改进CenterNet在遥感图像目标检测中的应用

田壮壮[1]，张恒伟[1]，王坤[1]，刘盛启[2]，邹前进[1]，赵镇[1]，陈育斌[1]

1. 电子信息系统复杂电磁环境效应国家重点实验室，洛阳 471000;
2. 国防科技大学 电子科学学院，长沙 410073

摘 要：为了提高遥感图像目标检测的效率及精度，本文提出了一种基于改进CenterNet的遥感图像目标检测方法。基于CenterNet的检测框架，该方法能够降低目标检测所需要的步骤，减少对锚框的依赖。而在CenterNet的基础上，所提方法通过采用带有转置卷积的ResNet作为骨干网络，降低了骨干网络的参数数量；然后针对训练用的热力图标签，提出了针对中心点设计的高斯核适用范围边长的计算方法；最后利用注意力机制，提高所提取特征中目标区域特征的有效性。在公开的高分辨率遥感图像上的实验结果表明，3种改进措施将目标检测的精度提高了4.0%，与此同时所需的检测时间降低为原来的31.9%。与其他对比方法相比，所提方法在精度和速度上均有一定的优势，表明所提方法在遥感图像目标检测中具有一定的实用性。

关键词：遥感图像，目标检测，深度学习，CenterNet，注意力机制

中图分类号：TP751/P2

引用格式：田壮壮,张恒伟,王坤,刘盛启,邹前进,赵镇,陈育斌.2023.改进CenterNet在遥感图像目标检测中的应用.遥感学报,27(12):2706-2715

Tian Z Z, Zhang H W, Wang K, Liu S Q, Zou Q J, Zhao Z and Chen Y B. 2023. Application of an improved CenterNet in remote sensing images object detection. National Remote Sensing Bulletin, 27(12):2706-2715[DOI:10.11834/jrs.20231638]

1 引 言

遥感图像目标检测作为遥感图像解译的重要组成部分，无论是在军事的目标探测、精确制导、战场态势分析等，还是在民用的地形检测、环境和灾害评估等方面，均有着广泛的应用场景。随着探测设备的增加，遥感图像数据量也不断增多，基于机器学习的目标检测方法逐渐受到人们的关注。

早期针对遥感图像目标检测任务的机器学习方法主要依赖先验知识进行人工设计（Cheng和Han，2016），需要针对不同类型的目标设计相应的特征提取和分类方法，因此难以满足更为复杂的检测需求。近年来，随着计算能力的提升，以卷积神经网络CNN（Convolutional Neural Network）为代表的深度学习方法凭借其能够从数据中学习有效特征的能力而得到快速发展。

基于卷积神经网络的目标检测方法可大致分为基于锚框的（anchor-based）方法和无锚框的（anchor-free）方法两大类。其中基于锚框的方法需要在检测时预设或生成一定数量的锚框作为回归候选区域的基础。此类方法包括基于区域的卷积神经网络R-CNN（Region-based CNN）（Girshick等，2016；Ren等，2015）、SSD（Single-shot Multi-box Detector）（Liu等，2016）、GA-FPN（Wang等，2019b）、YOLOF（Chen等，2021）等。这些方法将锚框内的特征作为分类和回归的基础，在一定程度上缩小了回归的范围区间，在降低回归难度的同时提高回归的精度。由于其精准度较高，因此在遥感图像目标检测中得到了广泛的使用（Wang等，2019a；Yang等，2021；Hong等，2022）。然而，由于基于锚框的方法通常需要设计锚框，并根据每个锚框内的特征分别进行分类和回归，因此步骤较为复杂，在一定程度上降低了方法的时效性。

收稿日期：2021-10-22；**预印本：**2022-04-23

基金项目：国家自然科学基金(编号:62001486)

第一作者简介：田壮壮,研究方向为遥感目标检测侦查的理论和应用。E-mail:tzz14@nudt.edu.cn

为了降低对锚框的依赖，许多学者近年来将注意力放在了无锚框的方法上，并取得了一定的成果。比如基于关键点的CornerNet（Law和Deng，2018）、CenterNet（Zhou等，2019）；基于密集预测的FCOS（Tian等，2019）、DDBNet（Chen等，2020）；基于Transformer的DETR（Carion等，2020）、UP-DETR（Dai等，2021）。这些方法分别通过不同的方式，去掉了预设锚框的步骤，并保持了较高的检测率。但与此同时，这些方法仍存在着一定的不足，比如基于关键点的方法受预测点准确性的影响较大，基于密集预测的方法仍需依靠对数量众多的框进行预测来保持对目标的覆盖，基于Transformer的方法则需要较长的训练时间才能达到较好的检测效果。

为了简化检测流程，提高检测效率，考虑到不同方法的优缺点，本文基于CenterNet对目标进行检测。CenterNet方法利用全卷积网络来直接回归目标的类别以及边界框的中心点、宽和高，实现对目标的检测。该方法在目标检测时步骤较少，并且无需依赖锚框。

对于遥感图像目标检测而言，原始CenterNet仍然存在着一些不足。首先，原始的CenterNet采用Hourglass（Newell等，2016）作为骨干网络。该网络通过多层的递归来构建对称的网络结构并进行特征提取。Hourglass最初用于人体姿态估计，这种递归式的结构有利于捕捉人体多个关键点之间的关联。但是对于遥感图像而言，其待检测目标的中心点之间此类的关联性较低。而且这种递归方法生成的网络具有较多的网络参数，在一定程度上增大了网络的训练难度，降低了推断速度。其次，CenterNet在生成高斯核适用范围边长的计算中，继承了CornerNet中的计算方法。虽然这种计算方法在CenterNet中也可以得到结果，但两种网络之间关键点的使用方式并不相同。因此，生成高斯核适用范围边长的计算方法与关键点使用方式之间的不匹配问题也在一定程度上限制了CenterNet的检测准确性。最后，遥感图像场景通常比较复杂，需要采用一定的机制在特征提取的过程中提高目标特征的有效程度，抑制背景噪声对目标检测的影响。

