第 60 卷 第 12 期/2023 年 6 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于注意力机制的航拍图像目标检测算法

白宗宝[1]，张俊举[1*]，高原[2]，胡友成[1]

[1]南京理工大学电子工程与光电技术学院，江苏 南京 210094；

[2]南京理工大学紫金学院电子工程与光电技术学院，江苏 南京 210023

摘要 针对现有基于水平视角图像的目标检测网络在无人机航拍图像上误检率和漏检率高的问题，提出一种基于改进注意力机制的航拍图像目标检测算法。首先，在Faster R-CNN主干网络输出端引入一种三叉戟注意力机制，分别提取三路池化层和三路扩张卷积层的多模式信息及多尺度特征信息进行压缩，实现特征通道和空间像素区域的权重再分配。其次，针对航拍图像目标的分类和边界框回归，引入一种双头检测机制，充分利用目标的语义信息和空间位置信息。在相关数据集上对所提算法进行评测，并与其他目标检测算法进行对比。结果表明所提算法的平均精度均值得到显著提升，在不同场景下的无人机航拍图像目标检测中获得更好的效果。

关键词 机器视觉；无人机；目标检测；注意力机制；双头检测机制

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP221025

Attention Mechanism-Based Object Detection Algorithm in Aerial Images

Bai Zongbao[1], Zhang Junju[1*], Gao Yuan[2], Hu Youcheng[1]

[1]*School of Electronic and Optical Engineering, Nanjing University of Science & Technology, Nanjing* 210094, *Jiangsu, China*;

[2]*School of Electronic and Optical Engineering, Nanjing University of Science and Technology ZiJin College, Nanjing* 210023, *Jiangsu, China*

Abstract A target detection algorithm for aerial images based on an improved attention mechanism is suggested to address the issue that the existing object detection network based on horizontal view images has a high false-positive rate and a high miss rate in aerial images. First, a trident channel and spatial attention module that extracts multi-mode and multi-scale characteristic map data of three-branch pooling layers and three-branch dilated convolution layers is added at the output of the Faster R-CNN backbone network so as to compress the data, thereby redistributing the weight of feature channels and spatial pixel regions. Second, a double-head detection mechanism is employed for the classification of the objects and bounding box regression in the aerial image to fully utilize the semantic and spatial location information. The suggested algorithm is further assessed on relevant datasets and contrasted with other object detection algorithms. The results indicate a significant enhancement of the mean average precision of the suggest algorithm, leading to better target detection for unmanned aerial vehicle images in various scenes.

Key words machine vision; unmanned aerial vehicle; object detection; attention mechanism; double-head mechanism

1 引 言

随着时代的发展，无人机凭借高度的机动性和安全性被广泛地运用到军用和民用领域。然而相较于水平视角的普通拍摄图像，无人机航拍图像除了拍摄角度和高度多变外，还具有拍摄视场大、目标占比小、检测细节缺失、易受光照因素影响、背景复杂度高等特点。这些特点导致针对无人机航拍图像的目标检测具有一定的难度。因此，如何设计一个针对航拍图像的目标检测算法是实现无人机自动化作业的重要问题之一。

近年来，基于深度学习的目标检测技术发展迅速，如双阶段检测方法R-CNN[1-4]、单阶段检测方法YOLO[5-7]和SSD[8]等、近年来一些舍弃先验框的新型

收稿日期：2022-03-16；**修回日期**：2022-04-17；**录用日期**：2022-06-13；**网络首发日期**：2022-06-24

基金项目：国家自然科学基金(61971386)

通信作者：*zj_w1231@163.com

检测方法 FCOS[9]和 CornerNet[10]等。这些算法在对自然场景图像进行目标检测时往往可以取得较好的效果，而由于无人机航拍图像的固有特性，将这些神经网络直接用于无人机目标检测时效果往往不理想。针对这一问题，2020 年刘芳等[11]提出了一种基于多尺度特征统合的自适应无人机航拍图像目标检测算法，构建轻量化深度残差网络（LResNet），并对空间尺寸一致的特征图进行加权融合操作，增强了网络的特征表达能力。2021 年，汪权等[12]提出了一种航拍图像绝缘子缺陷识别算法，通过增加网络的输出和改进网络的损失函数输出预测框，准确定位了物体位置。同一时期，许延雷等[13]提出了基于自适应阈值的改进 CenterNet 航拍图像目标检测算法，并通过数据增强手段降低了对航拍图像的误检率和漏检率。2022 年，张官荣等[14]提出了一种面向遥感图像小目标检测的轻量化 YOLOv3 算法，与其他方法相比，在损失 0.24% 精度的前提下，提升了 173% 的检测速度。

现阶段基于深度学习的无人机航拍图像目标检测算法大多采用提取多尺度信息、调整网络感受野的方式增强小目标特征，或是采用预处理手段进行数据增强，很少从图像三维特征图权重分配的合理性角度进行探究，而特征图中目标权重分配的合理性问题往往体现了网络对目标区域识别能力的高低。此外提取多尺度信息和调整网络感受野是为了增强小目标特征信息，其本质在一定程度上也是为了提升网络赋予目标区域的权重，尤其是小目标。更进一步，目前的算法的输出预测接口大多采用单一全连接层进行目标的分类和回归，缺少对航拍目标语义信息和空间位置信息的充分利用。

基于以上分析，本文从更好地提升网络赋予目标区域的权重以及如何充分利用目标语义信息和空间位置信息的角度出发，设计了一种三叉戟注意力模型（Tri-CSAM）以优化特征图的权重分布，同时利用双头检测机制（DH）优化网络预测结构，以期提升神经网络在航拍图像上的目标检测性能。

2　基于注意力机制的航拍图像目标检测算法

2.1　Faster R-CNN 算法简介

Faster R-CNN 是经典的双阶段目标检测算法，网络结构如图 1 所示。网络整体工作流程：首先利用特征提取网络进行下采样，得到特征图；然后将特征图送入区域生成网络（RPN），在 RPN 中对特征图进行前景与背景的二分类和初步边界框回归，同时生成约 2000 个候选区；再将候选区结合特征图送入 ROI pooling 层，经过进一步筛选和池化得到 7×7 大小的建议特征图；最后将建议特征图送入全连接层进行分类和边界框回归预测，并进行后处理。

