，完成对高动态空中红外目标的检测识别。

2 空间静态特征提取

高超声速飞行器对空中目标探测过程中，目标的尺度会在短时间内发生大幅度变化，如何对多尺度、多形态目标特征进行提取与预测十分关键。因此，文中设计加权双向循环特征金字塔结构用来预测空中目标多形态特征，同时引入可切换空洞卷积替代普通卷积，增大感受野的同时减少空域信息损失。加权双向循环特征金字塔结构如图 2 所示，左侧主干结构均采用空洞卷积生成不同尺度的特征图，右侧的双向特征金字塔模块用来融合多层特征。

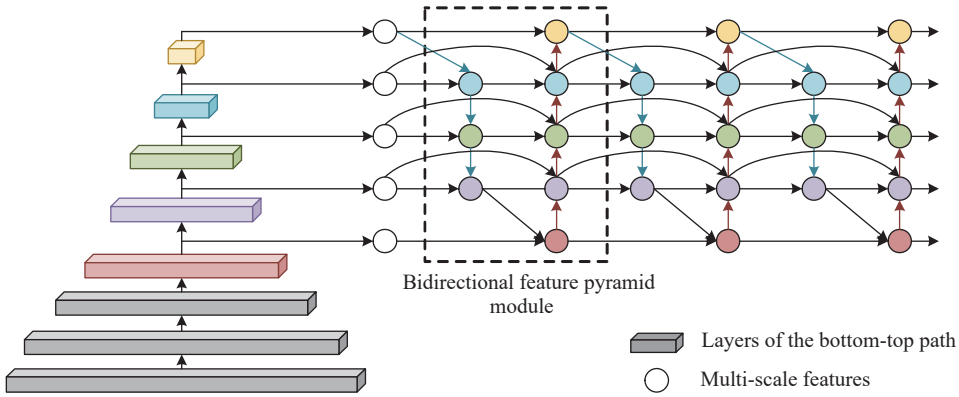

图 2 加权双向循环特征金字塔结构

Fig.2 Weighted bidirectional cyclic feature pyramid network

为提高模型效率，文中在原始特征金字塔上做了几点优化：

(1) 删除只有一个输入边的节点，因为只有一个输入边的节点没有进行特征融合，其对网络贡献较小，删除该节点则可简化网络，减小计算复杂度；

(2) 对处于同一层的输入与输出节点，添加一条额外的连接，在不增加计算成本的情况下融合更多特征；

(3) 构建双向特征融合路径，两条路径为一组，重复多次以实现更多高级特征融合。

同时，为了在不增加计算复杂度的情况下有效获取多尺度感受野，文中在主干网络中引入空洞卷积模块。一个空洞卷积模块由两个全局上下文模块和一个空洞卷积组件构成，同时设置了空洞率 r=3 的空洞卷积与 r=1 的常规卷积，对漏掉的像素点可使用标准卷积进行填补。空洞卷积模块的整体结构如图 3 所示。

可切换空洞卷积模块中的卷积操作可表示为 $y^{out}=Conv(x,w,r)$，其中 x 表示输入特征图，w 表示权重，与特征金字塔结构中使用的权重值保持一致，则完整的 SAC 组件可表示为 $S(x)\cdot Conv(x,w,1)+(1-S(x))\cdot Conv(x,w+\Delta w,r)$。其中，变换函数 $S(x)$ 由一个 5×5 的平均池化和一个 1×1 的卷积层构成，空洞率 r 默认设置为 3，卷积操作 $Conv(x,w,1)$ 和 $Conv(x,w+\Delta w,r)$ 采用锁定机制共享权重 w。将普通卷积转换为空洞卷积时，有更大空洞率的卷积权重就会丢失，所以对于空洞率为 r 的卷积层增加额外的可训练权重 Δw。