为了解决上述提到的问题，本文提出了一种基于改进CenterNet的遥感图像目标检测方法。首先，为了降低网络训练难度，所提方法将目标检测中常用的ResNet（He等，2016）作为骨干网络，并在骨干网络中采用转置卷积模块（Xiao等，2018），将ResNet生成的较小尺寸的特征图转换为同样保留语义但尺寸更大的特征图，避免过小特征图导致的关键点模糊或丢失。然后，通过分析CenterNet中关键点生成框和真实框之间的关系，提出了一种用于生成热力图标签的高斯核适用范围边长的计算方法，解决了原始计算方法与关键点使用方式的不匹配问题。最后，利用注意力机制（Zhu等，2019），构建不同元素之间的上下文关系，从而增强目标所在区域的特征，抑制背景区域的影响，进一步增强了提取特征的有效性。

2 模型方法

本文提出的遥感图像目标检测方法的流程图如图1所示，可大致分为骨干网络和检测网络两个部分。

骨干网络主要用于提取输入图像的特征。文中采用的骨干网络以ResNet为基础，通过增加带注意力机制的瓶颈模块（Bottleneck）来增强目标区域的特征，并利用转置卷积来扩大特征图。而在检测网络中，骨干网络提取的特征分别通过3个不同的卷积过程得到不同类别目标框中心点位置的热力图、目标框对应的宽高以及中心点位置的偏移量。其中，中心点位置的热力图用作目标框位置的粗预测；中心点位置偏移量用作对该位置进行微调，使其更加准确；宽高的预测则进一步明确了目标框的形状，最终得到不同类别目标框的检测结果。

下面将分别就骨干网络、检测网络以及训练时所需的损失函数进行详细介绍。

2.1 骨干网络

相较于CenterNet默认采用的用于人体姿态估计的Hourglass网络，本文采用了更常用于目标识别的ResNet作为骨干网络的基础。由于原始的ResNet输出特征图尺寸相较于原始图像的下采样尺度为32，容易导致较小尺寸目标的中心点随着特征图的缩小而难以表示甚至消失。因此本文采用带有转置卷积模块的ResNet作为CenterNet的骨干网络。

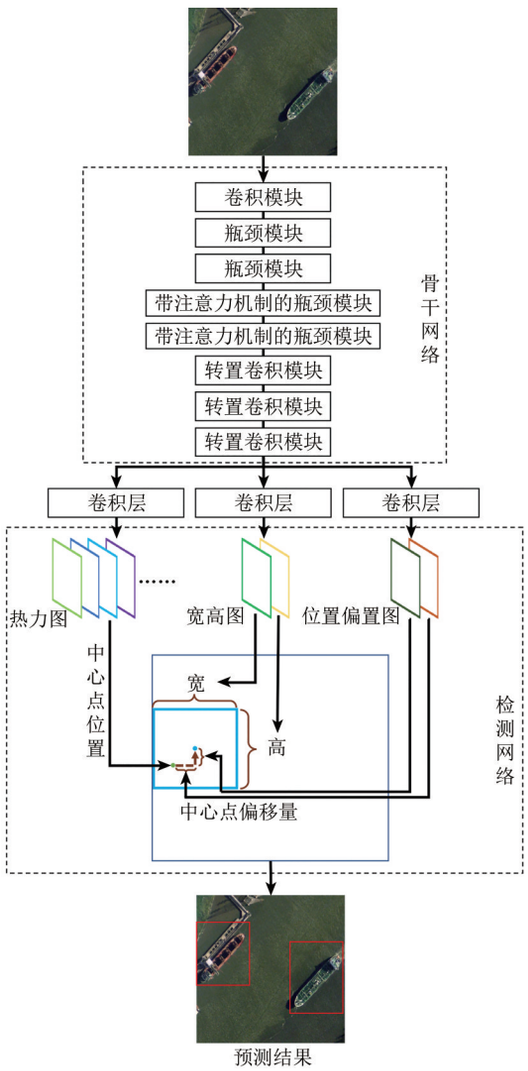

图1 所提遥感图像目标检测方法流程图

Fig. 1 The flow chart of the proposed object detection method for remote sensing images

传统卷积通过卷积核对输入特征图中每个区域内的值进行卷积运算，将结果作为输出特征值。其中，填充（Padding）操作通过在输入特征图的外围补零，来扩大卷积操作覆盖的区域；步长（Stride）则为每一步卷积操作之间的距离。因此，传统卷积的输出特征图大小通常小于输入特征图；填充越大，输出特征图越大；步长越大，输出特征图越小。转置卷积的目的是为了使输出特征图的大小大于输入特征图，其卷积核的运算过程虽然与传统卷积相同，但输入特征图的默认状态不同，填充以及步长的概念也与传统卷积相反。与传统卷积中输入特征图默认无补零不同，转置卷积在默认情况下就在输入特征图的外围补零，使得卷积核的感知区域恰好可以覆盖特征图中的一个点。这样使得转置卷积在默认情况下的输出特征图与输入特征图的大小一致。在转置卷积中，填充操作用以将补零的区域缩小；步长则是指输入特征图中相邻点之间设置的间隔，特征图相邻点之间空余的部分用零代替。因此，在转置卷积中，填充越大，输入特征图反而越小，输出特征图也随之减小；步长越大，输入特征图中相邻点的距离越大，输出特征图也随之增大。因为转置卷积可以扩大卷积操作的输出特征图，所以常被应用于解码器等结构之中，其过程如图2所示。

