文章编号：1003-0530(2022)12-2628-11

YOLOv5与Deep-SORT联合优化的无人机多目标跟踪算法

罗 茜 赵 睿 庄慧珊 罗宏刚
（华侨大学信息科学与工程学院，福建厦门 362021）

摘 要： 针对无人机平台下小目标检测性能差、目标尺度变化较大、复杂背景干扰等导致跟踪失败的问题，该文提出一种联合优化检测器YOLOv5(You Only Look Once)和Deep-SORT(Simple Online and Realtime Tracking with a Deep Association Metric)的无人机多目标跟踪算法。该算法使用改进的CSPDarknet53(Cross Stage Paritial Darknet53)骨干网络重新构建检测器中的特征提取模块，同时通过自顶向下和自底向上的双向融合网络设计小目标检测层，采用无人机航拍数据集训练更新优化后的目标检测网络模型，解决小目标检测性能差问题；在跟踪模块中，提出结合时空注意力模块的残差网络作为特征提取网络，加强网络感知微小外观特征及抗干扰的能力，最后采用三元组损失函数加强神经网络区分类内差异的能力。实验结果表明，优化后的目标检测的平均检测精度相比于原始YOLOv5提升了11%，在UAVDT数据集上相较于原始跟踪算法准确率与精度分别提高了13.288%、3.968%，有效减少目标身份切换频次。

关键词： 深度学习；目标跟踪；目标检测；无人机
中图分类号： TP391 **文献标识码：** A **DOI：** 10.16798/j.issn.1003-0530.2022.12.017

引用格式： 罗茜，赵睿，庄慧珊，等. YOLOv5与Deep-SORT联合优化的无人机多目标跟踪算法[J]. 信号处理，2022，38(12)：2628-2638. DOI：10.16798/j.issn.1003-0530.2022.12.017.
Reference format： LUO Xi，ZHAO Rui，ZHUANG Huishan，et al. UAV multi-target tracking algorithm jointly optimized by YOLOv5 and Deep-SORT[J]. Journal of Signal Processing，2022，38(12)：2628-2638. DOI: 10.16798/j.issn.1003-0530.2022.12.017.

UAV Multi-Target Tracking Algorithm Jointly Optimized by YOLOv5 and Deep-SORT

LUO Xi ZHAO Rui ZHUANG Huishan LUO Honggang
（Institute of Information Science and Engineering，Huaqiao University，Xiamen，Fujian 362021，China）

Abstract: Aiming at the problems of tracking failure caused by poor detection performance of small targets，large target scale changes，and complex background interference under the unmanned aerial vehicle platform，this paper proposed an unmanned aerial vehicle multi-target tracking algorithm that jointly optimized YOLOv5 (You Only Look Once) and Deep-SORT (Simple Online and Realtime Tracking with a Deep Association Metric). The algorithm used the improved CSP-Darknet53 (Cross Stage Paritial Darknet53) backbone network to reconstruct the feature extraction module in the detector. At the same time，the small target detection layer was designed by the top-down and bottom-up bidirectional fusion network. In the meanwhile，the optimized target detection network model was trained by the unmanned aerial vehicle aerial photography dataset，which solved the problem of poor detection performance of small targets. As for the tracking module，

收稿日期：2022-05-10；修回日期：2022-08-03
基金项目：福建省自然科学基金项目(2019J01055)

a residual network combined with the spatiotemporal attention module was proposed as a feature extraction network to enhance the network′s ability to perceive small appearance features and anti-interference. Finally, the triple loss function was used to strengthen the ability of the neural network to distinguish within-class differences. The experimental results show that the average detection accuracy of the optimized target detection is improved by 11% compared with the original YOLOv5, and the accuracy and precision of the UAVDT (The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking) data set are improved by 13.288% and 3.968% respectively compared with the original tracking algorithm, effectively reducing the target. Identity switching frequency.

Key words: deep learning; target tracking; target detection; unmanned aerial vehicle

1　引言

随着人工智能的发展与进步，无人机（Unmanned Aerial Vehicles, UAV）发展迅猛，由于无人机具有运动灵活、能耗低、适应性强、无人员伤亡风险等优势在军事和民用领域应用广泛[1-3]，例如用于军事上的敌情侦察、农业中的灌溉作业、企业数据检测等。此外，自 21 世纪以来，计算机视觉发展迅速，尤其是其中的视觉目标跟踪技术，发展越来越成熟，在很大程度上解决了人们生活中的智能化需求，例如：自动驾驶、智能安保等，提高了人们的生活质量，因此将目标跟踪技术与无人机进行融合展现了巨大的应用前景。

由于无人机可灵活移动，对地面观测的视角广阔，目标搜索范围较大，有助于采集更加全面的目标信息，与此同时，出现的干扰物体也会较多，这将导致目标与背景之间可区分性差、目标之间相互遮挡等问题的出现；此外，无人机受到地面高度的制约将导致图像中的目标多为小目标，且无人机由于自身的高速运动会频繁出现相机抖动、视角变换等现象，从而使得跟踪目标尺度变化较大；这些问题将影响无人机视角下的跟踪准确度与精度，且与其他场景下的跟踪相比，无人机平台下的跟踪目标身份切换次数也较多。在无人机视觉领域下，Bae S H 等人[4]首次提出利用轨迹置信度来解决目标遮挡问题的在线多目标跟踪算法，该算法虽然能够减小目标切换频次但降低了跟踪准确度以及精度。AL-SHAKARJI 等人[5]设计了一种三步级联数据关联方法，既保证了实时跟踪又保证了较高的跟踪精度，但与此同时增加了目标切换频次。Jin J 等人[6]提出一种新型的在线多目标跟踪网络，利用由 Siamese 网络提取到的外观信息与由光流和卡尔曼滤波器获取到的运动信息，将目标与现有轨迹关联起来得到跟踪结果，虽然能在一定程度上提高跟踪准确度，但降低了跟踪精度且加剧了目标漂移。由此可见，现有的无人机多目标跟踪算法难以平衡目标跟踪精度与目标漂移问题，实现稳定可靠的无人机多目标跟踪依旧面临着巨大的挑战。

