基于多尺度特征融合的自适应无人机目标检测

刘芳, 吴志威*, 杨安喆, 韩笑
北京工业大学信息学部, 北京 100022

摘要 针对无人机(UAV)航拍图像中目标占比较小、拍摄角度和高度多变等问题,提出了一种基于多尺度特征融合的自适应无人机目标检测算法。利用深度可分离卷积结合残差学习的优点,建立了轻量化特征提取网络。构建多尺度自适应候选区域生成网络,将空间尺寸一致的特征图按照通道维度进行加权融合操作,增强了特征对目标的表达能力,并利用语义特征指导网络在多尺度特征图上自适应生成与真实目标更加匹配的目标候选框。仿真实验表明,该算法有效提升了无人机航拍目标检测精度,具有较好的鲁棒性。

关键词 机器视觉; 无人机; 目标检测; 深度网络; 特征融合
中图分类号 TP391.4 **文献标志码** A **doi**: 10.3788/AOS202040.1015002

Multi-Scale Feature Fusion Based Adaptive Object Detection for UAV

Liu Fang, Wu Zhiwei*, Yang Anzhe, Han Xiao
Information Department, Beijing University of Technology, Beijing 100022, *China*

Abstract In the aerial image of unmanned aerial vehicle(UAV), the target is usually small, and the shooting angle and height are variable. To address the problems, we proposed an adaptive drone object detection algorithm based on the multi-scale feature fusion. First, lightweight feature extraction network was established using the advantages of deep separable convolution and residual learning. Second, a multi-scale adaptive candidate region generation network was constructed, and feature maps with the same spatial size were weighted and merged based on the channel dimensions, which enhance the feature expression ability to objects. Based on these multi-scale featured maps, the use of semantic features to generate target candidate frames can be more matchable with real objects. Moreover, simulation experiments demonstrate that this algorithm can effectively improve the accuracy of UAV detection and have better robustness.

Key words machine vision; unmanned aerial vehicle; object detection; deep network; feature fusion
OCIS codes 150.0155; 150.1135; 100.4996

1 引 言

随着科学技术的快速发展,无人机已被广泛应用到军事和民用领域,无人机航拍目标检测技术已成为人工智能和计算机视觉领域的热点课题。在多数情况下,无人机的拍摄视野很大,包含丰富的视觉内容,虽然它提供了更全面的场景信息,但是待检测的目标对象通常在图像中占比较小,且没有足够的检测细节,这导致了目前的目标检测算法效果不理想[1]。因此,准确高效地检测小目标是无人机航拍目标检测任务的关键问题之一。

近年来,基于深度学习的目标检测技术取得了巨大成功,不少学者开始利用深度学习方法进行无人机航拍目标检测。然而,从无人机所拍摄的数据中进行目标检测比正常自然场景中的物体检测要困难,这是因为前者除了目标较小之外,目标的外观和结构质量都很差,容易与噪声混淆,这给无人机航拍目标检测带来一定的挑战[2]。文献[3]将 Faster R-CNN 用于遥感航拍目标检测并进行验证实验,取得较好效果,但对小目标的检测效果不理想。文献[4]提出了适应空中目标检测任务特点和需求的 Faster R-CNN 改进策略,弥补了 Faster R-CNN 对弱小目标和被遮挡目标不敏感的缺陷并提升了检测精度。文献[5]提出一种具有横向连接的特征金字塔网络(FPN),利用多尺度特征和自上而下的结构实现目标检测。FPN 只利用顶层的特征进行检测,虽然信

收稿日期: 2019-12-25; **修回日期**: 2020-01-21; **录用日期**: 2020-02-21
基金项目: 国家自然科学基金(61171119)
* **E-mail**: wuzw_66@163.com

息丰富，但是经过层层池化，很多细节特征信息会丢失，而这些信息往往对小目标检测具有重要意义。文献[6]提出了一种基于视觉细节增强映射的无人机航拍目标检测方法，将可能包含运动目标的区域进行多分辨率放大，得到详细的目标信息，并重新排列到新的前景空间中进行视觉增强，最后将重新排列好的图片送入深度检测网络，该方法显著提高了对运动小目标的检测精度，但是增加了网络的计算量，降低了算法运行速度。综上所述，目前的无人机航拍目标检测算法往往对小目标无法实现准确高效的检测。

针对上述问题，为了在不损失检测效率的情况下提高对无人机航拍图像中小目标的检测精度，提出了一种基于多尺度特征融合的自适应无人机目标检测算法。首先，为了压缩特征提取网络模型结构并提高算法的计算效率，采用深度可分离卷积对标准卷积进行优化，将常规卷积神经网络中的卷积操作分成深度卷积层(Depthwise Convolution)和点卷积层(Pointwise Convolution)两部分，同时结合残差学习[7]的优点构建了轻量化深度残差网络(LResnet)。其次，在 LResnet 的基础上构建了多尺度自适应候选区域生成网络，利用反卷积级联结构增大深层级特征图分辨率以实现与其前一层特征图空间尺寸一致的特征图，将空间尺寸一致的特征图按照通道维度进行加权融合操作，很大程度地增强了特征对小目标的表达能力，并利用语义特征指导网络在多层级不同尺度特征图上自适应生成目标候选框。实验结果表明，该算法达到了较好的目标检测性能，与其他无人机航拍目标检测的算法对比，在保证算法速度的基础上，显著提高了对无人机航拍目标的检测精度。

