基于动态感受野的航拍图像目标检测算法

谢学立，李传祥，杨小冈，席建祥*，陈彤
火箭军工程大学导弹工程学院，陕西 西安 710025

摘要 针对现有基于平视图像目标检测算法在航拍图像上检测精度不高的问题，提出一种具有动态感受野的单阶段目标检测算法。该算法采用 SE-ResNeXt 作为特征提取网络，在 RetinaNet 结构中添加 bottom-up 短连接通路和全局上下文上采样模块，增强了检测层特征的结构性和语义性。构造具有动态感受野的检测子网络，动态选取适当尺度的感受野特征进行目标检测。在实景航拍数据集上进行实验评测，并与相关算法作对比，结果表明改进算法在数据集上表现良好，性能指标具有明显提升，即使在光线昏暗、下视视角、斜下视视角和密集目标等场景图像中，也具有较好的检测效果。

关键词 机器视觉；目标检测；航拍图像；RetinaNet；动态感受野
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/AOS202040.0415001

Dynamic Receptive Field-Based Object Detection in Aerial Imaging

Xie Xueli, Li Chuanxiang, Yang Xiaogang, Xi Jianxiang*, Chen Tong
College of Missile Engineering, Rocket Force University of Engineering, Xi'an, Shaanxi 710025, *China*

Abstract The accuracy of existing image-based methods for aerial imaging of flat-view images is limited. In this paper, a dynamic receptive field-based single-stage object detection algorithm is proposed to address this problem. First, the feature pyramid network is constructed by using SE-ResNeXt. This network is used as the backbone network to extract features efficiently. A bottom-up short connection path and a global context upsampling module are proposed to enhance the structural and semantic features of the detection layer. A dynamic receptive field-based detection subnet is designed to dynamically select the receptive field of an appropriate scale for object detection. Experimental evaluation is conducted on a realistic aerial dataset, and the results are compared with those of other related algorithms. The results show that the improved algorithm performs better on the dataset, and the performance score is evidently increased. It also exhibits good detection capability in scene images such as dim light, down view, oblique view, and dense objects.

Key words machine vision; object detection; aerial image; RetinaNet; dynamic receptive field
OCIS codes 150.1135; 110.2970; 100.4996; 100.3008

1 引 言

随着无人机摄影技术的发展，无人机航拍被广泛应用到安全监控[1]、森林防火[2]、交通巡查[3]等领域。航拍图像数据量迅速增长，促使航拍图像处理向智能化发展，基于航拍图像的目标检测迎来广阔的发展前景。

航拍图像目标检测是目标检测的重要分支之一。传统的目标检测算法主要采用滑动窗口范式，使用 HOG[4]、SIFT[5]、Haar[6]等人工设计的特征进行目标检测。由于传统算法未能利用深层语义的特征，其检测结果鲁棒性较差，泛化能力较弱。现阶段主流的目标检测算法都采用深度学习方案，深度学习具有自动提取特征的能力，有利于构建高性能的目标检测器。基于深度学习的目标检测算法可以分为两阶段算法和单阶段算法。两阶段算法以 Faster R-CNN[7]、R-FCN[8]等为主，包括区域提取和感兴趣区域分类两个阶段；单阶段算法以 YOLO 系[9-11]、SSD 系[12-14]和 RetinaNet[15]等为主，能够同时输出目标的位置信息和分类信息。一般认为：两阶段算法具有更高的检测精度，单阶段算法具有更快的运行速度。

收稿日期：2019-08-29；**修回日期**：2019-10-08；**录用日期**：2019-11-06
基金项目：国家自然科学基金(61867005，61763040，61703411，61503009，61574049)
* **E-mail**：xijx07@mails.tsinghua.edu.cn

当前目标检测的研究对象主要以人眼视角的平视图像为主，对俯视视角的航拍图像研究较少，而航拍图像观测到的目标特征类型和位置分布与平视图像有较大差异，直接将平视图像上的目标检测算法用于航拍图像，检测精度受限，必须作进一步改进以适应航拍图像特性。欧攀等[16]提出一种结合空间变换网络与 Faster RCNN 的目标检测框架，提高了算法对遥感图像目标旋转变化的检测鲁棒性。王俊强等[17]通过引入多特征融合改进 SSD，提高了 SSD 对航空图像小目标的检测性能。Liang 等[18]在 SSD 框架中添加由反卷积构建的多尺度特征辅助检测层，并引入空间上下文分析对检测结果进行推理，提高了无人机航拍图像小目标的检测精度。现阶段应用于航拍图像的检测算法，大都是通过多特征融合结构来提高小目标检测效果，很少从自适应感受野的角度进行研究。而感受野是卷积神经网络的重要属性，卷积层感受野的大小关系着目标提取特征的完整性与鲁棒性，尤其对小目标的影响更大。此外，多特征融合在一定程度上也是利用多卷积感受野来聚合上下文。当前大部分算法都是通过改进 SSD、YOLO 和 Faster RCNN 进行航拍图像目标检测，在速度和精度上均不能做到很好的权衡。RetinaNet 算法具有单阶段检测算法的运行速度的同时，还具有与两阶段检测算法相媲美的精度，更适合作为航拍图像目标检测的基线模型。

本文从感受野的角度出发，针对航拍图像目标尺度变化大，小目标数量较多，目标分布疏密度不均的特点，改进 RetinaNet，提出一种具有动态感受野（DRF）的单阶段目标检测算法，用于航拍图像的目标检测，称为 DRF-RetinaNet。与已有文献相比，本文主要的贡献点包括：

1）对典型航拍数据集进行数据分析，提出一组高效的锚框（anchor box）参数和与之相适的 Focal Loss 参数，以提高小目标的检测效果；

2）设计动态感受野检测子网络，动态调整检测特征的感受野，增强目标检测性能；

3）提出全局上下文上采样模块，提高网络中 top-down 特征融合的效益；

4）引入 bottom-up 短连接通路，丰富高层特征结构信息。

2　RetinaNet 介绍

RetinaNet 是由 Facebook AI 团队于 2018 年提出的一个高效的单阶段目标检测算法。算法采用 ResNet-FPN 作为骨干网络，并针对单阶段目标检测器存在的样本类别失衡问题，提出 Focal Loss 以调节难易样本的 loss 贡献度，有效缓解了样本失衡对检测性能的影响，达到了接近两阶段目标检测算法的检测精度。

RetinaNet 采用 ResNet 构建 FPN 特征提取网络，FPN 具有 top-down 结构，能将上层语义特征与底层结构特征融合。作者在 ResNet-FPN 的基础上，增加了 P6、P7 辅助检测层，构建了 P3-P7 五层检测特征层。RetinaNet 主要结构如图 1 所示。