传统跟踪算法大多采用概率密度和图像边缘特征作为跟踪标准，将概率梯度上升的方向作为目标搜索方向，这些算法虽然易部署，但其特征表示性能较差，无法处理复杂场景下的目标跟踪，而深度学习与图像处理的结合能够提高特征提取性能且处理速率远超于传统算法，因此基于深度学习的跟踪算法在性能上具有较大的优越性。其中，基于检测的跟踪[7]（Tracking-By-Detection）是目前使用最为广泛的跟踪算法之一。基于检测的跟踪主要分为两大类：一类是将检测模块和跟踪模块分别训练，再将二者关联起来进行目标跟踪；如 KIM 等人[8]提出一种新颖的多假设跟踪算法，通过卷积神经网络对每个目标进行外观建模，接着与假设轨迹进行最优匹配，该算法虽然在性能上相较原始算法有所提升，但速度仍不高。BEWLEY 等人提出 SORT[9]（Simple Online and Realtime Tracking），一经问世就引起了广泛关注，该算法因框架简单使得运行速度较快，但与此同时，算法抗遮挡能力较差，无法进行较长时间的稳定跟踪。另一类是将检测模块与跟踪模块集成到单一网络中进行多任务学习，同时完成目标检测与跟踪，例如 WANG 等人[10]提出了 JDE（Joint Detection and Embedding）算法。前者主要分为两个步骤：首先在目标检测模块中检测出单帧中的目标，提取其分类与定位信息，其次将这些信息输入到跟踪模块中，并提取目标的表观特征，最后将目标检测结果与跟踪结果通过选定的数据关联

方法进行匹配从而创建相应的轨迹；由此可知，这种方法需要对目标进行两次特征提取，实时性较差，但该方法可以针对每个任务分别训练最合适的模型，此外，跟踪模块首先根据检测到的目标边界框进行裁剪，然后进行特征提取有助于处理对象的比例变化。后者虽然只用单个网络就能同时进行分类、定位、与跟踪，但锚框较为粗糙，极易产生误检，尤其是在小目标及目标特征不够显著的情况下更易产生跟踪失败，虽然速度相对提升了，但跟踪精度与准确度要更低，即跟踪稳定性较差。而无人机视角下的图像具有视野范围大且背景杂乱、目标占据整个图像尺寸较小且特征不明显等特征，造成目标特征提取与模型建立困难，因此，与前者策略相比，后者策略难以在无人机平台下进行稳定的多目标跟踪。近年来，Transformer因其具有强大的自注意力层，在图像识别以及视频分析中具有广泛的应用。Zeng F等人[11]通过引入“track query”对整个视频中的跟踪实例进行建模，“track query”能够在帧间传输并更新从而完成目标跟踪任务。Cai J等人[12]基于Transformer设计了一种端到端的多目标跟踪框架，通过一个大的时空内存存储被跟踪对象ID，并根据跟踪需求自适应地提取和聚合内存中的有用信息来实现目标与轨迹之间的关联。Zhou X等人[13]首次提出了一种基于Transformer的全局多目标跟踪网络结构，利用Transformer对输入视频序列中的所有目标特征进行编码，并利用轨迹查询将这些目标分配给不同的轨迹。

基于以上分析，本文采用“Tracking-By-Detection”中两阶段跟踪策略，提出在无人机平台下，联合优化目标检测器与跟踪器的多目标跟踪算法。主要贡献如下：(1)针对无人机视角下小目标检测性能差、目标尺度变化较大问题，本文在YOLOv5的基础上进行改进，通过增加小目标检测层来提高对小目标检测精度，利用特征融合将不同尺度的特征进行多尺度加权融合以解决目标尺度变化大问题；在骨干网络引入Transformer结构，提高目标的定位精度。(2)针对复杂背景干扰、遮挡导致跟踪目标丢失问题，本文采用ResNet50作为外观特征提取的骨干网络，提高网络感知微小外观能力，并添加时空注意力模块从而有效地提取目标关键特征，引入Triple loss损失函数，加强区分类内差异能力。(3)通过在VisDrone2021[14]数据集上的大量实验证明，在无人机平台下，改进后的目标检测器的平均检测精确度比原始YOLOv5提高了11%；在UAVDT[15]数据集上跟踪准确度与精度分别提高了13.288%、3.968%，且在一定程度上减小了目标身份切换频次，能基本满足无人机平台下多目标跟踪稳定性需求。

2 本文算法

根据“Tracking-By-Detection”两阶段跟踪策略，本文通过联合优化检测模块与跟踪模块来解决无人机平台下易出现的目标漂移(ID Switch)，跟踪丢失等跟踪失败问题，主要由四个部分组成：目标检测、外观模型、目标运动预测模型以及数据关联，整个算法结构框架图如图1所示。首先将待检测视频传入到目标检测模块中进行目标检测，此时采用改进后的YOLOv5作为检测器，输出目标的检测框信