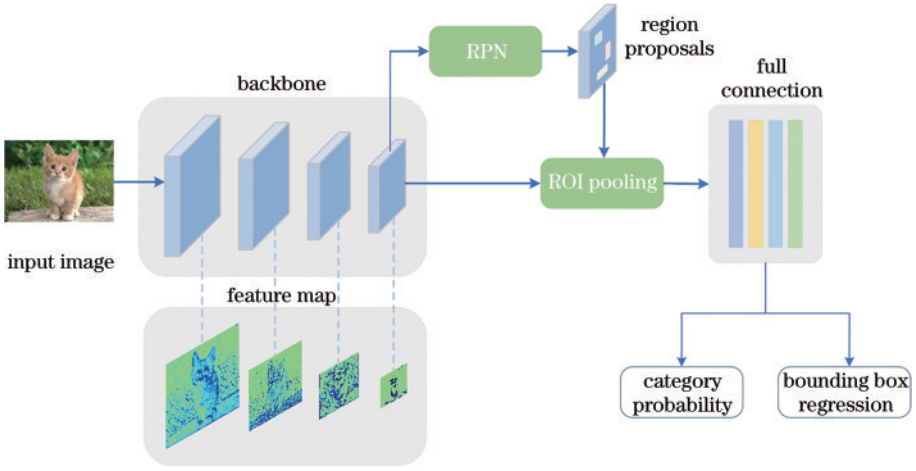

图 1　Faster R-CNN 的结构

Fig. 1　Structure of Faster R-CNN

2.2　三叉戟注意力机制

为了解决无人机航拍图像中目标占比小、拍摄角度和高度多变、复杂度高而引起的小目标误检率和漏检率高的问题，受计算机视觉中的注意力机制启发，本文设计了一种新型的结合通道注意力机制和空间注意力机制的三叉戟注意力模块，其组件主要包括通道注意力模块和空间像素注意力模块，图 2 为对应模块的结构图。

2.2.1　通道注意力模块

图像经过卷积网络的特征提取层后会得到多个通道特征图，这些特征通道往往对应着不同的特征响应，而普通卷积对所有特征通道的响应是相同的，即不同特征通道享有同等权重，这与实际任务需求是不相符的。图 3 展示了深度残差网络（ResNet-50）[15]不同网络节点的输出通道特征图，可以看出，不同通道对目标的特征响应具有差异性。本文基于 squeeze-and-excitation network（SENet）[16]，在全局平均池化的基础

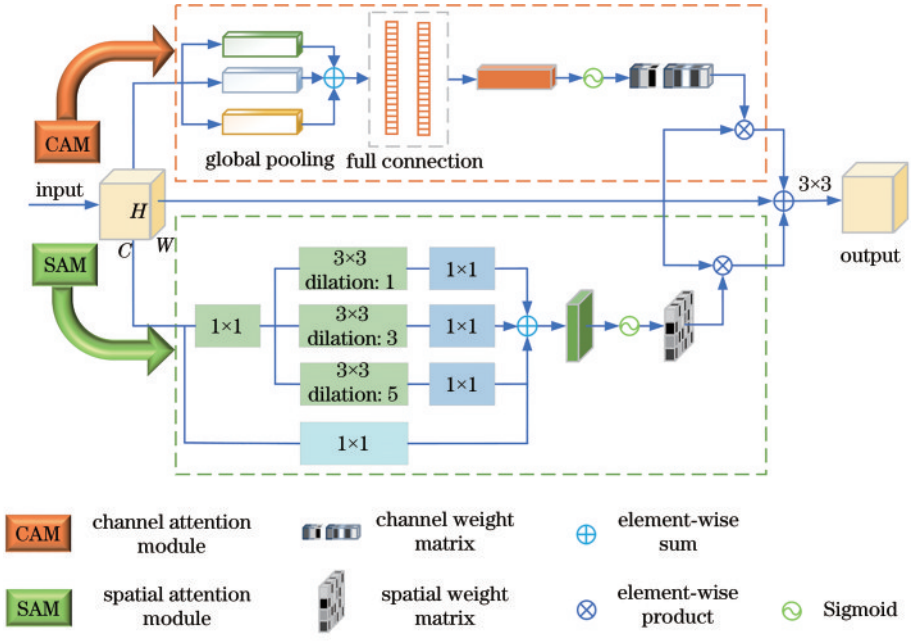

图 2　三叉戟注意力模块结构图

Fig. 2　Structure diagram of Tri-CSAM

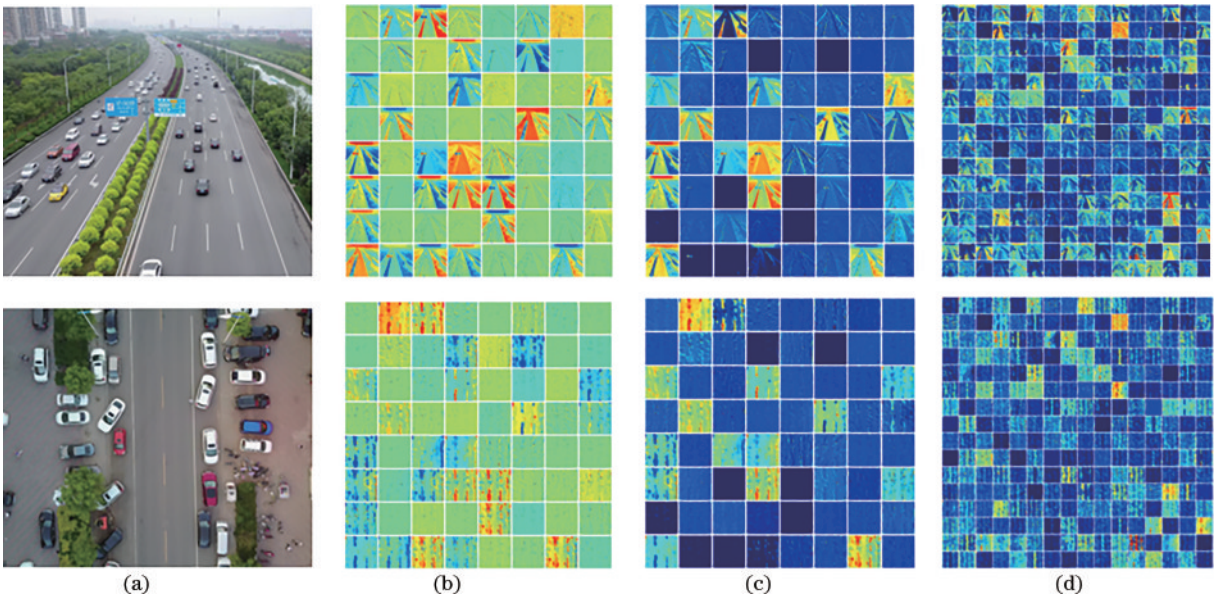

图 3　ResNet-50 不同节点通道特征图。(a)原图;(b)Conv1 卷积层;(c)Conv1 激活层;(d)Conv2_x 输出层

Fig. 3　Channel feature maps of ResNet-50 in different nodes. (a) Original image; (b) convolution layer of Conv1; (c) rectified linear unit of Conv1; (d) output layer of Conv2_x

