面向小目标检测的卫星视频跟踪算法

崔浩文[1,2]，许楚杰[1,2]，郑向涛[1]，卢孝强[1]

1. 中国科学院西安光学精密机械研究所 光谱成像技术实验室, 西安 710119;
2. 中国科学院大学, 北京 100049

摘 要： 遥感卫星视频中的目标小，场景变化多样，在遥感卫星视频上的进行多目标跟踪存在一定的困难。针对卫星视频中目标小的特点，提出了一种高分辨率遥感卫星视频的多目标跟踪算法，先检测疑似目标再进行轨迹关联。检测阶段构建小目标检测器，首先在主干网络中通过Transformer捕获全局的上下文信息，然后利用注意力机制增强目标特征，最后添加了一个预测小目标的分支；为使检测出的小目标与已有轨迹匹配，轨迹关联阶段，采用关注低置信度检测的关联算法。本文选取高分辨率遥感卫星视频进行实验，实验结果表明本文提出的方法在遥感卫星视频中的多目标跟踪数据集上的MOTA指标达到63.1%，相较于基准（baseline）模型提升13.5%，能够显著提升遥感卫星视频中多目标跟踪的性能。

关键词： 多目标跟踪，小目标检测，注意力机制，神经网络，高分大赛

中图分类号： TP701

引用格式： 崔浩文,许楚杰,郑向涛,卢孝强.XXXX.面向小目标检测的卫星视频跟踪算法.遥感学报,XX(XX):1-11
CUI Haowen, XU Chujie, ZHENG Xiangtao, LU Xiaoqiang. XXXX. Multi-Object Tracking by Detecting Small Objects in Satellite Video. National Remote Sensing Bulletin, DOI:10.11834/jrs.20232098]

1 引言

多目标跟踪旨在检测和估计视频中多个目标的时空轨迹，在视觉领域有着广泛应用，如安防监控、自动驾驶、智能交通等。随着遥感技术的发展，通过卫星平台获得地球表面运动目标的视频数据，实时观测地球表面目标的运动轨迹和状态，在城市规划、交通监控、军事侦察等发挥着重要作用。主流的多目标跟踪算法主要解决监控视频或移动设备拍摄视频下的目标跟踪问题，由于卫星视频与监控视频存在巨大差异，导致现有多目标跟踪算法应用在卫星视频上的性能较差。卫星视频和监控视频存在如下差异。（1）成像距离和视角不同，在相同图像分辨率下，卫星视频下的目标尺寸更小，目标的细节特征不明显，而监控视频下的目标细节特征更加显著，但是存在严重的目标间的遮挡问题；（2）卫星视频的背景更加复杂多样。遥感场景变化多样，目标的检测容易受背景干扰，如云雾、舰船运动产生的尾流等，导致误跟和漏跟的可能。

目前，主流多目标跟踪算法遵循基于检测的跟踪（Tracking-by-detection，TBD）范式，包含目标检测和帧间关联两个步骤。首先检测出每帧中目标可能出现的位置，然后根据检测出的候选目标位置建立时间上的关联匹配，实现目标运动轨迹关联。

随着目标检测技术的发展，许多方法利用更强的目标检测器来提高多目标跟踪的性能，如RetinaTrack（Lu等，2020）、CenterTrack（Zhou等，2020）、TransTrack（Sun等，2020）、ByteTrack（Zhang等，2021）等。这些检测器通常用于检测自然场景下的目标，因此在监控视频下仍然可以很好地检测目标，然而由于遥感影像与自然场景下的影像存在巨大差异，这些检测器直接应用在遥感场景下无法取得很好的性能。遥感图像中的小目标是导致检测性能下降的因素之一，

收稿日期： XXXX-XX-XX；**预印本：** XXXX-XX-XX

基金项目： 国家自然基金（编号：62271484）；国家杰出青年科学基金（编号：61925112）；陕西省创新能力支撑计划资助项目（编号：2020TD-015）

第一作者简介： 崔浩文，研究方向为计算机视觉。E-mail：cuihaowen20@mails.ucas.ac.cn

通信作者简介： 郑向涛，研究方向为计算机视觉。E-mail：zhengxiangtao@opt.ac.cn

而针对遥感图像中的小目标，目前有很多改进方案用于提高小目标的检测能力，如使用浅层特征（Van Etten等，2018）、生成对抗网络（Rabbi等，2020）、目标之间的度量（Xu等，2021；Wang等，2021）等。基于浅层特征检测小目标的方法最为简单直观，浅层特征能够保留小目标的特征，从而提高小目标的检测能力，但是会引入更大的计算量；基于生成对抗网络（Generative Adversarial Network，GAN）的方法通过GAN生成高质量的目标图像，增强了小目标的特征，这种方法同样会引入额外的计算；基于度量的方法，虽然不会引入额外的计算，但仍没有解决小目标可用特征少的问题。

数据关联阶段根据检测目标的特征计算目标和轨迹的相似度（特征相似度、IoU距离等），进而采取适当的匹配策略将检测目标和轨迹进行匹配。

常用于计算相似度的线索有目标的空间位置、运动信息以及外观特征。Bewley等（2016）提出Simple Online and Realtime Tracking（SORT），结合目标的位置和运动信息，基于卡尔曼滤波（Kalman Filter）（Kalman等，1960）预测轨迹在下一帧的位置，计算预测结果和检测结果的IoU距离作为相似度；Wojke等（2017）提出DeepSORT，在SORT的基础上加入重识别（ReID）模型，用于提取目标的外观特征，通过IoU相似度和外观特征相似度关联轨迹和检测目标；Zhou等（2020）使用目标和轨迹的中心位置来计算两者的相似度；Zhang等（2021）提出ByteTrack，在匹配过程中考虑低置信度的检测结果，只使用目标的运动信息和空间位置有效地缓解了遮挡以及小目标的问题；Du等（2022）提出StrongSORT，基于DeepSORT的架构，采用更强特征提取器和更加鲁棒的运动模型，此外还提出Appearance-Free Link model和Gaussian-Smoothed Interpolation模块，分别用于建模轨迹的全局联系和轨迹插值，提高关联的准确度。基于目标位置和运动信息的模型通常比较简单，但无法处理复杂的情况，如遮挡问题，适用于短时跟踪；而基于外观特征的匹配对遮挡问题更鲁棒，更适用于长时间的跟踪。

