激光与光电子学进展 先进成像

面向航摄图像目标检测的轻量级特征融合网络

樊强强[1]，史再峰[1,3*]，孔凡宁[1]，李少雄[1]，肖军[2]

[1]天津大学微电子学院，天津 300072；

[2]飞腾信息技术有限公司，天津 300459；

[3]天津市成像与感知微电子技术重点实验室，天津 300072

摘要 针对现有航摄图像目标检测算法中模型复杂、超参数多、检测精度较低的问题，提出一种面向航摄图像目标检测的轻量级多尺度特征融合网络。该网络采用Anchor-Free思想，通过逐像素预测的方式，减少了与Anchor相关的超参数；利用MobileNetV3作为特征提取网络并使用Ghost瓶颈模块优化多尺度特征融合网络，来降低网络的参数量和计算量；引入可变形卷积来构建可变形感受野模块，提高检测器对航摄图像目标形变的鲁棒性；同时采用标签分配策略SimOTA进行动态样本匹配，以缓解航摄图像目标分布密集、遮挡严重的检测问题。在数据集VisDrone2019-DET和NWPU VHR-10上对所提网络进行评估，检测精度AP^{50}分别达26.6%和94.4%，检测速度分别达59.9 frame/s和79.6 frame/s。与主流目标检测网络相比，所提网络在保持较高检测精度和速度的同时，具有较小的参数量和计算量，更适合应用于机载计算设备。

关键词 目标检测；Anchor-Free；可变形感受野块；特征融合；动态样本匹配

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP220859

Lightweight Feature Fusion Network for Object Detection in Aerial Photography Images

Fan Qiangqiang[1], Shi Zaifeng[1,3*], Kong Fanning[1], Li Shaoxiong[1], Xiao Jun[2]

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

[2]*Phytium Technology Co., Ltd., Tianjin* 300459, *China*;

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

Abstract The existing aerial photography image object detection algorithms have several problems, such as complicated models, too many hyperparameters, and poor detection accuracy. Therefore, this paper proposes a lightweight multiscale feature fusion network for object detection in aerial photography images. The proposed network employs the idea of Anchor-Free and reduces the hyperparameters related to Anchor through pixel-by-pixel prediction. First, MobileNetV3 is adopted as the backbone network for feature extraction, and the Ghost bottleneck module is used as the base block for multiscale feature fusion to reduce number of parameters and computational costs. Then, deformable convolution is introduced to construct a deformable receptive field block to improve the robustness of the detector to the deformation of aerial photography objects. Furthermore, the label assignment strategy SimOTA is employed for dynamic sample matching, which alleviates the problems of dense distribution and heavy occlusion of aerial photography objects. The proposed network is evaluated on VisDrone2019-DET and NWPU VHR-10 datasets. The detection accuracy AP^{50} of the proposed network reaches 26.6% and 94.4%, and the detection speed reaches 59.9 and 79.6 frame/s, respectively. Compared with other mainstream object detection networks, the proposed network has fewer parameters and computational costs while maintaining high detection accuracy and speed, making it more suitable for airborne computing devices.

Key words object detection; Anchor-Free; deformable receptive field block; feature fusion; dynamic label assignment

收稿日期：2022-03-02；**修回日期**：2022-04-09；**录用日期**：2022-05-18；**网络首发日期**：2022-05-28

基金项目：国家自然科学基金(62071326)

通信作者：*shizaifeng@tju.edu.cn

1 引 言

无人机航摄图像目标检测在交通监控、应急救援、农业生产等任务中应用广泛，已成为计算机视觉领域的重要研究方向之一[1]。当前，无人机航摄图像目标检测面临的挑战主要有两个方面：一是航摄图像目标具有尺度多变、分布密集、遮挡严重、尺寸偏小等特点，使得目标检测精度较差；二是无人机设备的计算能力有限，无法满足检测任务的实时性需求。

传统的无人机航摄图像目标检测采用滑动窗口的特征提取算法，主要有：方向梯度直方图特征(HOG)[2]、尺度不变特征变换(SIFT)[3]、加速鲁棒特征(SURF)[3]、Haar-like小波特征[4]等。尽管这些方法对航摄图像目标检测有一定的效果，但是设计这些特征复杂度高，在机载计算设备上难以进行实时检测。目前，研究人员把众多流行的卷积神经网络(CNN)应用在无人机航摄图像目标检测方面。Wang等[5]选择SSD[6]、Faster R-CNN[7]、RetinaNet[8]作为代表性的CNN目标检测器进行航摄行人检测，验证了基于CNN的目标检测器进行航摄图像目标检测的有效性。张瑞倩等[9]在Faster R-CNN和Cascade R-CNN的基础上，引入多尺度空洞卷积增大网络的感受野，提升航摄图像目标检测算法在尺度多变、遮挡等复杂场景下的检测精度。汪鹏等[10]对YOLOv3进行改进，在骨干网络上结合密集连接网络，加强特征传播，并引入Distance-IoU损失对边界框(bounding box)进行回归，提高对航摄图像目标的检测精度。刘芳等[11]提出一种多尺度自适应候选区域生成网络，利用反卷积级联结构加强底层和高层特征融合，并利用特征自适应分支增加候选框与真实目标的匹配度，来提高航摄图像目标检测精度。刘英杰等[12]对特征金字塔进行改进，通过添加并行分支强化对小目标的特征表达能力，并增加级联网络提高对小目标的定位能力。上述基于锚框(Anchor-Based)的航摄图像目标检测模型通过改进网络结构，优化锚框，在一定程度上提高了检测精度，但还存在以下问题：由于航摄图像上物体尺寸较小，大部分是背景区域，Anchor-Based模型进行均匀采样会导致正负样本极度不均衡，不利于模型收敛；锚框的生成通常需要做聚类分析，并且通过聚类分析得到的锚框不具有一般性，需要设计不同的锚框以适应不同的数据集[13]；需要引入大量的超参数来定义一系列长宽比和尺寸不一的锚框，这些超参数严重影响目标检测的速度和召回率等指标[14]。