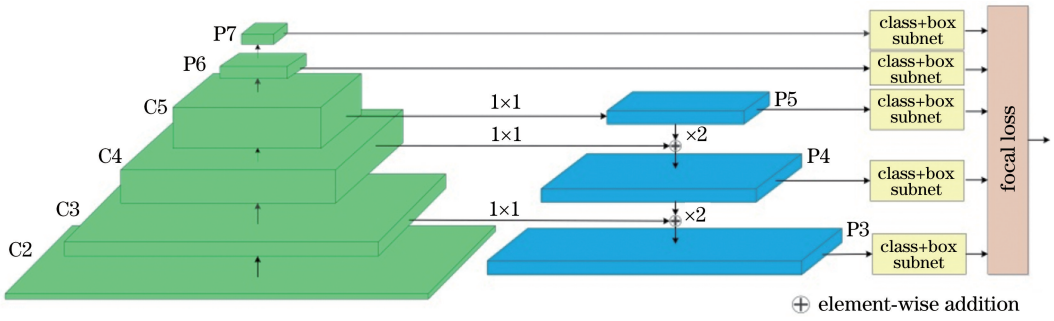

图 1　RetinaNet 结构图

Fig. 1　Structure of RetinaNet

采用分层检测的思想设计锚框，在低特征层上用小尺度锚来检测小目标，随着层级升高，逐级增大锚尺度使之适合更大目标的检测。为保证锚框的采样密度和与目标框的匹配率，在每个层级内设置{1/2，1，2/1}三种尺度和{2^0，$2^{1/3}$，$2^{2/3}$}三种长宽比的锚框。

分析了单阶段检测器精度不高的内在原因是存在样本类别不平衡。将 Focal Loss 函数作为分类损失函数，在交叉熵损失函数中添加动态调节因子，来减小易分类样本损失，增大难样本损失，从而引导优化器更关注于难样本的优化。Focal Loss 的定义为

$$F(p_t)=-\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{1}$$

$$p_t=\begin{cases}p, & y=1\\ 1-p, & \text{other}\end{cases}, \tag{2}$$

式中：y 为标签值；p 为模型预测值；α_t 为线性调节

因子，γ 为指数调节因子，两者属于超参数。

3　DRF-RetinaNet 算法

为使 RetinaNet 进一步适用于航拍图像目标检测，本文提出 DRF-RetinaNet 算法，其组件主要包括由 SE-ResNeXt-50 构成的 FPN 特征提取网络、bottom-up 短连接通路、GCU 融合模块和 DRF 检测子网络。

3.1　基于 SE-ResNeXt 的特征提取网络

SE-ResNeXt 由 ResNeXt[19] 模块和 SE[20] 模块组成。如图 2 所示，左侧为 ResNeXt 模块，右侧为 SE 模块，两模块组成一个基础的类残差模块。整个网络按照 ResNet 结构范式堆叠基础模块。本文采用 SE-ResNeXt 构建 FPN，作为检测特征提取网络。

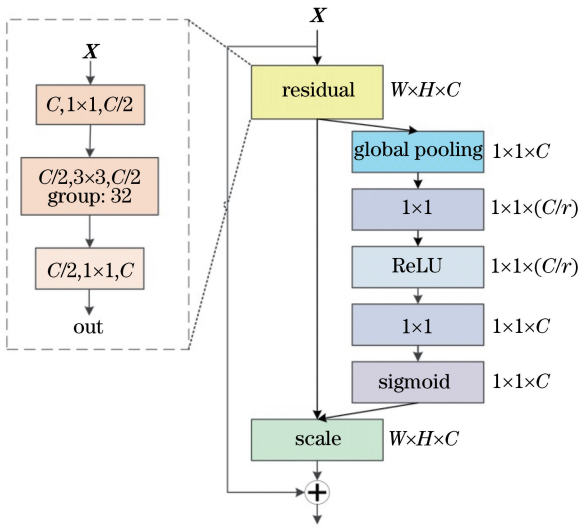

图 2　SE-ResNeXt 模块内部结构

Fig. 2　Internal structure of dual attention SE-ResNeXt module

ResNeXt 是结合 ResNet 与 Inception 思想设计的高效特征提取网络，采用了多分路卷积和残差连接，并巧妙融入了分组卷积，避免了经典 Inception 需要专门设计的问题。ResNeXt 模块为 ResNeXt 网络的基本堆叠单元。

SE 模块是一个轻量、有效的通道间注意力模块，用于标定卷积核各通道间的权重。SE 模块采用挤压-激励操作提取通道间注意力。设输入张量为 $\boldsymbol{X}\in\mathbb{R}^{W\times H\times C}$，挤压操作是对特征图中各个通道进行全局平均池化，得到一个长度为 $1\times1\times C$ 的全局信息描述符，该描述符表示各通道特征整体响应的相对强弱。设输入张量为 $\boldsymbol{X}\in\mathbb{R}^{W\times H\times C}$，挤压操作描述为

$$\boldsymbol{Z}=\sum_{k=1}^{C}\left[\frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\boldsymbol{X}(i,j)\right],\tag{3}$$

式中：$\boldsymbol{X}$ 表示输入特征图；$\boldsymbol{Z}$ 为全局特征描述符。

激励操作是采用全连接对全局信息描述符进行仿射变换，全面捕获通道依赖性。描述为

$$\boldsymbol{S}=\sigma\{f_{\mathrm{c}}'\{\delta[f_{\mathrm{c}}(\boldsymbol{Z})]\}\},\tag{4}$$

式中：f_{c} 为含有 C/r 个单元的全连接层；$\delta[\cdot]$ 表示 ReLU 激活函数；f_{c}' 表示含有 r 个单元的全连接层；$\sigma\{\cdot\}$ 表示 sigmoid 激活函数；$\boldsymbol{S}$ 为描述各通道权重赋值的张量，$\boldsymbol{S}\in\mathbb{R}^{1\times1\times C}$。

最后，将得到的通道间权向量与各通道对应相乘，输出 $\tilde{\boldsymbol{X}}$ 为

$$\tilde{\boldsymbol{X}}=\boldsymbol{X}_{\mathrm{a}}\cdot f_{\mathrm{e}}(\boldsymbol{S}_{\mathrm{a}}),\tag{5}$$

式中：$\boldsymbol{X}_{\mathrm{a}}$ 为输入特征图的单个通道特征；$\boldsymbol{S}_{\mathrm{a}}$ 为对应通道特征的权重；$f_{\mathrm{e}}(\cdot)$ 表示复制扩展操作。

3.2　Bottom-up 短连接通路

在 RetinaNet 中，检测层 P5 由 C5 特征层经过卷积得来。而 C5 层处于特征提取网络的顶层，经过了多层卷积和空间池化，其所含特征的语义性强而结构性弱。目标检测是目标定位与分类的多任务模型，需要同时包含足够语义信息和结构信息的特征才能提高检测的精度。

本文从 C3 层引入 bottom-up 短连接通路[21]，将底层 C3 特征与 P5 层进行融合，增强 P5 层特征的空间表达能力，使得在 top-down 过程中能向下传递更多结合语义指导的空间位置信息。采用“5×5_stride:4”卷积对 C3 特征层进行下采样处理，再与 P5 层进行“像素加”融合。连接示意图如图 3 所示。

