第 58 卷 第 20 期/2021 年 10 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

基于改进 CenterNet 的航拍图像目标检测算法

许延雷[1]，梁继然[1,2*]，董国军[3]，陈壮[1]

[1]天津大学微电子学院，天津 300072；

[2]天津市成像与感知微电子技术重点实验室，天津 300072；

[3]天津七一二通信广播股份有限公司，天津 300457

摘要 为提高航拍图像目标检测精度以及检测速度，提出了基于自适应阈值的改进 CenterNet 航拍图像目标检测算法。以目标的中心点作为关键点代替锚框进行分类和边界回归，设计自适应阈值预测分支对预处理结果进行筛选优化。同时设计了编码-解码结构的主干网络，通过可变形空洞卷积结构以及基于注意力机制的卷积连接结构，将浅层空间信息以及深层语义信息进行有效提取以及特征融合，提升了输出特征图质量。并通过结构化信息丢弃和利用误检、漏检目标构建新样本的方法实现数据增强，降低误检率及漏检率。在公开数据集 NWPU VHR-10 上进行实验，结果表明，与基于 ResNet-50 的 CenterNet 相比，本文算法的平均精度均值提升 5.17%，交并比为 0.50 和 0.75 的平均精度分别提升了 3.57%和 3.61%，检测速度达 45 frame·s^{-1}，取得了良好的检测精度和实时性的平衡。

关键词 图像处理；目标检测；卷积神经网络；自适应阈值；航拍图像

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2010013

Aerial Image Target Detection Algorithm Based on Improved CenterNet

Xu Yanlei[1], Liang Jiran[1,2*], Dong Guojun[3], Chen Zhuang[1]

[1] *School of Microelectronics, Tianjin University, Tianjin* 300072, *China*;

[2] *Tianjin Key Laboratory of Imaging and Sensing Microelectronic Technology, Tianjin* 300072, *China*;

[3] *Tianjin* 712 *Communication & Broadcasting Shareholding Co., Ltd., Tianjin* 300457, *China*

Abstract In order to improve the accuracy and speed of aerial image target detection, an improved CenterNet aerial image target detection algorithm based on adaptive threshold is proposed. The center point of the target is used as the key point to replace the anchor box for classification and boundary regression, and an adaptive threshold prediction branch is designed to screen and optimize the preprocessing results. At the same time, the encoding-decoding network structure is designed. Through the deformable cavity convolution structure and the convolutional block attention-connection structure based on the attention mechanism, shallow spatial information, and deep semantic information are effectively extracted and fused. In addition, data enhancement is realized by discarding structured information and building new samples with false and missing detection targets, so as to reduce false and missing detection rates. Experiments are performed on the open data set NWPU VHR-10, the results show that compared with CenterNet based on ResNet-50, mean average precision of proposed algorithm increased by 5.17%, and intersection of union of 0.50 and 0.75 are improved by 3.57% and 3.61%, respectively. The detection speed reaches 45 frame·s^{-1}, achieving good detection accuracy and real-time balance.

Key words image processing; target detection; convolutional neural network; adaptive threshold; aerial image

OCIS codes 100.3008; 100.4996; 100.2000

收稿日期：2020-12-02；**修回日期**：2020-12-29；**录用日期**：2021-01-06

基金项目：天津市科技重大专项与工程计划项目(19ZXZNGX00060)

通信作者：*liang_jiran@tju.edu.cn

1 引 言

无人驾驶飞机(无人机)[1],凭借高度的灵活性和机动性,被广泛应用于军事以及民用领域。利用无人机高空航拍进行目标检测,为军事侦察、车流监控、电路巡检等领域带来了极大便利。然而航拍图像目标检测存在背景复杂度高、小目标物体占比丰富以及不同图像中目标尺度变化剧烈等特点,导致目前应用在航拍图像上的目标检测算法检测精度较差。同时,无人机作为嵌入式移动设备,计算能力相对较弱,无法处理复杂算法却有高实时性的需求。因此如何更好地提升航拍图像目标检测精度同时保证高实时性成为了亟待解决的问题。

近年来基于卷积神经网络的目标检测算法被广泛应用于航拍图像目标检测[2-4],主要分为以基于区域的卷积神经网络(R-CNN)[5]为代表的双阶段检测算法[6]和以 YOLO[7]为代表的单阶段检测算法[8]。2016 年,Cheng 等[9]提出了基于区域以及旋转不变层的卷积神经网络(RICNN)模型,在现有 CNN 体系结构的基础上引入并学习新的旋转不变层来提高航拍小目标的检测能力,但检测速度较慢。2017 年 Deng 等[10]基于 Faster R-CNN 算法进行航拍检测同时提升了检测精度和速度,但依旧无法达到实时性需求。2019 年 Hu 等[11]使用改进的 YOLOv3 算法进行航拍图像检测,检测速度明显提升,但针对小目标的检测精度却远低于 Faster R-CNN。这些算法都基于 Anchor-Based 的思想,需要罗列大量的锚框作为待检测的候选区域,再基于这些区域进行分类和边界回归。锚框的生成需要人工设定大量的参数,如尺寸、宽高比、数量等,并进行额外后续处理如非极大值抑制[12]。这导致模型复杂度较高,计算量较大,同时小目标物体无法很好匹配锚框大小,进而发生漏检和误检,大量的锚框区域为负样本区域,这也造成了正负样本的不平衡。