轨迹和目标的匹配问题可以视为二分图匹配问题，通常采用匈牙利算法（Hungarian Algorithm）（Kuhn，1955）解决，随着深度学习的发展，基于深度神经网络计算匹配关系成为一种趋势。Pang等（2021）提出一个拟密集对比学习（Quasi-Dense Similarity Learning）学习目标的嵌入特征，通过双向的Softmax操作计算轨迹和检测目标的嵌入特征相似度，然后通过搜索最近邻完成匹配过程；Jiang等（2019）提出利用图神经网络来学习出轨迹和目标的匹配关系；Chu等（2021）提出一个图Transformer模型TransMOT，TransMOT将轨迹和检测结构建模成无向图，利用图Transformer编码器编码轨迹的时空信息，然后通过图Transformer解码器建立轨迹和检测的匹配关系。基于匈牙利算法的匹配方法简单高效，是多目标跟踪算法中主流的匹配算法，而通过深度神经网络计算出的匹配关系虽然准确性更高，但是其计算量更大，难以满足实时性的需求。

由于缺少高质量的公开卫星视频多目标跟踪数据集，现有的卫星视频多目标跟踪研究较少。Feng等（2021）提出Spaital Motion Information-Guided Network（SMTNet），用双分支的Long Short-Term Memory（LSTM）分别计算轨迹的运动特征以及空间特征，SMTNet基于已有的轨迹预测一个虚拟位

置，最后通过匈牙利算法将检测结果以及虚拟位置与轨迹匹配；Wu等（2021）使用Yolov3作为检测器，利用多粒度网络Multiple Granularity Network（MGN）提取更加丰富的目标外观信息，以提高关联的准确性；Wu等（2021）提出SFMFMOT，首先利用低速特征辅助检测网络检测运动目标，然后在关联阶段基于外观特征和空间信息匹配，最后利用运动特征消除静态误跟；He等（2022）提出一个联合检测与关联的模型TGraM，通过图卷积网络构建目标的时空关系，在训练过程基于多任务对抗梯度学习解决检测和ReID任务不一致的问题。

综上所述，本文提出卫星视频多目标跟踪算法，主要贡献如下：

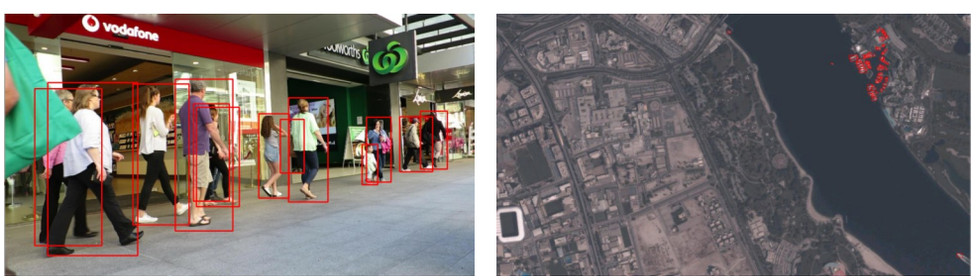
（1）针对卫星视频中的小目标检测问题，在检测网络中增加一个预测分支，提高预测特征图的分辨率，保留小目标的细节特征，此外还利用注意力机制进一步增强小目标的细节特征。

（2）利用Transformer的自注意力机制，编码全局的上下文信息，增强目标之间的联系，提高网络对于目标的响应，抑制复杂的背景。

（3）为确保检测出的目标能够与轨迹匹配，在关联阶段考虑低置信度的检测结果，从而提高跟踪性能。

2 研究方法或原理

现有多目标跟踪算法无法有效解决卫星视频的目标跟踪，如图1所示，卫星视频与监控视频存在显著差异：（1）卫星视频中的目标尺寸更小，可用特征少，给检测带来难度；（2）由于小目标和遮挡等问题导致目标置信度更低，使得关联难度更大。因此，本文提出一种卫星视频的多目标跟踪算法，如图2所示，（1）针对卫星视频中目标的特点设计小目标检测器YOLOS（YOLOX for small object），检测卫星视频中第T帧图像的目标。（2）采用一种两步关联策略，即首先将轨迹段与高置信度的检测匹配，然后匹配低置信度的检测，以确保由于尺寸小或者位于复杂背景中的目标能够与轨迹匹配。



（a） 监控视频 （b） 遥感卫星视频

（a） Surveillance Video （b） Remote Sensing Satellite Video

图1 监控视频和遥感卫星视频的差异

Fig.1 Differences between surveillance video and remote sensing satellite video

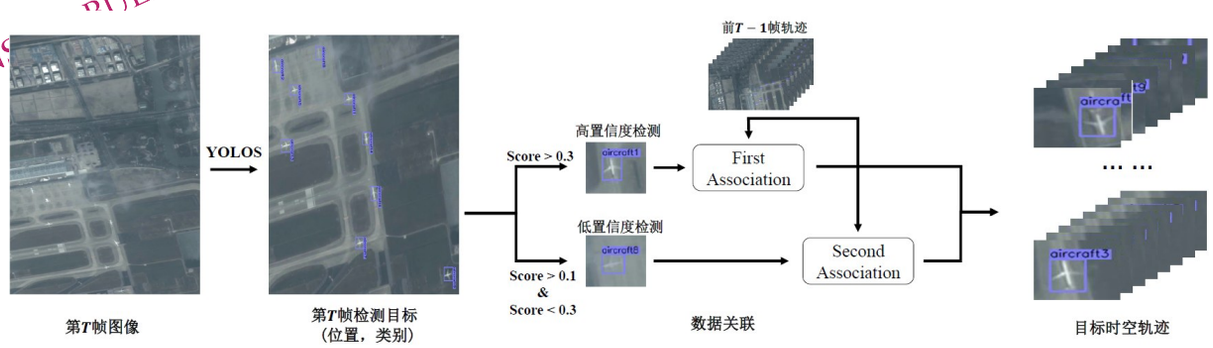


图2 多目标跟踪算法流程

Fig.2 Overview of proposed MOT method

2.1 小目标检测器

YOLOX是无锚框的一阶段目标检测算法，其主干网络为DarkNet53，颈部网络为PANet，YOLOX的检测头为解耦头，能进一步提高网络的检测性能。此外在YOLOX采用了一种更高效的SimOTA算法，在训练过程中自动为每个真值（ground-truth）分配正负样本，从而解决正负样本不均衡问题。