2　基于多尺度特征融合的自适应无人机目标检测算法

为了在不损失检测效率的情况下提高对无人机航拍图像中小目标的检测精度，提出一种基于多尺度特征融合的自适应无人机目标检测算法。算法模型的总体框架如图 1 所示。所提算法主要有两部分组成。第一个部分是用于特征提取的轻量化深度残差网络(LResnet)，结合残差学习的优点并将普通卷积操作分成深度卷积层和点卷积层两部分用于压缩网络参数，提高了网络的计算效率。第二部分是多尺度自适应候选区域生成网络，在 LResnet 四个层级中选取每层产生的具有相同大小的输出特征图的最后一层{C2、C3、C4、C5}，用 1×1 卷积把每层卷积的维度固定为 256，利用反卷积级联结构增大深层级特征图分辨率，实现与其前一层特征图空间尺寸一致，并将空间尺寸一致的特征图按照通道维度进行加权融合操作，得到目标表达能力更强的{P2、P3、P4、P5}四层不同尺度的特征图。在反卷积级联网络所产生的每一层特征上，都使用 GA-RPN (Guided Anchoring-Region Proposal Network)[8]根

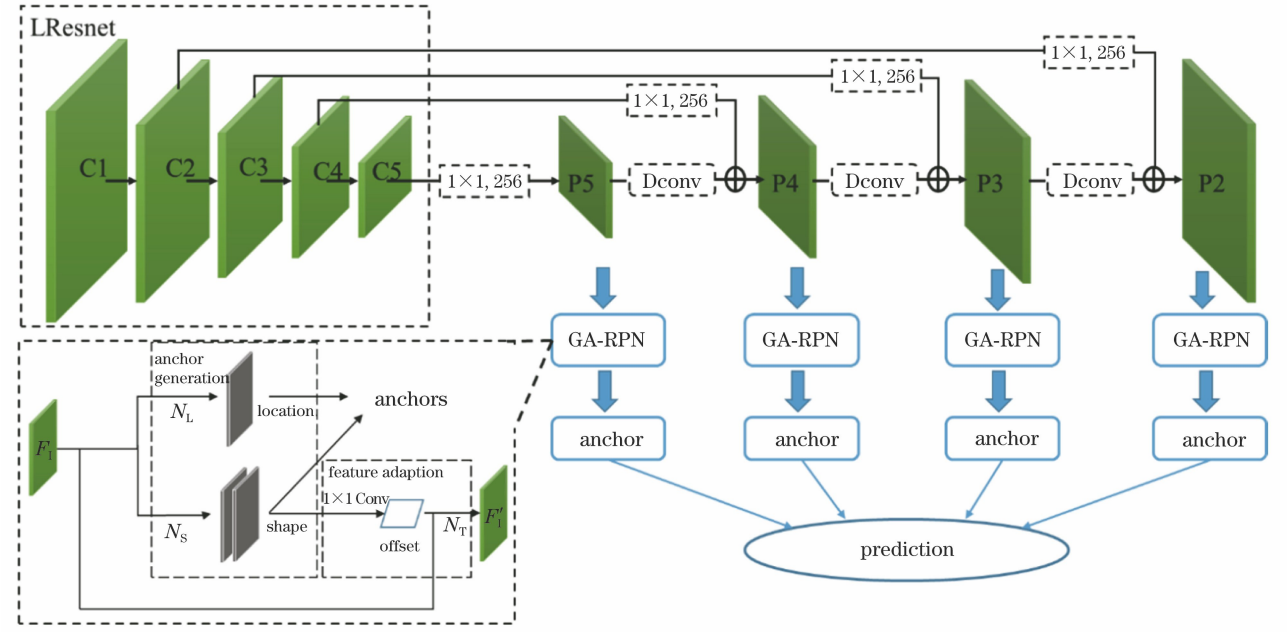

图 1　算法总体网络结构

Fig. 1　Framework of our algorithm

据语义特征自适应生成用于预测目标的候选框和相应的类别概率值，并通过非极大值抑制来获得最终的预测结果。

2.1　轻量化残差网络模型

卷积神经网络更深的层数以及更小的感受野，能够提高网络分类的准确性[9]。常规的卷积操作是将输入层的特征通道与卷积核进行卷积计算处理后相加输出，作为下一层的特征输入，计算过程如图 2(a)所示。但随着网络深度的增加，常规卷积参数量和计算量随网络层数加深而成倍增长[10]，导致模型尺寸增大，难以在计算资源受限的无人机平台应用。为了解决这一问题，采用深度可分离卷积对标准卷积进行优化，即将标准卷积分解成一个深度卷积和一个点卷积。深度卷积将输入图像特征的每一个通道单独分配一个卷积核，每个卷积核只负责对该通道的图像特征进行卷积操作。然后使用 1×1 的卷积核，将通过深度卷积得到的各个通道的输出结果进行压缩组合，其分解过程如图 2(b)所示。