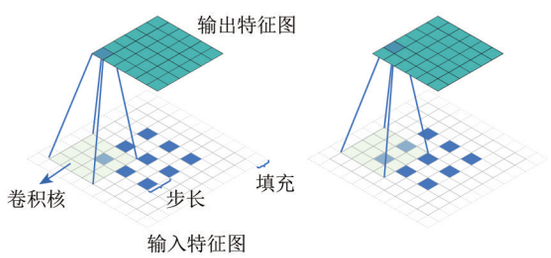

图2 转置卷积示意图

Fig. 2 The illustration of the transposed convolution

如图1所示，本文采用的骨干网络共包含1个卷积模块、4个瓶颈模块以及3个转置卷积模块，详细参数见表1。经过前5个层级的模块后，特征图的尺寸为原始图像的1/32。再经过3个转置卷积模块的上采样，最终得到的特征图的尺寸为原始图像的1/4。

表1 骨干网络详细参数

Table 1 Detail parameters of backbone network

模块层级	网络参数
1	卷积模块(7×7,通道数:64,步长:2,填充:3,池化:2)
2	瓶颈模块(通道数:64,步长:1)
3	瓶颈模块(通道数:128,步长:2)
4	带注意力机制的瓶颈模块(通道数:256,步长:2)
5	带注意力机制的瓶颈模块(通道数:512,步长:2)
6	转置卷积模块(4×4,通道数:256,步长:2,填充:1)
7	转置卷积模块(4×4,通道数:128,步长:2,填充:1)
8	转置卷积模块(4×4,通道数:64,步长:2,填充:1)

人眼通常会根据需求更加注意视觉场景中的特定部分，而忽视其他无关部分。由于本文采用的是基于关键点的方法，因此需要将特征图中的有效特征尽可能地集中于目标关键点所在的区域。为了达到这个效果，骨干网络中引入了注意力机制，以学习特征中每一个元素的重要程度，从而

对元素进行加权。注意力机制的核心问题就是如何求得每个元素的权重系数（Zhu等，2019）。

注意力机制中的主要内容包括查询（Query）项、键（Key）项以及值（Value）项。通过计算查询项和键项之间的相似性或相关性，计算键项对应的值项的权重值。除此之外，为了在注意力机制中有效地利用位置信息，往往在查询项和键项之中加入位置编码项来实现该目的。在本文中，采用特征图中的像素点作为注意力机制中的元素。查询项$\boldsymbol{Q}$、键项$\boldsymbol{K}$、相对位置编码项$\boldsymbol{J}$以及值项$\boldsymbol{V}$的计算过程为

$$\begin{cases}\boldsymbol{Q}=\mathrm{C}_q(\boldsymbol{U}_q)\\ \boldsymbol{K}=\mathrm{C}_k(\boldsymbol{U}_{kv})\\ \boldsymbol{J}=\mathrm{F}_x(\boldsymbol{E}_x)+\mathrm{F}_y(\boldsymbol{E}_y)\\ \boldsymbol{V}=\mathrm{C}_v(\boldsymbol{U}_{kv})\end{cases}\tag{1}$$

式中，$\boldsymbol{U}_q$和$\boldsymbol{U}_{kv}$分别表示查询项和键项的原始数据，C表示卷积层，F表示全连接层，$\boldsymbol{E}_x$和$\boldsymbol{E}_y$分别表示水平和垂直方向的位置编码。

在本文中，$\boldsymbol{U}_q$采用的是原始的输入特征图，$\boldsymbol{U}_{kv}$为经过2×2池化的特征图，池化的目的是为了减小维度以降低计算量。最终得到的输出特征图$\boldsymbol{U}'$为

$$\begin{cases}\boldsymbol{A}=\mathrm{softmax}\left((\boldsymbol{Q}+b_1)\times\boldsymbol{K}+(\boldsymbol{Q}+b_2)\times\boldsymbol{J}\right)\times\boldsymbol{V}\\ \boldsymbol{U}'=\boldsymbol{U}+\tau\mathrm{C}_a(\boldsymbol{A})\end{cases}\tag{2}$$

式中，b_1和b_2为查询项针对键项和位置编码项的偏置值，τ是用来调节注意力机制在输出特征中权重的可学习系数。

对于位置编码项，本文采用正余弦函数作为查询项和键项之间相对位置的编码方式。由于图像特征是二维的，因此需要对横纵坐标分别进行编码。具体而言，首先计算查询项和键项在横纵坐标上的距离；然后对每个通道对应的正余弦函数分配不同的波长值，分别计算正余弦函数的输出值；最后将正弦和余弦函数的输出值进行拼接，得到最终的输出结果。每个通道波长值的计算参考（Vaswani等，2017）得到。整个位置编码的运算过程为

$$\begin{cases}E_x(i,j,d)=\sin\left(\left(\rho_q^x(i)-\rho_k^x(j)\right)/\omega^{\frac{d}{d_m/4}}\right)\\ E_x(i,j,d+d_m/4)=\cos\left(\left(\rho_q^x(i)-\rho_k^x(j)\right)/\omega^{\frac{d}{d_m/4}}\right)\\ E_y(i,j,d)=\sin\left(\left(\rho_q^y(i)-\rho_k^y(j)\right)/\omega^{\frac{d}{d_m/4}}\right)\\ E_y(i,j,d+d_m/4)=\cos\left(\left(\rho_q^y(i)-\rho_k^y(j)\right)/\omega^{\frac{d}{d_m/4}}\right)\end{cases}\tag{3}$$

式中，$\rho_q^x(i)$和$\rho_k^x(j)$分别表示查询项第i个位置和键项第j个位置的横坐标，$\rho_q^y(i)$和$\rho_k^y(j)$表示它们的纵坐标。d_m表示输入特征的通道维度，d表示位置编码通道维度的索引。ω为决定正余弦函数波长的参数，此处被设为1000。

本文采用的是多头注意力机制（Multi-head Attention）。多头注意力机制将同一个组数据输入到多个不同的注意力头之中，并将多个注意力头的输出进行拼接，拼接后的输出值经过全连接层映射，作为整个多头注意力机制的输出。单个注意力头的流程如图3所示，其中，D、H、W分别表示特征图的通道数、高和宽，n_{bs}为同一批训练的样本数量，n_h为多头注意力机制中的注意力头数量，d_e为某一个注意力头中的特征通道数。