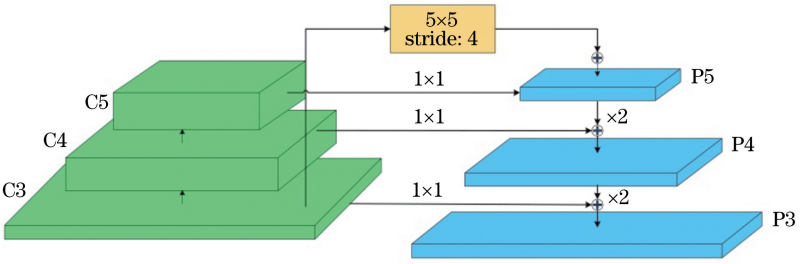

图 3　Bottom-up 短连接示意图

Fig. 3　Bottom-up short connection

3.3　全局上下文上采样融合模块

RetinaNet 采用 FPN 结构提取检测特征，FPN 通过 top-down 结构将上层特征与下层特征进行融合，实现语义信息的向下传播。Top-down 结构能够补充底层高分辨率特征所缺失的语义分类能力，从而提高了网络对小目标的检测能力。原始 RetinaNet 的 top-down 融合模块如图 4 所示。该模块中，高层特征经过双线性插值上采样，而后与低层特征进行“像素加”融合，最后经过 3×3 卷积去除上采样混叠。“像素加”融合模块表示为

$$\boldsymbol{P}=\varphi_{3\times3}[\varphi_{1\times1}(\boldsymbol{X}_{\mathrm{L}})+\theta_{\mathrm{up}}(\boldsymbol{X}_{\mathrm{H}})],\tag{6}$$

式中：$\boldsymbol{X}_{\mathrm{L}}$ 为低层特征；$\boldsymbol{X}_{\mathrm{H}}$ 为高层特征；$\varphi(\cdot)$ 表示卷积运算；$\theta_{\mathrm{up}}(\cdot)$ 表示上采样操作；$\boldsymbol{P}$ 表示融合后的输出特征。

“像素加”融合的先验约束要求输入特征图对应通道的特征具有语义类似，而高层特征经过多层卷积、池化和非线性激活后，其语义层次更加抽象，若直接将不同语义层级的特征图对应通道像素相加，会造成一定特征损失，影响语义信息传递效率。为更有效地利用高层语义特征，受文献[22-23]的启发，本文采用全局上下文上采样(GCU)模块进行 top-down 特征融合，其结构如图 5 所示。在“像素加”融合模块中加入一条高层特征注意力引导支路，在进行“像素加”融合之前对底层特征进行语义改造，减小了高低层特征之间的语义差异。GCU 模块表示为

$$\boldsymbol{P}_{\mathrm{g}}=\varphi_{3\times3}[\phi_{\mathrm{g}}(\boldsymbol{X}_{\mathrm{H}})\cdot\varphi_{1\times1}(\boldsymbol{X}_{\mathrm{L}})+\theta_{\mathrm{up}}(\boldsymbol{X}_{\mathrm{H}})],\tag{7}$$

式中：$\phi_{\mathrm{g}}(\cdot)$ 表示高层特征注意力编码运算。引入高层信息二次项有助于增强非线性表达能力。

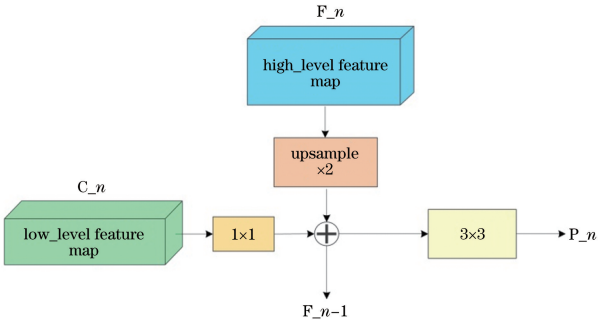

图 4　“像素加”融合模块结构图

Fig. 4　Structure of pixel-wise addition module

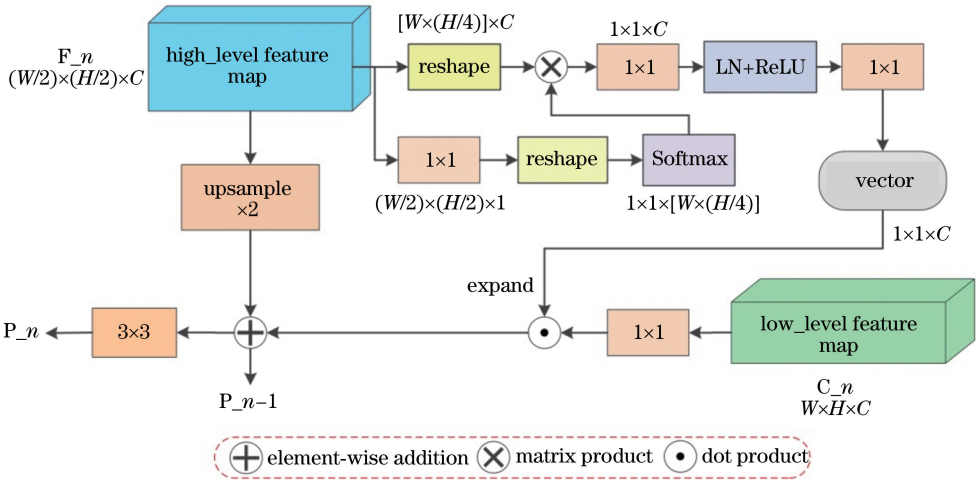

图 5　GCU 模块结构图

Fig. 5　Structure of GCU module

GCU 采用 GCNet[23] 进行高层特征注意力编码。首先采用精简的非局域模块代替全局平均池化层，进行高层特征各通道的全局上下文信息提取，为每个通道生成一个信息描述符；然后利用类似 SE 模块中的激励-挤压操作，对描述符向量进行仿射变换；最后使用生成的描述符向量指导底层特征进行语义改造。设输入特征图某个通道特征为 $\boldsymbol{x}$，该通道特征对应的输出为 $\boldsymbol{y}$，则高层特征注意力编码运算 ϕ_{g} 可表示为

$$\begin{cases}\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{W}_{v2}\times\delta[m]\\ m=\phi\left[\boldsymbol{W}_{v1}\times\sum\limits_{j=1}^{N_{\mathrm{p}}}\exp(\boldsymbol{W}_k\boldsymbol{x}_j)\Big/\sum\limits_{m=1}^{N_{\mathrm{p}}}\exp(\boldsymbol{W}_k\boldsymbol{x}_m)\right]\end{cases},\tag{8}$$

式中：$\boldsymbol{W}_k$ 和 $\boldsymbol{W}_v$ 表示线性转换矩阵，由 1×1 卷积实现；$\phi[\cdot]$ 表示层标准化(LN)操作；N_{p} 为该层所含特征的通道数。

3.4　DRF 检测子网络

不同尺度目标的检测所需的感受野不同，过大

的感受野会引入更多的噪声，干扰目标检测；过小的感受野，卷积核只能处理目标的局部信息，缺少上下文感知，也不利于目标的检测。文献[24]中指出特征感受野与检测目标的尺度间存在正相关关系。