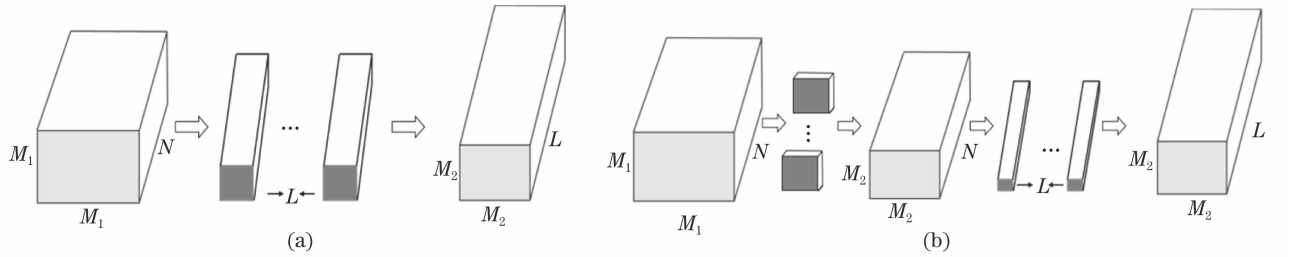

图 2　卷积分解示意图。(a)标准卷积过程；(b)分解后的卷积过程

Fig. 2　Schematic diagram of convolution decomposition. (a) Standard convolution process; (b) convolution process after decomposition

设输入特征图 F 的维度为 $M_1\times M_1\times N$，输出特征图 G 的维度为 $M_2\times M_2\times L$，卷积核的尺寸为 K。对于标准的卷积操作来说，经过一次完整卷积操作后的计算代价为 $K\times K\times N\times L\times M_2\times M_2$。

标准卷积操作只通过一步计算，将输入的图像特征进行卷积得到输出结果。本研究将标准卷积分解为两个步骤来进行操作。1) 对输入图像特征图 F 的 N 个通道中的每一个通道分配一个对应的卷积核，该卷积核大小与标准卷积层的卷积核大小一致，个数为 N，步长为 1 且进行 padding 操作，再进行卷积操作生成维度为 $M_2\times M_2\times N$ 的图像特征。2) 将 1)中通过深度卷积层所得到的图像特征作为点卷积层的输入进行计算。点卷积层所使用的卷积核大小为 1×1，个数为 L，则对于分离后的卷积，总计算代价为 $K\times K\times N\times M_2\times M_2+N\times 1\times 1\times L\times M_2\times M_2$。

对比他们的计算量比值为

$$\eta=\frac{K\times K\times N\times M_2\times M_2+N\times 1\times 1\times L\times M_2\times M_2}{K\times K\times N\times L\times M_2\times M_2}=\frac{1}{L}+\frac{1}{K^2}。\tag{1}$$

K 的取值一般为 3，而 $L\gg 9$，所以由(1)式可知，将标准卷积操作分解成深度卷积和点卷积两部分之后，其参数量和计算量约为标准卷积的 1/9。即分解后的卷积操作和标准卷积操作最后输出的特征图维度保持一致，且大幅度减少了网络参数量和计算量。

将普通卷积分解为深度卷积和点卷积两步后，虽然有效减少了网络模型参数计算量，但是使得网络层数大幅加深，在进行网络训练时容易出现梯度消失，导致模型训练难度变大，即出现“网络退化”现象[11]。而使用如图 3 所示的残差结构可以很好地减轻深层网络训练的负担，通过跳跃连接(Skip

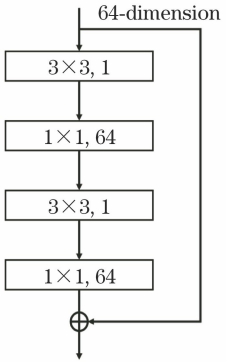

图 3　卷积神经网络残差模块结构图

Fig. 3　Convolutional neural network residual module structure diagram

Connection)的方式连接浅层网络与深层网络，相当于把底层的特征信息融合到高层中，可以保证输入的特征信息不流失，增强了特征对目标的表达能力，并且使得梯度能够很好地传递到浅层。

综上所述，构建了一个轻量化的深度残差网络模型(LResnet)用于提取航拍图像卷积特征。该模型网络结构参数如表1所示，其中Output Size表示输出特征尺寸，Kernel表示卷积核尺寸，Output Channels表示输出特征的维度。

表1　轻量化深度残差网络模型

Table 1　Lightweight deep residual network model

Layer	Type	Kernel	Output size	Number of output channels
X	input		224×224	3
Conv_1	Convolution	3×3,64 stride 2	112×112	32
Conv_2	Convolution	$\begin{bmatrix}3\times3,1\\1\times1,64\\3\times3,1\\1\times1,64\end{bmatrix}\times3$	56×56	64
Conv_3	Convolution	$\begin{bmatrix}3\times3,1\\1\times1,128\\3\times3,1\\1\times1,128\end{bmatrix}\times4$	28×28	128
Conv_4	Convolution	$\begin{bmatrix}3\times3,1\\1\times1,256\\3\times3,1\\1\times1,256\end{bmatrix}\times6$	14×14	256
Conv_5	Convolution	$\begin{bmatrix}3\times3,1\\1\times1,512\\3\times3,1\\1\times1,512\end{bmatrix}\times3$	7×7	512