尽管YOLOX取得良好的性能，但是其在小目标上的检测结果仍然比较低，在COCO test-dev数据集（Lin等，2014）上的AP指标仅为31.2%。因此，本文提出YOLOS来解决卫星视频中的小目标检测问题，其结构图如图3所示，红色标注为改进部分，（1）在YOLOX中增加一个预测分支，得到更高分辨率的特征图，从而更好地保留小目标的特征，此外，还利用CBAM增强小目标的细节特征，提高小目标的检测能力；（2）为了更好地检测位于复杂背景中的目标，利用Transformer在目标之间建立更加鲁棒的关联，进一步提高检测卫星视频中目标的能力。

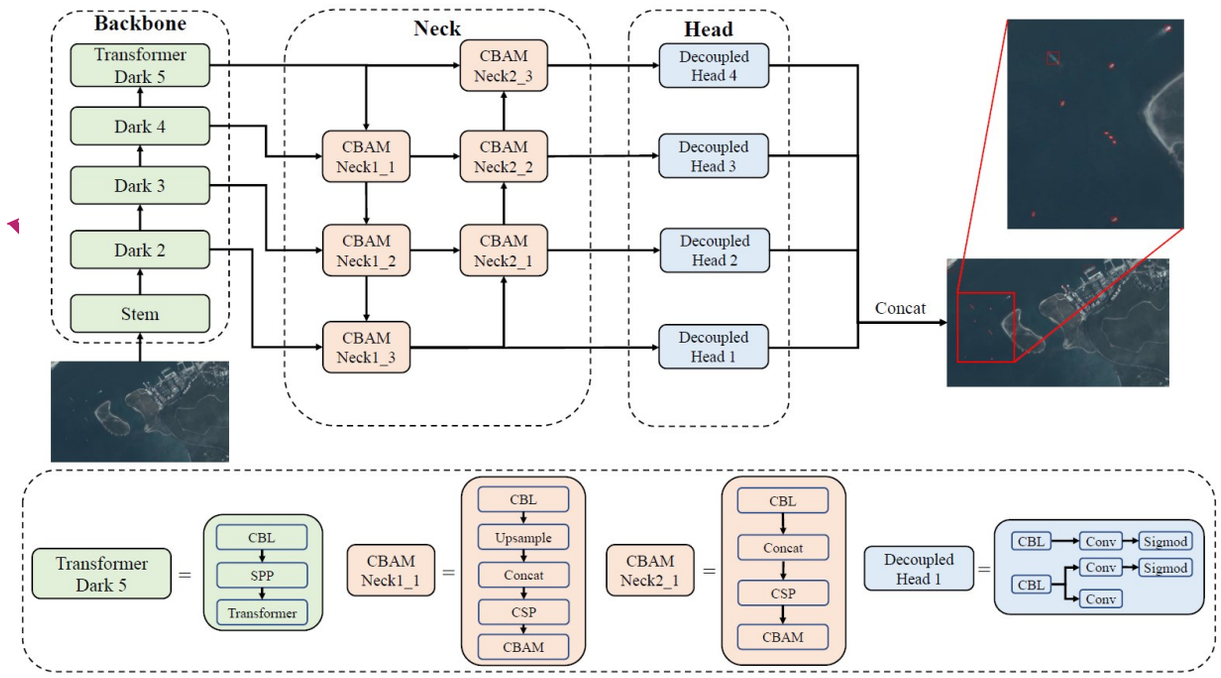


图3　YOLOS结构

Fig.3　Architecture of YOLOS

2.1.1　主干网络

当目标位于某些复杂背景中，如云雾、波浪、舰船运动产生尾流等，目标所在的局部区域很难为识别目标提供有效的信息，而图像中的目标存在相似性，如大小、形状、颜色等特征，因此利用全局的目标信息能够更好地识别位于复杂背景中的目标。受Transformer（Vaswani等，2017）的启发，本文利用Transformer中的编码器来为特征图提供全局的上下文信息，其结构如图4所示，Transformer编码器包含两个子层，一是多头注意力层（Multi-Head Attention），该层通过自注意力机制建模图像中不同位置的关系，二是多层感知机（MLP），用于变换维度，提高模型表达能力，两个子层都引入了层标准化（layerNorm）和Dropout操作，并通过残差结构连接。如图3所示，为了在特征图中融入全局的上下文信息，在主干网络的最后一个模块使用Transformer编码器，这样做一是可以减小使用Transformer带来的计算量和内存的增加，二是高层的特征图包含丰富的语义信息，借助Transformer的自注意力机制，加强全局目标间的联系，提高网络对位于复杂环境下目标的响应，增大目标与背景之间的差异，进而提高网络的检测能力。

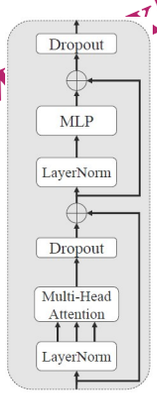


图4　Transformer编码器模块

Fig.4　Diagram of Transformer encoder

2.1.2　卷积注意力颈部网络

尽管使用Transformer能够间接提高网络的检测能力，但是卫星视频中存在大量的小目标，小目标可用特征信息少的问题没有得到处理，而且由于成像距离较大，卫星视频中的目标视觉特征都不明显，这就会导致大量的误检和漏检。因此，本文利用注意力机制来增强目标的特征，提升检测器整体的检测能力。YOLOX的每个颈部网络模块包含一个卷积层和CSP模块（Cross Stage Partial Module，跨阶段局部连接模块），在CSP模块前将多个特征图进行连接，本文在颈部网络中的每个模块最后添加了一个卷积注意力模块（CBAM）（Woo，2018），其结构如图3所示，CBAM是一个