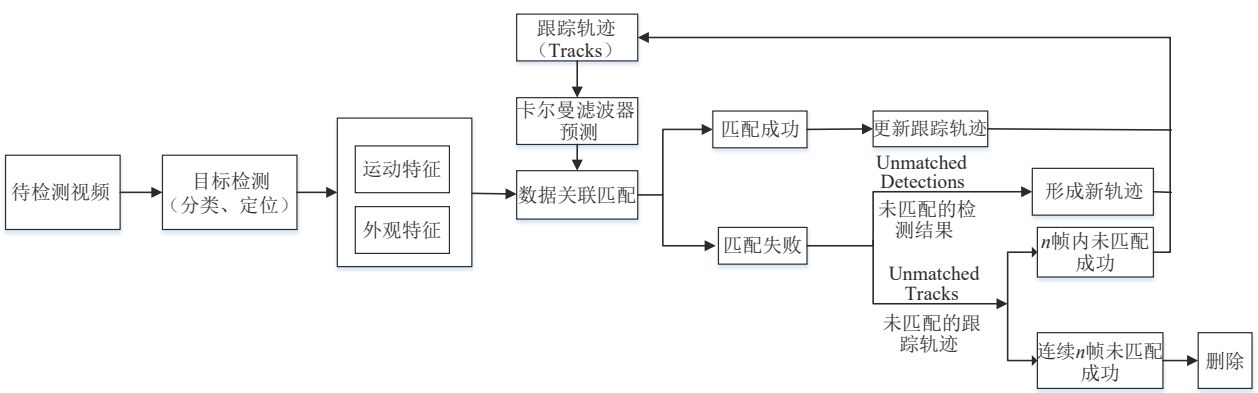


图1 算法结构框架图

Fig. 1 Algorithm structure framework diagram

息；其次通过跟踪模块中由残差块构造的特征提取网络提取目标的外观特征，同时利用跟踪模块中的预测模型输出目标的预测位置信息；最后将目标检测器的检测结果与跟踪模块的预测结果通过关联度量函数以及关联方法进行数据关联，根据关联结果得到最终的跟踪结果。

2.1　优化YOLOv5检测模型

多目标跟踪策略“Tracking-By-Detection”首先将输入图像送入检测器中的进行目标检测后再根据检测目标进行跟踪，跟踪性能在一定程度上依赖检测器效果，因此在跟踪的整个过程中，目标检测尤为重要。传统的目标检测算法主要流程为：首先在输入图像中采用滑动窗口对图像进行遍历滑动得到目标可能所在区域，得到候选框，其次提取候选框中的图像特征并转换为特征向量，常用的传统特征提取算法有梯度直方图、局部二值算法等，最后根据特征向量判别是否为目标对象以及对应的类别。传统的目标检测算法在候选框选择时会产生大量的冗余窗口进而产生冗余计算，影响整个算法的速度与性能，此外，传统算法只能提取低级特征，最终无法得到全局最优解。

近年来，计算机技术的高速发展解决了深度学习的计算复杂问题，促进了基于深度学习的目标检测算法的发展，相比于传统算法，其不仅提高了整个检测速度且能够以较高的精度识别目标，成为当前目标检测算法的主流研究方向。基于深度学习的目标检测算法根据检测过程中是否含有候选区域（region proposal）分支划分为：two-stage与one-stage检测算法两大类。two-stage检测算法在检测过程中含有region proposal，通过卷积神经网络对region proposal生成的候选框中的图像进行分类和定位，常见的算法有Faster R-CNN[16]、Libra R-CNN[17]；基于one-stage的检测算法使用回归的方法从空间上分割边界框和相关的类别概率，通过用一个单独的端对端网络完成目标的位置与类别的输出，显而易见整个算法的检测速度得到了很大的提升，与此同时，由于没有更精准的候选区域，该算法的检测精度会相应的降低，但随着计算机性能的提升与深度学习的不断发展，基于one-stage的检测算法网络结构得到不断的优化，性能得到不断的提高。one-stage检测算法中由YOLO[18]演变而来的YOLOv5不仅检测速度快且精度与two-stage中的Faster R-CNN相当，因此本文选用YOLOv5网络模型，并对YOLOv5网络结构进行改进优化后，将其作为多目标跟踪算法中的检测模块以达到更好的检测效果。

2.2　检测网络结构

虽然YOLOv5相对其他目标检测算法已取得较好的检测效果与实时性，但其对于小目标检测性能较差，而无人机视角下的目标多为小目标且目标尺度变化较大，对此，本文提出改进优化后的YOLOv5，对应的网络结构如图2所示。在骨干网络（Backbone）中第二层引出下采样的第四个尺度，利用FPN[19]（Feature Pyramid Networks，特征金字塔网络）结合PANet[20]（Path Aggregation Network，路径聚合网络）将四个不同分辨率的特征图进行特征融合，在预测部分增加小目标检测层进行微小物体检测，结合其余三个预测层，用四个不同感受野的预测层来提高对小目标检测精度，此外，采用自顶向下与自底向上的双向融合网络能较好地适应目标尺度变化。无人机平台下的背景复杂、目标与背景以及目标之间相互遮挡等问题影响目标定位的准确性，对此，本文在骨干网络最后一层将原始的BSP（Bottleneck and CSP）替换为Transformer[21]结构，利用Transformer捕获全局信息和上下文信息并通过其自注意力机制挖掘潜在的图像特征。Transformer结构如图2中的C3TR模块所示，其包含两个子层：multi-head attention layer（多头注意力层）和MLP（Multilayer Perception，多层感知机）全连接层；子层之间用残差结构连接，外加LayerNorm和Dropout层防止网络过拟合。