2.2 检测网络

在经过骨干网络得到特征图之后，需要通过检测网络来进一步预测目标框以及对应的类别。检测网络由3个全卷积的子网络组成，分别预测中心点的热力图、各个位置对应的宽高以及预测中心点因特征图与原始图像尺寸差异造成的位置偏差。

中心点热力图生成子网络由一个卷积核大小为3×3的卷积层、一个ReLU函数以及一个卷积核大小为1×1的卷积层组成。生成的热力图通道数为目标的类别个数，每一通道仅用来预测一个类别的关键点。另外两个子网络与热力图生成子网络类似，不同的是生成的宽高图以及位置偏置图的通道数均为2，分别预测宽和高以及中心点位置在纵横方向的偏置值。这些预测值与类别无关。检测过程如图1所示。

在预测的过程，首先寻找热力图中的极值点，并根据极值点对应的值的高低排序，得到固定数量的中心点。随后寻找这些中心点对应位置预测得到的宽高以及位置偏置值。最终得到的目标框左上角坐标$(\hat{x}_{lt},\ \hat{y}_{lt})$和右下角坐标$(\hat{x}_{rb},\ \hat{y}_{rb})$分别为

$$\begin{cases}\hat{x}_{lt}=(\hat{x}+\Delta\hat{x}-\hat{w}/2)\times s\\ \hat{y}_{lt}=(\hat{y}+\Delta\hat{y}-\hat{h}/2)\times s\\ \hat{x}_{rb}=(\hat{x}+\Delta\hat{x}+\hat{w}/2)\times s\\ \hat{y}_{rb}=(\hat{y}+\Delta\hat{y}+\hat{h}/2)\times s\end{cases}\tag{4}$$

式中，s是特征图相较于原始图像的下采样尺度，$(\hat{x},\ \hat{y})$为中心点位置，$(\hat{w},\ \hat{h})$为该位置对应的宽高预测值，$(\Delta\hat{x},\ \Delta\hat{y})$为该位置对应的位置偏置预测值。

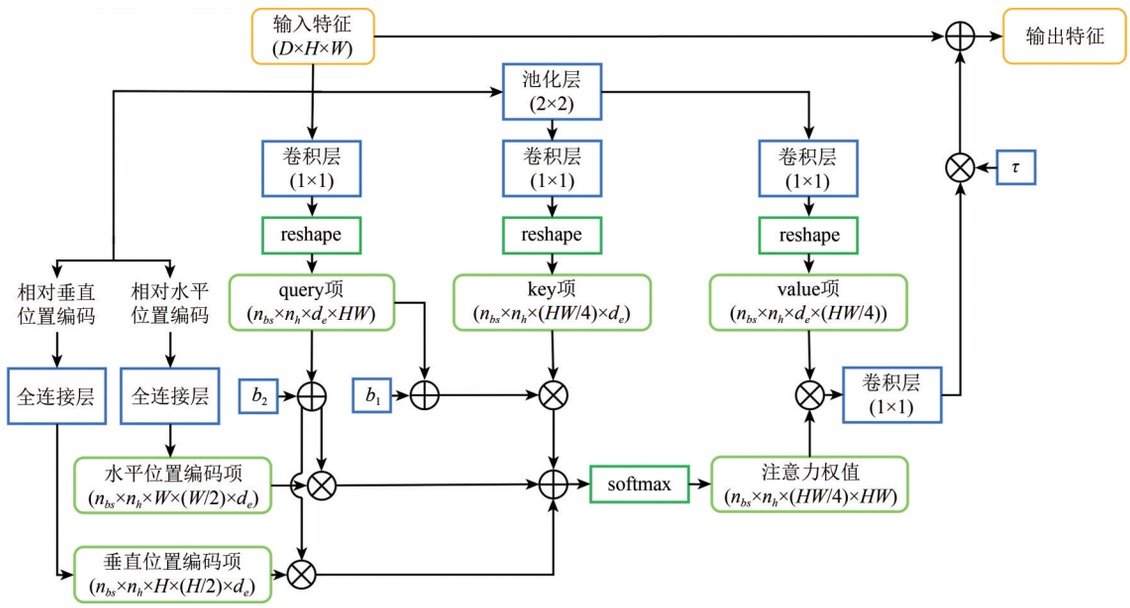

图3 注意力机制流程示意图

Fig. 3 The flow chart of the attention mechanism

2.3 损失函数

在网络训练中，需要基于目标框的真实坐标生成训练用的标签。假设训练集中目标框的标注格式为$(x_{lt}, y_{lt}, x_{rb}, y_{rb}, c)$，它们分别表示目标框左上角和右下角顶点的横纵坐标以及类别。CenterNet的检测网络有3部分，因此需要针对这3个部分分别生成训练用的标签。对于宽高的预测而言，直接利用两个定点的差值即可得到。位置偏置值则需要考虑下采样的影响。理想中的中心点位置应为$x_c = 0.5(x_{lt} + x_{rb})/s$以及$y_c = 0.5(y_{lt} + y_{rb})/s$，但是由于该位置坐标可能不是整数，因此需要进行向下取整操作来得到接近的中心点位置坐标。用$\lfloor\cdot\rfloor$表示该操作，则$(\tilde{x}_c, \tilde{y}_c) = (\lfloor x_c/s \rfloor, \lfloor y_c/s \rfloor)$。理想位置与取整后的差值，即为位置偏置值的标签。除了宽高和位置偏置值外，最重要的是热力图标签的生成。