上,引入全局最大池化和全局随机池化两种池化模式进行信息融合,提取输入特征图的不同模式信息,增强目标的特征传递性,减少特征信息损失,同时由于采取的是池化操作,因此参数量相对 SENet 并未发生变化,结构如图 2 上方矩形虚线框所示。

设输入特征张量 $\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$,其并行经过三种模式的全局池化,即全局平均池化(GAP)、全局最大池化(GMP)和全局随机池化(GSP),由此得到三个 $C\times 1\times 1$ 维度的通道特征量,该通道特征量初步表征了不同通道间的重要程度。三种池化过程可分别描述为

$$f_{\mathrm{GAP}}(\boldsymbol{X})=\sum_{k=1}^{C}\left[\frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}X(i,j)\right],\tag{1}$$

$$f_{\mathrm{GMP}}(\boldsymbol{X})=\sum_{k=1}^{C}\max\left[X(i,j)\right],i=1,\cdots,W,j=1,\cdots,H,\tag{2}$$

$$P_k(i,j)=\frac{X_k(i,j)}{\sum_{i=1}^{W}\sum_{j=1}^{H}X_k(i,j)},k=1,\cdots,C,$$

$$S_k=X_k\ \ \text{where}\ k\sim P(p_1,\cdots,p_C),\tag{3}$$

式中:f_{GAP} 和 f_{GMP} 分别表示全局平均池化函数和全局最

大池化函数；$X_k(i,j)$表示第k个通道坐标为(i,j)的像素值；$P_k(i,j)$表示第k个通道坐标为(i,j)处的像素对应概率，$\boldsymbol{P}_k$表示k通道对应的概率矩阵；$S_k=X_k$表示由对应概率矩阵所得到的通道特征量。

对生成的三个通道特征量进行加和操作，得到一个$C\times1\times1$维度的综合通道特征量，其融合了更多的模式信息，表达的各通道权重更为全面，同时避免网络输出过分依赖于某一特定特征，提升网络整体的鲁棒性。上述过程可描述为

$$\boldsymbol{F}_{\mathrm{GP}}=f_{\mathrm{GAP}}(\boldsymbol{X})+f_{\mathrm{GMP}}(\boldsymbol{X})+f_{\mathrm{GSP}}(\boldsymbol{X}),\tag{4}$$

式中：f_{GSP}表示全局随机池化函数。

SE网络会对池化后的通道特征量进行先降维再升维度的操作，以增加网络的非线性。然而这种先降维再升维的操作会损失更多的信息，不利于小目标的特征信息传递。因此，本文采用不降维操作。全局通道特征量直接通过全连接层，过程中特征量维度不变；接着使用Sigmoid函数进行非线性激活，得到特征图的通道权重系数；最后将输入特征图与通道权重系数相乘并与输入特征图进行跳跃连接，得到具有通道重要性差异的全新特征图。此过程可描述为

$$\boldsymbol{F}_{\mathrm{CAM}}=\sigma\left[\boldsymbol{W}_1(\boldsymbol{W}_0\cdot\boldsymbol{F}_{\mathrm{GP}})\right]\cdot\boldsymbol{X}+\boldsymbol{X}=\boldsymbol{X}\left\{1+\sigma\left[\boldsymbol{W}_1(\boldsymbol{W}_0\cdot\boldsymbol{F}_{\mathrm{GP}})\right]\right\},\tag{5}$$

式中：$\boldsymbol{W}_0$和$\boldsymbol{W}_1$表示全连接层权重系数矩阵；σ表示Sigmoid激活函数；$\boldsymbol{F}_{\mathrm{CAM}}$表示通道注意力网络输出特征图。

2.2.2　空间注意力模块

显著性目标一般存在于图像某个区域，该区域像素点对目标的显著值预测贡献大于其他区域像素点。对于航拍图像，其拍摄视场通常很大，而目标往往在图像中占比相对较小，这也导致了无人机航拍图像检测具有一定的挑战性。为了抑制图像中的复杂背景和噪声信息，增强网络对目标的关注度，引入多尺度空间注意力机制，其结构如图2下方矩形虚线框所示。

首先，输入特征图经过逐点卷积进行降维处理，生成三路维度降半的特征图和一路单一维度的特征图，其中维度降半的三路分支所采用的共享参数和逐点卷积的结构可以在不影响网络性能的前提下降低网络参数；为了在不减小特征图分辨率的情况下增大网络感受野，提取多尺度信息，增强网络对不同尺寸目标的敏感性，分别引入扩张率为1、3、5，卷积核大小为3×3的扩张卷积，组成三路并联的网络结构。相对于输入特征图，其真实感受野分别是1×1、3×3、7×7和11×11，图4为该结构各分支感受野示意图。

接着对三路维度降半的输出特征图采取加和操作，然后输入到Sigmoid激活函数得到最终的空间注意力权重矩阵。最后将权重矩阵与输入特征图相乘即可得到附带空间像素权重关系的特征图。设输入特征张量$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$，则上述整个过程可描述为

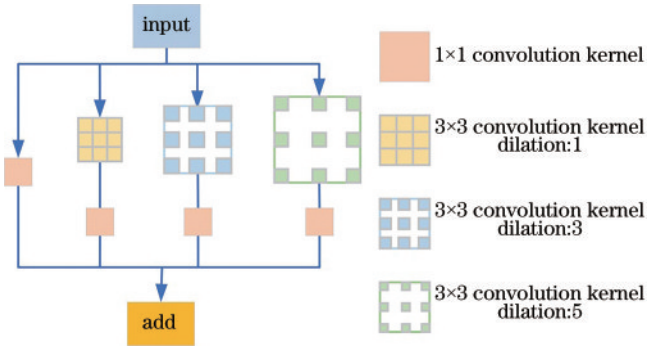

图4　多尺度感受野示意图

Fig. 4　Schematic of multi-scale receptive field

$$\boldsymbol{F}_{\mathrm{Conv1}}=f_{\mathrm{Conv1}}(\boldsymbol{X}),\ \boldsymbol{F}_{\mathrm{Conv2}}=f_{\mathrm{Conv2}}(\boldsymbol{X}),\tag{6}$$

$$\boldsymbol{F}_{\mathrm{SAM}}=\sigma\left\{\boldsymbol{F}_{\mathrm{Conv2}}+\sum_{k=0}^{2}f_{\mathrm{Conv3}}\left[f_{\mathrm{dia}}^{2k+1}(\boldsymbol{F}_{\mathrm{Conv1}})\right]\right\},\tag{7}$$

