FAGNet: 基于 MAFPN 和 GVR 的遥感图像多尺度目标检测算法

郑哲, 雷琳*, 孙浩, 匡纲要

(国防科技大学电子科学学院电子信息系统复杂电磁环境效应国家重点实验室 长沙 410073)
(alaleilin@163.com)

摘要: 针对大场景遥感图像内容复杂, 并且具有目标种类较多、尺度不一、方向多变等特点, 导致遥感图像中目标多类多尺度多方向的问题, 提出一种基于多尺度注意力特征金字塔网络(MAFPN)以及滑动顶点回归(GVR)机制的遥感图像目标检测方法. 首先利用骨干网络提取多层特征作为 MAFPN 的输入, MAFPN 结合特征融合和注意力机制, 在融合多个尺度的特征映射的基础上使用通道域注意力和空间域注意力机制来抑制噪声, 增强有效特征复用, 提高网络对目标多尺度特征的自适应性; 将 MAFPN 输出的融合特征图输入区域建议网络(RPN)生成感兴趣区域, 然后将其送入分类/回归网络; 在分类/回归网络中使用 GVR 机制在预测水平框的基础上增加 4 个顶点偏移比例参数和旋转因子, 将水平框转换为旋转框, 以减少边框中冗余区域, 使预测得到的旋转边框更贴合目标. 在 DOTA 公开数据集上与多种基于卷积神经网络的经典检测算法进行对比的实验结果表明, 该方法的平均检测精度得到显著提高, 能够更加准确地检测多个尺度以及多个方向的目标, 实现了多尺度目标的鲁棒性检测.

关键词: 遥感图像; 目标检测; 注意力机制; 特征融合; 卷积神经网络
中图法分类号: TP753 **DOI:** 10.3724/SP.J.1089.2021.18608

FAGNet: Multi-Scale Object Detection Method in Remote Sensing Images by Combining MAFPN and GVR

Zheng Zhe, Lei Lin*, Sun Hao, and Kuang Gangyao

(State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, College of Electronic Science, National University of Defense Technology, Changsha 410073)

Abstract: Remote sensing images of large scenes are complex, and have the characteristics of many categories of objects, different scales and changeable directions, which lead to the problem of multi-class, multi-scale and multi-oriented of objects in remote sensing images. A remote sensing image object detection method based on multi-scale attention feature pyramid network (MAFPN) and gliding vertex regression (GVR) mechanism is proposed. Firstly, multi-layer feature maps are extracted from backbone network as input of MAFPN, which combines feature fusion and attention mechanism. On the basis of fusing feature maps of multi-scale, channel attention and spatial attention mechanisms are used to suppress noise, enhance effective feature reuse, and improve the network's adaptability to object multi-scale features. The fusion feature map output by MAFPN is input to the region proposal network to generate the regions of interest, and then they will be sent to the classification regression network. The GVR mechanism is used in the object classification regression network and the four vertex offset ratio parameters and rotation factors are added on the basis of predicting the horizontal boxes, which converts the horizontal boxes into a rotating box to re-

收稿日期: 2020-08-26; 修回日期: 2021-02-27. 基金项目: 国家自然科学基金(61971426). 郑哲(1996—), 女, 硕士研究生, 主要研究方向为遥感图像目标检测、计算机视觉; 雷琳(1980—), 女, 博士, 教授, 硕士生导师, 论文通讯作者, 主要研究方向为多源遥感信息融合; 孙浩(1984—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为计算机视觉与人工智能; 匡纲要(1966—), 男, 博士, 教授, 博士生导师, 主要研究方向为遥感图像智能解译、SAR 图像处理.

duce the redundant area in the bounding boxes, makes the predicted rotating bounding boxes fit the object more closely. The experimental results on the DOTA public dataset, compared with many classical detection algorithms based on convolutional neural networks, show that the average detection accuracy of the proposed method is significantly improved, which can detect objects of multi-scales and multi-oriented more accurately, and achieve the robust detection of multi-scale objects.