2.2　多尺度自适应候选区域生成网络

无人机的空中拍摄范围很大，导致航拍图像中目标占比较小，没有足够的检测细节。利用深度卷积神经网络提取图像目标特征时，卷积神经网络的高层感受野较大，语义信息表征能力强，但是特征的几何信息表征能力弱，不利于小目标的检测；低层网络的感受野比较小，几何细节信息表征能力强，但是语义信息表征能力弱[12]。文献[13]中采用不同层级多种大小特征图来分级别预测目标，使用高层的特征预测普通目标而利用低层的特征预测小目标，这种方式虽然在预测小目标时考虑使用几何信息比较强的低层特征，但是因为缺乏高层语义特征，对小目标检测效果仍不理想。此外，传统的目标检测网络采用手工设计的固定尺寸候选框，针对不同的检测问题需要设计不同大小的候选框，而目标的尺寸千差万别，固定的设计将会阻碍检测精度提高[14]。并且为了保证有较高的召回率，需要在网络中生成大量的候选区域，其中许多候选区域与目标无关，会占用大量的计算资源。

为解决上述问题，所提算法基于LResnet框架，构建了一个多尺度自适应候选区域生成网络。通过反卷积级联结构在低层特征图中加权融入高层语义特征，增强了特征对目标的表达能力，并采用多层级不同尺度特征图用于目标预测，在反卷积级联网络所产生的每一层特征上，都根据图像特征预测候选框的位置和形状，生成稀疏且形状任意的候选框。

2.2.1　反卷积级联结构

图4显示了多尺度自适应候选区域生成网络中的反卷积级联结构。在LResnet中选择了多级特征映射$\{C_2、C_3、C_4、C_5\}$，对应于每个网络级最后一层的输出，具有该级网络最强的语义特征。它们相对于输入图像分别有{4 pixel，8 pixel，16 pixel，32 pixel}的步幅。通过对前一功能的横向连接对$\{C_2、C_3、C_4、C_5\}$进行反卷积运算并与上层特征加权融合，获得$\{P_2、P_3、P_4\}$，P_5层则直接由C_5经过1×1卷积得到。

具体而言，首先在高级特征图P_5上使用反卷积运算使得特征图大小与C_4一致，然后将它与相应的前级特征图C_4加权融合，得到一个新的特征图P_4。重复这个过程，直到生成与C_2大小一致的特征图P_2。另外，为了保持在前后特征图维度一致，在横向连接中增加1×1卷积核，并把每一层维度都固定为256。因为算法旨在提升对无人机航拍小目标的检测效果，高层特征语义信息表征能力强，而低层特征图拥有更多小目标的细节特征信息[15]，故在相同权值直接相加的基础之上，额外为6个不同的特征图分配加权数，加权融合公式为

$$\begin{cases}P_4=\alpha_1 D(P_5)+\alpha_2 C_4\\P_3=\alpha_3 D(P_4)+\alpha_4 C_3\\P_2=\alpha_5 D(P_3)+\alpha_6 C_2\end{cases},\tag{2}$$

式中：$D(\cdot)$为反卷积转化函数；α_1、α_2、α_3、α_4、α_5和α_6表示权重系数，取值分别为0.7，0.3，0.6，0.4，0.45，0.55，为避免特征信息冗余，各层融合的权重系数之和为1。反卷积级联结构各层参数如表2中所示。

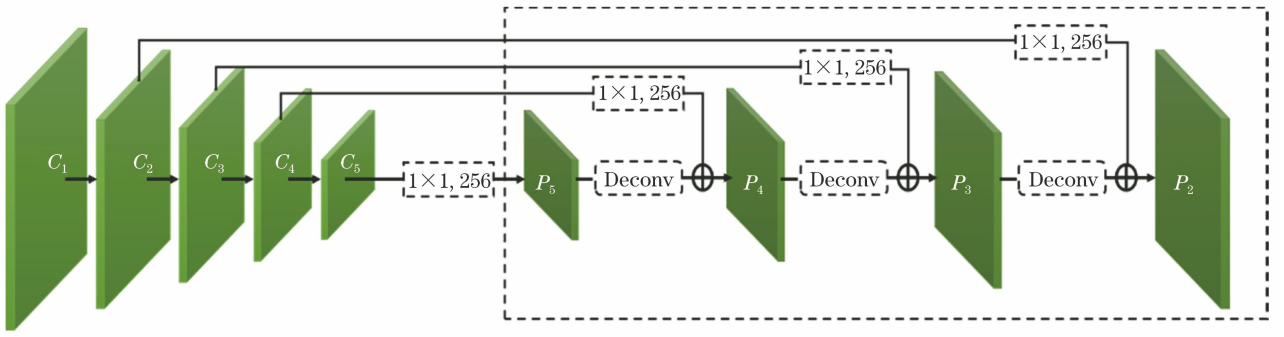

图 4　反卷积级联结构

Fig. 4　Deconvolution cascaded structure

表 2　反卷积各层参数

Table 2　Deconvolution layer parameters

Layer	Type	Kernel	Stride	Output size
h1	Deconvolution	3×3	1	14×14×256
h2	Deconvolution	3×3	1	28×28×256
h3	Deconvolution	3×3	1	56×56×256