式中：f_{Conv1}、f_{Conv2}、f_{Conv3}表示不同参数的逐点卷积；$\boldsymbol{F}_{\mathrm{Conv1}}$和$\boldsymbol{F}_{\mathrm{Conv2}}$表示经过逐点卷积压缩后的特征图，其中$\boldsymbol{F}_{\mathrm{Conv1}}\in\mathbf{R}^{\frac{C}{2}\times H\times W}$，$\boldsymbol{F}_{\mathrm{Conv2}}\in\mathbf{R}^{1\times H\times W}$；$f_{\mathrm{dia}}^{2k+1}$表示扩张率为$2k+1$的扩张卷积；$\boldsymbol{F}_{\mathrm{SAM}}$表示空间注意力网络的输出特征图。

2.3　双头检测机制

目标检测的本质是对目标进行分类和边界框回归。文献[17]针对这一特性进行了深入研究，结果表明被广泛用于目标检测网络中的卷积层和全连接层对目标的分类和回归具有不同的敏感性，卷积层往往对目标的位置信息更为敏感，全连接层则更关注目标的类别信息。而以往针对无人机航拍图像的目标检测网络都是将分类和回归耦合到同一个网络分支进行预测的，无法充分利用目标的语义信息和位置信息。因此，基于该研究，本文引入一种双头检测机制(DH)，对目标类别信息和位置回归信息进行解耦分离，并进行独立预测，整体结构如图5所示。

block 1和block 2为多个卷积结构组成的卷积块，结构[17]如图6所示。block 1结构为一个简单的残差结构。block 2模块分为两部分：第一部分是non-local neural networks结构[18]，目的是增强图像前景，联系全局信息，增强长距离区域间联系，提升回归预测效果；第二部分为残差结构。特征图经过block 1模块后图像深度由256提升至1024，将block 1的输出作为block 2的第一次输入，特征图连续经过两次block 2模块后维度保持不变，最后对该模块输出特征图进行全局最大池化，即得到类别概率和边界框回归参数。

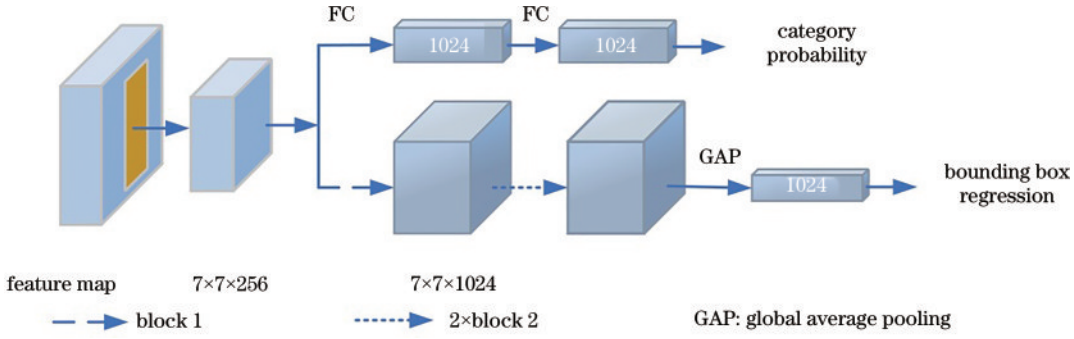

图 5　双头检测机制结构

Fig. 5　Structure of DH

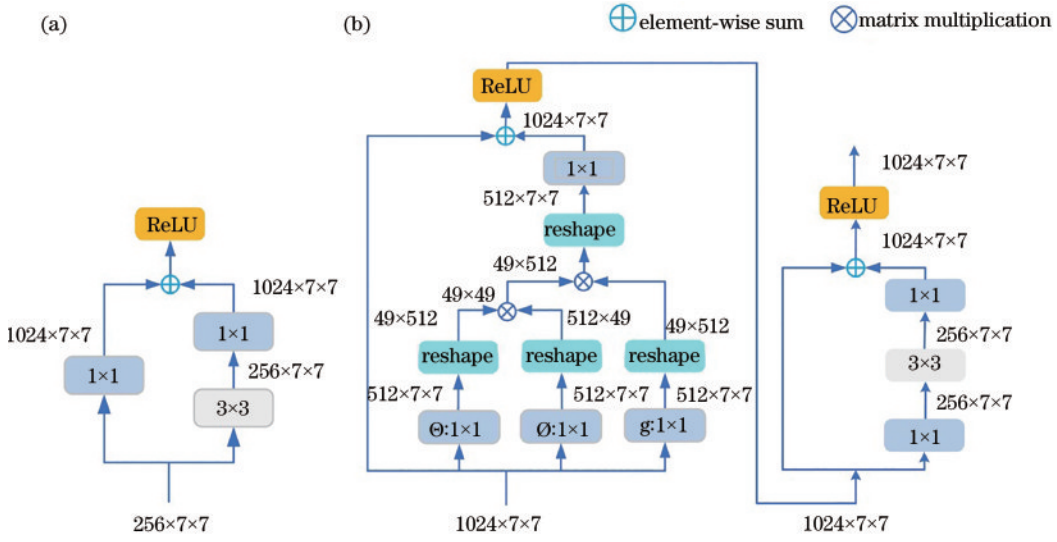

图 6　回归分支卷积块结构。(a)block 1结构；(b)block 2结构

Fig. 6　Structure of regression branch convolution module. (a) Structure of block 1; (b) structure of block 2

3　实验与分析

3.1　数据集处理与分析

为验证所提模型的效果，选取公开无人机航拍数据集 VisDrone[19]和 UAVDT[20]作为实验数据集。公开无人机航拍数据集 VisDrone 由天津大学 AISKYEYE 团队收集，数据集包含 10209 张图像，共 10 个预定义类别，分别为行人、人、汽车、货车、公共汽车、卡车、摩托车、自行车、三轮车和带棚三轮车。从 VisDrone 原始数据集中抽取 7000 张作为所提模型训练及验证的新数据集，其中训练集划分 5140 张，验证集划分 683 张，测试集划分 1287 张。UAVDT 目标检测数据集是从 10 个小时的航拍视频(约 80000 帧的代表帧)中提取并标注而来的，包含了 40735 张不同时间段、不同天气状况、不同拍摄高度及角度、不同车辆类别及不同遮挡情况的图像，主要关注对象是汽车、卡车及公共汽车。从该数据集中抽取 7000 张图片构成所提模型的数据集，其中训练集、验证集及测试集的划分比例与 VisDrone 数据集相同。