在原始的CenterNet中，利用高斯核将关键点分布到生成的热力图上。为了减小计算量，同时避免不同目标间的影响，需要将高斯核限制在一定范围内，通常是一个边长为l的正方形内。也即只在该正方形区域内进行高斯核的计算。热力图中坐标（x，y）处值的计算方法为

$$G(x,y) = \exp\left(-\frac{(x-\tilde{x}_c)^2 + (y-\tilde{y}_c)^2}{2\sigma^2}\right) \tag{5}$$

式中，σ是一个与目标大小相关的标准差，通常令$\sigma = l/6$。正方形区域外的热力图的值为0。

为了计算合适的边长l，原始CenterNet继承了CornerNet中的计算方法。该计算方法考虑了在如图4所示的3种极限情况下，关键点最边缘的点所构成的框与原始框之间的交并比IoU（Intersection over Union）仍然可以大于一定的阈值。虽然这种方法在CenterNet中可以得到结果，但是由于CenterNet是将中心点作为关键点，CornerNet则是将两个角点作为关键点，因此两种方法的3种极限情况并不相同。CenterNet的3种极限情况如图5所示。为了使高斯核适用范围边长与CenterNet的关键点使用方式更加一致，本文针对上述CenterNet的3种情况，提出了高斯核适用范围边长的计算方法。从左至右他们的IoU计算公式以及对应的边长值分别如式(6)—式(8)所示。

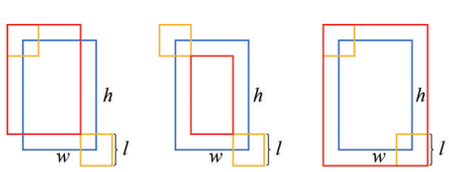

图4 CornerNet中的3种极限情况

Fig. 4 Three limiting cases in CornerNet

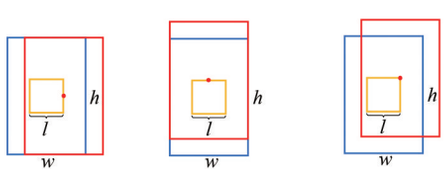

图5 CenterNet中的3种极限情况

Fig. 5 Three limiting cases in CenterNet

$$\begin{cases} \mathrm{IoU} = \dfrac{(2w - l)h}{(2w + l)h} \\ l = \dfrac{2w(1 - \mathrm{IoU})}{1 + \mathrm{IoU}} \end{cases} \tag{6}$$

$$\begin{cases} \mathrm{IoU} = \dfrac{w(2h - l)}{w(2h + l)} \\ l = \dfrac{2h(1 - \mathrm{IoU})}{1 + \mathrm{IoU}} \end{cases} \tag{7}$$

$$\begin{cases} \mathrm{IoU} = \dfrac{(2w - l)(2h - l)}{4wh - (2w - l)(2h - l)} \\ l = \dfrac{(1 + \mathrm{IoU})(w + h) + \sqrt{(1 + \mathrm{IoU})^2(w + h)^2 - 4(1 - \mathrm{IoU}^2)wh}}{1 + \mathrm{IoU}} \end{cases} \tag{8}$$

在确认了高斯核适用范围的边长后，即可根据式(5)和真实框的标签生成热力图。

在得到各个检测网络对应的标签之后，即可构建预测值和标签值之间的损失函数，对网络进行训练。其中，在热力图的训练中，热力图损失函数L_{hm}采用的是类似于Focal Loss的损失函数，其公式为

$$L_{\mathrm{hm}} = \frac{-1}{N}\sum_{x,\,y}\begin{cases} \left(1-\hat{G}(x,\,y)\right)^{\alpha}\log\left(\hat{G}(x,\,y)\right), & G(x,\,y) = 1 \\ \left(1-G(x,\,y)\right)^{\beta}\hat{G}(x,\,y)^{\alpha}\log\left(1-\hat{G}(x,\,y)\right), & \text{other} \end{cases} \tag{9}$$

式中，$\hat{G}(x,\,y)$为网络预测的热力图在$(x,\,y)$处的值，α和β是两个超参数，这里分别设置为2和4。N为关键点的个数。

对于目标的宽和高，以及目标在特征图上的中心位置与对应原图位置的偏移量，采用的是常见的L1损失函数。宽高的损失函数L_{wh}和中心点位置偏移的损失函数L_{os}分别为

$$\begin{cases} L_{\mathrm{wh}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\left|w_n - \hat{w}_n\right| + \left|h_n - \hat{h}_n\right| \\ L_{\mathrm{os}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\left|\Delta x_n - \Delta\hat{x}_n\right| + \left|\Delta y_n - \Delta\hat{y}_n\right| \end{cases} \tag{10}$$

最终得到的整体的损失函数是上述3个损失函数之和，即：

$$L_{\mathrm{all}} = \lambda_{\mathrm{hm}}L_{\mathrm{hm}} + \lambda_{\mathrm{wh}}L_{\mathrm{wh}} + \lambda_{\mathrm{os}}L_{\mathrm{os}} \tag{11}$$

式中，λ为各项的权重值。

3 实验和分析

3.1 实验数据

本文实验部分采用西北工业大学公布的高分辨率遥感图像数据集DIOR（Li等，2020）进行实验验证。该数据集共包含23463幅遥感图像，20个类别共190288个目标。类别包括飞机、机场、棒球场、篮球场、桥梁、烟囱、大坝、快速道路服务区、快速道路收费站、高尔夫球场、田径场、港口、立交桥、船舶、体育场、储油罐、网球场、火车站、车辆和风车。图像的空间分辨率在0.5—30 m，图像尺寸为800×800像素。为了保证训练集和测试集中图像的一致性，DIOR数据集在使用中被随机划分到两个数据集中。其中，训练集包含11725幅图像，测试集包含11738幅图像。

3.2 评价指标

为了更好地评价检测方法的表现，实验采用了常用的平均精度AP（Average Precision）作为评价指标。检测结果可以分为3类，分别是正确预测的正样本TP（True Positive），错误预测的正样本FP（False Positive）以及错误预测的负样本FN（False Negative）。用精度指标P表示所有检出目标中正确预测的比例，召回指标R表示正确预测的目标占应检出目标的比例，则：

$$\begin{cases} P = \dfrac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \\ R = \dfrac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \end{cases} \tag{12}$$