轻量级的注意力模块，通过空间注意力和通道注意力来对特征进行增强。卫星视频中大部分区域为城市、海洋等地理区域，使用CBAM模块能够让检测网络更好地关注目标所在的区域。

2.1.3 检测头

高层、低分辨率的特征图包含丰富的语义信息，但缺少细节信息，随着下采样操作小目标的细节特征会逐渐丢失，高分辨率的特征图能够保留小目标的细节，所以使用高分辨率的特征图对检测小目标是非常必要的。如图3所示，本文增加了一个预测分支（解耦头1），该预测分支的输入为卷积注意力颈部网络中的低层特征图，其下采样率为4，相较于其他预测分支能够得到高分辨率的特征图，该特征图对小目标更敏感，能够显著提升网络对卫星视频中小目标的检测能力。结合其他预测分支，四预测分支的结构也能缓解目标尺度变化带来的影响。

2.1.4 损失函数

每个解耦头包含回归和分类分支，分别用于回归目标的边界框和分类，对于回归边界框分支，采用GIoU损失函数（Rezatofighi等，2019），对于分类分支采用交叉熵损失函数。

2.2 关联算法

由于小目标尺寸小、细节特征不明显的特点，卫星视频中的小目标在检测阶段的预测置信度比较低，如果在数据关联阶段中将低置信度的检测结果视为背景，就会造成大量小目标的漏跟，显然在遥感卫星场景下是不适用的。因此，本文采用更加关注低置信度检测的弱数据关联算法Byte（zhang等，2021），其流程如算法1所示。

算法1：关联算法

输入：卫星视频 V，检测器 D，检测置信度阈值 τ_{high} 和 τ_{low}，卡尔曼滤波器 KF，轨迹初始化阈值 ε。

输出：卫星视频目标轨迹 $\mathcal{T}$

```
for frame f_T in V do
    Det_T ← 检测器 D 检测当前帧 f_T
    Det_T^high, Det_T^low ← 根据检测置信度阈值 τ_high 和 τ_low 划分检测结果
    𝒯 ← 基于 KF 预测轨迹在第 T 帧的位置
    基于 IoU 相似度，关联 Det_T^high 和 𝒯
    Det_T^remain ← 未匹配的高置信度检测结果
    𝒯_T^remain ← 未匹配的轨迹段
    基于 IoU 相似度，关联 Det_T^low 和 𝒯_T^remain
    𝒯_T^remain ← 未匹配的轨迹段
    从 𝒯 中删除 𝒯_T^remain
    𝒯 ← 基于阈值 ε 将 Det_T^remain 初始化为新的轨迹
return 𝒯
```

3 数据结果处理与分析

3.1 实验设置

（1） 实验数据和评价指标

本文所用的实验数据来自2021高分遥感图像解译大赛，使用的数据集为高分辨率光学卫星视频中多目标跟踪数据集（He等，2022）[1]，数据由吉林一号光学卫星采集，图像场景包括不少于15个国内外常用民用机场、港口等。该数据集中包含两类目标，飞机和舰船，训练集共80个由图像序列组成的视频，图像的分辨率为1080×1920，数据集示例如图5所示。由于训练集中有21个视频无目标标注，因此在实验中将有标注的59个视频的70%划分为训练集，用于训练模型，30%划分为测试集，用于测试模型的有效性。

实验测试使用的评价指标为MOTA（Bernardin等，2008），其计算公式如式（1）所示：

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t} \tag{1}$$

其中 FN_t 表示第 t 帧中目标漏检的个数，FP_t 表示第 t 帧中目标误检的个数，$IDSW_t$ 表示第 t 帧中目标ID发生切换的次数，GT_t 表示第 t 帧中真值（ground-truth）的个数。

（2） 对比方法

表1展示了对比实验选用的方法。Joint

Detection and Tracking（JDT）方法是指将检测和跟踪联合，进行端到端地学习训练。MSOT-CNN（Bahmanyar 等，2019）、Yolov3+MGN（Wu 等，2021）以及 DSFNet+SORT（Xiao 等，2021）是应用在遥感场景下的多目标跟踪算法，所有对比方法的参数设置都遵循原论文使用的参数。

3.2 实验细节

本文所提算法通过 Pytorch 框架实现，硬件环境为：Ubuntu18.04 操作系统，Intel Xeon 5220R CPU，NVIDIA RTX3090 GPU 显卡。

基于 SGD 优化器，YOLOS 在训练集上训练 90 个 epoch，初始学习率为每张图像 0.0000625，在训练过程中采用预热（warmup）和余弦学习率衰减策略。数据增强采用 Mosaic（Bochkovskiy 等，2020）和 MixUp（Zhang 等，2017），此外采用了强的旋转数据增强，即图像的旋转角度范围设为 $(-\pi, \pi]$，在训练的最后 20 个 epoch，关闭所有的数据增强。训练采用多尺度训练的方法，图像最长边包含的像素个数范围为 1120 到 1632，batch size 大小为 4。τ_{low} 和 τ_{high} 分别设置为 0.1 和 0.3，ε 设置为 0.6。