图 7 为两个数据集部分标签可视化结果，由图可以看出：航拍图像的拍摄视场通常较大，但目标占比小，部分目标还存在被遮挡的现象，比如建筑物或树木遮挡车辆和行人的情况；此外目标的背景光照环境较为复杂，存在强光或微光环境干扰的情况，这对目标检测网络的性能提出了更高的要求。总体而言，所选数据集图像背景复杂度较高，目标占比小，具有一定的检测难度，贴合实际应用情况。

3.2　算法实现细节

实验平台采用 i7-8700K 处理器，两张显存为 12G 的 NVIDIA TITAN XP 显卡，Ubuntu16.04 操作系统，PyTorch 深度学习框架。实验训练阶段训练轮次(epoch)设为 24，每次迭代批量大小(batch size)设为 2；初始学习率大小设置为 0.005，学习率衰减系数设置为 0.1，衰减轮次为第 16 和 22 轮；学习率预热策略选择“linear-warmup”，预热迭代次数为 500，预热起始学习率为 0.001；梯度下降方法选择“SGD”随机梯度下降法，权重衰减系数为 0.0001，动量为 0.9。

3.3　客观量化对比分析

为验证所提模型的有效性，将 Faster R-CNN (ResNet-50+FPN)[21]作为基线(baseline)模型进行实

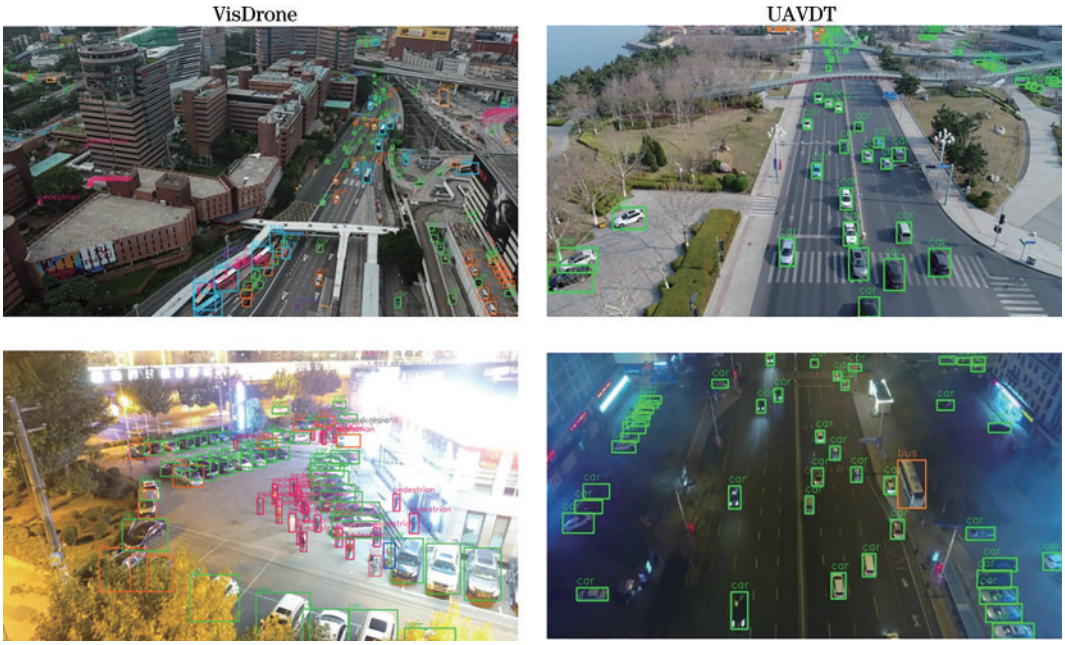

图 7　部分数据集标签可视化

Fig. 7　Label visualization of partial dataset

验，选用平均精度均值（mAP）、AP^{50}、AP^{75}、AP_S、AP_M和AP_L对模型进行评估。其中，mAP表示以0.05为步长，0.50到0.95之间的10个交并比（IoU）阈值下的平均精度均值，AP^{50}和AP^{75}分别表示IoU阈值为0.50和0.75时的精度均值，AP_S、AP_M和AP_L分别表示小、中、大尺寸目标的精度均值。

表1为所提模型在VisDrone数据集上的实验结果。从表1可知：对于G1模型，即仅添加Tri-CSAM模块的模型，相较基线模型，整体mAP提升1.07个百分点，并且在小目标上的AP值提升了1.38个百分点，这表明在注意力机制的作用下，网络的特征表达能力得到了增强，目标区域像素权重占比得到了有效提升，网络检测性能因此得到了提升；对于G2模型，即仅添加DH模块的模型，相较基线模型，mAP提升1.13个百分点，这表明相较以往的单分支预测目标分类和回归，双分支预测是更为有效的方法；相较于基线网络，所提模型的mAP提升了2.41个百分点。

表1　VisDrone数据集上的检测结果

Table 1　Detection results on VisDrone dataset　　unit: %

Model	Tri-CSAM	DH	AP^{50}	AP^{75}	AP_S	AP_M	AP_L	mAP
Baseline			40.12	23.52	9.23	35.66	44.42	23.22
G1	√		41.16	24.81	10.61	36.48	44.89	24.29
G2		√	42.03	25.27	10.33	36.71	45.25	24.35
Proposed model	√	√	42.91	25.74	11.25	36.82	44.97	25.63

为进一步验证所提模型的有效性和泛化能力，继续将所提模型在UAVDT数据集上进行实验，表2为对应实验结果。从表2可以得出：在分别添加了Tri-CSAM和DH模块后，mAP值分别获得了3.11个百分点和2.37个百分点的提升，这表明所提模块的有效性；在同时添加两种模块后，网络整体性能得到了更高的增益，mAP提升了4.51个百分点，这表明两种模块的结合对网络整体性能提升有很大的帮助。此外，在更换数据集进行实验时，所提模型依旧获得较好的性能，说明所提模型具有一定的普适性和泛化能力。总体而言，以上实验及分析结果表明：利用注意力机制对网络特征图进行权重重塑是一种提高网络检测性能的有效手段，利用通道与空间区域像素的权重重构可以实现网络聚焦于目标区域的目的，达到更优的检测效果；同时双头检测模型充分发挥了不同网络结构对目标语义信息和空间位置信息具有不同敏感程度的特点，实现了对目标的精准分类和回归；当模型同时使用这两种模块时可以达到更好的效果。

为了进一步验证所提模型的有效性，采用准确率-召回率（P-R）曲线对模型进行评估。图8为不同模型在不同IoU阈值下对行人类别的P-R曲线图，曲线越接近右上方说明模型效果越好。可以看出：所提模型在性能表现上优于基线模型，同时也优于G1和G2模型，表明所提模型在保持相同召回率的同时可以达到