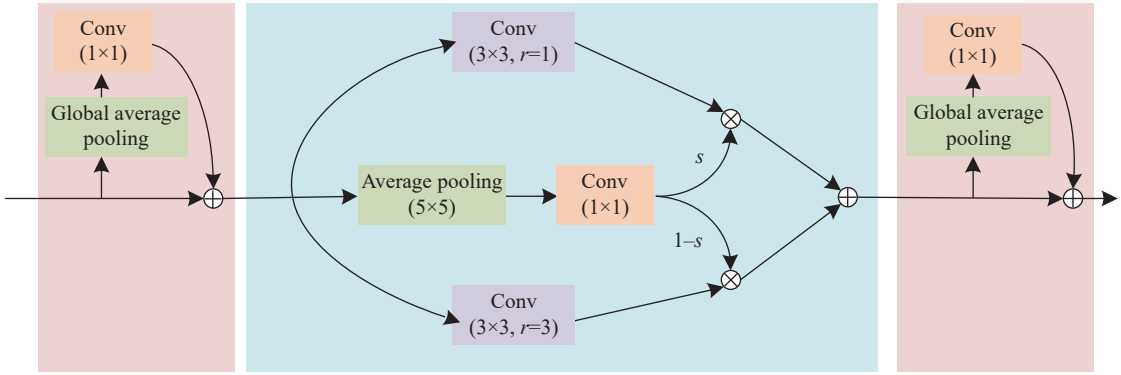

图 3 可切换空洞卷积模块

Fig.3 Switchable atrous convolution module

3 时序动态特征提取与时空特征融合

传统基于逐点匹配的稠密光流法计算量过大，在高超声速飞行器制导控制这一应用场景下，难以满足高实时性的需求。基于此考虑，文中采用特征点匹配的方法提取空中目标的角点信息，抑制静态背景噪声，将特征集中在运动区域中，生成掩膜图像。之后利用 LK 金字塔计算高动态目标的光流信息，并根据计算结果设计二维稀疏光流特征图，最后应用 3D 卷积提取时序特征。

3.1 目标特征角点提取

目标的角点位于目标边缘的交点处，在角点处任意方向上的微小移动都会导致梯度方向和幅值的大幅变化，因此，采用特征角点匹配的方法可以在减少计算量的同时有效替代基于逐个像素点匹配的稠密光流法。面向高动态空中目标检测场景时，由于背景复杂，真实目标角点信息会被淹没在大量的虚假目标角点中，从而降低特征角点提取的准确性。文中在 Shi-Tomasi 角点提取法[16] 的基础上通过图像掩模筛选出运动物体，使特征角点汇聚到运动区域上，避免大量错误计算，提升特征角点提取的效率。

获取图像掩模首先需计算连续两帧视频序列图像 I_{t-1} 和 I_t 的帧间差分：

$$D_t(x,y)=|I_t(x,y)-I_{t-1}(x,y)| \tag{1}$$

式中：(x,y) 为图像中像素坐标；$D_t(x,y)$ 为帧间差分图。

为方便后续计算，且为了尽可能在保留空中运动目标区域的同时筛去非运动区域目标，即噪声区域，文中设置了一个较大的阈值 λ，对帧间差分图进行二值化处理：

$$B_t(x,y)=\begin{cases}255, \text{if } D_t(x,y)>\lambda\\ 0, \text{ else}\end{cases} \tag{2}$$

式中：$B_t(x,y)$ 为二值化后的帧间差分图。由于阈值 λ 设置较大，为避免将真实目标错误掩盖，文中对 $B_t(x,y)$ 求局部最大值：

$$M_t=\max_{-k\leqslant x'\leqslant k,-k\leqslant y'\leqslant k,x'\neq 0,y'\neq 0} I_t(x+x',y+y') \tag{3}$$

其中，M_t 为最终生成的图像掩模，滤波核大小为 $(2k+1)\times(2k+1)$。

3.2 稀疏光流计算与特征图设计

光流法广泛应用于连续帧图像目标检测任务中，核心思想是对于 t 时刻图像上的像素点 $I(x,y)$，找到 $t+1$ 时刻该像素点在各个方向的位移量。在高动态场景中，LK 光流法的时间连续假设不再成立。基于此，文中应用 LK 金字塔结构对高动态运动目标光流进行处理，如图 4 所示。

当像素点高速运动时，对图像进行金字塔分层，每次将图像缩放为原始大小的一半，将分辨率高的图像置于金字塔底层，分辨率低的图像置于金字塔顶层。图像缩放的目的主要是减少像素的位移，从而使得 LK 光流法的时间连续假设得以满足。算法首先对顶层图像进行计算，将结果作为初始值传递至下一层，下一层的图像在此基础上计算光流和前后两帧间的仿射变化矩阵，再依次将这一层的光流和仿射矩阵继续向下传递，直至传递到最底层的原始图像层。通过自顶向下的迭代计算可实现对高速运动目标的光流求解，并根据计算结果设计二维稀疏光流特征图。