AP则是计算从$R = 0$到$R = 1$区间内精度P的平均值，更高的AP值意味着检测效果越好，反之亦然。由于AP是针对单个类别，而测试的数据集具有多个类别标签。因此采用多个类别AP的平均值用于多类目标检测效果的衡量，这个平均值常用mAP（Mean Average Precision）表示。

3.3 实验设置

本次实验中，所提方法的骨干网络基于ResNet-50。其中，如表1所示，转置卷积的卷积核大小为4×4，填充为1，步长为2。这样设计可以使得输出特征图的宽高是输入特征图的两倍。在网络的训练过程中，计算高斯核适用范围边长l时的边缘点所构成的框与原始框之间的IoU阈值设置为0.7。计算损失函数时，3个损失函数的权重值分别是$\lambda_{\mathrm{hm}} = 1$、$\lambda_{\mathrm{wh}} = 0.1$、$\lambda_{\mathrm{os}} = 1$。所提方法采用带动量的随机梯度下降算法进行训练。根据实验情况，训练在120代左右可保持相对稳定，因此训练代数设置为140。在测试时，先取中心点热力图中每个类别通道上排名前100的极值点，然后再将所有类

别取出的极值点排序，取前100个作为预测框的中心点。此外，为了减少CenterNet因热力图区域较大导致在大尺度目标检测中出现的重复框的情况，方法在预测网络的后端添加了非极大值抑制NMS（Non-Maximum Suppression）操作。所有实验均在搭载英伟达GTX 2080Ti显卡的Linux系统下完成。

3.4 与原始方法的对比

相较于原始CenterNet，本文所采用的方法主要包含以下3点改进。首先，采用带有转置卷积模块的ResNet-50取代默认的Hourglass-104作为骨干网络，减少了网络参数数量。然后，针对原本采用的CornerNet式的高斯核适用范围边长的计算方式不匹配CenterNet的问题，提出了以中心点预测为基础的边长求解方法。最后，为了使网络能够更加注意于目标所在的区域，引入了注意力机制，用以增强网络所提取的特征。

为了更好地比较4种方法，分别用CN-H表示采用Hourglass作为骨干网络的CenterNet，CN-R表示采用带有转置卷积模块的ResNet作为骨干网络的CenterNet，CN-RL是在CN-R的基础上修改了边长求解方式，CN-RLA则是在CN-RL的基础上进一步引入了注意力机制。4种方法除了上述不同点之外，其余的参数以及训练方式均保持一致，预测网络后段均添加了NMS操作。不同方法所得到的预测精度结果如表2所示，检测速度如表3所示，其中加黑数字表示几种方法在该项结果中的最优值。

实验结果表明，虽然Hourglass-104的参数数量要远多于ResNet-50，但CN-H在DIOR数据集上的整体表现并没有比CN-R好。与之相反，CN-R的mAP比CN-H高出约1.4%。除此之外，相较于CN-R平均每张图片34.9 ms的耗时，由于Hourglass网络参数过多，CN-H的测试时间达到了平均每张125.8 ms，约为CN-R耗时的4倍。在不同的高斯核适用范围边长生成方法上，CN-RL相较于CN-R提高了1.1%，其中在14个类上都表现更好。这说明本文提出的以中心点为基础的边长计算方法优于基于CornerNet式的边长计算方法，更符合CenterNet的需求。注意力机制的添加可以使得CN-RLA在CN-RL的基础上，将mAP再提升1.5%。但是也需要注意到，在飞机、储油罐、车辆等类别上，CN-RLA的结果反而要差于CN-H。这些类别的目标往往分布较为密集，目标之间存在一定的关联性。此时Hourglass的递归式结构或许能够更加有效地捕捉多个关键点之间的关联，从而提升提取特征的有效性。整体来看，本文所提出的改进后的CenterNet相较于原始的CenterNet在整体检测率上有4.0%的提升。所提方法得到的部分实验结果图如图6所示。

表2 不同改进部分对检测精度的影响

Table 2 Comparison of detection accuracy under different improvements

类别	检测方法			
	CN-H	CN-R	CN-RL	CN-RLA
飞机	**0.813**	0.782	0.772	0.759
机场	0.600	0.795	0.820	**0.840**
棒球场	0.792	0.750	0.764	**0.797**
篮球场	0.894	0.891	0.887	**0.902**
桥梁	0.413	0.407	**0.418**	0.414
烟囱	0.741	0.777	0.776	**0.798**
大坝	0.421	0.475	0.515	**0.563**
服务区	0.697	0.750	0.769	**0.788**
收费站	**0.732**	0.689	0.689	0.704
高尔夫球场	0.651	0.757	0.791	**0.805**
田径场	0.692	0.763	0.768	**0.771**
港口	0.365	0.397	0.392	**0.409**
立交桥	0.485	0.548	**0.555**	0.548
船舶	**0.756**	0.733	0.735	0.736
体育场	0.715	0.724	0.754	**0.797**
储油罐	**0.729**	0.637	0.638	0.653
网球场	**0.909**	0.895	0.889	0.891
火车站	0.440	0.519	0.521	**0.553**
车辆	**0.568**	0.478	0.487	0.505
风车	**0.852**	0.782	0.820	0.830
mAP	0.663	0.677	0.688	**0.703**

注：粗体表示当前指标的最优值。

表3 不同改进部分对检测速度的影响

Table 3 Comparison of detection speed under different improvements

参数	检测方法			
	CN-H	CN-R	CN-RL	CN-RLA
共耗时/s	1477	**410**	457	471
平均每张耗时/ms	125.8	**34.9**	38.9	40.1
每秒传输帧数	7.95	**28.63**	25.68	24.92