针对上述问题，本文提出了一种面向航摄图像目标检测的轻量级多尺度特征融合网络。该网络采用逐像素预测的方式预测目标的类别和位置，避免了与锚框相关的复杂计算和超参数的设定，提高了网络模型的推理效率。为了减少网络的参数量和计算量，采用基于深度可分离卷积和瓶颈结构的MobileNetV3[15]作为特征提取网络并利用Ghost瓶颈模块[16]优化路径聚合网络(PAN)。在感受野模块的基础上结合可变形卷积，构建了可变形感受野模块，提高了特征对形变目标的表达能力。在预测过程中，根据预测信息动态筛选候选框，实现候选框和真实框的精准匹配。在标准数据集上的实验结果表明，所提网络模型在保持较少参数量和计算量的前提下，取得了较高的检测精度和检测速度。

2 所提检测网络

2.1 网络结构

航摄图像目标由于自身特点和成像角度的差异导致标注框尺寸偏小、宽高比变化比较大，给Anchor的设定增加了难度，并且受背景因素的影响，候选区存在大量的负样本加剧了模型中正负样本不平衡的问题。FCOS模型[14]将目标检测任务转换为逐像素预测任务，直接对特征图上的每个像素点进行回归。具体来说，对于特征图上的某个像素点，当其在输入图像上的映射点落在真实框内部，则将该像素点视为一个正样本，并将真实框的类别标签赋予该像素点，回归过程则是计算像素点与真实框四个边的距离，输出为一个四维向量。对于远离目标中心的像素点，由于它们产生的边界框置信度比较低，定位效果较差，FCOS模型采用中心度分支来抑制这些低质量的边界框，进而提高检测性能。相比Anchor-Based模型，FCOS模型利用逐像素预测的方式，避免了与Anchor相关的复杂计算和超参数的设定，同时这种密集预测的方式大大增加了样本数量，对分布密集的航摄图像目标进行检测更具有优势。因此所提轻量级多尺度特征融合网络在预测网络部分采用逐像素预测的方式，以减少与Anchor相关的超参数。

所提轻量级多尺度特征融合网络的结构如图1所示。该网络由骨干网络、特征融合网络、预测网络三部分构成。骨干网络采用轻量级网络MobileNetV3对输入图像进行初步的特征提取。针对航摄图像尺度和形态多变的特点，在特征提取过程中基于可变形卷积和感受野模块，构建了可变形感受野模块(D-RFB)，D-RFB充分融合了可变形卷积采样的细节信息和空洞卷积采样的不同感受野下的语义信息，有利于提高网络模型识别形变和多尺度目标的能力。特征融合部分，利用Ghost-PAN结构对不同尺度的上下文信息进行融合，Ghost-PAN采用计算效率高的Ghost瓶颈模块替换PAN中的标准卷积，目的是在不损失检测精度的情况下，降低网络模型的参数量和计算量。预测部分，采用逐像素预测的方式，通过解耦头(Decoupled Head)[17]在L_3、L_4和L_5三个尺度的特征图上进行分类和回归。

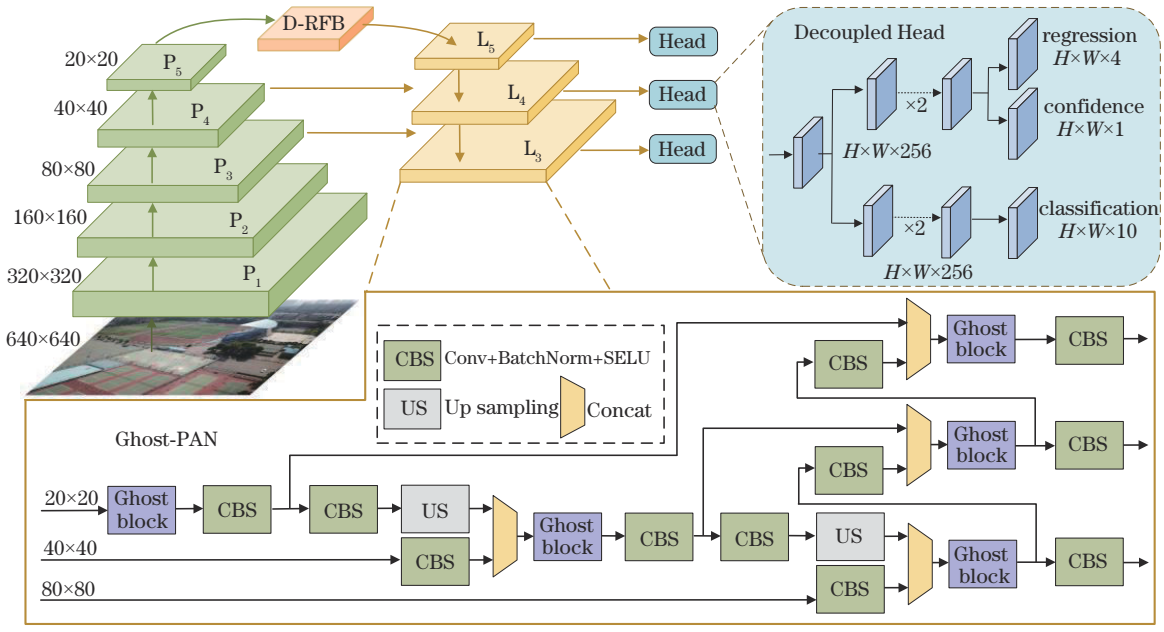

图 1　网络整体结构

Fig. 1　Overall network architecture

2.2　骨干网络

基于CNN的深层骨干网络在提高检测准确率方面取得了跨越式发展，但是大量的参数和计算成本使其难以直接应用在资源有限的无人机设备上。MobileNetV3网络[15]采用具有线性瓶颈的残差结构，大大减少了模型的计算量和参数量；激活函数选用h-swish，避免了大量的指数运算；引入SE通道注意力机制，并通过神经架构搜索算法获得卷积核和通道的最佳数量，在精度和速度方面表现优异。因此采用MobileNetV3[15]作为骨干网络，对航摄图像进行初步的特征提取，具体网络结构如表1所示。Bneck为由深度可分离卷积和残差结构设计的基础瓶颈块；SE代表瓶颈块中使用的通道注意力机制；NL为激活函数的类型，包括HS(h-swish)和RE(ReLU)两种；s代表步长。