2.2.2　自适应候选区域生成

设输入图像为 I，则 anchor 在图像上的概率分布函数可以表示为

$$P(x,y,w,h\,|I)=P(x,y\,|I)p(w,h\,|x,y,I),\tag{3}$$

式中：(x,y) 是 anchor 的中心坐标；w 和 h 是 anchor 的宽度和高度。即 anchor 在图像 I 上的分布函数可以分解为 anchor 中心点的概率分布函数和在确定中心点条件下 anchor 形状的概率分布函数。根据(3)式，GA-RPN 将 anchor 的生成分成两个分支，一个分支用于预测 anchor 的中心位置，另一个分支用来预测 anchor 的形状。这种网络设计本质上不同于传统的 anchor 生成方案，因为每个预测中心位置的点只对应一个 anchor 而不是对应一组设定好的 anchor。这种生成方式大大降低了 anchor 的生成数量，而且允许生成任何长宽比例的 anchor 形状，可以更好地检测小目标和特殊长宽比例的物体。

如图 5 所示，anchor 的中心预测分支网络 N_{L} 产生一个与输入特征图 F_{I} 尺寸大小相同的概率图，$P(i,j\,|F_{\mathrm{I}})$ 表明在特征图 (i,j) 位置可能出现目标物体的概率，对应于图像 I 中的坐标为 $[(i+1/2)s,(j+1/2)s]$，其中 s 是特征图的步长。N_{L} 分支网络使用 1×1 的卷积网络来获得目标的置信图，之后利用 Sigmoid 函数将其转化为概率值。根据生成的概率图，通过选择相应的概率值大于预先定义的阈值的位置(对比实验中取值为 0.05)，从而确定目标可能存在的区域。

在确定目标的可能位置之后，下一步是确定可能存在于每个位置的目标的形状。形状预测分支 N_{S} 网络分支包含一个 1×1 大小的卷积层，可以产生两通道映射，包含 dw 和 dh 的值。具体来说，在输入特征图 F_{I} 中，形状预测分支将预测每个位置的最佳形状 (w,h)，因为 w 和 h 的范围可能很大，需要经过变换输出 dw 和 dh，变换公式为

$$\begin{cases}w=\lambda\times s\times\exp(\mathrm{d}w)\\ h=\lambda\times s\times\exp(\mathrm{d}h)\end{cases}。\tag{4}$$

根据 dw 和 dh 可以映射出 w 和 h，其中 λ 是经验尺度因子(本文实验中取值为 8)。该非线性变换映射可以将 $[0,1000]$ 映射到 $[-1,1]$，使得形状预测分支计算更加简单、稳定。

在传统的 RPN 中，anchor 在整个特征图中都是一致的，即在每个位置上的形状和尺度都是固定的，因此特征图可以学习连续的表达方式。而根据图像特征自适应生成的 anchor 形状是根据位置的不同而变化的。在此条件下，为了取得较好的特征表达效果，较大 anchor 的特征编码较大区域的内容，较小的 anchor 特征抽取较小区域的内容，故采用 anchor 特征自适应分支网络 N_{T} 将特征进行转化，该分支网络采用 3×3 的可变形卷积层[16]来实现，即

$$f_i'=N_{\mathrm{T}}(f_i,w_i,h_i),\tag{5}$$

式中：f_i 是第 i 个位置的特征；w_i,h_i 是对应的 anchor 形状。即如图 5 所示，首先从形状预测分支的输出预测偏移量，之后在原始特征图上使用可变形卷积来获得 f_i'。

2.3　多任务损失函数

因为采用自适应生成候选区域，所以在传统的分类损失 L_{cls} 和回归损失 L_{reg} 的基础上，增加了

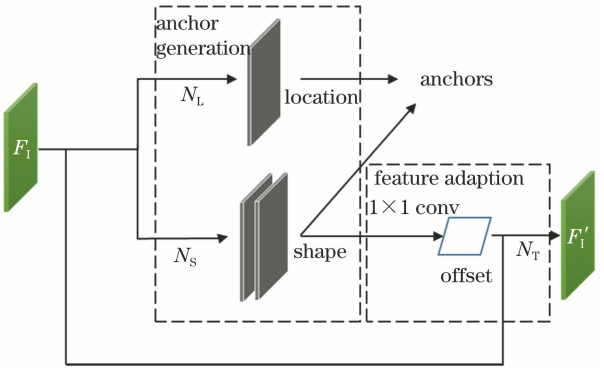

图 5　自适应候选区域生成

Fig. 5　Adaptive candidate region generation

anchor 定位损失函数 L_{loc} 和 anchor 形状损失函数 L_{shape}。总的目标损失函数可以表示为

$$L=L_{cls}+L_{reg}+\beta_1 L_{loc}+\beta_2 L_{shape}, \tag{6}$$

式中：β_1 和 β_2 为多任务损失函数加权系数，取值分别为 1 和 0.1。

为了训练 anchor 定位分支所生成的中心位置尽可能地和真实目标中心重合，让网络在自适应生成候选区域时，更多的 anchor 出现在真实目标的中心附近。本研究把每个真实目标框都分为三个部分：1)得到真实目标框 (x_g,y_g,w_g,h_g) 的特征映射 (x_g',y_g',w_g',h_g')；2)在目标特征映射区域内定义两个区域 $(x_g',y_g',\delta_1 w_g',\delta_1 h_g')$ 和 $(x_g',y_g',\delta_2 w_g',\delta_2 h_g')$，其中 δ_1、δ_2 取值分别为 0.2 和 0.5，$(x_g',y_g',\delta_1 w_g',\delta_1 h_g')$ 为中心区域，$(x_g',y_g',\delta_2 w_g',\delta_2 h_g')$ 以内不含中心区域的位置为忽略区域，其余部分为外围区域；3)把中心区域作为正样本，外围区域作为负样本，利用 Focal Loss[17] 来训练定位分支。中心区域是目标特征中心位置的小部分区域，相对于整个目标而言，这部分特征对角度变化的敏感度低，利用这部分特征检测模型，能够学习旋转不敏感的特征表示，这在一定程度上减小了无人机航拍角度变化对检测性能的影响。