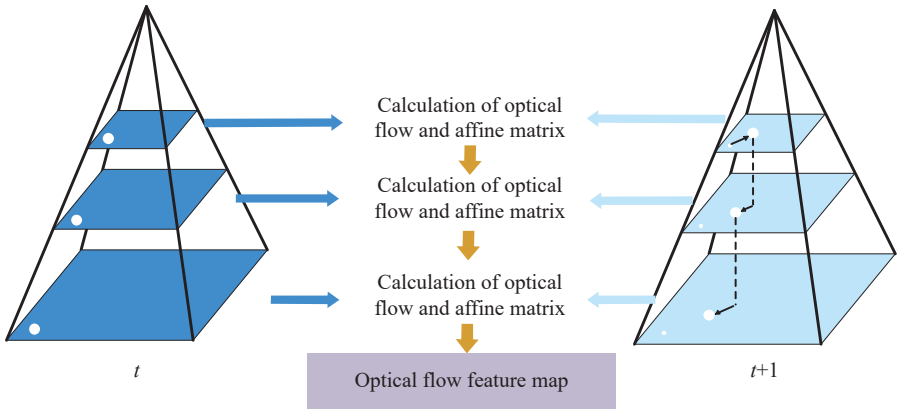

图 4 金字塔 LK 光流

Fig.4 Pyramid LK optical flow

3.3 时序特征提取

在二维光流特征图的基础上，文中利用如图 5 所示的 3D 卷积模块对特征图进行卷积运算，提取目标动态时序特征。

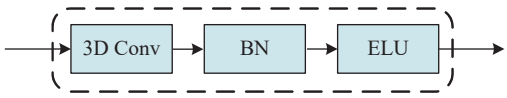

图 5 3D 卷积模块

Fig.5 3D convolution module

每组卷积模块由 3D 卷积算子、批量归一化 (Batch Normalization, BN) 层和 ELU(Exponential Linear Units) 激活函数组成。其中，BN 层在训练时对每个 batch 中特征图的每个通道进行 0 均值、1 标准差的归一化，ELU 激活函数能够使得神经元的平均激活值趋近于 0，并且对噪声更具有鲁棒性，有利于提取高动态目标特征。

4 实验分析

依靠高超声速飞行器搭载的红外探测器探测目标时，由于相对速度很大，目标在短时间内会发生较大位移，同时大小、形态明显变化。由于条件限制，实验室无法利用高超声速飞行器对空中红外目标进行实时检测。基于此考虑，文中构建了一个包含 1 500 张大小为 640×512 的常速运动红外无人机 (UAV) 连续帧图像序列，并选取多帧间隔、包含多尺度、多形态目标图像作为实验测试集，背景包括建筑、树木、云朵等，以模拟复杂背景空中目标检测场景。为验证文中方法的有效性，选取了三组测试集中包含建筑、空中飞鸟 (点噪声)、云层等干扰的连续帧图像进行对比实验，对比算法包括 C3D[17]、TSN[18]、ECO[19]、3DLocalCNN[20]、TAda[21]。

文中根据识别准确率、实时性和计算资源评估算法性能，如表 1 所示。其中识别准确率 Accuracy=(TP+TN)/(TP+TN+FP+FN)，即所有正确预测为正样本的数据与正确预测为负样本的数据数量占总样本的比值；算法实时性指标 FPS(Frames Per Second) 表示网络每秒可处理图像帧数；算法运行占用资源 Run memory 以 GB(Gigabyte) 计算。

结合对连续帧图像的识别结果以及表 1 可以看

表 1 不同算法在自建数据集上识别效果对比

Tab.1 Comparison of detection performance of different algorithms on self-built dataset

Method	Accuracy	Speed/FPS	Run memory/GB
C3 D[17]	82.31%	25.9	2.32
TSN[18]	85.73%	23.3	3.58
ECO[19]	86.57%	27.6	3.14
3DLocalCNN[20]	85.78%	21.6	2.79
TADa[21]	87.41%	29.1	4.01
Proposed method	89.87%	27.0	2.19

出，TSN、ECO、3DLocalCNN、TADa 四种方法可以较好地识别无人机目标，但存在对空中点噪声以及云层背景的大量误检，虚警率很高；C3D 方法对于背景噪声的抑制较好，但无法对连续帧图像中的目标实时跟踪，存在丢帧的现象，识别准确率低。文中提出的基于深度空时域特征融合的目标识别方法能够有效抑制复杂背景中的噪声信息，大幅降低虚警率；保持实时性的同时目标识别准确率达到了 89.87%，优于现有基于时空域特征融合的目标识别算法。