RetinaNet 基于 anchor 机制实现目标检测，通过在 P3～P7 检测层上设置尺度逐层加倍的锚框，以实现目标多尺度分层检测。此外，为进一步提高锚框对检测目标的匹配率，在各层级内添加额外的小尺度锚框来提高采样密度。RetinaNet 各检测层特征的感受野大小随网络结构而固定，但先验的锚框尺度是变化的。对于小目标，可能存在层级感受野与锚框严重不匹配的情况。

基于此，本文设计了一个具有动态感受野的检测子网络，用于不同检测层的目标尺度适应检测，检测子网络结构如图 6 所示。首先使用 DRF 模块调节检测层特征感受野，然后采用两条对称的带有瓶颈结构的支路完成特征的进一步抽象处理，最后利用卷积层进行类别和位置信息输出。在不同层级检测层，DRF 模块能够动态选择目标尺度相关的感受野，进行目标检测。使用瓶颈结构可以实现参数压缩的同时，进一步提高特征的非线性表达能力。

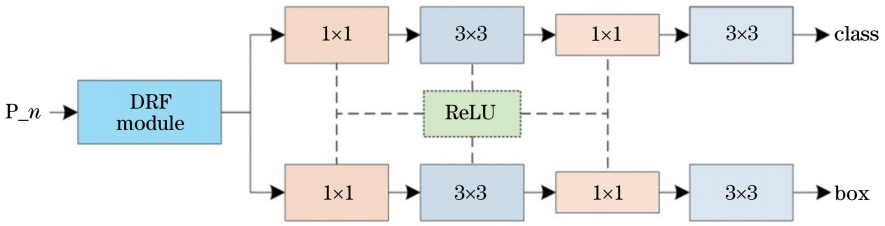

图 6　检测子网络结构图

Fig. 6　Structure of object detection subnet

DRF 模块如图 7 所示。模块由两个相同的子模块串联而成，并采用瓶颈结构设计来降低参数。DRF 子模块具有一定的感受野调节能力，主要由多支路卷积和通道选择模块组成。

1）采用空洞卷积设计三条卷积支路，获三种感受野特征，并以通道叠加方式进行聚合，此时特征图通道内具有三种感受野特征。设输入特征图 $\boldsymbol{X}\in\mathbb{R}^{W\times H\times C}$，此过程可表示为

$$\hat{\boldsymbol{X}}=\varphi_{1\times1}(\boldsymbol{X}),\hat{\boldsymbol{X}}\in\mathbb{R}^{W\times H\times C_1},\tag{9}$$

$$\boldsymbol{X}'=C_{\mathrm{at}}[\varphi_1(\hat{\boldsymbol{X}})_{W\times H\times C_2},\varphi_2(\hat{\boldsymbol{X}})_{W\times H\times C_2},\varphi_3(\hat{\boldsymbol{X}})_{W\times H\times C_2}],\boldsymbol{X}'\in\mathbb{R}^{W\times H\times C_1},\tag{10}$$

式中：$\varphi_{1\times1}$表示 1×1 卷积；$C_{\mathrm{at}}[\cdot]$表示 concatenate 通道堆叠操作；$\varphi_1(\cdot)$，$\varphi_2(\cdot)$，$\varphi_3(\cdot)$表示生成三种尺度感受野的多分支卷积运算。

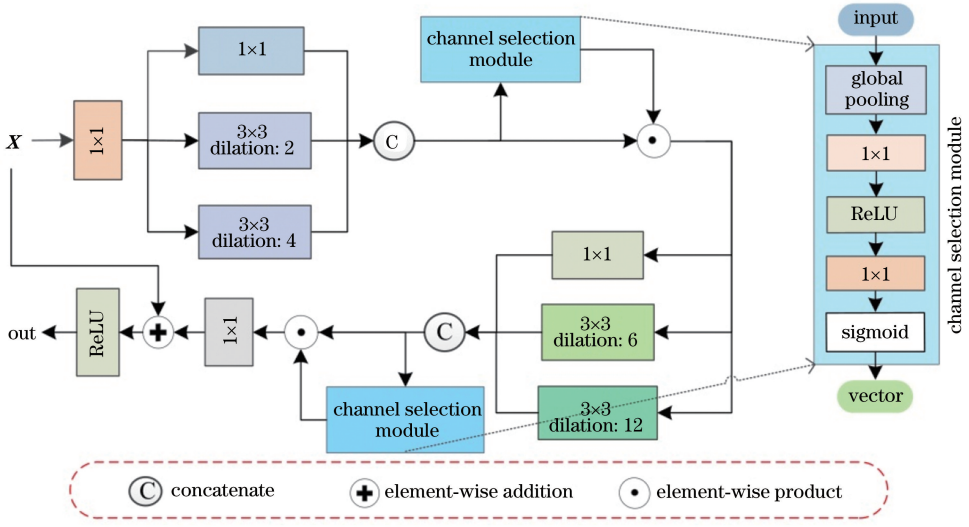

图 7　DRF 模块结构图

Fig. 7　Structure of DRF module

2）使用通道选择模块对多感受野特征图进行通道标定，生成通道调节权重向量。选用 SE 模块生成通道权重向量。由于通道标定采用软注意力机制，不同尺度感受野间将进行软切换，充分构建了上

下文信息。

$$\boldsymbol{V}=C(\boldsymbol{X}'),\ \boldsymbol{V}\in\mathbb{R}^{1\times1\times C_1},\tag{11}$$

式中：$C(\cdot)$表示类 SE 模块运算；$\boldsymbol{V}$ 表示通道调节权重向量。

3）使用 1×1 模块进行通道融合，并采用残差连接保证模块的有效性，输出一个具有动态感受野的特征图，此过程可表示为

$$\boldsymbol{Y}'=\varphi'_{1\times1}(\boldsymbol{X}'\cdot\boldsymbol{V}),\ \boldsymbol{Y}'\in\mathbb{R}^{W\times H\times C},\tag{12}$$

$$\boldsymbol{Y}=\delta[\boldsymbol{Y}'+\boldsymbol{X}],\ \boldsymbol{Y}\in\mathbb{R}^{W\times H\times C}。\tag{13}$$

动态感受野主要由(10)～(13)式实现，(10)式将 3 种感受野的特征结合成一张特征图，(11)式通过自监督的方式生成通道选择向量，(12)式实现权重向量与 1×1 卷积结合，这样在整合通道时，能够根据权重标定不同通道即不同感受野特征的响应大小。

$$Y'_{i,j,l}=w_l\cdot\sum_{k=1}^{C_1}(X'_{i,j,k}\cdot V_k)=\sum_{k=1}^{C_1}[X'_{i,j,k}\cdot(V_k\cdot w_l)],\tag{14}$$

式中：$X'_{i,j,k}$表示输入特征的像素坐标；w_l 表示1×1卷积核的权值；C_1表示特征通道数量。

DRF 模块通过串联两个子模块，增大了感受野的调节范围。多个 DRF 子模块串联时可能会出现感受野偏向混叠问题，例如，感受野 1，3，5 的 DRF 子模块和感受野 1，5，7 的 DRF 子模块串联，可调节的感受野为 1，5，7，3，7，9，5，9，11，此时 5，7，9 三种感受野出现概率为 1，3，11 的两倍，这种不均匀先验分布会使得感受野调节偏向于出现概率大的感受野尺度，降低了 DRF 模块感受野的调节能力。