3.3 消融实验结果分析

我们采用高分辨率光学卫星视频中多目标跟踪数据集的测试集进行实验，以验证我们提出的各项改进对于跟踪性能的影响，结果如表 2 所示。在增加一个预测分支（解耦头 1）后，跟踪性能有一个明显的提升，MOTA 指标从 49.6% 增加到 52.0%；在使用了强的旋转数据增强后，MOTA 指标提升非常大，我们的分析是遥感图像中的目标会呈现出各种不同的角度，使用强的旋转数据增强能够使网络学习到目标在不同方向的特征，从而提高网络的泛化性能；在上述基础上增加 CBAM 注意力机制后，增强了目标的特征，MOTA 指标也有一定程度的提升；通过 Transformer 将全局的上下文信息融合到特征中也能提升跟踪性能。由于在检测网络中增加了一个预测分支以及使用了 Transformer，我们提出的方法相较于 Baseline，处理速度有所降低，Baseline 的处理速度能够达到 15FPS，我们的方法仅为 10FPS 左右。

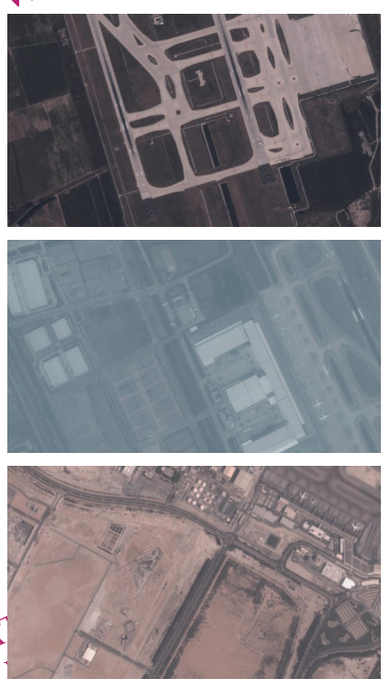
为了更直观地展示各项改进的有效性，如图 6 所示，我们可视化了网络最后一层的特征响应图，图中红色越深的区域代表网络对于该区域的响应值越高。图 6（a）为测试集中的某一帧图像，其中红色框表示目标；图 6（b）至（e）依次展示了解耦头 1 到解耦头 4 特征图的可视化结果，其特征图的分辨率依次减小，可视化结果表明，随着预测特征图的分辨率变大，网络对于单个小目标的响应值更高，更容易检测出小目标；图 6（f）和（g）表示分别表示在增加 CBAM 和 Transformer 后，解耦头 1 特征图的可视化结果，结果表明，在使用了 CBAM 注意力后，只有目标所在位置的响应值高，网络对小目标的注意权重更大，这有利于小目标的检测；在增加 Transformer 编码上下文信息后，特征图中目标与背景之间的差异更大，进一步提高小目标的检测能力。

我们还验证了弱数据关联的有效性，我们采用了两个实验，一是只使用高置信度的检测结果与轨迹进行匹配，二是同时考虑高置信度和低置信度的检测结果，其结果如表 3 所示。由表 3 可以看出在卫星视频场景下直接忽略低置信度的检测是不合理的，会导致部分小目标无法匹配轨迹，造成跟踪精度的降低。

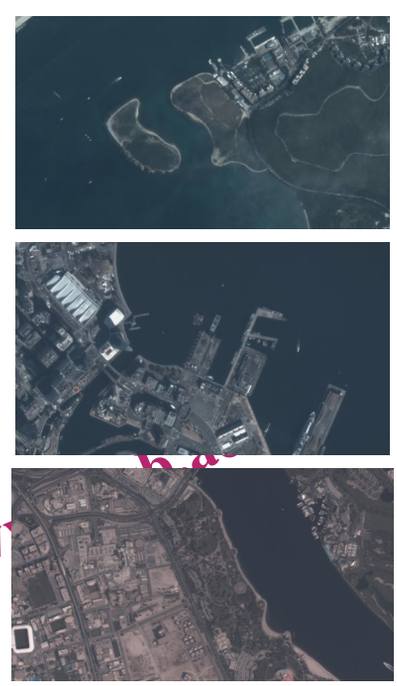
我们选取了 SORT（Bewley 等 2016）、DeepSORT（Wojke 等 2017）、MOTDT（Chen 等，2018）三种数据关联方法与 Byte 进行比对。为了公平地比较不同轨迹关联的差异，检测阶段都使用我们提出的改进 YOLOX，对比结果如表 3 所示。表 4 结果说明，采用 Byte 的 MOTA 指标最高，而且 SORT、MOTDT、Byte 在 MOTA 指标相近的情况下，Byte 的 IDF1 指标更高，表明 Byte 的关联准确性更高。此外 DeepSORT 和 MOTDT 在匹配时使用了 ReID 模型，这两种方法的 MOTA 指标都低于不使用 ReID 模型的 SORT 和 Byte，这是因为在卫星视频中，不同目标之间的外观特征差异小，以及目标与复杂背景之间的差异不突出，使用目标的外观特征会损害跟踪器的性能，因此在关联阶段需要根据卫星视频中目标的特点针对性地设计外观特征提取器，或者注重利用目标的时空、运动等信息提高关联的准确性。

如表 5 所示，我们还验证了检测与关联对跟踪性能的影响，检测阶段我们分别采用 YOLOX 和 YOLOS，检测阶段分别采用 SORT 和 Byte。表 5 的结果说明，我们在检测上的改进能够极大提升卫星视频多目标跟踪的性能，而不同的关联算法对

跟踪性能的影响相对较小。我们推测这是由于遥感视频成像的特点造成的，即以鸟瞰视角观测到的目标运动模式相对简单，使得关联阶段的难度低于自然场景的监控视频，卫星视频中很少会出现目标之间的遮挡、非线性的目标模式等问题，所以在卫星视频中采用SORT这种较为简单的关联算法也能取得较高的跟踪性能。