为验证所提出的基于深度学习的目标识别方法相比传统方法的优势，选取 PSTNN[1]、NRAM[2]、TDLMS[3] 和 Top-hat[4] 四种传统方法对图 6 中的三组连续帧图像进行测试，实验结果如图 7 所示。

由传统方法无人机目标识别结果可以看出，PSTNN 误检较少，但只能滤出无人机发动机、旋翼等高温位置，无法整体检出目标，当目标与背景重叠时目标检测效果差；NRAM 同样无法整体检测出无人机目标，且当背景中存在大量高温物体时，检测效果差；TDMLS 能以较高准确率提取出运动目标，但存在明显的运动轨迹，影响识别效果；Top-hat 能将目标准确滤出，但存在大量误检，虚警率过高。

以上对时空域融合方法与传统方法的分析证明了文中方法在高超声速飞行器制导场景中的有效性，满足高动态下红外目标智能检测识别的需求。

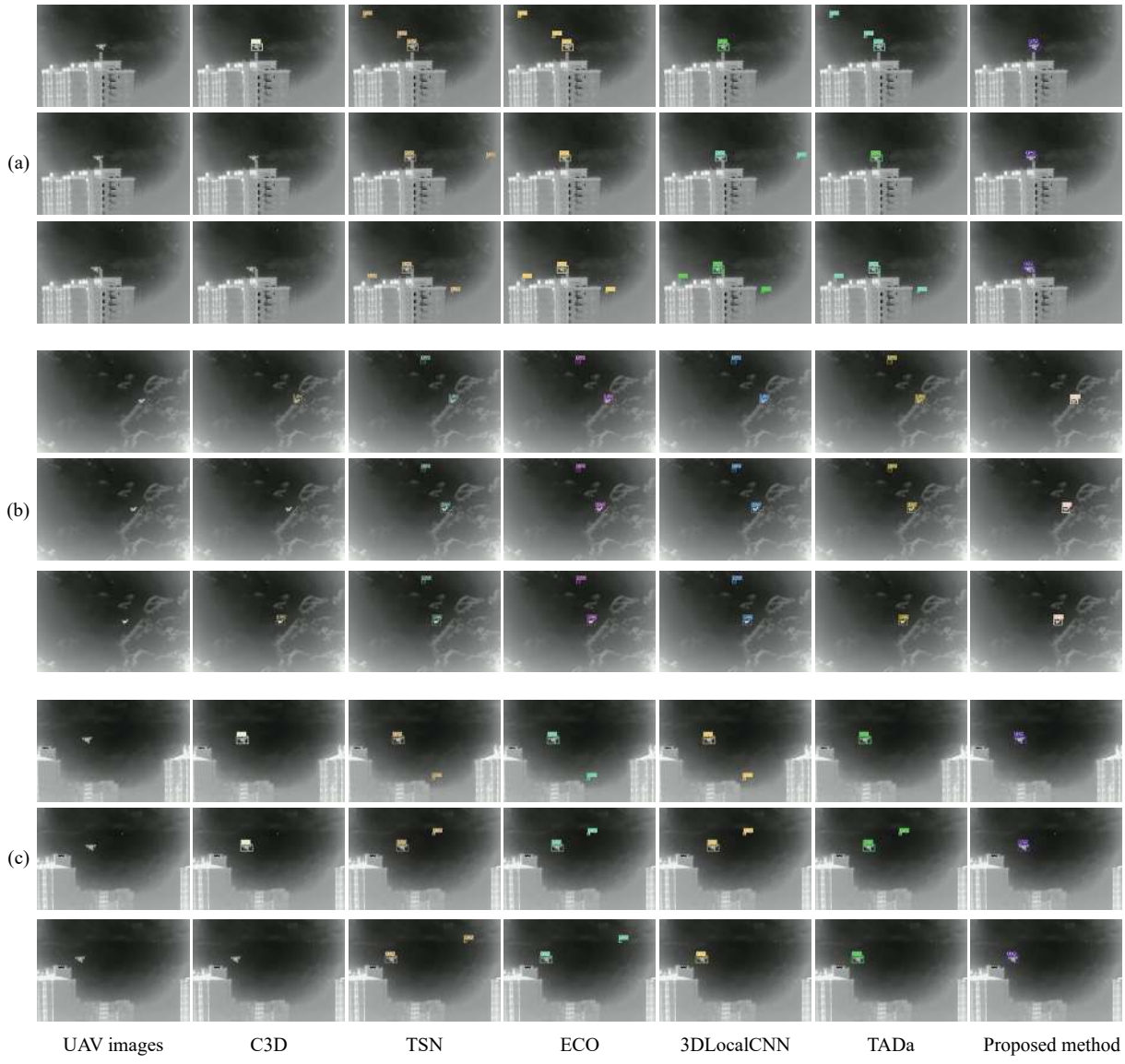

图 6 三组连续帧无人机目标识别结果对比

Fig.6 Comparison of target recognition results of UAV in three consecutive frames