当前提出了一种全新的检测思路,即 Anchor-Free[13-14],通过预测像素点和真实标签之间的置信度进行分类回归。2018 年 Law 等[15]提出了 CornerNet,以目标边界的左上角点和右下角点为关键点来预测边界框,不再需要锚框,检测精度和速度都有一定提高,但是由于小目标物体嵌入向量的误差问题,CornerNet 在航拍图像目标检测上精度提升并不明显。2019 年 Zhou 等[16]提出了 CenterNet,直接利用目标物体中心点的特征信息进行目标分类和边界回归,在 COCO 数据集中取得了较好的检测精度和实时性的平衡,但应用于航拍图像中,小目标物体的中心点信息和空间轮廓信息随着分辨率的降低而大量丢失,无法进行有效地后处理,检测精度仍然不足。

本文针对航拍图像检测的难点,以 CenterNet 为基准模型,通过增加自适应阈值预测分支对检测结果进行筛选优化,在主干网络中设计可变形空洞卷积结构和基于注意力机制[17]的卷积连接(CBA-connection)结构。并通过结构化信息丢弃以及利用误检、漏检目标构建新样本的方式实现数据增强。本文模型在保证高实时性的同时显著提升了航拍目标的检测精度,在无人机等嵌入式移动设备上有较高的应用价值。

2 基于自适应阈值的改进 CenterNet 算法

2.1 算法简述

针对航拍图像目标尺度和形态多变且背景复杂度高的问题,本文设计的航拍图像目标检测算法以预测目标中心点作为关键点来进行分类和回归,将真实标注的矩形框信息进行重定义,更加关注于真实目标区域,中心点则更加能够反映不同尺度和形态的目标特征,减少了背景因素的影响,有利于提升检测精度。

检测流程如图 1 所示,输入图像经由主干网络进行特征提取后输出四个预测分支,关键点热力图用于预测中心点作为关键点的概率,表示为 $\hat{Y}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$,目标尺寸用于预测目标的宽和高,表示为 $\hat{S}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$,中心点偏移误差用于补偿下采样引起的中心点离散化误差 $\hat{O}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$。增加自适应阈值预测分支,用于预测自适应阈值 $\hat{T}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times C}$,优化回归结果。其中 W 和 H 分别为图像的宽和高,R 是尺寸缩放比例,C 是输出通道数也

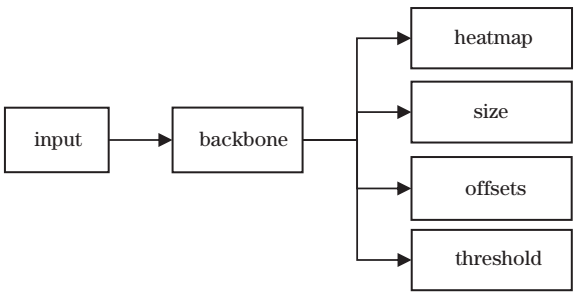

图 1 算法流程图

Fig. 1 Algorithm flowchart

是关键点的类别数。随后基于关键点进行目标分类预测和中心点坐标信息回归，再基于偏移误差对中心点坐标进行补偿，最后结合目标宽高尺寸预测物体的边界框。

首先，在关键点热力图预测分支中，设输入图像为 $I\in\mathbb{R}^{W\times H\times 3}$，对于真实标注的关键点 p，将其通过高斯核函数，映射到热力图 $\hat{Y}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$ 上，可以描述为

$$Y_{xyc}=\exp\left[-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2}\right],\quad(1)$$

式中：Y_{xyc} 表示真实关键点 p 的高斯核，从目标中心到边缘由 1 到 0 递减；$(\tilde{p}_x,\tilde{p}_y)$ 为分辨率降低后的像素点位置；σ_p^2 是与目标宽高尺寸相关的标准差。$\hat{Y}_{xyc}$ 表示预测目标中心点作为关键点的类别置信度；$\hat{Y}_{xyc}=1$ 表示坐标 (x,y) 对应类别 C；$\hat{Y}_{xyc}=0$ 表示背景。

首先，在中心点预测的损失函数中引入焦点损失，降低高置信度样本的损失，优化样本类别不均衡以及样本分类难度不平衡问题，可描述为

$$L_k=-\frac{1}{N}\cdot\sum_{xyc}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}\log(\hat{Y}_{xyc}), & \text{if } Y_{xyc}=1\\ (1-Y_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha}\log(1-\hat{Y}_{xyc}), & \text{otherwise}\end{cases},\quad(2)$$

式中：超参数 α、β 的引入减少了易分类样本的损失，使模型更关注于不易分类和错分类的目标，在本实验中设置为 2.0 和 4.0；N 为图像关键点数量，用来将损失值归一化处理。

其次，在中心点偏移误差预测分支中，由于数据的离散，图像经过下采样缩放后的关键点位置产生偏移，偏移量设为 $\hat{O}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$，因此在预测关键点位置时额外增加了偏移量的预测 $\hat{O}_p$，损失函数采用 L1 损失，可描述为

$$L_{OFF}=\frac{1}{N}\sum_{p=1}^{N}\left|\hat{O}_{\tilde{p}}-\left(\frac{p}{R}-\tilde{p}\right)\right|,\quad(3)$$

式中：p 为真实标注的关键点；R 为缩放倍数；$\tilde{p}$ 为下采样缩放后的关键点；所有类别共享同一个 L_{OFF}，且只在关键点的位置处增加偏移量。