表 1　骨干网络的详细结构

Table 1　Detailed structure of backbone network

Input size	Operator	Exp size	Output size	SE	NL	s
640×640×3	Conv2d		320×320×16		HS	2
320×320×16	Bneck, 3×3	16	320×320×16		RE	1
320×320×16	Bneck, 3×3	64	160×160×24		RE	2
160×160×24	Bneck, 3×3	72	160×160×24		RE	1
160×160×24	Bneck, 5×5	72	80×80×40	1	RE	2
80×80×40	Bneck, 5×5	120	80×80×40	1	RE	1
80×80×40	Bneck, 5×5	120	80×80×40	1	RE	1
80×80×40	Bneck, 3×3	240	40×40×80		HS	2
40×40×80	Bneck, 3×3	200	40×40×80		HS	1
40×40×80	Bneck, 3×3	184	40×40×80		HS	1
40×40×80	Bneck, 3×3	184	40×40×80		HS	1
40×40×80	Bneck, 3×3	480	40×40×112	1	HS	1
40×40×112	Bneck, 3×3	672	40×40×112	1	HS	1
40×40×112	Bneck, 5×5	672	20×20×160	1	HS	2
20×20×160	Bneck, 5×5	960	20×20×160	1	HS	1
20×20×160	Bneck, 5×5	960	20×20×160	1	HS	1

2.3　多尺度自适应感受野模块

不同于自然场景图像，无人机的飞行高度多变，导致航摄图像的目标尺度变化剧烈，另外受到拍摄角度的影响，目标会产生一定程度的变形。如果只采用固定大小和尺寸的标准卷积对无人机航摄图像进行采样，将无法准确提取航摄图像的目标特征。可变形卷

积网络(DCN)[18]通过增加偏移量来学习目标的几何形变特征,其卷积核形状和大小能够根据采样目标的尺度和形状得到动态调整,具有自适应的感受野。标准卷积采用形状规则的网格$\boldsymbol{R}$在输入特征图进行上采样,并对采样值赋予权值$\boldsymbol{w}$,然后进行加权求和,则输出特征图$\boldsymbol{y}$上$\boldsymbol{p}_j$点的运算过程为

$$y(\boldsymbol{p}_j)=\sum_{\boldsymbol{p}_n\in\boldsymbol{R}}w(\boldsymbol{p}_n)x(\boldsymbol{p}_j+\boldsymbol{p}_n),\tag{1}$$

式中:$\boldsymbol{x}$代表输入特征图;$\boldsymbol{y}$代表输出特征图;$\boldsymbol{p}_n$代表$\boldsymbol{R}$中的采样点的位置;$\boldsymbol{p}_j$代表输出特征图某个点的位置。可变形卷积在采样网格$\boldsymbol{R}$中增加了偏移量$\{\Delta\boldsymbol{p}_n\mid n=1,2,\cdots,N\}$,其中$N$为采样点的数目,则输出特征图$\boldsymbol{y}$上$\boldsymbol{p}_j$点的运算过程变为

$$y(\boldsymbol{p}_j)=\sum_{\boldsymbol{p}_n\in\boldsymbol{R}}w(\boldsymbol{p}_n)x(\boldsymbol{p}_j+\boldsymbol{p}_n+\Delta\boldsymbol{p}_n)。\tag{2}$$

受可变形卷积网络[18]的启发,本文在感受野模块(RFB)[19]的基础上引入可变形卷积,将感受野模块中的3×3标准卷积替换为3×3可变形卷积,5×5卷积层替换为两个堆叠的3×3可变形卷积,以此构建了一个D-RFB。其自适应地提取形变和不均匀分布的高级语义信息,以实现多尺度的航摄图像目标检测。

设计的可变形感受野模块的结构如图2所示,包括多分支可变形卷积层和多分支空洞卷积层两部分。多分支可变形卷积层模仿人类视觉感受野机制,采用3个不同感受野的可变形卷积来提取形变和多尺度特征,使得采样过程不再局限于固定的卷积核,而是能够根据航摄图像目标的形变信息和尺度信息进行自适应卷积核形状调整;多分支空洞卷积层采用空洞卷积改变了到中心的采样距离,扩大了卷积分支的感受野,从根本上缓解了下采样过程中航摄图像目标细节特征损失的问题。

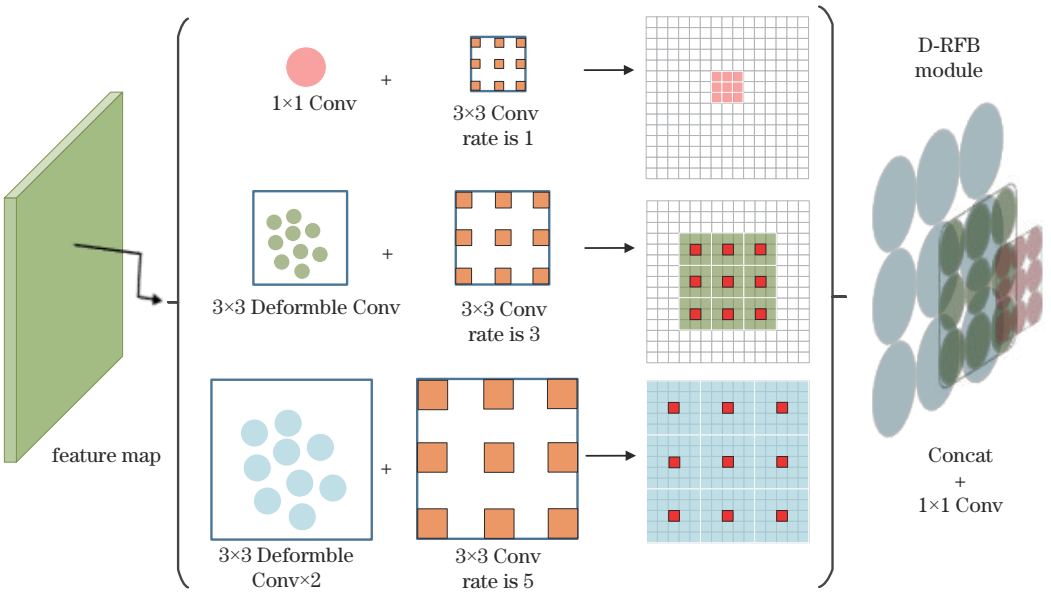

图2　可变形感受野模块

Fig. 2　Deformable receptive field block

2.4　特征融合Ghost-PAN结构

深层特征具有大的感受野,包含丰富的上下文信息;浅层特征分辨率大,具有许多细节特征信息和准确的位置信息。为了实现不同尺度的特征提取,提升检测精度的同时保持较高的检测速度,在PAN的基础上引入Ghost瓶颈模块[16],设计了Ghost-PAN结构。Ghost-PAN使用Ghost瓶颈模块替换PAN结构中的标准卷积,在保持较低计算量的情况下,实现具有丰富语义信息的深层特征和细节较多的浅层特征的融合。