2.3　优化Deep-SORT

Deep-SORT[22]在SORT基础上使用更加可靠的关联度量和关联方法，能够有效地进行长时间跟踪并在很大程度上减少了跟踪过程中的身份转换（identity switches，IDs）。关联度量的选取以及关联方法是跟踪模块的主要任务，目前的关联度量方法主要是利用外观相似度、运动状态信息、位置关系等建立；关联方法主要分为两大类：离线和在线关联，离线关联根据全部时间序列中的目标信息，全局优化进行目标和轨迹的关联，而在线关联仅使用当前

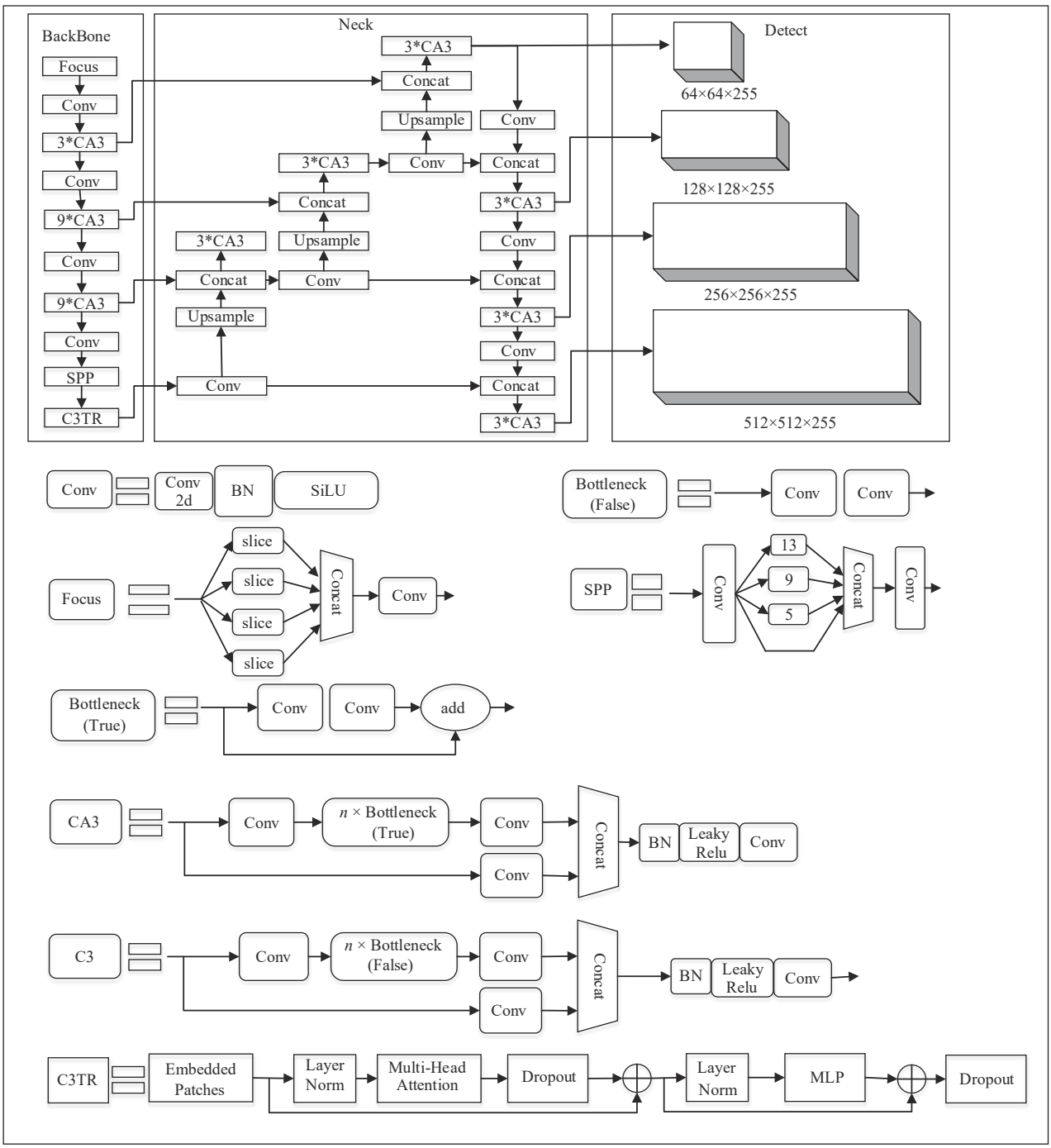


图2 优化目标检测网络结构图

Fig. 2 Optimization target detection network structure diagram

帧与历史帧信息进行关联。在无人机平台下,多目标跟踪算法只能逐帧采集图像,因此在线关联方法更加适用于无人机视角下的多目标跟踪。Deep-SORT联合目标运动状态信息与外观特征进行关联度量,以此达到相对稳定的多目标跟踪状态。

Deep-SORT利用卡尔曼滤波器预测速度快且精度较高的特点来估计目标下一帧的状态:

$$(x, y, g, h, x', y', g', h') \tag{1}$$

其中,(x, y)为目标边界框中心坐标;g为目标边界框的长宽比值;h为边界框高度;(x', y', g', h')表示

为对应参数的速度。得到预测信息后，采用马氏距离对目标检测位置与预测的目标状态进行相似性运动匹配：

$$\boldsymbol{d}^{(1)}(i,j)=(\boldsymbol{d}_j-\boldsymbol{y}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{d}_j-\boldsymbol{y}_i) \tag{2}$$