最后，在目标尺寸预测分支中，假设 $[x_1^{(p)},y_1^{(p)},x_2^{(p)},y_2^{(p)}]$ 为目标 p 的真实边界框角点坐标，则目标的真实尺寸大小为 $S_p=[x_2^{(p)}-x_1^{(p)},y_2^{(p)}-y_1^{(p)}]$，设目标尺寸的预测量 $\hat{S}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$，则 p 点的尺寸表示为 $\hat{S}_p$，损失函数可描述为

$$L_S=\frac{1}{N}\sum_{p=1}^{N}|\hat{S}_p-S_p|。\quad(4)$$

2.2　自适应阈值预测分支

在多分类检测任务中，通过对不同分类的置信度进行比较获得最终的分类结果，然而这样的处理对于低置信度预测情况难以区分，容易产生误检、漏检以及错分类的现象。对于希望得到更高准确率而非召回率的应用场景，可以采取阈值筛选的方式对预测结果进行处理。然而设定的阈值来源于经验，会对算法结果产生较大的影响，无论是手动调参还是网格搜索都难以达到最优的筛选效果。另外，采取相同的阈值对不同目标的空间位置进行筛选预测并不合理，没有考虑到样本以及不同分类的特性。

因此，增加自适应阈值预测分支，对每一个像素点进行自适应二值化，二值化阈值由网络学习得到，这样最终的输出图对于阈值有较好的鲁棒性。在通过阈值进行边界框构建的后处理过程中，自适应地考虑样本特性和不同分类目标的预测难度，能够更好地筛选预测目标，提升后处理对分类以及边界回归的筛选能力，增强了中心点预测、中心点偏移以及目标尺寸预测三者之间的关联性。一般来说，通过固定阈值 t 来描述缩放后的像素点 $\tilde{p}$ 的二值化过程，表示为

$$B=\begin{cases}1, & \text{if } \hat{p}\geqslant t\\ 0, & \text{otherwise}\end{cases},\quad(5)$$

式中：B 表示二值化结果，但是上述公式不可微，无法随网络进行训练优化，为解决此问题，引入近似阶跃函数执行近似二值化，可描述为

$$\hat{B}_p=\frac{1}{1+\exp[-k(\tilde{p}-\hat{T}_p)]},\quad(6)$$

式中：$\hat{B}_p$ 表示近似二值化的结果；$\hat{T}_p$ 表示自适应阈值；k 为放大因子，该函数与标准二值化函数类似，并且可以微分，能够在训练中与网络一起进行优化，在训练中梯度随着 k 的增加而增大，梯度的分离不仅可以帮助区分前景与背景，还可以有效区分图像目标聚集区域的小目标物体。

本文将(6)式作为热力图上像素点 $\tilde{p}$ 的自适应激活函数，用来预测中心点的置信度，但是阈值 $\hat{T}$ 是无法进行训练的，本文引入交并比(IoU)损失函

数将关键点的自适应阈值预测这一步骤加入到网络中一起训练。通过中心点预测结合偏移预测和目标尺寸预测回归得到目标边界框 $\tilde{B}_{\mathrm{BOX}}=(\tilde{b}_x,\tilde{b}_y,\tilde{b}_w,\tilde{b}_h)$，再同真实标签的目标边界框 $B_{\mathrm{BOX}}=(b_x,b_y,b_w,b_h)$ 进行交并比计算，得到 IoU 损失，在此过程中阈值 $\hat{T}$ 也得到了训练，可描述为

$$L_{\mathrm{B}}=-\ln\frac{\mathrm{Intersection}(B_{\mathrm{BOX}},\tilde{B}_{\mathrm{BOX}})}{\mathrm{Union}(B_{\mathrm{BOX}},\tilde{B}_{\mathrm{BOX}})}。\tag{7}$$

2.3　整体损失及边界回归

整体损失(L_{T})包括热力图中心点的损失、中心点偏移损失、目标尺寸损失以及自适应阈值损失，可描述为

$$L_{\mathrm{T}}=L_{\mathrm{k}}+\lambda_{\mathrm{S}}L_{\mathrm{S}}+\lambda_{\mathrm{OFF}}L_{\mathrm{OFF}}+\lambda_{\mathrm{B}}L_{\mathrm{B}},\tag{8}$$

式中：λ_{S}、λ_{OFF} 和 λ_{B} 为调节其损失对权重影响的参数，在本文中设为 0.1、1.0 及 0.2，其中自适应阈值损失的权重是以 0.1 为步长进行权重微调后的最优结果，其他参数根据 CenterNet 基础算法进行设置。整个网络在每个关键点位置计算得到 5 个预测值，分别为关键点类别 C、偏移量(δx，δy)、尺寸(w，h)，所有输出都基于同一特征提取网络。

为得到准确的目标边界框预测信息，将热力图上所有响应点与其相邻的 8 个点进行比较，若该点的响应值不小于周围临近点则保留作为热力图的峰值点，共保留 100 个最高的峰值点，作为初步预测的目标中心点也是热力图的关键点。定义 $\hat{P}_C$ 为预测得到的 C 类别的 N 个中心点的集合，将其表示为关键点

$$\begin{aligned}\hat{P}=&\left[(\hat{x}_i,\hat{y}_i)\right]_{i=1}^{N},\\&(\hat{x}_i+\delta\hat{x}_i-\hat{w}_i/2,\hat{y}_i+\delta\hat{y}_i-\hat{h}_i/2,\\&\hat{x}_i+\delta\hat{x}_i+\hat{w}_i/2,\hat{y}_i+\delta\hat{y}_i+\hat{h}_i/2),\end{aligned}\tag{9}$$

式中：$(\delta\hat{x}_i,\delta\hat{y}_i)=\hat{O}_{\hat{x}_i,\hat{y}_i}$ 是预测的中心点偏移量；$(\hat{w}_i,\hat{h}_i)=\hat{S}_{\hat{x}_i,\hat{y}_i}$ 是预测的宽高尺寸，得到如(9)式所示的边界框角点坐标。