和传统的 anchor 设计不同，在训练形状预测分支时因为不知道准确的 anchor 大小，所以无法利用传统方式计算交并比(IOU)损失，为了解决这个问题，枚举不同比例和大小的 anchor 选取最大值作为最终的 vIOU(a_{wh},G)，即

$$\mathrm{vIOU}(a_{wh},G)=\max \mathrm{IOU}(a_{wh},G) \tag{7}$$

式中：IOU(•)是传统 IOU 的定义；G 代表真实目标框；a_{wh} 表示 anchor 变量。本研究枚举了 9 对常见的不同比例和大小的 anchor 作为 a_{wh}，并使用最大值作为最终的 vIOU(a_{wh},G)。L_{shape} 定义为

$$L_{shape}=l_1\left[1-\min\left(\frac{w}{w_g},\frac{w_g}{w}\right)\right]+l_1\left[1-\min\left(\frac{h}{h_g},\frac{h_g}{h}\right)\right], \tag{8}$$

式中：l_1 为 smooth L_1 损失函数。

3　实验结果及分析

实验平台采用 i7-7700 处理器，NVIDIA GTX1080Ti 显卡，内存为 16G、Ubuntu16.04 操作系统。本研究所采用的实验数据来自 VisDrone 无人机目标检测数据集[18]，数据包括市区、乡村、公园、马路等多种自然场景，由无人机平台在不同位置、不同高度拍摄获得。VisDrone 无人机目标检测数据集包含 10209 张静态图像，其中包含 6471 张训练集，548 张验证集，3190 张测试集。数据集中标记有 10 个预定义的类，即行人、人、汽车、货车、公共汽车、卡车、摩托车、自行车、带棚三轮车和三轮车(其中“行人”是指具有站立或行走姿势的人，“人”是具有其他姿势的人)。因为无人机的飞行高度和摄像头方向发生不断的变化，该数据集中的大多数目标尺度和拍摄角度变化较大，且小目标占比较多，部分数据中目标分布密集，是一个多尺度、多角度、多背景的大型无人机目标检测数据集。

3.1　目标检测定性结果分析

为验证所提算法在实际任务中的有效性，抽取无人机航拍目标检测较为困难的实际应用场景图像进行测试，评估算法检测结果并进行可视化分析。可视化检测结果展示如图 6 所示。在图 6(a)所示的两幅无人机航拍图像中，目标占比都非常小，除了一些外观结构非常差的小目标之外，本文算法可以检测到绝大多数目标，说明所提算法利用多尺度自适应候选区域生成网络将低层特征和高层特征进行多级加权融合，并利用语义特征指导网络来自适应生成目标候选框的方法，显著提高了航拍小目标检测精度。当遇到图 6(b)所示的目标数量众多且发生遮挡的情况下，算法依然可以有较好的检测效果，说明所提算法采用多尺度特征进行目标预测，很大程度上增强了特征表达能力。此外，如图 6(c)所示，算法受光照变化的影响很小，在夜间灯光照射的昏暗环境下依然能够有很好的检测性能，这也证明了所提算法可以应对多种环境变化，能满足实际任务需要，有很好的泛化能力。

3.2　算法可行性验证分析

为验证特征提取网络的有效性，将 LResNet 与

图 6　算法在不同情况下可视化检测结果展示。(a)小目标检测结果展示;(b)密集目标检测结果展示;(c)光照变化目标检测结果展示

Fig. 6　Visualization detection results of the proposed algorithm in different situations. (a) Small target detection results; (b) dense target detection results; (c) detection results of target under different illuminations

ResNet 进行对比实验,通过该实验数据来检验所提算法所用特征提取网络的性能。为了较快地验证算法的可行性,实验在同等条件下随机选取 VOC2012 数据集[19]中 13700 张图像作为训练集,3425 张图像作为测试集。

实验结果如表 3 所示。从表中可以看出,LResent 的网络模型大小仅为 10.2 MB,约为 ResNet-50 网络模型大小的 1/10,但在相同条件下与 ResNet 在 VOC2012 上的分类精度仅相差 0.7%。而且 LResNet 模型在运行阶段所需的显存占用率也大大减小,仅需要使用 546 MB 显存,约是 ResNet-50 网络显存占用率的 20%。这表明 LResNet 在损失极少检测精度的情况下,极大地减少了网络参数量,并大大降低了算法运行时内存占用率。

表 3　特征提取网络比较

Table 3　Feature extraction network comparison

Model	Size /MB	Ratio /%	Accuracy /%
Resnet	97.7	—	81.3
LResnet	10.2	10.4	80.6

为了验证所提算法中反卷积级联结构(以下简称 DC 模块)和 GA-RPN 在无人机航拍目标检测算法中的有效性,基于 VisDrone 目标检测数据集,设计如表 4 所示的几组对比实验。测试实验在 VisDrone-test-dev 数据集(1610 张无人机航拍图像,包含 VisDrone 数据集中的各类情况)进行,将 Faster-Rcnn[20](Resnet50+RPN)设置为基线对比网络,采用平均精度均值(mAP)、平均精度(AP,包括 IOU 为 0.50 和 0.75 的 AP,记为 AP^{50} 和 AP^{75})评