其中，$\boldsymbol{y}_i$代表第i条轨迹的目标预测位置，$\boldsymbol{d}_j$为第j个目标检测边界框，$\boldsymbol{S}_i$为卡尔曼滤波器预测得到的当前帧协方差矩阵，得到二者之间的马氏距离后，利用χ^2分布，通过设置阈值来判断第j个目标检测边界框是否与第i条轨迹相似：

$$\boldsymbol{b}_{i,j}^{(1)}=1[\boldsymbol{d}^{(1)}(i,j)\leqslant\boldsymbol{t}^{(1)}] \tag{3}$$

式中，$\boldsymbol{t}^{(1)}$为χ^2对应95%置信度阈值，本文中取为9.4877；若$\boldsymbol{d}^{(1)}(i,j)$小于阈值，$\boldsymbol{b}_{i,j}^{(1)}$取1，表示关联成功，否则取0，表示关联失败。为了防止丢失目标重新进入视野而出现身份转换现象的发生，对一段帧数内目标特征向量进行保留，由集合$\boldsymbol{R}_k$表示：

$$\boldsymbol{R}_k=\{\boldsymbol{r}_k^{(i)}\}(k\in[1,L_k]) \tag{4}$$

式中，L_k为保留的目标特征数量，本文设定为100。对于检测结果与轨迹的外观特征的度量，Deep-SORT[22]采用最小余弦距离测量二者之间的特征距离：

$$d^{(2)}(i,j)=\min\{1-\boldsymbol{r}_j^{\mathrm{T}}\boldsymbol{r}_k^{(i)}|\boldsymbol{r}_k^{(i)}\in\boldsymbol{R}_i\} \tag{5}$$

式中，$d^{(2)}(i,j)$为计算第i个跟踪器临近N个成功关联的轨迹与第j个被检测目标特征向量间的最小余弦距离。$\boldsymbol{r}_j^{\mathrm{T}}\boldsymbol{r}_k^{(i)}$表示归一化后的第$j$个检测目标与第$i$个轨迹特征向量之间的最小余弦距离。与$b_{i,j}^{(1)}$类似，设定阈值$t^{(2)}$，根据$d^{(2)}(i,j)$判断二者是否关联成功：

$$b_{i,j}^{(2)}=1[d^{(2)}(i,j)\leqslant t^{(2)}] \tag{6}$$

得到运动与外观信息的匹配关联度量后，将二者机制融合构造关联度量函数：

$$c_{i,j}=\lambda d^{(1)}(i,j)+(1-\lambda)d^{(2)}(i,j) \tag{7}$$

其中λ为不同关联度量比例系数，$c_{i,j}$越小表示第i个跟踪轨迹与第j个检测目标越相似。联合考虑运动特征与外观特征来判断第i个轨迹与第j个被检测目标是否关联成功：

$$b_{i,j}=\prod_{m=1}^{2}b_{i,j}^{(m)} \tag{8}$$

最后根据关联度量结果利用匈牙利匹配算法得到最优跟踪轨迹。

无人机存在大量的运动不确定性，如风力影响无人机稳定飞行而产生的相机抖动以及无人机自身的飞行速度等，从而导致在无人机视角下相同目标相邻两帧的马氏距离仍然很大，即使在匹配正确的情况下，也可能会误判为非同一目标，最终造成匹配失败。因此，关联度量需侧重于外观特征信息，此时，跟踪模块中的外观建模性很大程度上决定了最终跟踪结果。随着计算机视觉的高速发展，基于深度学习的特征提取网络展现出其性能的优越性与特殊性，与传统方法相比，它能够更加完整、可靠地学习与提取物体的特征。Deep-SORT采用11层神经网络输出128维目标特征向量，该网络仅能提取较为明显的特征，对于一些细微特征提取能力很差，而无人机视角下的目标多为小目标，目标之间的差异在图像中可能并不明显。因此，为了提高原始网络的特征提取能力，选择用ResNet50网络输出2048维目标特征向量，加强网络对细微特征提取能力，此外，在该网络中增加时空注意力机制以提高网络重识别能力，改善无人机视角下目标长期被遮挡而造成跟踪丢失问题，最后引入Triple loss损失函数增强网络区分类内差异的能力，最终提高算法跟踪准确度与精度，并减小身份切换频次。

3　实验与分析

3.1　数据集

实验过程中采用VisDrone2021数据集训练目标检测网络，最终在UAVDT数据集上验证所提出算法的跟踪性能。VisDrone2021[14]数据集由天津大学机器学习和数据挖掘实验室AISKYEYE团队创建，包含了无人机拍摄的65228帧和10209张静态图像组成的400个视频序列，覆盖了14个不同城市的对象（行人、车辆、自行车等10种类别对象）。UAVDT[15]数据集由100个视频序列组成，共8000帧图片，涵盖了不同场景下无人机拍摄的车辆图像，包括广场、高速公路、主干道、路口等场景，除此之外，该数据集收集了从低到高三个不同高度下的拍摄图像并包含了视角变换、相机运动、光照变化、背景干扰等航拍难点。

3.2　评价指标

为了更好地评估优化后算法的性能，本文使用目标检测典型评价指标AP（Average Precision）和

mAP(mean Average Precision)来评估优化前后目标检测器的性能,采用CLEAR MOT[23]评价指标衡量多目标跟踪算法的性能。

AP指目标检测精确率与召回率绘制的Precision-Recall(PR)曲线与x轴围成的面积,其中精确度与召回率表达式分别为式(9)、(10)所示:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