3　网络模型设计及训练优化策略

3.1　可变形空洞卷积模块

航拍图像目标受环境干扰较大，其尺度、姿态等特征多变，例如在不同航拍角度下的行人目标存在形态特征上的显著差异，使用传统卷积核按照规定格点进行采样时，范围为固定的矩形感受野，难以适应目标形变。针对此问题，本文在空洞空间金字塔池化的基础上，应用可变形卷积来代替原始的 3×3 卷积核进行空洞卷积计算，设计了可变形空洞卷积模块(D-ASPP)。可变形卷积又使得网络不再局限于矩形区域采样，而是聚焦于目标的尺度和形变信息，能够更加灵活地对目标信息进行多尺度融合。

传统卷积使用规则网格 R 在输入特征图 $\boldsymbol{x}$ 上进行采样，并以权值 $\boldsymbol{w}$ 对采样值进行加权并求和，则输出特征图 p_i 点的值可描述为

$$\boldsymbol{y}(p_i)=\sum_{p_n\in\mathbb{R}}\boldsymbol{w}(p_n)\boldsymbol{x}(p_i+p_n),\tag{10}$$

式中：用 p_i 表示输出特征图 $\boldsymbol{y}$ 上某个点的位置；p_n 表示标准采样格点的位置。在可变形卷积中引入采样点位置偏移，并用 $\{\Delta p_n\mid n=1,2,\cdots,N\}$ 表示，其中 N 为采样点个数，则输出特征图上 p_i 点的表达式变为

$$\boldsymbol{y}(p_i)=\sum_{p_n\in\mathbb{R}}\boldsymbol{w}(p_n)\boldsymbol{x}(p_i+p_n+\Delta p_n)。\tag{11}$$

设计的可变形空洞卷积结构如图 2 所示，包括一个 1×1 卷积核进行通道降维，三个不同空洞率的可变形卷积进行多尺度采样，通过不规则卷积提升

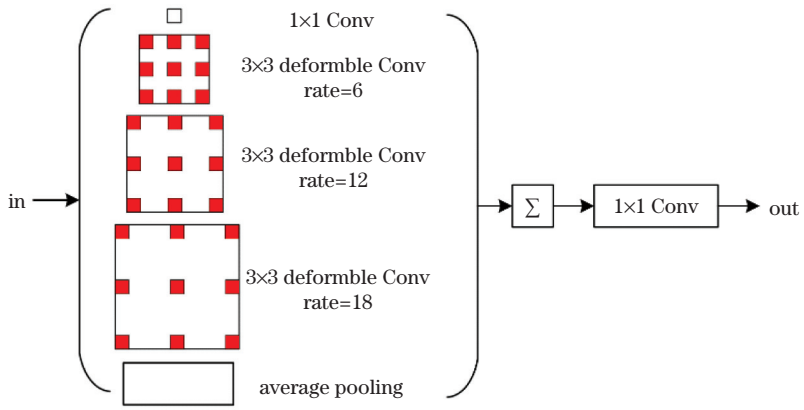

图 2　可变形空洞卷积结构示意图

Fig. 2　Schematic diagram of deformable cavity convolution structure

对目标物体的形变适应能力并充分结合空洞卷积的优势，增加权重为 0 的空洞点，在不增加卷积核大小和滑动步长的情况下，扩大感受野范围。并添加了均值池化层，将特征图进行全局平均池化并获取全局信息，最后通过 1×1 卷积核进行多尺度特征融合，在扩大感受野的同时保证了特征图的输出分辨率不受影响，可以理解为局部下采样的过程。

3.2 CBA-connection 结构

网络模型随着下采样层数的加深，携带的特征信息更加丰富，而这些在不同空间和通道维度上的信息并不都是有效的，也有复杂的背景信息。在原始网络中，特征信息对最终预测的贡献能力相同，这就使小目标检测极易受到复杂的背景干扰，导致目标的轮廓细节不明显。因此本文设计了 CBA-connection 结构，如图 3(a)所示，引入注意力机制为空间域和通道域赋予权重。

首先将浅层特征图以及深层特征图，通过注意力模块进行空间域和通道域维度信息的自适应加权。再通过反卷积[18]的方式进行上采样，恢复到与浅层网络输出特征图相同的分辨率，并将分辨率相同的两个特征图进行跳跃连接，对目标检测贡献较大的高级语义特征和低级空间特征进行融合。

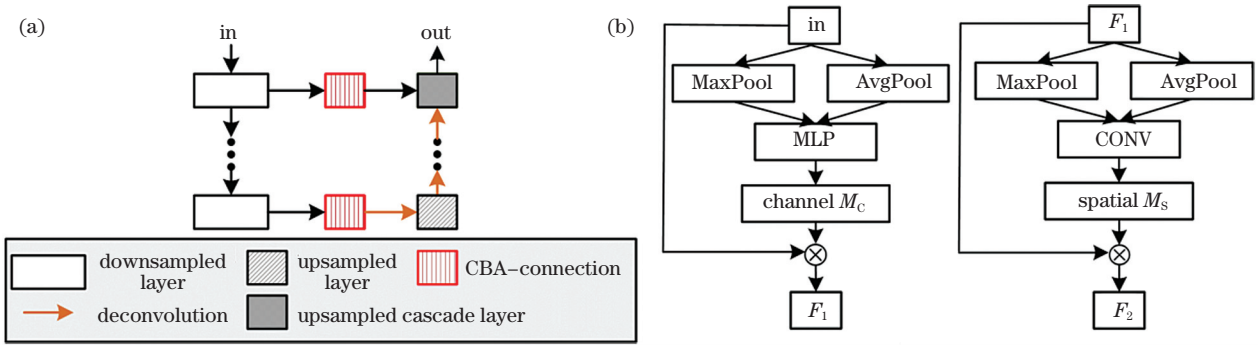

图 3　CBA-connection。(a)整体结构示意图；(b)注意力模块流程图

Fig. 3　CBA-connection. (a) Overall structure diagram; (b) attention module flowchart