（a） 飞机
（a） plane



（b） 船
（b） ship

图5 数据集示例
Fig.5 Examples of dataset

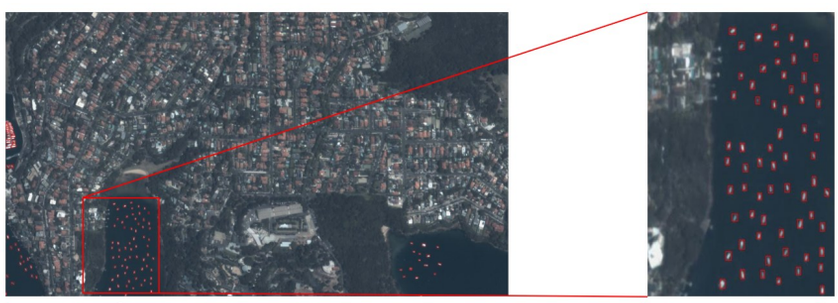
表1 对比方法
Table 1 Comparative methods

方法类别	对比方法
JDT	CenterTrack(zhou等，2020年)
	FairMOT(Zhang等，2021年)
	TransTrack(Sun等，2021年)
	TraDeS(Wu等，2021年)
TBD	MSOT-CNN(Bahmanyar等，2019年)
	Yolov3+MGN(Wu等，2021年)
	DSFNet+SORT(Xiao等，2021年)
	ByteTrack(zhang等，2021年)
	StrongSort(Du等，2022年)

表2　检测器的消融实验(↑表示越高越好，↓表示越低越好)

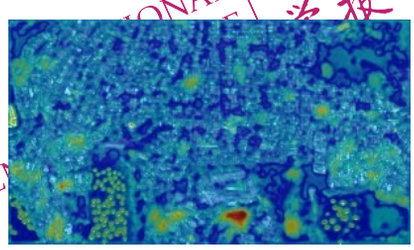
Table 2　Ablation Study on detector(↑indicates the higher metrics is better，↓indicates the lower metrics is better)

Baseline	Decoupled Head	Rotate Augmentation	CBAM	Transformer	MOTA(%) ↑	FP ↓	FN ↓	IDSW ↓
√					49.6	3561	32627	4
	√				52.0	3286	31168	1
√	√	√			61.1	6494	21369	53
√	√	√	√		62.8	5534	21184	1
√	√	√	√	√	63.1	4539	21888	42



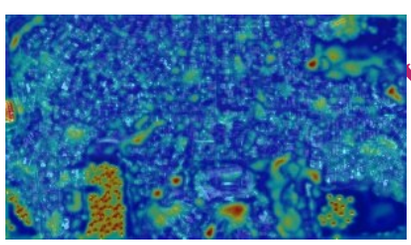
(a)　图像

(a)　Image



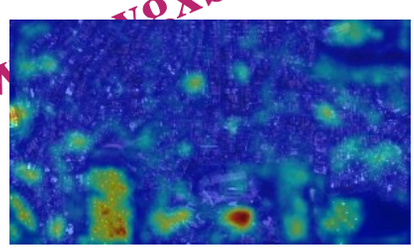
(b)　解耦头1

(b)　Decoupled Head 1



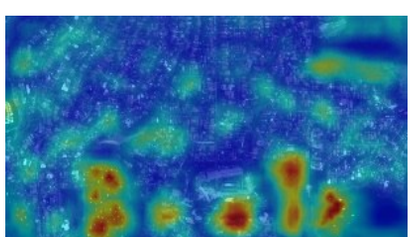
(c)　解耦头2

(c)　Decoupled Head 2



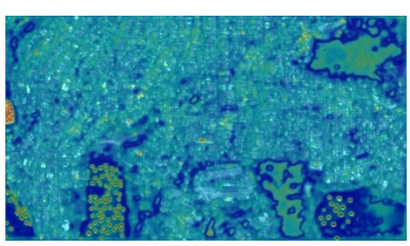
(d)　解耦头3

(d)　Decoupled Head 3



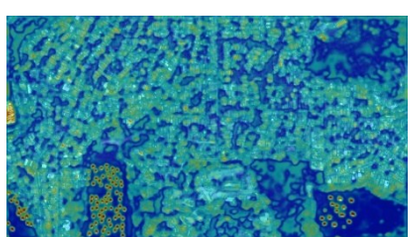
(e)　解耦头4

(e)　Decoupled Head 4



(f)　使用卷积注意力模块

(f)　w/ CBAM



(g)　使用Transformer

(g)　w/ Transformer

图6　特征图可视化结果

Fig.6　Visualization results of feature maps

3.4　对比实验结果分析

我们将提出的方法与其他多目标跟踪算法进行比较，对比结果如表6所示。表6的结果表明，我们提出的方法在MOTA和IDF1指标上都优于其他方法。基于公式1和表6可知，在卫星视频的多目标跟踪中，FN和FP对于MOTA指标的影响更大，即检测器的好坏更容易影响跟踪器的性能，如MSOT-CNN是基于单目标跟踪的方法，在复杂的背景下容易跟丢，导致FP过高；DSFNet更加注重检测运动目标，无法精准的检测出静止目标，导致FN过高。此外，相较于未使用ReID特征的方法，使用ReID特征关联的方法在卫星视频的跟踪

中没有展现出其优势，如FairMOT和CenterTrack、StrongSORT和ByteTrack，这是因为卫星视频中目标之间的外观特征差异小，基于外观特征更容易产生匹配错误。在遥感场景和监控场景下进行多目标跟踪存在明显的差异，如ByteTrack在MOT17数据集上MOTA指标高达80.3%，而在卫星视频数据集中MOTA仅为49.6%，在卫星视频下进行多目标跟踪存在更大的挑战。

表3 数据关联的消融实验(↑表示越高越好，↓表示越低越好)

Table3 Ablation Study on data association(↑indicates the higher metrics is better，↓indicates the lower metrics is better)

Method	MOTA(%)↑	IDF1(%)↑
Low Score Detections w/o	58.7	76.5
Low Score Detections w	63.1	78.0

表4 数据关联的对比(↑表示越高越好，↓表示越低越好)

Table 4 Comparison of data association(↑indicates the higher metrics is better，↓indicates the lower metrics is better)

Method	MOTA(%)↑	IDF1(%)↑
SORT	63.0	76.3
DeepSORT	53.4	70.2
MOTDT	62.9	77.2
Byte	63.1	78.0

表5 与基准方法的对比(↑表示越高越好，↓表示越低越好)