其中,TP为正确检测出的目标,FN为未检测目标,FP为错检目标。mAP是指所有检测目标类AP的平均值,即:

$$\text{mAP} = \frac{\sum_{i=1}^{n} \text{AP}_i}{n} \tag{11}$$

多目标跟踪CLEAR MOT[23]评价指标主要包含以下参数:

ML(Mostly Lossed):与GT(Ground Truth)小于20%时间内都匹配成功的跟踪数;

MT(Mostly Tracked):与GT(Ground Truth)在80%时间内都匹配成功的跟踪数;

MOTP(Multiple Object Tracking Precision)[23]:

$$\text{MOTP} = 1 - \frac{\sum_{t,i} d_t^i}{\sum_t c_t} \tag{12}$$

其中d_t^i表示第t帧下检测目标O_i与其匹配的跟踪器预测的目标位置之间的距离,c_t表示第t帧的成功匹配数,MOTP(多目标跟踪精度)与目标检测精度有关,反映跟踪的定位精确度,数值越接近1表示精度越高。

MOTA(Multiple Object Tracking Accuracy)[23]:

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + \text{fp}_t + \text{mme}_t)}{\sum_t g_t} \tag{13}$$

其中m_t为t帧时刻漏检数,fp_t(false positive)为t帧时刻误报数量,mme_t(mismatches)为t帧时刻错误匹配数量,MOTA(多目标跟踪准确度)与目标检测精度无关,衡量跟踪算法在检测目标与保持轨迹的性能,数值越接近1表示性能越好。

3.3 实验与结果分析

本文的网络训练与验证平台为Intel(R) Core(TM) i7-6700K CPU和1060Ti GPU。首先验证改进后的YOLOv5性能的优越性,选取VisDrone2021中的目标检测数据集训练检测器,实验结果如表1所示。由于最终在UAVDT数据集上跟踪的目标只有Car、Bus、Truck三大类,所以在目标检测网络中关注该三类检测的平均精度,以及所有类的mAP,表格中的mAP@.5:.95表示在不同的交并比阈值(从0.5至0.95,步长为0.05)下的平均mAP。由表1可知,原始YOLOv5网络模型对于无人机平台下的小目标检测性能较差且不同类别目标检测精度差异较大,Car的AP达到64.5%而Truck的AP仅有14.3%,所有类的mAP只有21.0%。

表1 YOLOv5改进前后性能对比结果

Tab. 1 Performance comparison results before and after YOLOv5 improvement

Network models	Car AP (%)	Bus AP (%)	Truck AP(%)	mAP (%)	mAP@.5:.95 (%)
YOLOv5	64.5	42.1	14.3	21.0	10.3
YOLOv5_1	**72.9**	**53.1**	**35.1**	**32.0**	**17.1**

本文优化后的检测器网络模型训练结果如表中YOLOv5_1所示,其训练结果mAP为32.0%,相比原始网络提高了11.0%,且所有类别的AP值都得到大幅度的提高,说明本文改进的目标检测网络模型在检测无人机视角下的图像具有明显的优势,给无人机视角下的多目标跟踪提供了一个良好的检测器,避免因漏检、错检等检测器性能导致后续的跟踪失败问题。

为了进一步验证所提出的无人机平台下多目标跟踪算法的性能,本文在UAVDT中选取白天、黑夜、雾天以及不同角度和不同高度等复杂场景下的数据集进行验证,实验结果如表2所示,此外,将本文算法与其他两种多目标跟踪算法的结果进行对比,对比结果如表3所示。表2中Deep-SORT1表示优化后的跟踪器,且由表2可知,优化后的目标检测器(YOLOv5_1)与原始跟踪器(Deep-SORT)相结合能够明显降低目标身份切换频率,与原始跟踪算法相比降低了162次,且漏检数(False Negative,FN)减小了15593帧,这表明检测器的性能能够直接影响跟踪算法的整体性能;当将改进后的目标检测器与原始跟踪器结合时,由于目标检测器精度提高而原

表2　本文算法在UAVDT数据集中的跟踪结果

Tab. 2　Tracking results of the algorithm in UAVDT dataset

Tracking Model	MOTA (%)	MOTP (%)	FN(帧)	FP(帧)	IDs
YOLOv5+Deep-SORT	23.237	71.332	102840	157750	1096
YOLOv5_1+Deep-SORT	15.692	70.894	87247	199230	934
YOLOv5+Deep-SORT1	31.217	71.541	120900	112720	872
ours	**36.525**	**75.300**	75750	140640	**819**

表3　目标跟踪算法性能对比结果

Tab. 3　Target tracking algorithm performance comparison results

Tracking Model	MOTA(%)	MOTP(%)	FN(帧)	FP(帧)	IDs
CMOT[4]	27.178	75.112	146420	98915	2920
SORT[9]	33.163	76.71	166490	57440	3918
ours	**36.525**	**75.300**	75750	140640	**819**