Ghost模块首先使用标准卷积得到通道数较少的本征特征图,然后在本征特征图的基础上使用计算量较少的线性运算来生成Ghost特征图,最后对本征特征图和Ghost特征图进行级联得到输出特征图。假设输入和输出特征图的高和宽为H和W,输入特征图的通道数为N_1,输出特征图的通道数为C,卷积核大小为$K\times K$,不考虑偏置项,标准卷积的计算量为

$$N_{\mathrm{SC}}=H\times W\times N_1\times C\times K\times K。\tag{3}$$

Ghost模块的计算量为

$$N_{\mathrm{GM}}=H\times W\times N_1\times\frac{C}{S}\times K\times K+(S-1)\times\frac{C}{S}\times H\times W\times D\times D,\tag{4}$$

式中:$D\times D$为线性运算的卷积核大小;S为输出特征图通道数和本征特征图通道数的比值,且$S\geqslant 1$;$K=D$。

则标准卷积和Ghost模块的计算量之比为

$$\frac{N_{SC}}{N_{GM}}=\frac{H\times W\times N_1\times C\times K\times K}{H\times W\times N_1\times\frac{C}{S}\times K\times K+(S-1)\times\frac{C}{S}\times H\times W\times D\times D}\approx\frac{N_1\times S}{N_1+S-1}\approx S。\tag{5}$$

Ghost瓶颈模块的结构如图3所示，Ghost瓶颈模块由两个Ghost模块堆叠而成。第一个Ghost模块用于增加特征层的通道数，第二个Ghost模块用于压缩特征层的通道数以匹配捷径连接。引入Ghost瓶颈模块到PAN结构中作为多层之间特征融合的基础模块，不仅降低了PAN结构的计算量，而且避免了大量使用1×1卷积导致网络深度过浅、感受野不足的问题，使得网络更适合航摄图像的目标检测。

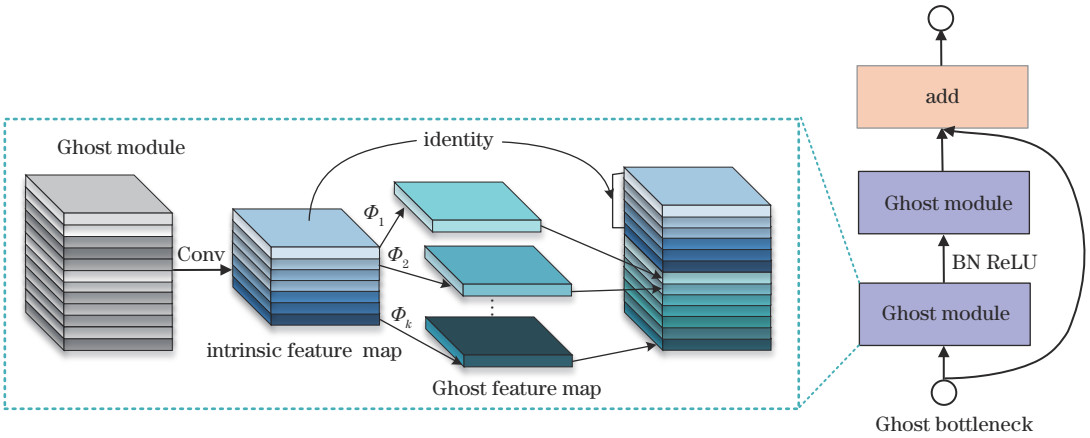

图3　Ghost瓶颈模块

Fig. 3　Ghost bottleneck module

2.5　SimOTA标签匹配和损失函数

现有的目标检测模型通常预定义一个交并比(IoU)作为阈值来划分正负锚框。由于航摄图像目标分布密集且遮挡严重，固定的阈值难以适应不同的检测场景，需要根据检测目标重新设计样本匹配策略或者调整参数。针对此问题，本文采用SimOTA[17]进行动态标签匹配，使得标签匹配不再只依赖静态的先验信息，而是能根据当前网络输出的预测信息进行动态最优匹配。SimOTA的流程如表2所示。在使用SimOTA方法对预测框和真实框进行匹配之前，为了提高匹配效率，首先通过中心先验确定候选区域，初步筛选出n个预测框作为候选框；然后计算这n个候选框和真实框的损失和IoU，得到代价矩阵$\boldsymbol{E}_{cost}$和参数k，对于每个真实框，选择$\boldsymbol{E}_{cost}$最小的k个候选框作为正样本。

经过标签匹配后，所提模型采用广义交并比(GIoU)损失函数计算真实框和对应的正样本预测框之间的位置损失。GIoU损失的计算公式为

$$L_{GIoU}=1-\frac{|A\cap B|}{|A\cup B|}+\frac{|E/(A\cup B)|}{|E|},\tag{6}$$

式中：A代表目标的真实框；B代表目标的预测框；E代

表2　SimOTA标签匹配实施流程

Table 2　Implementation flow of SimOTA label assignment

Algorithm 1: simplify optimal transport assignment (SimOTA)
Input: n is the number of initial selected candidate boxes C, m is the number of ground truth objects in image $\boldsymbol{Y}$, P_j^{class} is predicted class score for candidate box a_j, P_j^{box} is predicted bounding box for a_j (j=1, 2, …, n), G_i^{class} is ground truth class for ground truth g_i, G_i^{box} is bounding box for g_i (i=1, 2, …, m), ε=3
Output: get k candidate boxes as positive samples of g_i
1 calculate class loss: L_{ij}^{class}=BCELoss (P_j^{class}, G_i^{class})
2 calculate regression loss: L_{ij}^{reg}=GIoULoss (P_j^{box}, G_i^{box})
3 calculate cost: $c_{ij}=L_{ij}^{class}+\varepsilon L_{ij}^{reg}$
4 select the top10 candidate boxes with the highest IoU for each g_i
5 sum these 10 IoU and take integers to get the top k for each g_i
6 for i=1 to m do
7　　select the top k candidate boxes with the least cost within a fixed center region for g_i
8　　if a candidate box a_j matches multiple ground truths then select the least cost ground truth matching a_j
9　　else a_j is selected as a positive sample of g_i