Table 5 Comparsion of baseline(↑indicates the higher metrics is better，↓indicates the lower metrics is better)

Method	MOTA(%)↑	IDF1(%)↑	FP↓	FN↓	IDSW↓
YOLOX+SORT	49.4	68.2	3783	32505	8
YOLOX+Byte	49.6	68.3	3561	32627	4
YOLOS+SORT	63.0	76.3	6251	20156	2
YOLOS+Byte	63.1	78.0	4539	21888	42

表7展示了2021高分遥感图像解译大赛高分辨率光学卫星视频视频中多目标跟踪赛道的结果，结果表明，我们提出的多目标跟踪算法具有一定的优越性，并且我们的检测器只在比赛给定的训练集上训练，未使用额外的数据。

表6 与其他方法的对比(↑表示越高越好，↓表示越低越好)

Table 6 Comparison of other methods(↑indicates the higher metrics is better，↓indicates the lower metrics is better)

Method	MOTA(%)↑	IDF1(%)↑	FP↓	FN↓	IDSW↓
MSOT-GNN	−10.6	44.7	40561	38678	89
DSFNet+SORT	−6.9	15.6	11997	64655	11
TransTrack	15.1	26.5	832	59959	138
FairMOT	37.4	62.8	12099	32797	44
Yolov3+MGN	43.9	59.4	8297	31646	323
CenterTrack	45.2	56.3	2259	36314	729
TraDeS	46.7	64.1	9161	28985	126
StrongSORT	46.9	55.4	12683	23696	1687
ByteTrack	49.6	68.3	**3561**	23627	**4**
Ours	**63.1**	**78.0**	4539	**21888**	42

表7 2021高分大赛多目标跟踪赛道前5名结果

Table 7 Top 5 results of MOT on 2021 Gaofen Challenge

Team	MOTA(%)
思源致远队	70.7231
OPTCV(Ours)	66.7951
RADI-逐星	65.2419
跟踪track	64.4516
AHU_MMIC	62.3512

4 结论

针对在高分辨率光学卫星视频中进行多目标跟踪的困难，本文在检测阶段进行改进，首先借助Transformer中的自注意力机制在特征图中融合全局的上下文信息，辅助小目标的检测，然后利用注意力机制进行特征增强，让网络更关注目标所在的区域，最后通过添加一个预测分支，使用高分辨率的特征图来检测小目标。此外在关联阶段，由于数据中存在的大量小目标导致检测出的目标置信度较低，同时考虑高置信度和低置信度的检测结果，保证检测到的小目标能够与轨迹关联。我们将高分辨率光学卫星视频中的多目标跟踪数据集划分成训练和测试集，并将我们提出的方法在测试集中验证其有效性，通过实验表明我们提出的方法更加有效。

此外，本文提出的方法仍存在局限性，首先为了提高跟踪的准确性能，牺牲了模型的运行效

率，难以达到实时性的需求；其次在关联阶段未充分考虑遥感场景下多目标跟踪的特点，如由云雾、隧道等因素引起的目标遮挡问题，目标之间的相对关系等。本文所提方法对于舰船的跟踪效果仍不理想，未来我们会更加关注遥感场景下舰船的多目标跟踪问题。

志　谢　此次高分辨率光学卫星视频中多目标跟踪数据集来自2021高分遥感图像解译软件大赛。在此表示衷心的感谢！

参考文献(References)

Bahmanyar R, Azimi S and Reinartz P. 2019. Multiple Vehicle and People Tracking in Aerial Imagery using Stack of Micro Single-Object-Tracking CNNs. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 163-170 [DOI:10.5194/isprs-archives-XLII-4-W18-163-2019]

Bernardin, K and Stiefelhagen, R. 2008. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. Journal on Image and Video Processing, 1-10 [DOI:10.1155/2008/246309]

Bewley A, Ge Z, Ott L, Ramos F and Upcroft B. 2016. Simple Online and Realtime Tracking. Proceedings of the IEEE International Conference on Image Processing, 3464-3468 [DOI:10.1109/ICIP.2016.7533003]

Bochkovskiy A, Wang CY and Liao HY. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934v1

Chen L, Ai H, Zhuang Z and Shang C. 2018. Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification. Proceedings of the IEEE International Conference on Multimedia and Expo, 1-6 [DOI: 10.1109/ICME.2018.8486597]

Chu P, Wang J, You Q, Ling H and Liu Z. 2021. Transmot: Spatial-Temporal Graph Transformer for Multiple Object Tracking. arXiv:2104.00194.

Du Y, Song Y, Yang B and Zhao Y. 2022. StrongSORT: Make DeepSORT Great Again. arXiv:2202.13514.

Feng J, Zeng D, Jia X, Zhang X, Li J, Liang Y and Jiao L. 2021. Cross-Frame Keypoint-Based and Spatial Motion Information-Guided Networks for Moving Vehicle Detection and Tracking in Satellite Videos. ISPRS Journal of Photogrammetry and Remote Sensing, 117: 116-130 [DOI:10.1016/j.isprsjprs.2021.05.005]

Ge Z, Liu S, Wang F, Li Z and Sun J. 2021. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430v2.

He Q, Sun X, Yan Z, Li B and Fu K. 2022. Multi-Object Tracking in Satellite Videos With Graph-Based Multitask Modeling. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-13 [DOI: 10.1109/TGRS.2022.3152250]

Jiang X, Li P, Li Y and Zhen X. 2019. Graph Neural Based End-to-End Data Association Framework for Online Multiple-Object Tracking. arXiv:1907.05315v1.