注意力模块的网络结构如图 3 (b)所示，设输入特征图为 $\boldsymbol{F}\in\mathbb{R}^{H\times W\times C}$，首先输入通道注意力模块，通过均值池化和最大值池化在空间域对特征图进行压缩，得到两个不同的空间背景描述 $\boldsymbol{F}^{C}_{\mathrm{avg}}$ 和 $\boldsymbol{F}^{C}_{\mathrm{max}}$，分别表示平均池化特征和最大池化特征，再输入共享的多层感知器(M_{MLP})并进行逐元素求和合并，得到通道注意力图 $\boldsymbol{M}_C(\boldsymbol{F})\in\mathbb{R}^{1\times 1}$，计算过程可描述为

$$\begin{aligned}\boldsymbol{M}_C(\boldsymbol{F})&=S\{M_{\mathrm{MLP}}[A_{\mathrm{AvgPool}}(\boldsymbol{F})]+\\&M_{\mathrm{MLP}}[M_{\mathrm{MaxPool}}(\boldsymbol{F})]\}=\\&S\{\boldsymbol{W}_1[\boldsymbol{W}_0(\boldsymbol{F}^{C}_{\mathrm{avg}})]+\boldsymbol{W}_1[\boldsymbol{W}_0(\boldsymbol{F}^{C}_{\mathrm{max}})]\},\end{aligned}\tag{12}$$

式中：$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 为 MLP 的权重；S 表示 sigmoid 函数。空间注意力模块则通过相似的池化操作来提取特征图的通道信息，得到 $\boldsymbol{F}^{S}_{\mathrm{avg}}$ 和 $\boldsymbol{F}^{S}_{\mathrm{max}}$，并聚合在一起通过卷积运算得到空间注意力图 $\boldsymbol{M}_S(\boldsymbol{F})\in\mathbb{R}^{H\times W}$，计算过程可描述为

$$\begin{aligned}\boldsymbol{M}_S(\boldsymbol{F})&=S\{\boldsymbol{f}^{7\times 7}[A_{\mathrm{AvgPool}}(\boldsymbol{F});M_{\mathrm{MaxPool}}(\boldsymbol{F})]\}=\\&S[\boldsymbol{f}^{7\times 7}(\boldsymbol{F}^{S}_{\mathrm{avg}};\boldsymbol{F}^{S}_{\mathrm{max}})],\end{aligned}\tag{13}$$

式中：$\boldsymbol{f}^{7\times 7}$ 表示 7×7 的卷积层，经过注意力模块依次得到一维的通道注意力图 $\boldsymbol{M}_C(\boldsymbol{F})\in\mathbb{R}^{1\times 1\times C}$ 以及二维的空间注意力图 $\boldsymbol{M}_S(\boldsymbol{F})\in\mathbb{R}^{H\times W\times 1}$，整体推导过程可描述为

$$\begin{cases}\boldsymbol{F}_1=\boldsymbol{M}_C(\boldsymbol{F})\otimes\boldsymbol{F}\\\boldsymbol{F}_2=\boldsymbol{M}_S(\boldsymbol{F}_1)\otimes\boldsymbol{F}_1\end{cases}。\tag{14}$$

引入注意力机制有效筛选了对目标检测贡献程度大的关键信息，提升了输出特征图的表达能力。最后将得到的同分辨率特征图进行跳跃连接，这一做法在不影响网络预测性能的基础上，为小目标检测提供了更加精细和有效的特征信息，如梯度等，细化了检测结果的中心位置预测和偏移值回归。

3.3 主干网络结构

所设计的编码-解码结构的主干网络如图 4 所示，下采样包括 4 倍、8 倍残差层以及 D-ASPP，上采样由 CBA-connection 结构构成。将尺寸为 512 pixel×512 pixel 的航拍图像输入 ResNet 进行下采样，残差层通过跳跃连接的方式将输入信息和输出信息充分融合，很大程度上保存了小目标的空间几何特征，同时有效避免了深层网络的梯度消失。随后用 D-ASPP 模块代替 16 倍以及 32 倍下采样残差层，进行局部下采样，在不降低下采样分辨率的同时有效提取了多尺度感受野的特征信息，在获得深层语义信息的同时从根本上缓解了下采样过程中小目标空间信息损失问题。最后连接 CBA-connection 结构，经由注意力模块进行空间域和通道域的加权，然

后通过反卷积进行上采样，恢复到 4 倍下采样时的分辨率，再和第一个残差单元经由注意力模块的输出特征图进行同分辨率的跳跃连接，充分融合了具有更多空间信息的浅层特征和具有丰富语义信息的深层特征，以得到空间轮廓信息丰富、目标表达能力更强的高分辨率输出特征图。