表包含 A 和 B 的最小矩形框。

受复杂背景的影响，航摄图像目标检测存在大量的负样本和困难样本，而交叉熵损失函数赋予正负样本、难易样本的权重一样，不利于模型训练。针对正负样本不平衡问题，Focal loss[8]在交叉熵损失函数中赋予正样本权重因子 α，以调整正负样本的损失。同时，针对难易样本不平衡问题，Focal loss在交叉熵损失函数中又添加了一个调制因子 γ，以降低简单样本的损失，使模型专注于训练困难样本。基于此，采用Focal loss计算置信度损失，计算公式为

$$L_{\mathrm{conf}}=-\alpha\hat{C}_i^j\left(1-C_i^j\right)^{\gamma}\ln C_i^j-(1-\alpha)\left(1-\hat{C}_i^j\right)\left(C_i^j\right)^{\gamma}\ln\left(1-C_i^j\right),\tag{7}$$

式中：$\hat{C}_i^j$ 代表真实框的置信度；C_i^j 代表预测框的置信度。

类别损失采用交叉熵损失函数计算，可以表示为

$$L_{\mathrm{class}}=-\hat{P}_i^j\ln P_i^j-\left(1-\hat{P}_i^j\right)\ln\left(1-P_i^j\right),\tag{8}$$

式中：$\hat{P}_i^j$ 表示真实框的类别分数；P_i^j 表示预测框的类别分数。

联合了位置损失、置信度损失和类别损失的复合损失函数的计算公式为

$$L=\lambda L_{\mathrm{GIoU}}+L_{\mathrm{class}}+L_{\mathrm{conf}},\tag{9}$$

式中：λ 为权重参数，$\lambda=5$。

3 分析与讨论

3.1 实验环境、数据集与训练设置

实验中模型训练环境和测试环境为：Windows10系统、Intel Core i9-9900KF CPU @3.60 GHz、GPU型号为NVIDIA GeForce RTX 3090 24 GB、PyTorch 1.7.0版本、CUDA版本11.0。

使用VisDrone数据集[1]和NWPU VHR-10数据集[20]对所提目标检测网络模型进行评估。VisDrone数据集由无人机拍摄，涵盖了不同地区的多种自然场景，小目标、遮挡目标占比丰富。VisDrone数据集标注了10类常见目标，包含10209张静态图像，其中训练集有6471张图像，验证集有548张图像，测试集有3190张图像（本文测试集为VisDrone-DET-test-dev，共包含1610张图像）。NWPU VHR-10数据集包含目标的图像有650张，利用亮度调节、对比度变化、旋转等方法对数据集进行扩增，扩增后训练集有1560张图像，测试集和验证集有390张图片。

在ImageNet数据集上对网络进行预训练[15]，然后在VisDrone和NWPU VHR-10数据集上使用随机梯度下降（SGD）对网络进行训练，SGD动量为0.9，权重衰减为0.0005，置信度阈值设置为0.5，非极大值抑制的IoU阈值设置为0.3。在前5个epoch中，学习率从0.00001线性增加到0.0001，后续的145个epoch中，学习率由余弦学习率调度。参考文献[8]，损失函数中 α 设置为0.25，γ 设置为2.0。

3.2 在VisDrone数据集上的实验

为了验证所提模型进行航摄图像目标检测的有效性，对所提模型与二阶段模型Faster R-CNN[7]、一阶段模型YOLOv4[21]、无锚检测模型CenterNet[22]、轻量级检测模型YOLOv4-tiny[21]在VisDrone数据集上进行对比实验。采用AP（IoU阈值为0.5∶0.05∶0.95内的10个mAP的平均值）、AP^{50}（IoU阈值为0.5的mAP）、AP^{75}（IoU阈值为0.75的mAP）、参数量、十亿次浮点计算数（BFLOPs）和速度这6个评价指标对不同模型进行定量分析。其中Faster R-CNN模型的输入图片大小为600×1000像素，其余模型输入图片大小为960×960像素。

表3为所提模型与主流模型在VisDrone数据集上的评估结果。从表3可以看出：所提模型在检测精度AP、AP^{50}和AP^{75}方面的表现优于模型Faster R-CNN、CenterNet、YOLOv4-tiny，接近深度神经网络模型YOLOv4；Faster R-CNN由于利用手工设计锚框并只在单一深层特征上进行预测，不能很好地适应航摄图像目标尺度多变、小目标众多的实际情况，因此在VisDrone数据集上的检测效果较差；YOLOv4模型得益于深层特征提取网络和聚类分析设定锚框，在VisDrone数据集上取得了较好的检测精度，但是其设定的锚框不具有一般性，并且需要引入大量的超参数来定义这些尺寸不一的锚框，严重影响了检测速度。在参数量、计算量以及检测速度方面，所提模型仅次于轻量级检测模型YOLOv4-tiny，优于其他对比模型。值得注意的是，所提模型的参数量和计算量仅为YOLOv4的12.1%和11.6%，但是AP只比YOLOv4低1.7个百分点。这表明所提模型采用多尺度自适应感受野模块提取形变和多尺度特征，同时使用Ghost-