Key words: remote sensing image; object detection; attention mechanism; feature fusion; convolutional neural networks

目标检测是遥感图像解译分析中一个重要研究方向,随着对地观测技术的发展,针对遥感图像目标检测的研究越来越受到重视.遥感图像目标检测旨在自动、准确、高效地找出遥感图像中的感兴趣目标,并同时给出目标的类别和位置,这在军事作战、国防建设、城市规划和环境监测等许多实际应用中发挥着至关重要的作用.近年来,随着深度卷积神经网络在计算机视觉中的深入应用^[1],卷积神经网络成为了目标检测的有力工具,基于卷积神经网络的方法不仅可以拟合大量复杂数据,还能自动学习图像中最有益于目标检测的特征,充分挖掘图像信息,因此在目标检测问题上深度卷积神经网络方法比传统方法更具优势.

目前的目标检测框架可分为 2 类.(1) 两阶段方法.这类基于区域预测的方法在深度特征的基础之上预测得到若干候选目标区域,然后将其输入区域分类/回归网络得到目标的类别和位置.基于区域预测的方法代表性的有 Faster R-CNN^[2], R-FCN^[3], Mask R-CNN^[4], Cascade R-CNN^[5]等. Han 等^[6]在 Faster R-CNN 的基础上进行改进,提出 R-P-Faster R-CNN 算法,有效地改善了小目标的检测效果,提高了算法的检测性能.邓志鹏等^[7]提出基于多尺度形变特征的遥感图像目标检测框架,由区域预测网络和区域鉴别网络两部分组成,有效提升了检测精度. Li 等^[8]提出一种新的检测框架,其在 Faster R-CNN 中的区域建议网络(region proposal networks, RPN)中引入多角度锚框,可以有效地处理遥感图像中目标任意方向分布的问题,同时为了解决遥感图像中目标外观模糊,提出了一种双通道特征融合网络.姚群力等^[9]设计了多尺度特征金字塔 EFPN 并引入聚焦分类损失作为分类损失函数,有效地提高了多尺度遥感目标的检测能力. Tang 等^[10]提出一种超区域建议网络 HRPN 寻找疑似车辆目标区域,并使用硬负样本挖掘进一步提高了检测精度. Yang 等^[11]提出一种基于旋

转区域的船舶检测框架 R-DFPN,该方法主要使用密集连接特征金字塔增强特征的传播,促进特征复用,使用旋转锚框解决密集排列目标的检测困难问题. Zhu 等^[12]提出一种结合级联思想的机场检测方法,通过改进 RPN 和检测网络的损失函数提高机场的检测精度. Liu 等^[13]提出的旋转边框可以有效地处理任意方向的车辆、舰船和飞机等目标的检测问题.这类基于区域预测的算法检测精度较高,但由于是两阶段方法,检测速度仍然较慢.(2) 单阶段方法.这类方法无需提取候选目标区域,直接从深度特征图中进行目标分类和位置回归,得到目标检测结果.这类方法的检测精度与第 1 类方法相比略逊一筹,但检测速度较快,具有代表性的算法是 YOLO^[14]系列、SSD^[15]系列、RetinaNet^[16]以及 CornerNet^[17]等.为实现遥感图像中实时的目标检测,已有研究人员将单阶段方法应用到遥感领域.例如, Tang 等^[18]借鉴了 SSD 检测算法的思想从而实现更快速的车辆目标检测. Liu 等^[19]以 YOLOv2^[20]网络为基础设计了一个可以检测任意方向船舶的检测框架.

以上方法都具有卓越的性能,但遥感图像场景复杂,目标种类较多、不同目标的尺度可能较小也可能较大,并且存在密集排布的现象.因此,该领域仍存在如下问题.

(1) 遥感图像中目标尺度差异较大,尤其是较小尺度的目标,在经过多次池化操作后,小尺度目标的信息损失严重,所以小目标的检测效果有待进一步提升.

(2) 遥感图像中感兴趣目标分布密集,如车辆和舰船等,可能存在排列紧密以及分布杂乱的情况,导致在同一感兴趣区域(region of interest, RoI)中可能存在多个目标实例,且极易引入背景噪声,致使检测精度降低.