为避免感受野分布不平衡问题，需要对串联的多支路卷积进行专门设计，使可选择的感受野均匀分布在可调节区间。据此，本文设置两个子模块的调节感受野分别为 1，5，9 和 1，13，25，这样的设计使串联所能形成的感受野均匀分布于区间[1，33]中，呈现为公差为 4 的 9 种递增等差感受野，消除了感受野调节的先验偏向。

4　实验与分析

4.1　数据集处理与分析

VISDrone[25]数据集由天津大学 AISKYEYE 团队收集，现公开有 7018 张航拍图像，包括行人、小车、三轮车、遮阳篷三轮车等共 10 类目标的标注信息，具有斜下视和下视、白天和夜晚等多个场景图像（部分示例见图 8），是一个充满挑战的数据集。本文将 VISDrone 数据集中 10 类目标合并为 7 类，其中 people 和 pedestrian 合为 people，car 和 Van 合为 car，tricycle 和 awning-tricycle 合为 tricycle，构成 VISDrone-g 数据集。将 VISDrone-g 数据集的 7018 张图像划分为训练集、验证集和测试集，分别为 5018，1000，1000 张，并转换成 COCO[26]标准数据集格式，以便使用 COCO 工具包进行算法评测。

图 8　VISDrone-g 数据集部分示例

Fig. 8　Partial sample of VISDrone-g dataset

VISDrone-g 由无人机实景航拍图像构成，于航拍图像目标特性十分契合。本文针对 VISDrone-g 训练集中真实标注的信息进行统计分析，最终获得关于航拍图像目标的尺度和形状先验信息，以便于设计合适的锚框参数。对数据集的统计如图 9 所示。图9(a)统计了以目标框面积的开方为度量的图像目标尺度分布特性，图 9(b)统计了目标框的长宽比例分布。从图 9(a)可以看出，目标尺度分布在 32 分界线以左的数量较大，按照 COCO 对小目标的定义，面积小于 32×32 的物体为小目标，故小目标检测是该数据集检测的关键。从图 9(b)可以看出目标的长宽比例主要集中在 0.5，1.0，1.5，2.0，2.5，3.0 附近，以等高宽型和细高型框为主。

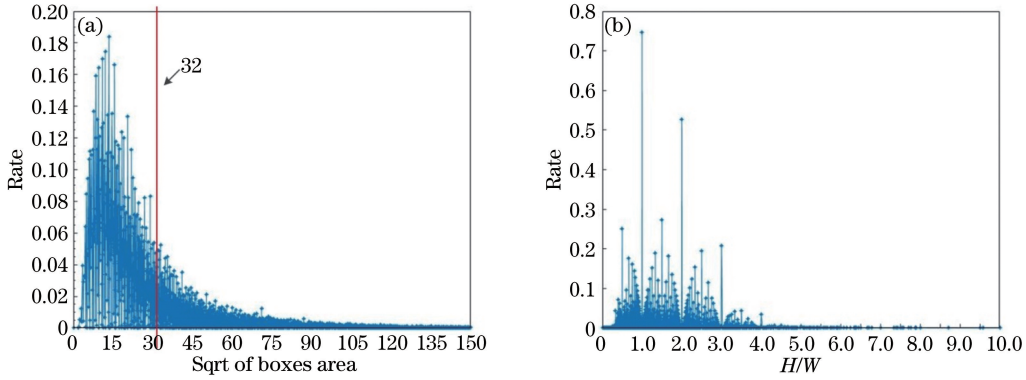

图 9　VISDrone-g 统计分析。(a)目标尺度分布特性；(b)目标框的长宽比例分布特性

Fig. 9　Statistical of the VISDrone-g. (a) Object scale distribution characteristics; (b) object frame length and width proportional distribution characteristics

4.2　实验实现细节

采用深度学习框架 pytorch-0.4 进行神经网络搭建，实验环境为 Windows 10，实验平台配置为：CPU(Xeon E5-1650 v4，3.6 GHz)；GPU(NVIDIA TITAN-X，12 GB)；运行内存为 32 GB；固态硬盘容量为 512 GB。使用 GPU 进行神经网络训练与测试。

参数设置：本文主要的超参数主要包括预设锚框的大小、长宽比和 Focal Loss 的调节因子。为匹配更多的小目标，将锚框的 base_size 设置为 16，同时在各特征像素点处生成 12 个长宽比为{0.5，1，2，3}，尺度比为$\{2^{-1/2},2^{0},2^{1/2}\}$的锚框，增大锚的采样密度。对 RetinaNet 的锚框匹配策略进行修改，保证每个锚框只匹配交并比(IOU，η_{IOU})最大的真值框。为提高锚框对小目标的匹配率，将正样本阈值下调为 0.4，负样本阈值调整为 0.3。Focal Loss 参数设为：$\alpha_{\mathrm{t}}=0.25$，$\gamma=3.0$。

实施细节：使用 SE-ResNeXt-50 在 COCO 目标检测数据集上的训练权重作为模型的预训练参数。采用 Adam＋SGD 训练策略，先使用 Adam 以 1×10^{-5}的学习率训练 10 个 epoch 使其初步收敛，再使用 SGD 以 1×10^{-6}的学习率训练 20 个 epoch，进一步降低损失。在训练阶段，保持长宽比，将输入图像放缩至短边为 608。使用 Focal Loss 作为分类损失函数，使用 Smooth L1Loss 作为边框回归损失函数。采用 Soft-NMS[27]作为后处理，设定 IOU 阈值为0.45，过滤掉重复检测的目标框。数据增强采取图像标准化、随机水平翻转和随机旋转等操作，翻转和旋转角度范围为$(-\pi/36,\pi/36)$。

4.3　实验结果及分析

4.3.1　评价准则

为全面评测算法的性能，采用 COCO 目标检测评价体系、算法运行时间和 $F_{1-\mathrm{score}}$对算法性能进行评测。COCO 评价体系包含 AP、AR_1和 $\mathrm{AP}^{\mathrm{small}}$等共 12 种度量指标，对检测器的位置检测性能和类别预测性能进行全方位评价。该指标体系中：P 表示准确率(precision)，为所有预测框中被正确识别的比率；R 表示召回率(recall)，为所有真值中被正确检测到的比率。AP、AR 分别表示 P、R 在不同 IOU 阈值条件下所有类别的平均值。其中，指标 AP 综合了 IOU 从 0.50 到 0.95 之间共 10 个阈值下的多类别平均准确率，是最有说服力的指标之一。AP_{50}是 PASCAL VOC 评价指标，计算了 IOU 在0.50下的多类别平均准确率，同理，AP_{95}是指 IOU 在0.75下的多类别平均准确率。此外，COCO 评价体系还包括算法对不同尺度目标的检测性能。按照 COCO 评价体系定义，目标面积 $S<32^2$为小目标，$32^2\leqslant S<96^2$为中等目标，$S\geqslant96^2$为大目标。通过