表3 不同模型在VisDrone数据集上的评估结果对比

Table 3 Comparison of evaluation results of different models on VisDrone dataset

Model	Backbone	AP /%	AP^{50} /%	AP^{75} /%	Parameters/10^6	BFLOPs	Speed /(frame·s^{-1})
Faster R-CNN[7]	VGG16		15.2				20.4
CenterNet[22]	ResNet50	12.4	22.7	12.4	32.67	246.01	45.2
YOLOv4[21]	CSPDarknet53	16.8	31.2	16.7	64.36	321.30	28.8
YOLOv4-tiny[21]	Tiny Darknet	10.6	19.8	10.4	6.06	36.99	65.2
Proposed model	MobileNetV3	15.1	26.6	15.5	7.79	37.35	59.9

PAN对深层特征和浅层特征进行融合，并利用SimOTA来动态筛选候选框，在保持较低参数量和计算量的情况下，很大程度上增强了对各类目标的特征表达能力，提高了航摄图像目标检测精度。

P-R曲线描述精确率和召回率的关系，P-R曲线包围的面积越大，表征着模型的检测效果越好。图4(a)展示了所提模型对VisDrone数据集上10类目标的P-R曲线，10类目标中公共汽车和小轿车具有相对较大的像素，因此对它们的检测精度较高，P-R曲线包围的面积也较大。图4(b)和图4(c)展示了不同模型对公共汽车和小轿车两类目标的P-R曲线，由图4(b)和图4(c)可见，所提模型对公共汽车和小轿车这两类目标的检测效果优于Faster R-CNN、CenterNet和YOLOv4-tiny这3种检测模型，接近深度神经网络模型YOLOv4。

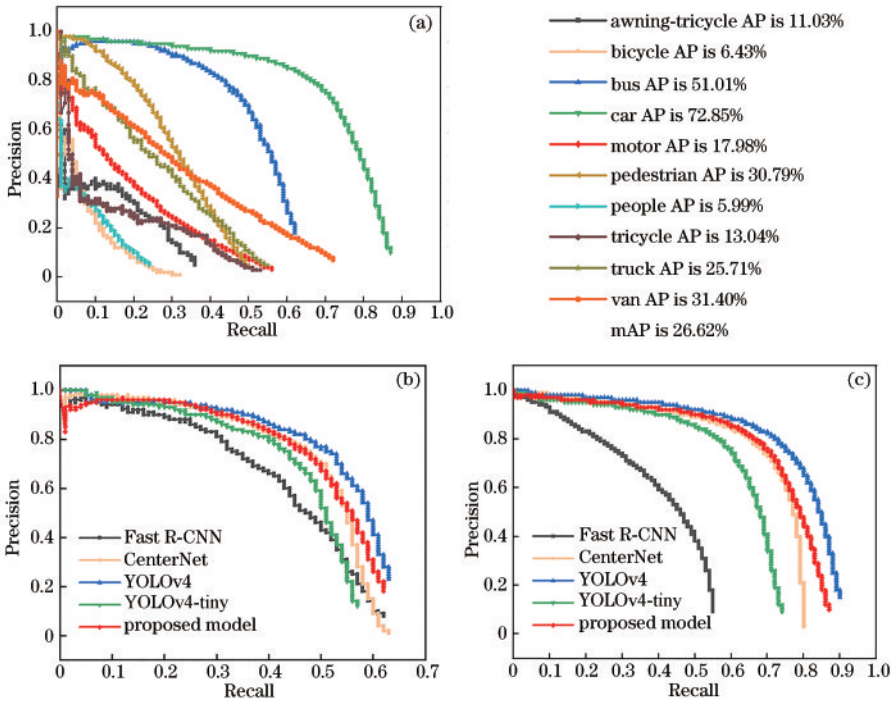

图4　P-R曲线对比结果。(a)所提模型对十类目标的P-R曲线；(b)对公共汽车的P-R曲线；(c)对小轿车的P-R曲线

Fig. 4　Comparison result of P-R curve. (a) P-R curves of the proposed model for ten classes of objects; (b) P-R curve for the bus; (c) P-R curve for the car

为了验证所设计的D-RFB和Ghost-PAN结构的有效性，以及各模块对检测效果的提升，在VisDrone数据集上进行了消融实验，结果如表4所示。引入D-RFB后，所提模型的检测精度AP^{50}提升了1.5个百分点，并且与原始的RFB相比，D-RFB使所提模型的检测精度AP^{50}从22.6%提升到23.4%，这说明可变形卷积对航摄图像目标的特征表达能力更强，融合了可变形卷积的感受野模块更适用于尺度和形态多变的航摄图像目标的检测；引入Ghost-PAN结构后，所提模型的检测精度AP^{50}从20.1%提升到23.4%，同时与原始的PAN结构相比，轻量化的Ghost-PAN结构提高了所提模型的推理速度，而检测精度并没有大幅度下降；引入SimOTA和Focal loss后，模型的正负样本不均衡问题得到进一步的解决，因而检测精度AP^{50}分别提升了1.8个百分点和1.4个百分点。

图5为所提模型与Faster R-CNN、YOLOv4、

表4　消融实验结果

Table 4　Results of ablation study

MobileNetV3+ Decoupled Head	D-RFB	RFB	Ghost-PAN	PAN	SimOTA	Focal loss	AP^{50}/%	Speed/(frame·s^{-1})
√							18.6	72.3
√	√						20.1	68.2
√		√	√				22.6	61.2
√	√			√			23.6	54.3
√	√		√				23.4	59.9
√	√		√		√		25.2	59.7
√	√		√		√	√	26.6	59.9

YOLOv4-tiny模型在VisDrone数据集上不同场景下的检测效果。可以看到：Faster R-CNN和YOLOv4-tiny模型在小目标、多尺度、密集、遮挡场景下的漏检现象比较严重；YOLOv4-tiny模型在光照变化场景下存在误检，如图5(c)中矩形框标注部分所示；YOLOv4模型和所提模型在这些场景下可以检测到绝大多数目标，不存在误检的情况。从定性分析的角度，所提模型在VisDrone数据集上的检测效果仅次于YOLOv4模型，优于Faster R-CNN和YOLOv4-tiny。