注：粗体表示当前指标的最优值。

图6 所提方法在部分图像上的检测结果

Fig. 6 The detection results of the proposed method on partial images

3.5 与其他方法的对比

为了更加量化地评估所提方法的检测表现，实验将所提出的方法与其他常见的方法进行比较。这些方法包括FPN（Lin等，2017a）、RetinaNet（Lin等，2017b）、FCOS（Tian等，2019）、FoveaBox（Kong等，2020）。其中FPN和RetinaNet是经典的基于锚框的检测方法，而FCOS和FoveaBox是另两种近两年提出的无锚框的目标检测方法。

4种对比方法均采用ResNet-50作为骨干网络，并采用特征金字塔结构进行不同大小特征图的提取。4种方法均采用带动量的随机梯度下降算法进行训练。根据实验情况，训练在10代左右可保持相对稳定，因此4种方法的训练代数统一设置为12。其中，FPN中区域生成网络的锚框尺寸为8，锚框宽高的比例分别是{0.5，1.0，2.0}，感兴趣区域池化方式为RoI Align，池化后的特征图大小为7×7。RetinaNet中，锚框的尺寸分别是{2^2，$2^{7/3}$，$2^{8/3}$}，锚框宽高的比例分别是{0.5，1.0，2.0}。FCOS中没有使用锚框，但其将不同尺寸特征图可回归的目标大小限制在了一定范围内，范围分别是{1-64，65-128，129-256，257-512，>512}。与FCOS类似，FoveaBox也对不同尺寸特征图可响应的目标大小进行了限制，范围分别是{1-64，32-128，64-256，128-512，512-2048}。

这些方法与所提方法在不同类别上的AP值如表4所示。此外，实验还比较了不同方法的检测速度，结果如表5所示。其中加黑数字表示几种方法在该项结果中的最优值。

实验结果显示，包括所提方法在内的无锚框检测方法已经赶上甚至超越基于锚框的FPN和RetinaNet方法，这表明无锚框方法可以在不借助锚框的情况下对目标进行检测分类。在具体类别上，所提方法在多数类别下都有着较好的检测效果，但在桥梁、田径场、立交桥等目标长宽较大的类别上仍落后于基于锚框的FPN方法。这表明所提方法对长宽较大的目标框进行回归预测时仍存在一定的难度。整体来看，所提方法的mAP值达到了70.3%，优于其他比较方法。除了检测精度外，所提方法在检测速度上也具有显著的优势，具有一定的实用性。

表4 所提方法与其他检测方法的检测精度比较

Table 4 Comparison of detection accuracy under different methods

检测方法	FPN	RetinaNet	FCOS	FoveaBox	CN-RLA
飞机	0.607	0.661	0.611	0.669	**0.759**
机场	0.770	0.784	0.826	0.796	**0.840**
棒球场	0.749	0.744	0.766	0.767	**0.797**
篮球场	0.879	0.878	0.876	0.876	**0.902**
桥梁	**0.464**	0.372	0.428	0.427	0.414
烟囱	0.802	0.802	**0.806**	0.798	0.798
大坝	0.600	0.631	**0.641**	0.606	0.563
服务区	0.767	0.767	0.791	**0.811**	0.788
收费站	**0.706**	0.588	0.672	0.664	0.704
高尔夫球场	0.773	0.799	**0.820**	0.745	0.805
田径场	**0.832**	0.764	0.796	0.803	0.771
港口	0.460	0.366	0.464	**0.500**	0.409
立交桥	**0.600**	0.566	0.578	0.576	0.548
船舶	**0.751**	0.687	0.721	0.731	0.736
体育场	0.674	0.690	0.648	0.715	**0.797**
储油罐	0.611	0.431	0.634	0.600	**0.653**
网球场	0.873	0.857	0.852	0.869	**0.891**
火车站	0.588	0.584	**0.628**	0.545	0.553
车辆	0.450	0.407	0.438	0.427	**0.505**
风车	0.879	0.851	0.875	**0.883**	0.830
mAP	0.692	0.661	0.694	0.690	**0.703**

注：粗体表示当前指标的最优值。

表 5 所提方法与其他检测方法检测速度比较

Table 5 Comparison of detection speed under different methods

参数	检测方法				
	FPN	RetinaNet	FCOS	FoveaBox	CN-RLA
共耗时/s	685	658	604	714	**471**
平均每张耗时/ms	58.4	56.1	51.5	60.8	**40.1**
每秒传输帧数	17.14	17.84	19.43	16.44	**24.92**

注:粗体表示当前指标的最优值。

4 结 论

本文针对遥感图像目标检测方法的效率问题，采用了改进CenterNet来减少目标检测所需要的步骤，降低对于锚框的依赖。所提方法通过采用带转置卷积模块的ResNet降低了网络的参数量，加快检测速度；提出了针对中心点预测方式的高斯核适用范围边长的计算方法，解决了原有计算方法与CenterNet的不匹配问题，提升了热力图的标签质量，更加有利于训练；通过引入注意力机制，对所提取的特征进行增强，使其更加注意于目标区域。

在公开的DIOR数据集上的实验结果表明，所提出的改进CenterNet在具有较快检测速度的情况下同时具备较高的检测精度，表明CenterNet等无锚框方法在不采用锚框等基准信息的情况下也能够取得较好的检测效果，为下一步无锚框方法在遥感目标检测上的应用提供了思路和参考。实验结果也同时表明，所提方法对分布密集的目标以及长宽较大的目标的预测仍有不足，未来将针对以上类型目标进行重点研究。

参考文献(References)

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers. arXiv: 2005.12872

Chen Q, Wang Y M, Yang T, Zhang X Y, Cheng J and Sun J. 2021. You only look one-level feature//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 13034-13043 [DOI: 10.1109/CVPR46437.2021.01284]