计算在三种尺度目标的 AP 和 AR 值，评测算法对不同尺度目标的检测性能。各指标具体说明见图 10。$F_{1-\text{score}}$综合考虑了准确率和召回率，是一个全面的衡量指标，该值越大算法性能越好。$F_{1-\text{score}}$可表示为

$$F_{1-\text{score}}=\frac{2PR}{P+R}。\tag{15}$$

基于上述评价准则，本文设置了 3 组实验对算法进行评测。

```
Average Precision (AP):
  AP                  % AP at IoU=.50:.05:.95 (primary challenge metric)
  AP^IoU=.50          % AP at IoU=.50 (PASCAL VOC metric)
  AP^IoU=.75          % AP at IoU=.75 (strict metric)
AP Across Scales:
  AP^small            % AP for small objects: area < 32^2
  AP^medium           % AP for medium objects: 32^2 < area < 96^2
  AP^large            % AP for large objects: area > 96^2
Average Recall (AR):
  AR^max=1            % AR given 1 detection per image
  AR^max=10           % AR given 10 detections per image
  AR^max=100          % AR given 100 detections per image
AR Across Scales:
  AR^small            % AR for small objects: area < 32^2
  AR^medium           % AR for medium objects: 32^2 < area < 96^2
  AR^large            % AR for large objects: area > 96^2
```

图 10　COCO 目标检测评价指标详解[26]

Fig. 10　Detailed explanation of COCO object detection and evaluation indexes[26]

4.3.2　Focal Loss 调参实验

RetinaNet 提出 Focal Loss 解决类别不均衡问题，取得了显著效果。Focal Loss 定义如(1)式。在 VISDrone-g 数据集中，小目标占大多数，能匹配的正样本锚框较少，且小目标所含的特征信息较弱，属于难分类样本。本文设置的锚框相较于原 RetinaNet，单层级内先验框的数量更多且更密集，一定程度上加重了正负样本数量失衡。因此需要更大的调节系数来缓解类别失衡。本文设置多组参数进行实验，实验结果如表 1 所示，随着参数增大，AP、AP_{50}和 $F_{1-\text{score}}$呈现先变大后减小趋势。选取其中最大 AP、AP_{50}对应的 $\alpha_t=0.25$，$\gamma=3.0$ 作为最终参数。该组参数相较于原始 RetinaNet，调节幅度更大，与前文的分析预测一致。

表 1　Focal Loss 参数实验表

Table 1　Focal Loss parameter tuning

α_t	γ	AP /%	AP_{50} /%	$F_{1-\text{score}}$
0.20	2.0	23.93	38.18	47.24
0.25	2.0	24.37	39.95	48.43
0.25	3.0	25.14	42.62	52.47
0.30	3.0	24.82	41.23	50.72

4.3.3　模块消除实验

本实验是通过加减关键模块来分析各模块对算法性能的影响，本文对 SE-ResNeXt，bottom-up，GCU 和 DRF 检测子网络等组件进行消除实验。其中 GCU 模块和原“像素加”模块作对比，DRF 检测子网络和原卷积检测子网络作对比。从表 2 可以看出：随着组件的添加，算法的性能逐步提升，与 RetinaNet* 相比，DRF-RetinaNet 的 AP 提升 6.17 个百分点，AP_{50}提升 13.97 个百分点。且同时加入 GCU 和 bottom-up 组件带来的涨点，要比两者单独带来的涨点之和更高，这说明两个组件间存在促进互补关系。Bottom-up 短连接给顶层特征带来了空间信息，GCU 融合模块提高了高层特征向底层特征的传输效率，引入这两个模块能增强 FPN 的特征信息流，有利于提取更有效的目标检测特征。

表 2　模型组件性能对比

Table 2　Performance comparison of model components

Module	Whether or not it contains					
RetinaNet*	√					
SE-ResNeXt		√	√	√	√	√
Bottom-up			√		√	√
GCU				√	√	√
DRF detection subnet						√
AP /%	18.97	20.22	21.33	22.05	23.17	25.14
AP_{50} /%	28.65	30.64	32.78	34.33	37.83	42.62

Note: * indicates that anchor parameters have been adjusted according to section 4.2.

4.3.4　算法对比实验

本实验将 DRF-RetinaNet 与 RetinaNet、参数调整的 RetinaNet、SSD、Faster R-CNN、R-FCN 和

YOLO v3 等 6 种算法进行对比，其中 RetinaNet、SSD 和 Faster R-CNN 采用 mmdetection 库（https://github.com/open-mmlab/mmdetection）实现，R-FCN、RFB-Net 和 YOLO v3 采用 GitHub-pytorch 版本实现。各算法均在训练集上训练至收敛，并在测试集进行测试。

表 3 为各检测算法在测试集上的性能表现，可以看出 DRF-RetinaNet 算法的性能提升明显，AP 达到 25.14%，AP_{50} 达到 42.62%，算法的处理速度约为 9 frame/s。调整锚框参数后的 RetinaNet 算法（RetinaNet*）性能获得提升，AP 达到 18.97%比原始 RetinaNet 增加 2.62 个百分点，说明 anchor-based 目标检测方法对图像类型具有较强的敏感性，因此在进行不同类型的检测任务前，需对相关数据进行分析，以确定合适的锚框参数。

表 4 为各算法对不同尺度目标的检测效果，可以看出 DRF-RetinaNet 对于小目标的检测准确率 AP^{small} 达到 13.62%，AP^{medium} 达到 40.34%，AP^{large} 达到 55.95%，$F_{1-score}$ 达到 52.47，显著超过其他算法，同时对其他尺度的目标检测性能也有很大提升。且调整锚框参数后的 RetinaNet 对小目标的检测效果具有一定的提高，RetinaNet 的 AP^{small} 较原始 RetinaNet 的 AP^{small} 增加了 2.55 个百分点，说明本文基于数据集统计分析设计的锚框参数适合航拍图像目标分布特性。DRF-RetinaNet 在速度上保持单阶段检测器的速度，在精度上具有明显优势，是一个有效的航拍图像目标检测算法。

表 3　各种算法性能对比

Table 3　Performance comparison of each algorithm

Method	Input size	Basebone Network	AP /%	AP_{50} /%	AP_{75} /%	AR_1 /%	AR_{10} /%	AR_{100} /%	Time /ms
Faster R-CNN	600	Resnet-50	16.72	24.32	14.15	4.21	12.47	16.65	137
R-FCN	600	Resnet-101	19.35	30.18	19.52	5.65	18.73	22.56	178
SSD	512	Vgg-16	12.23	17.29	11.54	3.71	11.22	15.41	54
RFB-Net	512	Resnet-50	14.87	22.17	12.06	4.34	13.15	17.38	75
YOLO v3	416	Darknet-53	14.75	21.86	12.17	4.12	12.93	17.41	67
RetinaNet	608	Resnet-50	16.35	23.18	13.92	4.85	14.75	18.36	85
RetinaNet*	608	Resnet-50	18.97	28.65	17.42	4.92	17.25	20.52	88
DRF-RetinaNet	608	SE-ResNeXt-50	25.14	42.62	24.71	7.82	24.22	31.24	103

Note: * indicates that anchor parameters have been adjusted according to section 4.2.