价指标来进行定量分析。

对比表4中的方法①和方法②可以看出，算法将参数量较多的Resnet50替换成轻量化残差网络后，mAP相较与Faster-RCNN仅下降了约0.11个百分点，说明利用深度可分离卷积优化后的网络模型对检测效果影响甚微，但却大大减少了网络的参数量。对比方法②和方法③可以看出，在网络中加入反卷积级联结构后，mAP提升了约2.5个百分点，AP50提升约3个百分点，说明所提算法利用反卷积级联结构将低层特征图中加权融入高层语义特征，所产生的多层加权融合特征对目标的表达能力更强，采用多层级特征信息预测目标，更适用于随着无人机飞行高度变化的多尺度航拍目标的检测。方法④相比方法③将RPN替换成GA-RPN，在IOU为75的情况下AP增加了3.5个百分点，反映出利用语义特征自适应生成用于预测目标的候选框比人为设计的目标候选框更为匹配，也表明本文所提出的目标检测框架具有较好的分类能力和较高的边框回归精度。

表4　算法各模块有效性对比试验

Table 4　Effectiveness test of each module for different methods　%

Method	mAP	AP^{50}	AP^{75}
①Faster-RCN(Resnet50+RPN)	18.63	35.87	17.86
②LResnet+RPN	18.52	35.75	17.44
③LResnet+DC+RPN	21.03	38.46	18.03
④LResnet+DC+GA-RPN(ours)	**22.12**	**38.76**	**21.53**

算法在VisDrone测试数据集中对每个类别的检测结果如表5所示，对比指标为各类目标的AP。从表中可以看出，可以看出，由于Faster-RCNN算法采用单一深度特征和人工设计的候选框进行目标预测，并不能很好适应无人机航拍数据集中目标尺度多变、小目标较多的实际情况，因此对大多类目标检测效果较差。而相较于Faster-RCNN，本文所提算法除“人”和“公共汽车”外其他类别的AP都有1～7个百分点的提升。其中对于“行人”、“自行车”和“摩托”这几种在图像中占比较小的目标，平均检测精度都有了明显的提高，表明所提算法采用多尺度自适应候选区域生成网络所产生的不同尺度的融合特征，同时使用加权融合后的多种尺度特征进行目标预测，并利用语义特征指导网络来自适应生成目标候选框，很大程度上增强了对各类目标的特征表达能力，提高了航拍目标检测精度。

表5　VisDrone测试数据集各类别检测结果对比

Table 5　Comparison between the results of ten categories from ours model and Faster-RCNN on VisDrone dataset　%

Method	Pedestrian	Person	Bicycle	Car	Van	Truck	Tricycle	Awn	Bus	Motor
Faster-RCNN	18.34	**7.62**	6.76	43.31	27.53	19.95	10.13	7.65	**36.87**	8.79
Ours	**22.43**	7.61	**8.56**	**50.18**	**34.63**	**24.34**	**14.11**	**9.08**	36.25	**14.88**

3.3　主流无人机目标检测算法对比实验

对比实验阶段对不同主流的目标检测网络通过相同的数据训练得到的模型进行评价，主要比较不同目标检测算法在无人机低空飞行时对地面目标的检测识别能力，并验证本文算法的检测性能。对比算法包含RetinaNet、FPN、YOLOv3[21]、CornerNet[22]。在实验中，分别采用mAP、AP^{50}、AP^{75}、帧率(FPS)评价指标来进行检测准确度和检测速度的定量分析。

表6中结果表明，在无人机航拍数据上与主流目标检测算法对比，从目标检测精度指标来看，本文算法的AP值有了较大提高，mAP达到了22.12%，比YOLOv3高出1.82%。AP^{50}是评价算法分类能力的有效指标，而AP^{75}则能够体现检测框架对边界框位置回归的能力[23]，通过对比实验可以看出，本文所提的算法在IOU为0.75的情况下取得了21.53%的检测精度，相较于YOLOv3取得了5.73%的增益，这也表明本文所提出的目标检测框架表现出更好的分类能力和更高的边框回归精度。检测精度的明显提高，其主要原因是多尺度自适应候选区域生成网络通过反卷积级联结构在低层特征图中加权融入高层语义特征，并采用多层级不同尺度特征图用

表6　主流目标检测算法无人机航拍数据对比试验

Table 6　Comparison test of UAV aerial data with mainstream object detection algorithm

Method	mAP /%	AP^{50} /%	AP^{75} /%	Frame rate /(frame·s^{-1})
FPN	16.51	32.20	14.91	6
YOLOv3	20.30	**44.12**	15.80	**44**
RetinaNet	11.81	21.37	11.62	11
CornerNet	17.41	34.12	15.78	13
Ours	**22.12**	38.76	**21.53**	24

于目标预测，这大大提升了随着无人机视角和飞行高度变化的各类目标的检测效果。除此之外，在反卷积级联网络所产生的每一层特征上，都根据图像特征预测候选框的位置和形状，生成稀疏而且形状任意的候选框，这种候选框不受固定尺寸限制，与真实目标框更加匹配，直接提升了边框回归的精度。本文算法检测速度相比双阶段的目标检测算法FPN、RetinaNet等有了明显的提升，虽未达到YOLOv3的速度水平，但依然有24 frame·s^{-1}的检测速度，其主要原因是LResnet大大降低了网络的参数量，提升了算法的运行效率。