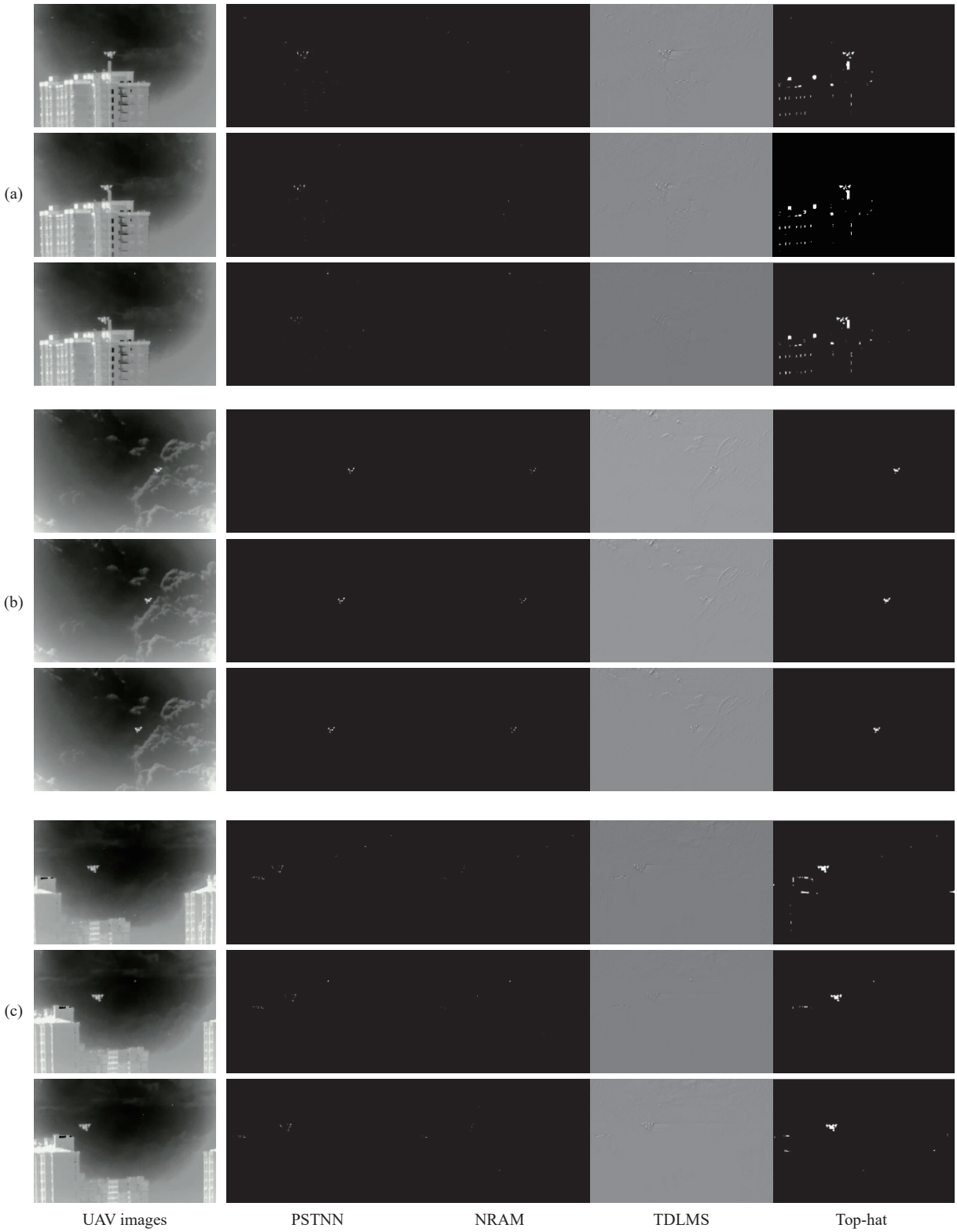

图 7 传统方法无人机目标识别结果对比

Fig.7 Comparison of UAV target recognition results by traditional methods

5 结 论

文中针对在复杂背景下对高动态空中红外目标智能检测识别的需求，设计并实现了基于深度空时域特征融合的高动态空中多形态目标检测方法，通过分析典型空时域融合算法与基于噪声抑制、目标增强的

传统方法对红外空中目标的识别结果，证明文中方法可有效对复杂背景噪声进行抑制并提取高动态目标特征，对空中红外目标检测准确率达到 89.87%，同时具有较快的检测速度。

参考文献：

[1] Jiang Taixiang, Huang Tingzhu, Zhao Xile, et al. Multi-dimensional imaging data recovery via minimizing the partial sum of tubal nuclear norm [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[2] Zhang Landan, Peng Lingbing, Zhang Tianfang, et al. Infrared small target detection via non-convex rank approximation minimization joint l2, 1 norm [J]. ***Remote Sensing***, 2018, 10(11): 1821.

[3] Hadhoud M M, Thomas D W. The two-dimensional adaptive LMS (TDLMS) algorithm [J]. ***IEEE Transactions on Circuits and Systems***, 1988, 35(5): 485-494.

[4] Bai Xiangzhi, Zhou Fugen. Analysis of new top-hat transformation and the application for infrared dim small target detection [J]. ***Pattern Recognition***, 2010, 43(6): 2145-2156.

[5] Zhao Lu, Xiong Sen. Target recognition based on multi-view infrared images [J]. ***Infrared and Laser Engineering***, 2021, 50(11): 20210206. (in Chinese)

[6] Tang Peng, Liu Yi, Wei Hongguang, et al. Automatic recognition algorithm of digital instrument reading in offshore booster station based on Mask-RCNN [J]. ***Infrared and Laser Engineering***, 2021, 50(S2): 20211057. (in Chinese)

[7] Beery S, Wu G, Rathod V, et al. Context R-CNN: Long term temporal context for per-camera object detection [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 13075-13085.

[8] Li Jingyu, Yang Jing, Kong Bin, et al. Multi-scale vehicle and pedestrian detection algorithm based on attention mechanism [J]. ***Optics and Precision Engineering***, 2021, 29(6): 1448-1458. (in Chinese)

[9] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. ***Advances in Neural Information Processing Systems***, 2014, 27: 568-576.