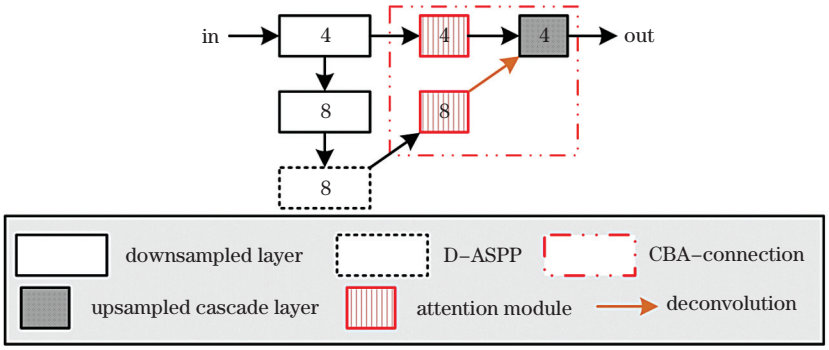

图 4　主干网络结构示意图

Fig. 4　Schematic diagram of backbone network structure

3.4　数据增强

为增强模型效果和泛化能力，本文将模型在 COCO 数据集上进行预训练，再对训练数据进行结构化信息丢弃以及利用误检、漏检目标构建新样本的方式达到数据增强的目的。

结构化信息丢弃通过生成与原图相同分辨率的掩模与原图相乘，均匀删除部分正方形区域，实现了特定区域的信息丢弃，强化模型在训练过程中对于目标局部特征的认知，弱化对于目标全部特征的依赖，同时避免过拟合发生。在实际训练中检查遮挡格点与目标的交并比，并进行修正，确保遮挡面积不超过目标的 10%，并控制掩码的遮挡比例同时对遮挡格点不采取全黑的处理，而是在通道均值上下随机振荡，达到数据增强目的的同时尽量不破坏原始图片的像素值分布。

为了使训练出的模型更加鲁棒，减少误检和漏检，采取利用误检和漏检目标构建新样本的数据增强方法。在训练过程中将训练数据与真实标签进行对比，将 IoU 小于 0.3 的预测结果作为误检目标进行筛选和裁剪。利用负样本作为背景，将误检和漏检目标通过随机的透视变换后与背景融合，构建为全新的样本加入训练集。

4　实验结果及分析

4.1　实验平台及数据集

实验平台采用 i7 处理器，NVIDIA TITAN XP 显卡，内存为 12 G，Ubuntu16.04 操作系统。选取 NWPU VHR-10[19] 公开数据集进行实验，NWPU VHR-10 来源于 10 级地理遥感图像，经过数据翻转和随机裁剪等增强操作将样本扩大 8 倍，其中训练集由 3840 张图片组成，验证集和测试集各有 1280 张图片。涵盖了无人机在不同背景下的多种目标种类，也包含多种拍摄角度下的密集目标场景，同时小目标占比丰富。

4.2　有效性对比实验

分别采用 ResNet-50、DLA-34、ResDcn-18 三种基础主干网络，通过相同的数据和训练策略进行对比实验。对比在 NWPU VHR-10 数据集中的损失曲线，如图 5 所示，选择前 35 个训练周期的损失进行对比。可以看到，本文模型在训练数据集上收敛速度最快且效果最好，并有继续收敛的趋势，收敛效果与 DLA-34 网络相近，优于 ResNet-50 和 ResDcn-18 网络，说明本文模型能够有效收敛，较好拟合 NWPU VHR-10 航拍数据集。

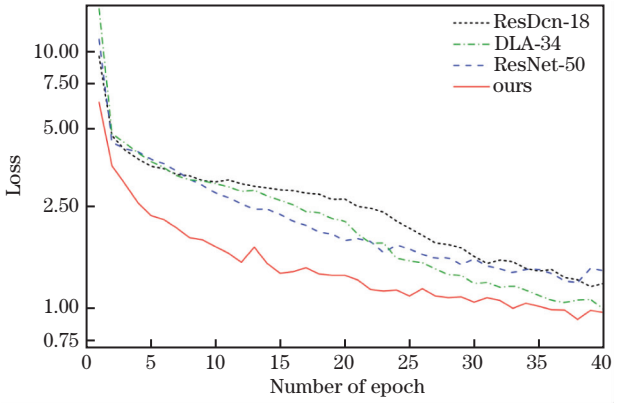

图 5　不同模型的损失曲线图

Fig. 5　Loss graph of different models

采用平均精度均值（mAP）、IoU 为 0.50 和 0.75 的平均精度（AP^{50}、AP^{75}）以及帧率（FPS）这 4 种评价指标进行检测精度和检测速度的定量分析，其中 AP^{50} 常用来评价检测算法的目标分类能

力，AP^{75} 则表示检测算法预测边界框的精确程度。与基础网络的对比实验结果如表 1 所示。

表 1 不同主干网络检测效果对比

Table 1 Comparison of detection effects of different backbone networks

Backbone	mAP /%	AP^{50} /%	AP^{75} /%	FPS /(frame·s^{-1})
ResNet-50	20.05	42.05	19.75	65
DLA-34	22.50	**45.88**	20.50	55
ResDcn-18	14.01	36.00	15.25	**131**
Ours	**25.22**	45.62	**23.36**	45

从表 1 看出，本文模型在检测精度方面效果良好，尤其是边界回归精度显著提升。模型的 mAP 相较 ResNet-50 基础网络提升 5.17%，相较 DLA-34 提升 2.72%，相较 ResDcn-18 提高了 11.21%。AP^{75} 标准上比表现最好的 DLA-34 模型提升接近 3%，但 AP^{50} 相比 DLA-34 略有不及，这是因为 DLA-34 网络的特点是深度特征融合，每一个下采样层都和上采样的同分辨率特征图进行了融合，取得了较优的分类能力，但是其边界回归能力不及本文模型。模型检测速度为 45 frame·s^{-1}，由于增加了注意力模块以及生成了新的预测分支，模型推理的计算量相应有所提升，但依旧能够满足实时性需求。综上所述，本文模型在检测精度和检测速度上取得了较好平衡。