Kalman, RE. 1960. A New Approach to Linear Filtering and Prediction Problems. Journal of Fluids Engineering, Transactions of the ASME, 82 (1): 35-45 [DOI:10.1115/1.3662552]

Kuhn, HW. 1955. The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly,2 (1-2): 83-97 [DOI:10.1002/nav.3800020109]

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick CL. 2014. Microsoft COCO: Common Objects in Context. Proceedings of the European Conference on Computer Vision, 740-755 [DOI:10.1007/978-3-319-10602-1_48]

Lu Z, Rathd V, Votel R and Huang J. 2020. Retinatrack: Online single stage joint detection and tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 14668-14678 [DOI:10.1109/CVPR42600.2020.01468]

Pang J, Qiu L, Li X, Chen H, Li Q, Darrell T and Yu F. 2021. Quasi-Dense Similarity Learning for Multiple Object Tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 164-173 [DOI:10.1109/CVPR46437.2021.00023]

Rabbi J, Ray N, Schubert M, Chowdhury, S and Chao D. 2020. Small-Object Detection in Remote Sensing Images with End-to-End Edge-Enhanced GAN and Object Detector Network. Remote Sensing, 12(9):1432 [DOI:10.3390/rs12091432]

Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I and Savarese S. 2019. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 658-666 [DOI: 10.1109/CVPR.2019.00075

Sun P, Cao J, Jiang Y, Zhang R, Xie E, Yuan Z, Wang C and Luo P. 2020. TransTrack: Multiple Object Tracking with Transformer. arXiv:2012.15460v2.

Van Etten A. 2018. You Only Look Twice: Rapid Multi-Scale Object Detection in Satellite Imagery. arXiv:1805.09512.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems, 5999-6009 [DOI:10.48550/arXiv.1706.03762]

Wang J, Xu C, Yang W and Yu L. 2021. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. arXiv:2110.13389.

Wojke N, Bewley A and Paulus D. 2018. Simple Online and Realtime Tracking with a Deep Association Metric. Proceedings of the IEEE International Conference on Image Processing, 3645-3649 [DOI:10.1109/ICIP.2017.8296962]

Woo S, Park J, Lee JY and Kweon IS. 2018. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision, 3-19 [DOI:10.1007/978-3-030-01234-2_1]

Wu J, Cao C, Zhou Y, Zeng X, Feng Z, Wu Q and Huang Z. 2021. Multiple Ship Tracking in Remote Sensing Images Using Deep Learning. Remote Sensing, 13(18): 3601 [DOI: 10.3390/rs13183601]

Wu J, Cao J, Song L, Wang Y, Yang M and Yuan J. 2021. Track to De-

tect and Segment: An Online Multi-Object Tracker. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 12347-12356 [DOI: 10.1109/CVPR46437.2021.01217]

Wu J, Su X, Yuan Q, Shen H and Zhang L. 2021. Multivehicle Object Tracking in Satellite Video Enhanced by Slow Features and Motion Features. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-26 [DOI: 10.1109/TGRS.2021.3139121]

Xiao C, Yin Q, Ying X, Li R, Wu S, Li M, Liu L, An W and Chen Z. 2022. DSFNet: Dynamic and Static Fusion Network for Moving Object Detection in Satellite Videos. IEEE Geoscience and Remote Sensing Letters, 19: 1-5 [DOI: 10.1109/LGRS. 2021. 3124222]

Xu C, Wang J, Yang W and Yu L. 2021. Dot Distance for Tiny Object Detection in Aerial Images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1192-1201 [DOI: 10.1109/CVPRW53098.2021.00130]

Zhang Y, Sun P, Jiang Y, Yu D, Yuan Z, Luo P, Liu W and Wang X. 2021. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. arXiv:2110.06864v2

Zhang Y, Wang C, Wang X, Zeng W and Liu W. 2021. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. International Journal of Computer Vision, 129: 3069-3087 [DOI:10.1007/s11263-021-01513-4]

Zhou X, Koltun V and Krähenbühl P. 2020. Tracking Objects as Points. Proceedings of the European Conference on Computer Vision, 474-490 [DOI:10.1007/978-3-030-58548-8_28]

Multi-Object Tracking by Detecting Small Objects in Satellite Video

CUI Haowen[1,2], XU Chujie[1,2], ZHENG Xiangtao[1], LU Xiaoqiang[1]

1.Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China;
2.University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Multi-object tracking aims at locating the position of objects and estimating the trajectory of objects in remote sensing satellite videos, which has attracted much attention, such as security monitoring, motion analysis, and intelligent transportation. Compared with surveillance videos, remote sensing satellite videos contain smaller objects and a larger background, which is difficult to detect the foreground object. In addition, remote sensing satellite videos are extremely large, which requires massive computation and storage. Multi-object tracking in remote sensing satellite videos faces higher real-time requirements. Based on the mentioned problems, a multi-object tracking method for remote sensing satellite videos is proposed in this paper, which adopts tracking-by-detection paradigm. First, the backbone added Transformer to capture the global context information in the detection stage so that the detector can distinguish the objects and background. Then, the attention mechanism is used to enhance objects' features, which can make the proposed method pay more attention to the region of objects. Finally, an extra prediction branch is added to the network to get a high-resolution feature map, which retains details of small objects and is beneficial to small object detection. Due to the small objects and occlusion in remote sensing satellite videos, the confidence of hard positive samples is quite low. In the data association stage, to associate detected small objects with the existing trajectories better, an association strategy is adopted to consider high and low confidence detections simultaneously. To verify the effectiveness of the proposed method, ablation experiments and comparison experiments are carried out on the remote sensing satellite videos dataset. The proposed method achieves 63.1% MOTA and 78.0% IDF1. It can be seen that the proposed method has the best performance, which reflects its advantage of multi-object tracking in remote sensing satellite videos. The proposed method won second place in the multi-object tracking challenge of the 2021 Gaofen challenge. The proposed method is dedicated to solving the difficulty of small object tracking in remote sensing satellite videos, and some helpful methods for small object tracking are used. Experimental results show that the proposed method can improve the performance of multi-object tracking in remote sensing satellite videos.

Key words: multi-object tracking, small object detection, attention, neural network, gaofen challenge

Supported by Supported by the National Natural Science Foundation of China (No. 62271484), the National Science Fund for Distinguished Young Scholars (No. 61925112) and the Innovation Capability Support Program of Shaanxi (No. 2020TD-015)