4　结　论

针对无人机拍摄图像中目标占比较小、拍摄角度和高度多变等问题提出了一种基于多尺度特征融合的自适应无人机目标检测算法。利用深度可分离卷积结合残差学习的优点建立了轻量化特征提取网络，在此基础上，构建多尺度自适应候选区域生成网络，将空间尺寸一致的特征图按照通道维度进行加权融合操作，增强了特征对目标的表达能力，并利用语义特征指导网络自适应生成与真实目标更加匹配目标候选框。所提方法在VisDrone无人机图像目标公开数据集上取得了22.15%的mAP、21.53%的AP^{75}，实验结果表明，反卷积级联结构所产生的多层加权融合特征对目标的表达能力更强，使用多层特征信息更有利于对无人机航拍小目标的检测，利用语义特征自适应生成用于预测目标的候选框比人为设计的目标候选框更为匹配。所提算法虽然设计了轻量化残差网络，降低了特征提取网络的计算量，但使用多尺度自适应候选区域生成网络增加了算法的计算代价，下一步将继续优化网络结构，以期进一步提高无人机目标的检测精度和速度。

参　考　文　献

[1] Vaddi S, Kumar C D, Jannesari A. Efficient object detection model for real-time UAV applications[EB/OL]. (2019-05-30)[2019-12-15]. https://arxiv.org/abs/1906.00786.

[2] Pei W, Xu Y M, Zhu Y Y, et al. The target detection method of aerial photography images with improved SSD[J]. Journal of Software, 2019, 30(3): 738-758.
裴伟, 许晏铭, 朱永英, 等. 改进的SSD航拍目标检测方法[J]. 软件学报, 2019, 30(3): 738-758.

[3] Wang J C, Tan X C, Wang Z H, et al. Faster R-CNN deep learning network based object recognition of remote sensing image[J]. Journal of Geo-Information Science, 2018, 20(10): 1500-1508.
王金传, 谭喜成, 王召海, 等. 基于Faster R-CNN深度网络的遥感影像目标识别方法研究[J]. 地球信息科学学报, 2018, 20(10): 1500-1508.

[4] Feng X Y, Mei W, Hu D S. Aerial target detection based on improved faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进Faster R-CNN的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[5] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 2117-2125.

[6] Li J, Dai Y R, Li C C, et al. Visual detail augmented mapping for small aerial target detection[J]. Remote Sensing, 2018, 11(1): 14.

[7] 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 (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[8] Wang J Q, Chen K, Yang S, et al. Region proposal by guided anchoring[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 2965-2974.

[9] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 1-9.

[10] Howard A, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2019-12-15]. https://arxiv.org/abs/1704.04861.

[11] Ma Q, Zhu B, Cheng Z D, et al. Detection and recognition method of fast low-altitude unmanned aerial vehicle based on dual channel[J]. Acta Optica Sinica, 2019, 39(12): 1210002.
马旗, 朱斌, 程正东, 等. 基于双通道的快速低空无人机检测识别方法[J]. 光学学报, 2019, 39(12): 1210002.

[12] Ren Y, Zhu C R, Xiao S P. Small object detection in optical remote sensing images via modified faster R-CNN[J]. Applied Sciences, 2018, 8(5): 813.

[13] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]// Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[14] Li C H, Ru L, He L Y. The region proposal network target detection method with variable anchor box [J/OL]. Journal of Beijing University of Aeronautics and Astro: 1-8[2019-12-15]. https://doi.org/10.13700/j.bh.1001-5965.2019.0531.
李承昊, 茹乐, 何林远. 一种可变锚框候选区域网络的目标检测方法[J/OL]. 北京航空航天大学学报: 1-8[2019-12-15]. https://doi.org/10.13700/j.bh.1001-5965.2019.0531.

[15] Wu T S, Zhang Z J, Liu Y P, et al. A lightweight small object detection algorithm based on improved SSD [J]. Infrared and Laser Engineering, 2018, 47(7): 0703005.
吴天舒, 张志佳, 刘云鹏, 等. 基于改进SSD的轻量化小目标检测算法[J]. 红外与激光工程, 2018, 47(7): 0703005.

[16] Dai J F, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks[EB/OL]. (2016-05-20)[2019-12-15]. https://arxiv.xilesou.top/abs/1605.06409.

[17] 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. New York: IEEE, 2017: 2980-2988.

[18] Zhu P F, Wen L Y, Bian X, et al. Vision meets drones: a challenge[EB/OL]. (2018-04-20)[2019-12-15]. https://arxiv.org/abs/1804.07437.

[19] Everingham M, Eslami S M A, van Gool L, et al. The PASCAL visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[20] 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.

[21] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2019-12-15]. https://arxiv.org/abs/1804.02767.

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

[23] Yao Q L, Hu X, Lei H. Objectdetection in remote sensing images using multiscale convolutional neural networks[J]. Acta Optica Sinica, 2019, 39(11): 1128002.
姚群力, 胡显, 雷宏. 基于多尺度卷积神经网络的遥感目标检测研究[J]. 光学学报, 2019, 39(11): 1128002.