表 2　UAVDT 数据集上的检测结果

Table 2　Detection results on UAVDT dataset　　unit: %

Model	Tri-CSAM	DH	AP^{50}	AP^{75}	AP_S	AP_M	AP_L	mAP
Baseline			42.31	26.61	10.08	35.78	48.69	24.65
G1	√		45.87	29.79	12.58	40.36	50.93	27.76
G2		√	44.91	29.26	11.98	39.88	51.36	27.02
Proposed model	√	√	46.39	30.02	12.46	42.26	51.73	29.16

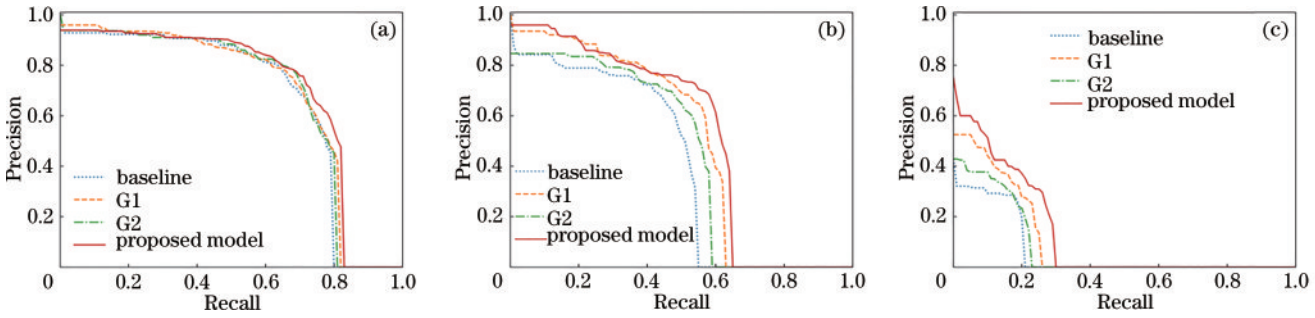

图 8　不同 IoU 阈值下的 P-R 曲线。(a)IoU 为 0.50;(b)IoU 为 0.75;(c)IoU 为 0.90

Fig. 8　P-R curves under different IoU thresholds. (a) IoU is 0.50; (b) IoU is 0.75; (c) IoU is 0.90

更高的精准度;并且随着 IoU 阈值的提升,所提模型的检测性能也呈上升趋势,这说明在三叉戟注意力机制和双头检测模型的共同作用下,所提模型在更高精度要求下可以达到更好的效果。

3.4　主观视觉对比分析

为验证所提模型的有效性,进一步从主观视觉上对模型检测效果进行分析。如图 9 所示,选取三张不同拍摄场景及拍摄角度的图像进行网络热力图分析。具体而言,图 9(a)和图 9(b)输入图像的拍摄角度都接近 45°,但拍摄场景差异性很大;图 9(b)和图 9(c)具有相似的拍摄场景,但拍摄角度差异较大,其中图 9(c)的拍摄角度接近 90°。通过实验对比可以发现:基线模型基本上可以聚焦于目标区域,但同时也会被复杂背景环境干扰,对背景的关注偏多,比如基线模型在图 9(b)上更多地聚焦于左上方的住宅区;G1 模型对背景环境的关注减少了,更多地聚焦在目标点区域附近,提升了网络对目标的特征信息捕捉能力及目标区域权重;G2 模型聚焦的范围变小,即对目标的关注不再是一个模糊的区域范围,而是缩小到目标本身,完成了更精准的分类,但背景环境对网络的影响仍然存在,只是相较基准网络已经有所减弱;所提模型在注意力机制和双头检测机制的共同作用下可以更好地克服不同场景、不同目标尺寸、不同拍摄角度所带来的影响,聚焦于目标个体,提取更多、更准确的目标特征信息,抑制复杂的背景信息,提高目标权重,实现更好的目标检测效果,提升了鲁棒性。

为进一步验证所提模型的有效性,选取三张不同光照条件和拍摄角度的无人机航拍图像进行测试,并对模型检测结果进行可视化分析。图 10 为不同模型在白天、傍晚和夜间航拍图上的检测结果可视化图。通过观察可以发现:对于图 10(a),基线模型存在将区域 2 中的公共汽车误检为卡车的现象,同时对区域 1 中的行人和区域 3 中的车辆存在漏检现象;对于图 10(b),基线模型对区域 1~4 中的密集小目标行人以及区域 5 中的车辆小目标存在漏检现象;在图 10(c),基线模型对区域 1 和区域 2 的行人目标存在漏检现象。在分别添加 Tri-CSAM 和 DH 模块后,检测结果有一定提升,漏检率和误检率有一定程度下降,比如在图 10(a)的区域 2 中,G1 模型成功检测到公共汽车目标,未发生误检现象;在图 10(a)区域 1 中,G2 模型检测到小目标行人,降低了漏检率;但这些网络仍然存在部分小目标漏检现象,比如图 10(b)区域 1~4 中的行人小目标等。通过以上对比分析可以发现,在注意力模块的作用下,网络在复杂光照环境下的检测性能有了一定程度的提升,这些提升主要依赖于模块对网络焦点的合理重塑。这种焦点的重塑使网络对背景环境中的嘈杂信息有一定程度的抑制作用,对复杂光照环境下的目标有了更强的聚焦能力。与此同时可以发现,在引进双头检测机制后,网络在不同光照环境下的鲁棒性也有所增强,这些提升主要是从卷积网络和全连接网络的差异性方面获得的,具体来说,卷积网络更专注于对目标的位置检测,而全连接网络则更擅长对目标的分类。总体而言,所提注意力机制和双头检测机制都可以在一定程度上提升网络在不同光照环境下的检测效果,同时利用这两种机制,网络整体性能可以获得更高的增益,在不同光照环境下,网络具有更好的鲁棒性。

3.5　主流无人机目标检测算法性能对比

为验证所提模型的优势,对所提模型与其他无人机航拍图像目标检测模型进行对比。对比模型包括 Mask R-CNN、Cascade R-CNN[22]、YOLOv3、DPN[23]、