表 4　算法对不同尺度目标检测效果对比

Table 4　Performance comparison of algorithms for different scales object detection

Method	AP^{small} /%	AP^{medium} /%	AP^{large} /%	AR^{small} /%	AR^{medium} /%	AR^{large} /%	$F_{1-score}$
Faster R-CNN	7.14	24.42	36.73	10.62	26.75	41.41	33.15
R-FCN	9.85	26.13	40.25	14.57	32.71	47.79	40.67
SSD	5.85	20.03	34.07	7.63	24.97	38.68	26.41
RFB-Net	6.62	22.18	34.28	9.55	25.77	40.82	33.13
YOLO v3	6.25	22.26	36.17	9.72	25.72	40.27	32.73
RetinaNet	7.27	23.95	36.72	10.31	26.63	42.23	32.69
RetinaNet*	9.82	25.35	38.31	14.93	31.91	44.82	37.92
DRF-RetinaNet	13.62	40.34	55.95	17.42	49.97	61.53	52.47

Note: * indicates that anchor parameters have been adjusted according to section 4.2.

图 11 展示了 DRF-RetinaNet 与 RetinaNet* 的可视化对比结果。三组图中，待检测目标尺度变化区间较大，第一组图目标尺度相对较大，而其他两组图的目标尺度较小。从三组图中可以看出 RetinaNet* 存在较多的漏检现象，而 DRF-RetinaNet 能够更好地适应尺度变化，具有更好的

小目标检测性能，同时对重叠人群的检测效果也有明显提高。

从测试集中选取 4 组典型场景图像进行效果展示。图 12 为暗场景环境下目标的检测效果，图 13 为密集目标的检测效果，图 14 为斜下视图像目标检测效果，图 15 为下视图像目标检测效果。图中方框为算法检测结果。实验效果表明本文算法能够较好地检测出不同场景中特定目标，尤其是对车辆、人等多样本的目标具有较高的检测精度。但对于重叠遮挡的物体会出现误识别，比如正在骑车的人容易识别，但所骑的车和车的类别很难识别，尤其当目标尺度很小时，检测难度更大。总体来说，DRF-RetinaNet 目标检测算法对于光线、视角、目标尺度等因素变化具有一定视觉鲁棒性。

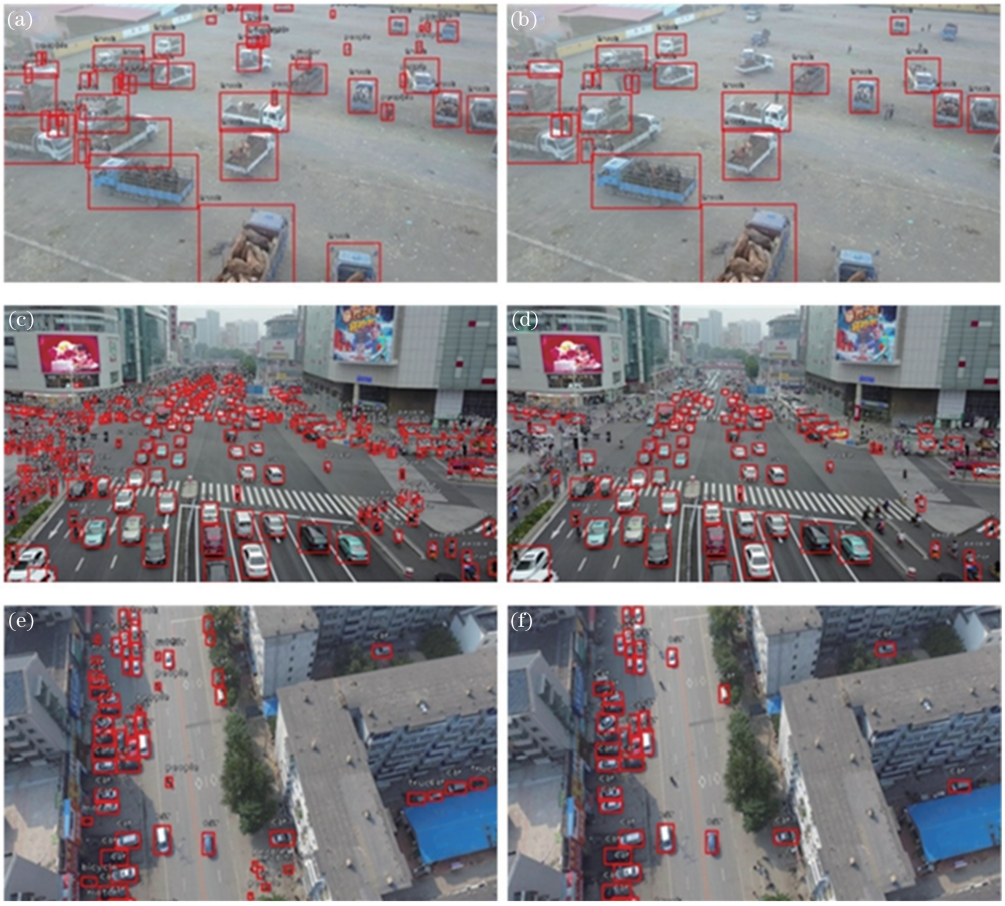

图 11　DRF-RetinaNet 和 RetinaNet* 的可视化对比。(a)(c)(e) DRF- RetinaNet 检测结果；(b)(d)(f) RetinaNet* 检测结果

Fig. 11　Visual contrast between DRF-RetinaNet and RetinaNet* . (a)(c)(e) DRF-RetinaNet′s detection result; (b)(d)(f) RetinaNet′s detection result