为进一步验证本文模型的有效性，以 CenterNet(ResNet-50)为基础，引入 D-ASPP 得到 Ours+ 模型，并在新的模型基础上增添 CBA-connection 结构得到 Ours++模型，最后再增加自适应阈值预测分支，得到 Ours+++模型，依次引入不同模块后的实验结果如表 2 所示。相较基础网络，Ours+的 mAP 提升 2.52%，这证明了利用 D-ASPP 结构代替 ResNet-50 中 32 倍下采样残差层后，特征提取能力显著增强。原因是虽然高倍下采样能进一步提取深层的语义信息，但是针对航拍小目标物体，这一过程损失的空间信息对特征表达能力影响更大，而 D-ASPP 结构有效保留了目标更为精细的空间轮廓信息，边界回归能力明显增强。Ours++的 mAP 较 Ours+提升了 1.84%，这证明了经过注意力机制加权和特征图级联后，浅层信息和深层信息得到充分结合和有效利用。在注意力机制作用下，融合特征图中权重较低的背景等信息被充分过滤，更有利于提取小目标物体的有效信息，加强特征融合的有效性。而引入自适应阈值预测分支后，虽然在 AP^{50} 指标上有轻微浮动，但 mAP 和 AP^{75} 进一步提升，这说明了自适应阈值预测的方式考虑了不同样本的分类特性，对不同置信度的目标物体进行了有效筛选，提升了模型的边界回归能力。

表 2 依次引入所设计结构后的有效性对比

Table 2 Comparison of effectiveness after successively introducing the designed structure

Backbone	mAP /%	AP^{50} /%	AP^{75} /%
ResNet-50	20.05	42.05	19.75
Ours+(D-ASPP)	22.57	43.63	20.97
Ours++(CBA-connection)	24.41	45.71	22.13
Ours+++(Threshold)	25.22	45.62	23.36

不同目标种类的评估结果如表 3 所示，以 ResNet-50 为基准，各种类目标的检测精度都有了相应提升，对于飞机、船只等小目标种类上提升效果良好，符合算法设计对于小目标检测任务的优化目标。但是本文方法在网球场目标检测上效果不明显，这是因为 NWPU VHR-10 数据集中网球场样本较少的缘故，网络没有对网球场目标的特征进行有效学习。

表 3 NWPU VHR-10 数据集不同目标种类的识别精度评估

Table 3 Evaluation of the recognition accuracies of different target types in the NWPU VHR-10 data set

Target type	AP /%	Target type	AP /%
Airplane	29.77	Basketball court	17.69
Ship	26.94	Ground track field	30.64
Storage tank	32.32	Harbor	29.31
Baseball diamond	31.29	Bridge	17.16
Tennis court	16.18	Vehicle	20.90

4.3 主流算法对比实验

与其他主流目标检测算法在相同数据集上进行对比实验，如表 4 所示，其中用 ResNet^{+} 代表本文的主干网络。可以看出，所提模型有较为明显的优势，检测精度超过 YOLOv3、CornerNet 和 RetinaNet 三种常用的单阶段检测算法，尤其是体现目标回归框精确程度的 AP^{75} 超过 YOLOv3 将近 4%，与 Faster R-CNN 进行对比，检测精度略有不足。但 Faster R-CNN 作为双阶段检测算法在速度上明显较慢，达不到实时检测的要求，而本文算法的检测速度高达 45 frame·s^{-1}，超过了 YOLOv3，在检测速度方面表现最佳。

从检测精度方面分析，Faster R-CNN 由于双阶段的检测特性，检测精度较高，YOLOv3 虽然也在每个检测格点生成了 9 个锚框，但由于划分网格的精细程度不够，针对小目标物体和目标聚集区域的检测能力明显不足，CornerNet需要通过目标两个

表 4　不同检测算法在 NWPU VHR-10 数据集上的检测效果对比

Table 4　Comparison of the detection effects of different detection algorithms on the NWPU VHR-10 data set

Method	Backbone	mAP /%	AP^{50} /%	AP^{75} /%	FPS /(frame·s^{-1})
Faster R-CNN	ResNet-101	**27.23**	**48.57**	**24.50**	9
YOLOv3	DarkNet-53	22.49	45.22	19.50	41
RetinaNet	ResNet-50	16.88	30.26	16.05	19
CornerNet	Hourglass-104	19.14	39.90	17.79	22
Ours	ResNet^{+}	25.22	45.62	23.36	**45**

角点进行分类预测判断，然而航拍图像目标密集且尺寸较小，在众多角点都距离相近时，无法准确判定哪两个角点属于同一物体，将导致许多错误的边界框预测。本文模型直接通过中心点作为热力图的关键点进行回归预测，主干网络能够提取有效的特征信息同时保证输出热力图的特征分辨率，能较为准确地针对小目标物体进行分类预测。同时，偏移误差分支预测以及目标尺寸分支预测较好补偿了中心点位置，提高边框回归的精度，但对比双阶段检测算法仍有一定的差距。从检测速度分析，仅通过中心点信息进行分类预测和边界回归，本文模型就消除了锚框的限制，且相较其他算法，主干网络的模型复杂度较低，检测速度有了大幅提升。

4.4　航拍图像目标检测效果展示

改进后的 CenterNet 算法对复杂环境下的航拍目标检测效果如图 6 所示。图 6(a)为训练中通过结构化信息丢弃生成的新样本数据，图 6(b)为利用误检、漏检目标构建的新样本数据。图 6(c)为弱光环境，图 6(d)为强光环境。如图 6 所示，本文模型的检测效果良好，有效克服了光线明度的干扰，能检测到绝大部分的目标物体，并且边界框位置较为精准。图 6(e)～(f)展示的是目标聚集区域的检测情况，大部分小目标物体也都能够被有效检测且正确分类。