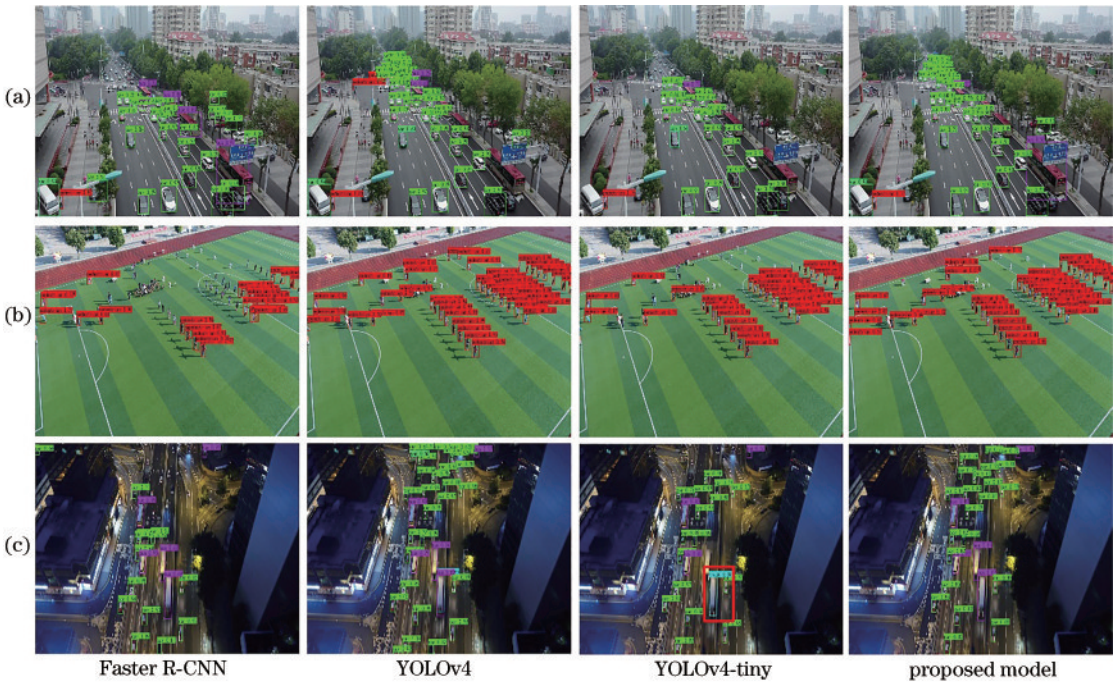

图5　不同模型在不同场景下的检测结果对比。(a)多尺度、遮挡场景；(b)小目标、密集场景；(c)光照变化场景

Fig. 5　Comparison of detection results of different models in different scenarios. (a) Multi-scale, occluded scene; (b) small object, dense scene; (c) illumination change scene

3.3　在NWPU VHR-10数据集上的实验

相比VisDrone数据集，NWPU VHR-10数据集上的目标尺寸和特征差异较大，尺度变化也更为明显。为了测试所提模型的多尺度检测能力和泛化能力，对所提模型与主流模型在NWPU VHR-10数据集上进行对比实验。其中Faster R-CNN模型的输入图片大小设定为600×1000像素，其余模型的输入图片大小设定为640×640像素，具体实验结果如表5所示。从表5可以看出，在NWPU VHR-10数据集上，所提模型的精度AP和AP^{75}最佳，AP^{50}仅次于深度神经网络模型YOLOv4，远高于轻量级检测模型YOLOv4-tiny、二阶段模型Faster R-CNN和无锚检测模型CenterNet。所提模型在高IoU阈值时的检测精度高于所有对比模型，其主要原因是所提模型在预测过程中通过中心先验确定候选区域，然后根据代价矩阵筛选正样本，提高了预测框与真实框的拟合度，进而提升了对边界框的回归精度。

表5　不同模型在NWPU VHR-10数据集上的评估结果对比

Table 5　Comparison of evaluation results of different models on NWPU VHR-10 dataset

Model	Backbone	AP /%	AP^{50} /%	AP^{75} /%	Parameters /10^6	BFLOPs	Speed /(frame·s^{-1})
Faster R-CNN[7]	VGG16		81.8				20.9
CenterNet[22]	ResNet50	45.4	84.1	40.7	32.67	109.34	55.3
YOLOv4[21]	CSPDarknet53	58.0	96.2	62.7	64.36	142.80	44.2
YOLOv4-tiny[21]	Tiny Darknet	29.8	72.9	18.0	6.06	16.44	84.3
Proposed model	MobileNetV3	59.2	94.4	64.9	7.79	16.60	79.6

表6详细列出了所提模型和其他主流模型在NWPU VHR-10数据集上对10个目标类别的检测精度。从表6可以看出，所提模型对飞机、棒球场、桥梁、田径场的检测精度均高于其他检测模型，对像素较大的物体(飞机、棒球场、田径场、储油罐、网球场)的检测精度高于95%，对像素偏小的物体(舰船、汽车)的检测精度也达90%以上。这表明所提模型具有较强的多尺度检测能力和泛化能力，适合用于对航摄图像目标的检测。

表 6　不同模型在 NWPU VHR-10 数据集上对 10 个目标类别的评估结果

Table 6　Evaluation result of different models on NWPU VHR-10 dataset for 10 classes of objects　　unit: %

Target category	Faster R-CNN	CenterNet	YOLOv4	YOLOv4-tiny	Proposed model
Airplane	97.71	99.81	99.79	99.30	99.95
Baseball diamond	94.14	91.87	95.81	90.71	98.63
Basketball court	78.38	78.45	98.39	65.47	89.49
Bridge	72.56	81.52	87.13	34.85	89.07
Ground track field	96.98	71.77	97.18	73.85	99.21
Harbor	84.01	75.08	96.75	49.18	89.68
Ship	72.80	87.67	96.52	90.37	93.02
Storage tank	81.83	92.01	96.91	86.49	96.65
Tennis court	83.44	85.33	99.95	77.12	96.40
Vehicle	56.17	77.51	93.89	61.54	91.47
mAP	81.82	84.10	96.23	72.89	94.36