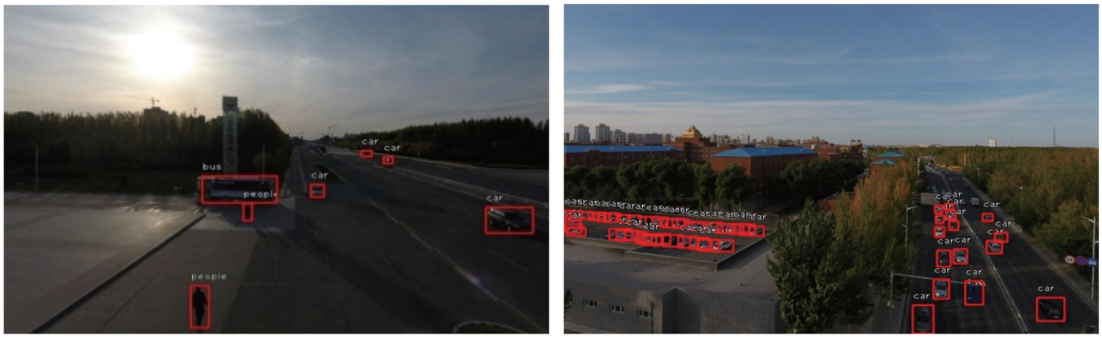

图 12　暗场景目标检测效果

Fig. 12　Detection results of dim light

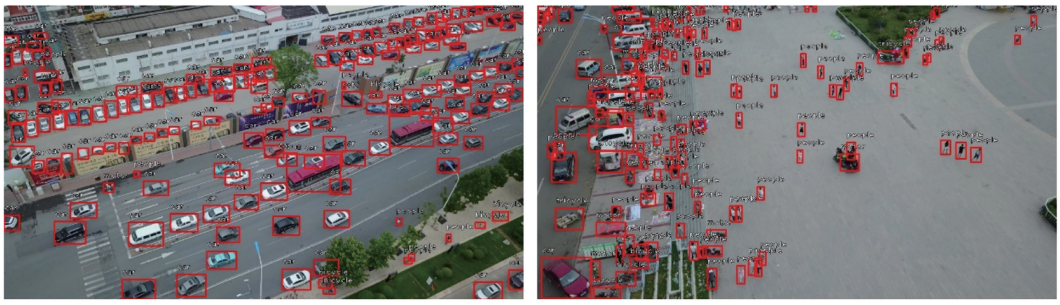

图 13　密集小目标检测效果

Fig. 13　Detection results of dense objects

图 14　斜下视图像目标检测效果

Fig. 14　Detection results of oblique view

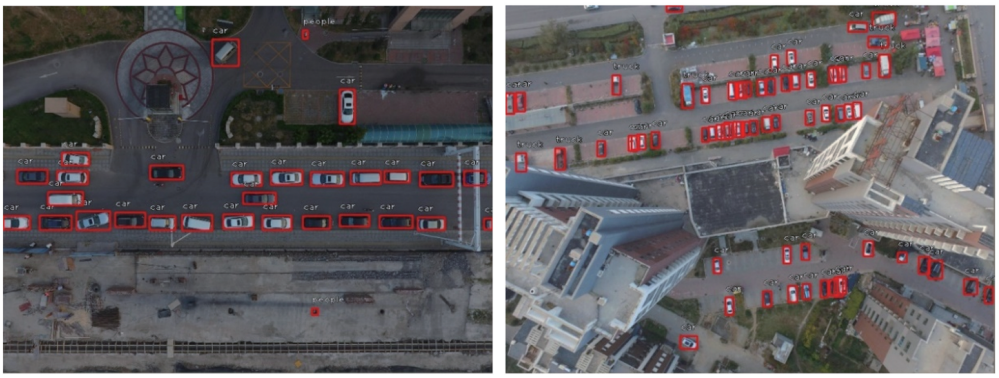

图 15　下视图像目标检测效果

Fig. 15　Detection results of down view

5　结　　论

针对现有目标检测算法在无人机航拍图像上检测精度不高的问题，改进了 RetinaNet 算法以适应于航拍图像目标检测。基于对典型实景航拍图像数据集的统计分析，发现小目标检测是航拍图像目标检测精度提升的一个瓶颈，基于此，调整锚框的形状与尺度以适应航拍图像目标，算法使用 SE-ResNeXt 提取高质量检测特征，引入 bottom-up 连接提升高层特征的结构信息，提出 GCU 模块代替“像素加”模块，增强语义特征向下传播的效率，设计基于 DRF 模块的检测子网络，实现了检测层尺度内感受野动态调整。在实景航拍数据集上进行实验评测，结果显示 DRF-RetinaNet 算法具有明显的性能优势，对不同场景航拍图像均有较好的检测效果。但算法对于小类别样本和小尺度目标的检测精度还有待进一步提高。下一步计划使用能保持高分辨率的特征提取网络，加入针对小目标检测的数据增强方法，以进一步提高检测性能。

参　考　文　献

[1]　Aguilar W G, Luna M A, Moya J F, et al. Pedestrian detection for UAVs using cascade classifiers with meanshift[C]// 2017 IEEE 11th International Conference on Semantic Computing (ICSC), January 30-February 1, 2017, San Diego, CA, USA. New York: IEEE, 2017: 509-514.

[2] Yuan C, Liu Z X, Zhang Y M. UAV-based forest fire detection and tracking using image processing techniques[C]// 2015 International Conference on Unmanned Aircraft Systems (ICUAS), June 9-12, 2015, Denver, CO, USA. New York: IEEE, 2015: 639-643.

[3] Xu Y Z, Yu G Z, Wang Y P, et al. Car detection from low-altitude UAV imagery with the faster R-CNN[J]. Journal of Advanced Transportation, 2017, 2017: 2823617.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 8588935.

[5] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[6] Viola P, Jones M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[7] 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. Canada: NIPS, 2015: 91-99.

[8] Dai J, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks[C]// Advances in Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. Canada: NIPS, 2016: 379-387.

[9] 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, 2016: 779-788.

[10] 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, 2017: 6517-6525.

[11] Redmon J, Farhadi A. Yolov3: an incremental improvement[J/OL]. (2018-04-08)[2019-08-28]. https://arxiv.xilesou.top/abs/1804.02767.

[12] 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.

[13] Fu C Y, Liu W, Ranga A, et al. Dssd: deconvolutional single shot detector[J/OL]. (2017-01-23)[2019-08-28]. https://arxiv.xilesou.top/abs/1701.06659.

[14] Li Z, Zhou F. FSSD: feature fusion single shot MultiBox detector[J/OL]. (2018-05-17)[2019-08-28]. https://arxiv.xilesou.top/abs/1712.00960.

[15] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2999-3007.

[16] Ou P, Zhang Z, Lu K, et al. Object detection in of remote sensing images based on convolutional neural networks[J]. Laser & Optoelectronics Progress, 2019, 56(5): 051002.
欧攀, 张正, 路奎, 等. 基于卷积神经网络的遥感图像目标检测[J]. 激光与光电子学进展, 2019, 56(5): 051002.

[17] Wang J Q, Li J S, Zhou X W, et al. Improved SSD algorithm and its performance analysis of small target detection in remote sensing images[J]. Acta Optica Sinica, 2019, 39(6): 0628005.
王俊强, 李建胜, 周学文, 等. 改进的SSD算法及其对遥感影像小目标检测性能的分析[J]. 光学学报, 2019, 39(6): 0628005.

[18] Liang X, Zhang J, Zhuo L, et al. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[19] Xie S N, Girshick R, Dollar P, et al. Aggregated residual transformations for deep neural networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5987-5995.

[20] 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, 2018: 7132-7141.

[21] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 8759-8768.

[22] Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation[J/OL]. (2018-11-25)[2019-08-28]. https://arxiv.xilesou.top/abs/

1805.10180.

[23] Cao Y, Xu J, Lin S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond [J/OL]. (2019-04-25) [2019-08-28]. https://arxiv.xilesou.top/abs/1904.11492.

[24] Li Y, Chen Y, Wang N, et al. Scale-aware trident networks for object detection[J/OL]. (2019-08-20) [2019-08-28]. https://arxiv.xilesou.top/abs/1901.01892.

[25] Zhu P, Wen L, Bian X, et al. Vision meets drones: a challenge [J/OL]. (2018-04-23) [2019-08-28]. https://arxiv.xilesou.top/abs/1804.07437.

[26] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M] // Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[27] Bodla N, Singh B, Chellappa R, et al. Soft-NMS: improving object detection with one line of code[C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 5562-5570.