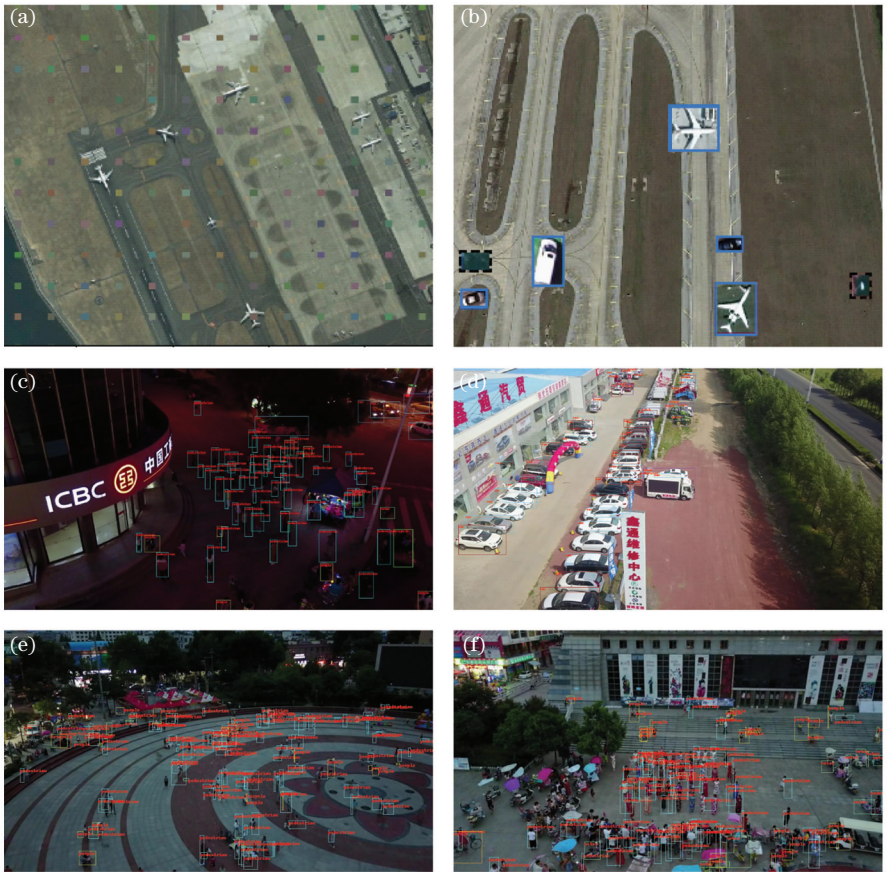

图 6　航拍图像目标检测效果展示。(a)结构化信息丢弃；(b)误检、漏检目标构建新样本；(c)夜晚场景；(d)强光场景；(e)(f)聚集区域

Fig. 6　Aerial image target detection effect display. (a) Structured information dropout; (b) false and missed detection targets construct new samples; (c) night scene; (d) strong light scene; (e) (f) gathering area

5 结 论

本文提出了一种基于自适应阈值的改进CenterNet航拍图像目标检测方法。通过增加自适应阈值预测分支提升了模型的边界回归能力。在基于ResNet-50的主干网络中设计D-ASPP结构以及CBA-connection结构，提升了特征提取能力。通过结构化信息丢弃和利用误检、漏检目标构建新样本的数据增强方式，提升了样本质量。与基于ResNet-50的CenterNet相比，mAP提升5.17%，AP^{50}提升3.57%，AP^{75}提升3.61%，检测速度达45 frame·s^{-1}，在检测精度与实时性方面取得了良好平衡，能更好适配无人机嵌入式移动设备。但是随着模型复杂度的提高，检测速度有所降低，进一步改进网络结构在不降低检测速度的基础上提升检测精度仍是下一步的研究重点。

参 考 文 献

[1] Tijtgat N, Van Ranst W, Volckaert B, et al. Embedded real-time object detection for a UAV warning system[C]// 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2110-2118.

[2] Xie B, Zhu B, Zhang H W, et al. Gradient clustering algorithm based on deep learning aerial image detection[J]. Laser & Optoelectronics Progress, 2019, 56(6): 061007.
解博, 朱斌, 张宏伟, 等. 基于深度学习航拍图像检测的梯度聚类算法[J]. 激光与光电子学进展, 2019, 56(6): 061007.

[3] Liu Y J, Yang F B, Hu P. Parallel FPN algorithm based on Cascade R-CNN for object detection from UAV aerial images[J]. Laser & Optoelectronics Progress, 2020, 57(20): 201505.
刘英杰, 杨风暴, 胡鹏. 基于Cascade R-CNN的并行特征金字塔网络无人机航拍图像目标检测算法[J]. 激光与光电子学进展, 2020, 57(20): 201505.

[4] 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.
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.

[5] 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.

[6] Ren S Q, He K M, 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.

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

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

[9] Cheng G, Zhou P C, Han J W. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7405-7415.

[10] Deng Z P, Sun H, Zhou S L, et al. Fast multiclass object detection in optical remote sensing images using region based convolutional neural networks[C]// 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), July 23-28, 2017, Fort Worth, TX, USA. New York: IEEE Press, 2017: 858-861.

[11] Hu Y Y, Wu X J, Zheng G D, et al. Object detection of UAV for anti-UAV based on improved YOLO v3[C]// 2019 Chinese Control Conference (CCC), July 27-30, 2019, Guangzhou, China. New York: IEEE Press, 2019: 8386-8390.

[12] Neubeck A, van Gool L. Efficient non-maximum suppression[C]// 18th International Conference on Pattern Recognition (ICPR'06), August 20-24, 2006, Hong Kong, China. New York: IEEE Press, 2006: 850-855.

[13] Zhou X Y, Zhuo J C, Krähenbühl P. Bottom-up object detection by grouping extreme and center points[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 850-859.

[14] 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.

[15] Law H, Deng J. CornerNet: detecting objects as paired keypoints[M] // Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11218: 765-781.

[16] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points[EB/OL]. (2019-08-16)[2020-04-02] https://arxiv.org/abs/1904.07850.

[17] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[18] Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional networks[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 2528-2535.

[19] Cheng G, Han J W, Zhou P C, et al. Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 98: 119-132.