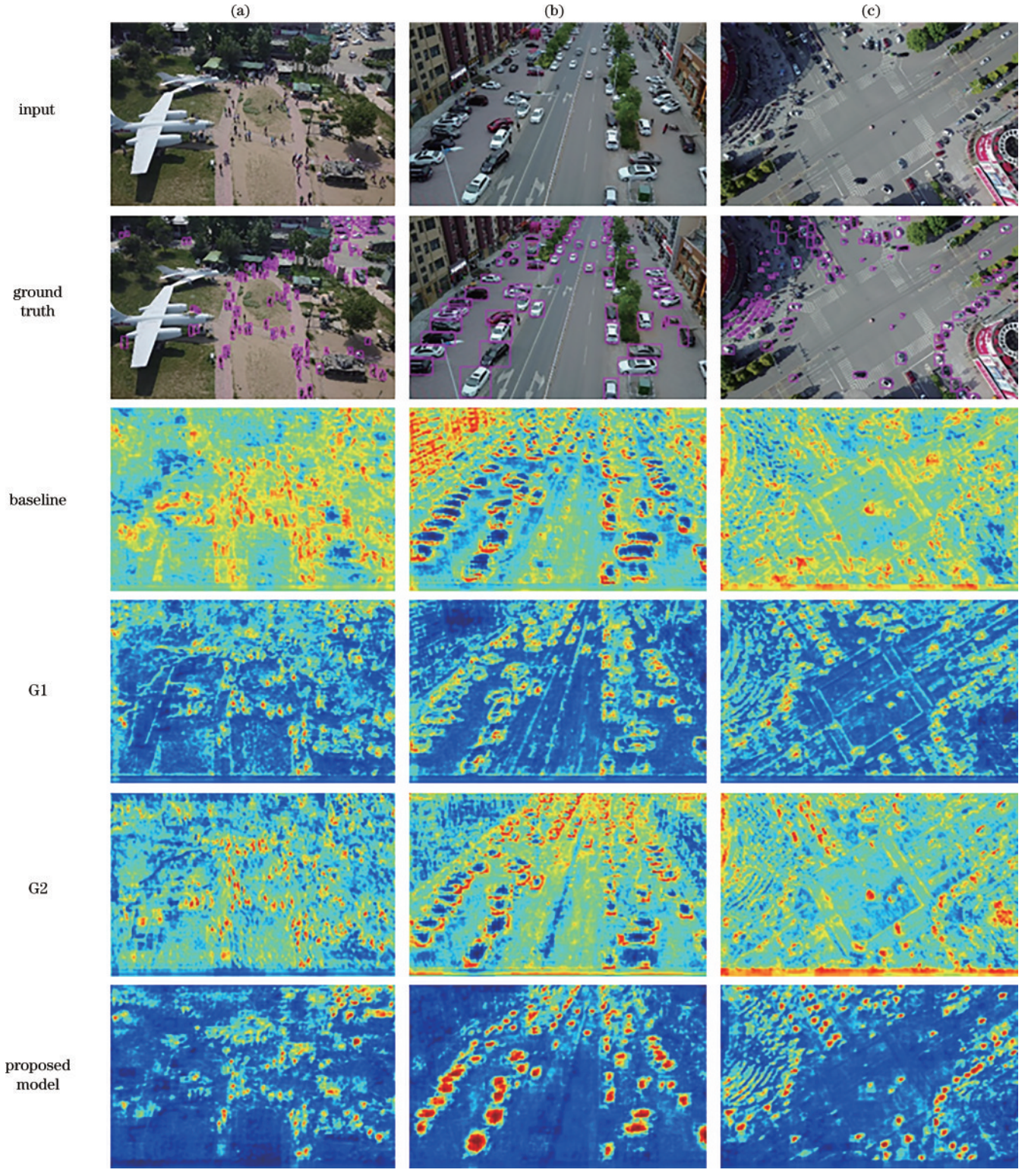

图 9 网络热力图可视化结果。(a)45°航拍视角小目标行人场景;(b)45°航拍视角中等目标车辆场景;(c)90°航拍视角小目标车辆场景

Fig. 9 Visualization results of network heat map. (a) Small target pedestrian scene of 45° aerial view; (b) medium target vehicles scene of 45° aerial view; (c) small target vehicles scene of 90° aerial view

TridentNet[24]、CornerNet 和 LResNet,采用 mAP、AP^{50}、AP^{75}和帧率进行综合评估,表 3 为对比结果。从检测精度来看:在 AP^{50}指标上,所提模型低于第一名 DPN 的 50.01% 和第二名 TridentNet 的 43.29%,居第三位;但 mAP 达 25.63%,比对比模型分别提高了 2.41 个百分点、1.47 个百分点、9.54 个百分点、5.80 个百分点、0.54 个百分点、3.12 个百分点、8.22 个百分点和 3.31 个百分点;同时 AP^{75}指标也优于其他模型,这说明随着检测精度的提升,所提模型的检测精度呈现上升趋势,并且 mAP 表现得更好。所提模型在引入新