[10] Zhang Hongying, An Zheng. Human action recognition based on improved two-stream spatiotemporal network [J]. ***Optics and Precision Engineering***, 2021, 29(2): 420-429. (in Chinese)

[11] Donahue J, Hendricks L A, Guadarrama S, et a1. Long-term recurrent convolutional networks for visual recognition and description [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 2625-2634.

[12] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2012, 35(1): 221-231.

[13] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6299-6308.

[14] Wu Haibin, Wei Xiying, Liu Meihong, et al. Improved YOLOv4 for dangerous goods detection in X-ray inspection combined with atrous convolution and transfer learning [J]. ***Chinese Optics***, 2021, 14(6): 1417-1425. (in Chinese)

[15] Zhang Ruiyan, Jiang Xiujie, An Junshe, et al. Design of global-contextual detection model for optical remote sensing targets [J]. ***Chinese Optics***, 2020, 13(6): 1302-1313. (in Chinese)

[16] Shi J, Tomasi C. Good features to track [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994: 593-600.

[17] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks [C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2014: 1725-1732.

[18] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition [C]// European conference on computer vision. Springer (ECCV), 2016: 20-36.

[19] Zolfaghari M, Singh K, Brox T. Eco: Efficient convolutional network for online video understanding [C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 695-712.

[20] Huang Zhen, Xue Dixiu, Shen Xu, et al. 3D local convolutional neural networks for gait recognition [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021: 14920-14929.

[21] Huang Ziyuan, Zhang Shiwei, Pan Liang, et al. TAda! Temporally-adaptive convolutions for video understanding [C]//International Conference on Learning Representations (ICLR), 2022.