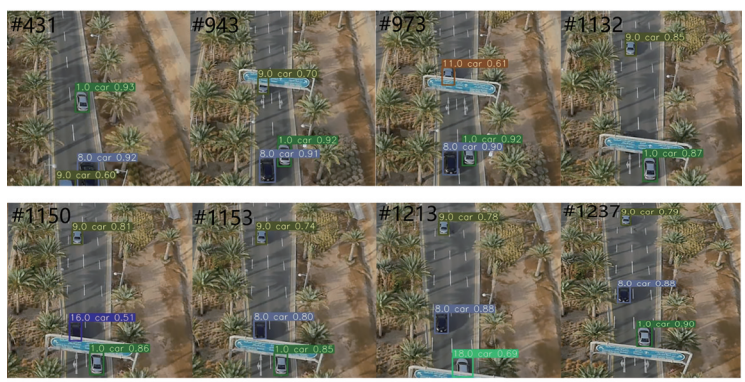
始跟踪器特征提取能力不足，在很大程度上影响了跟踪准确度与精度，造成最终跟踪准确度与精度的下降。因此需对Deep-SORT进行改进，本文通过优化Deep-SORT特征提取网络提高MOTA与MOTP值，且大幅度减小IDs。联合优化后的跟踪算法性能如表中ours所示，相比于原始跟踪算法，MOTA与MOTP分别提高了13.288%，3.968%，IDs减小了277次；除此之外，原始检测器与优化后的跟踪器结合后的跟踪性能也能够得到相应的提升，虽然性能没有联合优化算法优越，但相比于原始跟踪算法，提高了跟踪准确度和精度且改善了IDs，进一步表明改进后的跟踪器的鲁棒性和优越性。表3中选取了具有代表性多目标跟踪算法与本文算法进行对比，由表可知CMOT[4]、SORT[9]算法难以平衡目标跟踪精度与目标漂移问题，而本文算法在目标跟踪精度、准确度以及目标切换频次上达到较好的效果。

综上所述，本文所提出的多目标跟踪算法相较于原始跟踪算法不仅能够提高跟踪准确度和精确度，还能够降低目标身份切换频次，能够在无人机平台下进行稳定可靠的多目标跟踪。

为了更加直观地展示跟踪效果并验证所提出算法的鲁棒性，本文选取了UAV123[24]数据集上的视频进行展示，实验效果如图3所示。原始跟踪算法跟踪效果如图3(a)所示，在视频中的第431帧，未能识别出刚进入视野中的左边白色车辆，直至后两帧才检测到该目标并标记ID为“9.0 car”，在后续跟踪过程中由于背景遮挡，导致跟踪失败，使得目标ID在第973帧时切换为“10.0 car”；而左边白色车辆发生了两次身份切换。将原始检测器替换为YOLOv5_1后的跟踪算法跟踪效果如图3(b)所示，整体上的检测精度较前者算法检测精度要更高，且抗遮挡能力较强，如图中第943帧与第1150帧所示，即使在目标被遮挡的情况下依旧能够识别出被检测对象，改善了原始跟踪算法因背景复杂导致跟踪丢失问题；此外，由于Transformer具有捕获全局信息和上下文信息的特点，进一步增强了目标重新识别的能力，图3(b)中虽然三辆车在跟踪过程中都发生一次IDs，但后续又能成功重新识别为原来的ID，如第431帧与第1237帧所示，三辆车的ID相同。将跟踪器替换为Deep-SORT1后的跟踪算法跟踪效果如图3(c)所示，由于优化后的跟踪器增强了网络提取细微特征的能力以及区分类内差异的能力，与原始跟踪算法相比，能够改善跟踪过程中的失帧问题，如图中第1156帧所示，相比于图3(a)，图3(c)中并没有出现跟踪丢失问题，且整个跟踪过程中上目标切换过程帧数较少。联合优化后的跟踪算法可视化实验结果如图3(d)所示，整个车辆跟踪过程中除右边白色车辆发生一次IDs外，其余车辆并没有发生身份切换，并且能在目标大部分被遮挡的情况下成功地重识别，整体目标检测精度较原始跟踪算法要高。此外，本文算法针对雾天公路真实场景下的车辆进行跟踪实验，实验仿真结果如图4所示，该场景下的目标多为小目标且伴有非目标遮挡以及视线干扰，由实验结果可知，本文算法能够解决无人机平台下小目标检测性能差、目标尺度变化较大、复杂背景干扰的跟踪问题，如图4中‘32.0 car’，在整个跟踪过程中仅发生一次IDs，且在第351帧时仍能够识别被浓雾严重遮挡的小目标‘32.0 car’；目标‘11.0 car’在跟踪全程中仅发生短暂的目标丢失后又能够重识别为原来的ID。

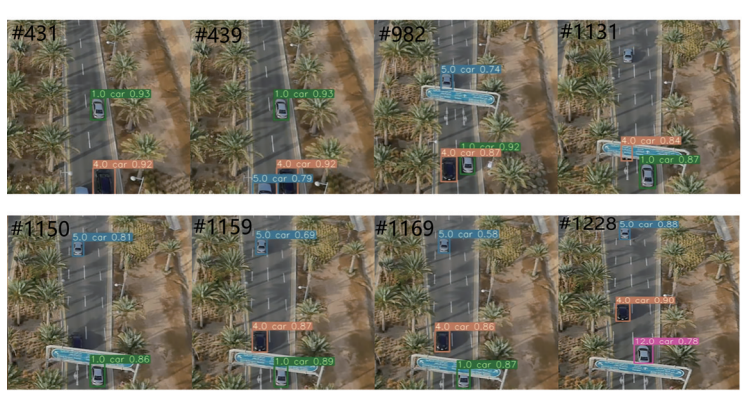
综上所示，本文所提出的联合优化的多目标跟踪算法能够在无人机平台下进行稳定可靠的多目标跟踪。