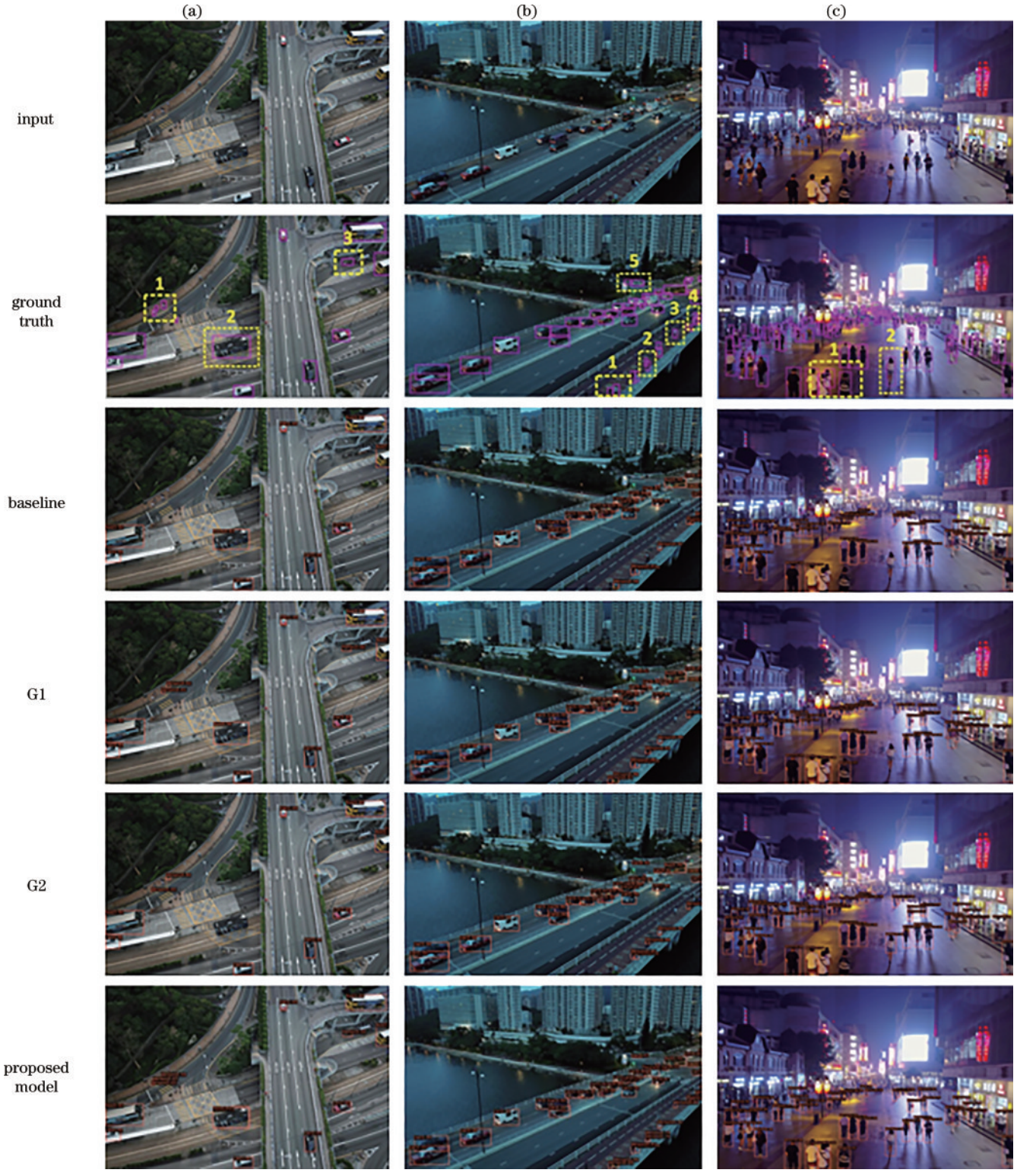

图 10　航拍图像检测结果可视化。(a)正常光照环境;(b)弱光环境;(c)强光环境

Fig. 10　Visualization results of aerial image detection. (a) Normal light scene; (b) week light scene; (c) strong light scene

型注意力机制后,有效提升了目标区域像素所占权重,增加了重要通道贡献,优化了网络模型聚焦点,提升了目标检测效果。与此同时,由于双头检测机制的引入,网络在目标分类和回归任务上的性能也得到了一定程度的提升。在模型运算速度上,所提模型虽未达到YOLOv3水平,但在池化层、逐点卷积和参数共享方法的作用下,和Faster R-CNN模型基本持平。这表明,所提模型在不显著增加计算量的前提下,可以一定程度上增强网络性能,提升了航拍图像目标检测精度。

表 3 不同算法的性能对比

Table 3 Performance comparison of different algorithms

Model	mAP / %	AP^{50} / %	AP^{75} / %	Speed / (frame·s^{-1})
Faster R-CNN	23.22	40.12	23.52	19
Mask R-CNN	24.16	41.16	23.87	11
Cascade R-CNN	16.09	31.91	15.01	13
YOLOv3	19.83	38.23	20.63	30
DPN	25.09	50.01	21.83	
TridentNet	22.51	43.29	20.50	
CornerNet	17.41	34.12	15.78	16
LResNet	22.32	39.63	23.17	
Proposed model	25.63	42.91	25.74	17

4 结 论

针对无人机航拍图像目标检测误检率和漏检率高的问题，提出了一种基于改进注意力机制的航拍图像目标检测算法。一方面，利用不同池化层提取多模式信息，优化目标通道权重，提升网络对目标通道响应差异的区分能力；另一方面，利用扩张卷积提取多尺度信息，提升目标区域像素权重。与此同时，引入双头检测机制，实现目标分类和位置回归的独立输出。实验结果表明，与其他算法相比，所提模型的mAP得到了很大的提升。所提模型可以有效提升特征图中的目标权重，充分利用目标语义信息和空间位置信息，改善无人机航拍图像因拍摄角度、高度、目标占比及光照等因素所导致的目标误检率及漏检率高的问题。虽然在三叉戟注意力网络中使用了池化、参数共享以及逐点卷积等操作，但多尺度的信息提取仍然增加了一定的参数量，下一步的研究重点主要围绕如何在不降低检测速度的基础上提升检测精度的内容。

参 考 文 献

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[2] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[3] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems, December 7-12, 2015, Montreal, Quebec, Canada. New York: Curran Associates, 2015: 91-99.

[4] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

[5] 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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[6] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[7] Redmon J, Farhadi A. Yolov3: an incremental improvement[EB/OL]. (2018-04-08)[2022-02-04]. https://arxiv.org/abs/1804.02767.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[9] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9626-9635.

[10] Law H, Deng J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(3): 642-656.

[11] 刘芳, 吴志威, 杨安喆, 等. 基于多尺度特征融合的自适应无人机目标检测[J]. 光学学报, 2020, 40(10): 1015002.
Liu F, Wu Z W, Yang A Z, et al. Multi-scale feature fusion based adaptive object detection for UAV[J]. Acta Optica Sinica, 2020, 40(10): 1015002.

[12] 汪权, 易本顺. 基于Gaussian YOLOv3的航拍图像绝缘子缺陷识别[J]. 激光与光电子学进展, 2021, 58(12): 1210022.
Wang Q, Yi B S. Insulator defect recognition in aerial images based on Gaussian YOLOv3[J]. Laser & Optoelectronics Progress, 2021, 58(12): 1210022.

[13] 许延雷, 梁继然, 董国军, 等. 基于改进CenterNet的航拍图像目标检测算法[J]. 激光与光电子学进展, 2021, 58(20): 2010013.
Xu Y L, Liang J R, Dong G J, et al. Aerial image target detection algorithm based on improved CenterNet[J]. Laser & Optoelectronics Progress, 2021, 58(20): 2010013.

[14] 张官荣, 陈相, 赵玉, 等. 面向小目标检测的轻量化YOLOv3算法[J]. 光学学报, 2022, 59(16): 1610008.
Zhang G R, Chen X, Zhao Y, et al. Lightweight YOLOv3 algorithm for small object detection[J]. Acta Optica Sinica, 2022, 59(16): 1610008.

[15] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[16] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-

7141.

[17] Wu Y, Chen Y P, Yuan L, et al. Rethinking classification and localization for object detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10183-10192.

[18] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[19] Zhu P F, Du D W, Wen L Y, et al. VisDrone-VID2019: the vision meets drone object detection in video challenge results[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 227-235.

[20] Du D W, Qi Y K, Yu H Y, et al. The unmanned aerial vehicle benchmark: object detection and tracking[EB/OL]. (2018-03-26)[2022-04-02]. https://arxiv.org/abs/1804.00518.

[21] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[22] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 6154-6162.

[23] Chen Y P, Li J N, Xiao H X, et al. Dual path networks [EB/OL]. (2017-07-06)[2021-02-04]. https://arxiv.org/abs/1707.01629.

[24] Li Y H, Chen Y T, Wang N Y, et al. Scale-aware trident networks for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6053-6062.