Chen R, Liu Y, Zhang M D, Liu S, Yu B and Tai Y W. 2020. Dive deeper into box for object detection//16th European Conference on Computer Vision. Glasgow: Springer: 412-428 [DOI: 10.1007/978-3-030-58542-6_25]

Cheng G and Han J W. 2016. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 117: 11-28 [DOI: 10.1016/j.isprsjprs.2016.03.014]

Dai Z G, Cai B L, Lin Y G and Chen J Y. 2021. UP-DETR: unsupervised pre-training for object detection with transformers//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 1601-1610 [DOI: 10.1109/CVPR46437.2021.00165]

Girshick R, Donahue J, Darrell T and Malik J. 2016. Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1): 142-158 [DOI: 10.1109/TPAMI.2015.2437384]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hong M B, Li S W, Yang Y C, Zhu F Y, Zhao Q J and Lu L. 2022. SSPNet: scale selection pyramid network for tiny person detection from UAV images. IEEE Geoscience and Remote Sensing Letters, 19: 8018505 [DOI: 10.1109/LGRS.2021.3103069]

Kong T, Sun F C, Liu H P, Jiang Y N, Li L and Shi J B. 2020. FoveaBox: beyound anchor-based object detection. IEEE Transactions on Image Processing, 29: 7389-7398 [DOI: 10.1109/TIP.2020.3002345]

Law H and Deng J. 2018. CornerNet: detecting objects as paired keypoints//15th European Conference on Computer Vision. Munich: Springer: 765-781 [DOI: 10.1007/978-3-030-01264-9_45]

Li K, Wan G, Cheng G, Meng L Q and Han J W. 2020. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 159: 296-307 [DOI: 10.1016/j.isprsjprs.2019.11.023]

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017a. Feature pyramid networks for object detection//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Lin T Y, Goyal P, Girshick R, He K M and Dolár P. 2017b. Focal loss for dense object detection//2017 IEEE International Conference on Computer Vision. Venice: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//14 European Conference on Computer Vision. Amsterdam: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//14th European Conference on Computer Vision. Amsterdam: Springer: 483-499 [DOI: 10.1007/978-3-319-46484-8_29]

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press: 91-99

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 6000-6010

Wang C, Bai X, Wang S, Zhou J and Ren P. 2019a. Multiscale visual attention networks for object detection in VHR remote sensing images. IEEE Geoscience and Remote Sensing Letters, 16(2): 310-314 [DOI: 10.1109/LGRS.2018.2872355]

Wang J Q, Chen K, Yang S, Loy C C and Lin D H. 2019b. Region proposal by guided anchoring//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 2960-2969 [DOI: 10.1109/CVPR.2019.00308]

Xiao B, Wu H P and Wei Y C. 2018. Simple baselines for human pose estimation and tracking//15th European Conference on Computer Vision. Munich: Springer: 472-487 [DOI: 10.1007/978-3-030-01231-1_29]

Yang X, Yan J C, Feng Z M and He T. 2021. R3Det: refined single-stage detector with feature refinement for rotating object. Proceedings of the AAAI Conference on Artificial Intelligence, 35(4): 3163-3171 [DOI: 10.1609/aaai.v35i4.16426]

Zhou X Y, Wang D Q and Krähenbühl P. 2019. Objects as points. arXiv: 1904.07850

Zhu X Z, Cheng D Z, Zhang Z, Lin S and Dai J F. 2019. An empirical study of spatial attention mechanisms in deep networks//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 6687-6696 [DOI: 10.1109/ICCV.2019.00679]

Application of an improved CenterNet in remote sensing images object detection

TIAN Zhuangzhuang[1], ZHANG Hengwei[1], WANG Kun[1], LIU Shengqi[2], ZOU Qianjin[1], ZHAO Zhen[1], CHEN Yubin[1]

1.State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System (CEMEE), Luoyang 471000, China;
2.College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract: Nowadays, object detection methods based on deep learning are widely used in the interpretation of remote sensing images. The anchor-based methods usually need to design the anchor boxes first, which requires more detection steps and time cost. This study proposed an object detection method of remote sensing images based on the improved CenterNet. The method can simplify the object detection process and improve efficiency.

The CenterNet uses a fully convolutional network to directly predict the heat map of the center points, widths, and heights of the corresponding objects, and the position offsets of the center points. The heat maps are used to generate the rough positions of the objects, and the offsets can fine-tune the positions to make them more accurate. The widths and heights further constitute the shape of the object boxes. The different heat maps decide the object categories. On the basis of CenterNet, the proposed method first adopts the ResNet with transposed convolution as the backbone network. The transposed convolution can expand the output feature maps, and ResNet can reduce the number of parameters in the backbone network compared with the Hourglass network. Second, the proposed method defines the length of Gaussian kernel under three limit conditions between the predicted and real boxes in CenterNet. The Gaussian kernel is applied to generate the heat map label, which is used for network training. Finally, the multi-head attention mechanism is introduced into the backbone network to learn the importance of each element in the feature maps. The weights assigned to the elements reflect their effectiveness, which makes the effective features concentrate in the regions of the object key points as much as possible.

The experiments use mean Average Precision (mAP) to evaluate the object detection results on the multiple categories. All the experiments are conducted at the DIOR dataset. The results show that the CenterNet using the ResNet with transposed convolution is 1.4% higher than that using the Hourglass. The proposed calculation of the length of the Gaussian kernel can increase the mAP by 1.1%. The addition of attention mechanism can further improve the mAP by 1.5%. At the same time, the proposed method reduces the time cost by 31.9% compared with the conventional method. The experimental results show that the proposed method can improve detection accuracy without sacrificing the detection speed. The ablation experiments of different parts also show that the ResNet with transposed convolution, the designed calculation method of the length of the Gaussian kernel, and the attention mechanism can effectively improve the mAP. The comparison with other methods also proves that the proposed method is practical.

Key words: remote sensing image, object detection, deep learning, CenterNet, attention mechanism

Supported by National Natural Science Foundation of China (No. 62001486)