(a) YOLOv5+Deep-SORT



(b)YOLOv5_1+Deep-SORT

(c) YOLOv5+Deep-SORT1



(d) ours

图 3 跟踪效果对比结果图

Fig. 3 Tracking effect comparison results

图4　复杂场景跟踪结果图

Fig. 4　Complex scene tracking results

4　结论

针对无人机视角下的多目标跟踪，本文提出了一种联合优化目标检测器与跟踪器的多目标跟踪算法，实现了无人机平台下多目标可靠稳定跟踪。本文通过在YOLOv5原始网络中增加小目标检测层进行特征融合解决小目标检测效果差问题，利用Transformer捕获全局信息和上下文信息能力挖掘图像潜在信息，改善航拍图像中因背景复杂干扰、遮挡等导致漏检和错检问题；在跟踪模块，采用ResNet50作为特征提取网络并添加时空注意力模块来更好地提取外观显著性特征，加强网络感知微小外观特征及抗干扰的能力，引入三元组损失函数加强神经网络区分类内差异的能力，整体上增强了跟踪器的鲁棒性。实验结果表明，与原始算法相比，本文算法在无人机平台下不仅能够以较高精度检测目标，而且跟踪精度与准确度都有所提升，抗遮挡能力较强，能够适应复杂环境中的目标跟踪，解决了无人机平台下因小目标难检测、背景复杂、目标相互遮挡等导致的跟踪失败问题，基本满足无人机平台下稳定可靠跟踪条件，具有实际应用价值。

参考文献

[1] GAO Ming, JIN Lisheng, JIANG Yuying, et al. Manifold Siamese network: A novel visual tracking ConvNet for autonomous vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(4): 1612-1623.

[2] YFANTIS E A. A UAV with autonomy, pattern recognition for forest fire prevention, and AI for providing advice to firefighters fighting forest fires[C]//2019 IEEE 9th Annual Computing and Communication Workshop and Conference. Las Vegas, NV, USA. IEEE, 2019: 409-413.

[3] 杨建秀，谢雪梅，石光明，等. 特征信息增强的无人机车辆实时检测算法[J]. 信号处理, 2022, 38(5): 901-914. YANG Jianxiu, XIE Xuemei, SHI Guangming, et al. Real-time UAV vehicle detection based on enhanced feature information [J]. Journal of Signal Processing, 2022, 38(5): 901-914.(in Chinese)

[4] BAE S H, YOON K J. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. IEEE, 2014: 1218-1225.

[5] Al-SHAKARJI N M, BUNYAK F, SEETHARAMAN G, et al. Multi-object tracking cascade with multi-step data association and occlusion handling[C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Auckland, New Zealand. IEEE, 2018: 1-6.

[6] JIN J, LI X, LI X, et al. Online multi-object tracking with Siamese network and optical flow[C]//2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC). Beijing, China. IEEE, 2020: 193-198.

[7] ZHANG Yifu, SUN Peize, JIANG Yi, et al. ByteTrack: multi-object tracking by associating every detection box [EB/OL]. 2021: arXiv: 2110.06864 [cs. CV]. https://

doi.org/10.48550/arXiv.2110.06864.

[8] KIM C, LI Fuxin, CIPTADI A, et al. Multiple hypothesis tracking revisited[C]//2015 IEEE International Conference on Computer Vision. Santiago, Chile. IEEE, 2015: 4696-4704.

[9] BEWLEY A, GE Zongyuan, OTT L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing. Phoenix, AZ, USA. IEEE, 2016: 3464-3468.

[10] WANG Z, ZHENG L, LIU Y, et al. Towards real-time multi-object tracking[C]//European Conference on Computer Vision. Glasgow US: Springer, 2020: 107-122.

[11] ZENG F, DONG B, WANG T, et al. Motr: End-to-end multiple-object tracking with transformer [EB/OL]. 2021: arXiv: 2105.03247 [cs. CV]. https://doi. org/10.48550/arXiv.2105.03247.

[12] CAI J, XU M, LI W, et al. MeMOT: Multi-object tracking with memory[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA. IEEE, 2022: 8090-8100.

[13] ZHOU X, YIN T, KOLTUN V, et al. Global tracking transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA. IEEE, 2022: 8771-8780.

[14] ZHU P, WEN L, DU D, et al. Detection and tracking meet drones challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 1-1.

[15] DU D, QI Y, YU H, et al. The unmanned aerial vehicle benchmark: Object detection and tracking[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 370-386.

[16] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[17] PANG Jiangmiao, CHEN Kai, SHI Jianping, et al. Libra R-CNN: Towards balanced learning for object detection [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 821-830.

[18] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 779-788.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. IEEE, 2017: 936-944.

[20] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 8759-8768.

[21] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. 2020: arXiv: 2010.11929 [cs.CV]. https://arxiv.org/abs/2010.11929.

[22] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric [C]// 2017 IEEE International Conference on Image Processing. Beijing, China. IEEE, 2017: 3645-3649.

[23] BERNARDIN K, STIEFELHAGEN R. Evaluating multiple object tracking performance: The CLEAR MOT metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008: 1-10.

[24] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for UAV tracking[C]//Computer Vision – ECCV 2016, 2016: 445-461.

作者简介

罗 茜 女,1998年生,江西宜春人。华侨大学信息科学与工程学院硕士研究生,主要研究方向为无线通信与目标跟踪。
E-mail: xiluo@stu. hqu. edu. cn

赵 睿 男,1980年生,江苏扬州人。华侨大学信息科学与工程学院副教授,博士,通信工程系主任,IEEE会员,近年主要研究方向为通信信号处理和机器学习。
E-mail: rzhao@hqu. edu. cn

庄慧珊 女,1996年生,福建莆田人。华侨大学信息科学与工程学院硕士研究生,主要研究方向为大数据挖掘。
E-mail: 906931238@qq. com

罗宏刚 男,1996年生,山西晋中人。华侨大学信息科学与工程学院硕士研究生,主要研究方向为时序InSAR城市沉降监测。
E-mail: 1274722836@qq. com