图 6 展示了所提模型在 NWPU VHR-10 数据集上对飞机、储油罐、舰船、棒球场、田径场、篮球场、网球场、港口、桥梁、车辆 10 类目标的检测结果，可以看到所提模型对形变和尺度变化的目标具有良好的检测效果。在 VisDrone 和 NWPU VHR-10 两个数据集上的实验验证了所提模型在参数量、计算量和检测速度的均衡上具有优势，同时还对复杂场景下的航摄图像目标具有良好的检测精度和鲁棒性。

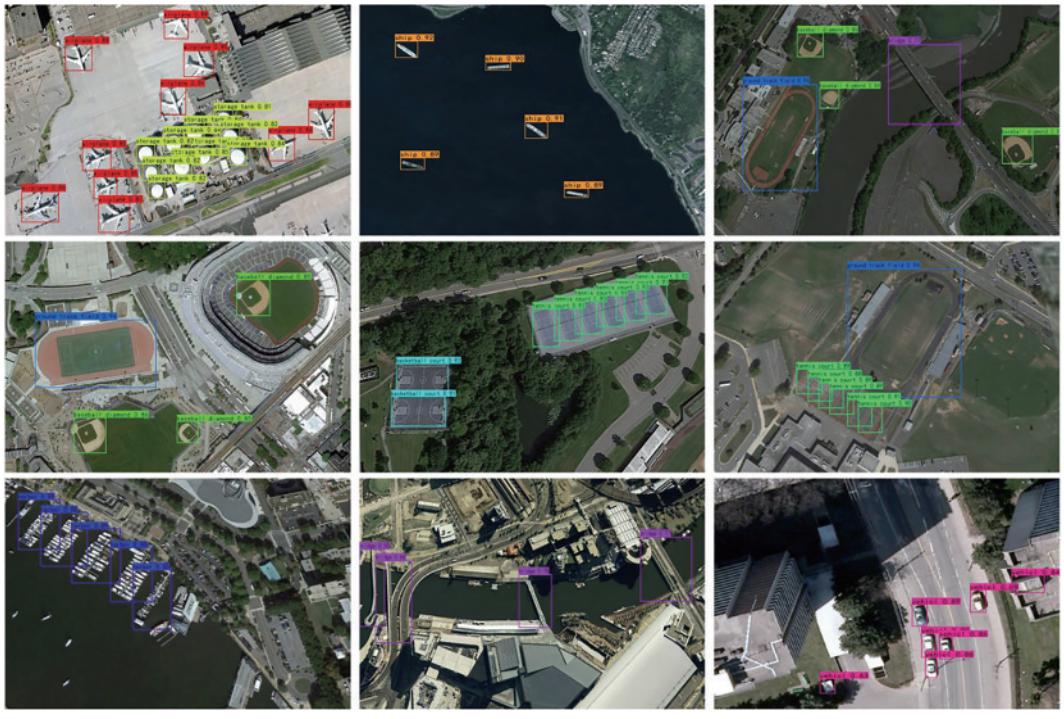

图 6　所提模型在 NWPU VHR-10 数据集上的检测结果

Fig. 6　Detection results of the proposed model on NWPU VHR-10 dataset

4　结　　论

提出了一种面向航摄图像目标检测的轻量级多尺度特征融合网络。该网络通过逐像素预测的方式，减少了模型中与 Anchor 有关的超参数；采用基于深度可分离卷积和瓶颈结构的 MobileNetV3 作为特征提取网络，并使用 Ghost 瓶颈模块替换特征融合网络中的标准卷积，减少了网络的参数量和计算量；利用融合可变形卷积的感受野模块加强特征提取，增强了网络识别形变和多尺度目标的能力；在预测过程中根据预测信息动态筛选候选框，并引入 Focal loss 使模型专注于训练正样本和困难样本，缓解了正负样本不平衡的问题。在 VisDrone 和 NWPU VHR-10 数据集上的实验结果表明，与主流的检测模型相比，所提模型在计算量、检测速度和检测精度上取得了良好的均衡，可以更好地适配机载计算设备。

参 考 文 献

[1] Zhu P F, Wen L Y, Du D W, et al. Detection and tracking meet drones challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 7380-7399.

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

[3] Micheal A A, Vani K. Comparative analysis of SIFT and SURF on KLT tracker for UAV applications[C]//2017 International Conference on Communication and Signal Processing (ICCSP), April 6-8, 2017, Chennai, India. New York: IEEE Press, 2017: 1000-1003.

[4] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, December 8-14, 2001, Kauai, HI, USA. New York: IEEE Press, 2001: 511-518.

[5] Wang X L, Cheng P, Liu X C, et al. Fast and accurate, convolutional neural network based approach for object detection from UAV[C]//44th Annual Conference of the IEEE Industrial Electronics Society, October 21-23, 2018, Washington, DC, USA. New York: IEEE Press, 2018: 3171-3175.

[6] 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.

[7] 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.

[8] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[9] 张瑞倩, 邵振峰, Portnov Aleksei, 等. 多尺度空洞卷积的无人机影像目标检测方法[J]. 武汉大学学报·信息科学版, 2020, 45(6): 895-903.
Zhang R Q, Shao Z F, Portnov A, et al. Multi-scale dilated convolutional neural network for object detection in UAV images[J]. Geomatics and Information Science of Wuhan University, 2020, 45(6): 895-903.

[10] 汪鹏, 辛雪静, 王利琴, 等. 基于YOLOv3的光学遥感图像目标检测算法[J]. 激光与光电子学进展, 2021, 58(20): 2028006.
Wang P, Xin X J, Wang L Q, et al. Object detection algorithm of optical remote sensing images based on YOLOv3[J]. Laser & Optoelectronics Progress, 2021, 58(20): 2028006.

[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] 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.
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.

[13] 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.

[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] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, Seoul, Korea (South). New York: IEEE Press, 2019: 1314-1324.

[16] Han K, Wang Y H, Tian Q, et al. GhostNet: more features from cheap operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1577-1586.

[17] Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. (2021-08-06)[2021-10-20]. https://arxiv.org/abs/2107.08430.

[18] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 764-773.

[19] Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11215: 404-419.

[20] 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.

[21] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2021-10-25]. https://arxiv.org/abs/2004.10934.

[22] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points[EB/OL]. (2019-08-06)[2021-12-21]. https://arxiv.org/abs/1904.07850.