(3) 遥感图像中目标的方向具有随机性,且有些目标的纵横比较大,针对这类纵横比较大的目

标的检测,若使用水平边框进行训练和测试,对于本身处于倾斜角度的目标的预测结果很有可能会被非极大值抑制(non-maximum suppression, NMS)错误地抑制,从而导致漏检问题。

针对以上问题,本文在 Faster R-CNN 的基础上提出基于注意力机制的多尺度任意方向目标检测算法 FAGNet(feature pyramid network + attention + gliding vertex network),其主要包含3个部分:多尺度注意力特征金字塔网络(multi-scale attention feature pyramid network, MAFPN), RPN 以及分类/回归网络。MAFPN 首先基于 ResNet^[21]骨干网络设计了密集连接特征金字塔网络以提取具有更丰富语义信息的深层特征图,与自然场景图像相比,遥感图像目标尺度差异增大,且由于分辨率和拍摄角度等不同,导致遥感图像目标检测难度变大,为了使网络能够更加关注有效区域(目标区域),MAFPN 在密集特征金字塔中的多个尺度的融合特征图上加入空间域注意力机制和通道域注意力机制,使二者互为补充,降低无效信息的干扰,突出目标信息,在增强特征复用的基础上有效地选择对目标任务有益的特征信息,赋予其更大的权重,增强网络对多尺度特征的表达能力。将最终得到的特征图送入 RPN,根据所设置的锚框尺度、比例和滑动步长得到一系列矩形区域,经过初步的分类和筛选得到疑似含有目标的候选区域,将其送入分类/回归网络。在回归阶段引入了滑动顶点回归机制(gliding vertex regression, GVR),在边框回归过程中除了生成水平边框的4个参数(x, y, w, h)还将生成5个参数($\alpha_1, \alpha_2, \alpha_3, \alpha_4, r$),通过这5个参数将水平边框转换为倾斜边框,从而使边框更加贴合目标形状。本文采用裁剪、旋转、翻转等方法对 DOTA 数据集^[22]进行整理和扩充,并进行大量实验。经实验证明,本文方法可以有效提高多类多尺度多方向目标的检测性能。

1 相关工作

1.1 多尺度目标检测策略

由于图像中不同目标的尺度差异较大的问题,近年来相继出现了如 FPN^[23], SNIPER^[24], TridenNet^[25]以及 M2Det^[26]等一些性能优良的多尺度目标检测算法。FPN 利用深度卷积神经网络内在的多尺度结构和金字塔分级结构构造一种自底向上、自顶向下、横向连接的架构,横向连接将具有相同尺寸的特征图进行融合,并在不同层次的融合特征

图上进行预测。这种方法可以融合高低层的特征信息,使融合特征更具鲁棒性,但融合操作也相对弱化了高层和底层的特征信息。SNIPER 是一种多尺度训练算法,将原图像缩放到多个尺度,然后在每个尺度的图像上以一定的滑动距离获取一系列同样大小的切片,若切片完全覆盖若干个真值框,则被选择为正样本块,并使用原训练集训练得到的 RPN 初步筛选得到不含真值框的负样本块。这些正负样本块将代替原图像送入网络用于训练。该算法通过生成多个尺度的切片,并适当选择正负样本块,有效提高了多尺度目标的检测性能。TridenNet 主要是通过对感受野的控制改善多尺度目标的检测性能,大的感受野对于大尺度目标的检测效果更好,小的感受野对于小目标更加友好,因此,该算法构造了一个3支并行的结构,3条支路共享权重,并利用不同的膨胀率使每条支路的感受野大小不同,3条支路分别负责大、中、小相应尺度范围内的样本的训练,这种分支训练策略有效提高了不同尺度目标的检测性能,避免了极端尺度目标造成的负面影响。M2Det 提出一种更有效的多级特征金字塔结构来提取图像特征。首先对骨干网络提取的特征进行融合,通过细化 U 形模块和特征融合模块提取更具代表性的多尺度特征,然后通过尺度特征聚合模块融合多级特征,每个特征图都由多个层次的特征组成,最后分层进行检测得到最终检测结果。该算法主要利用了多个 FPN 结构,使得到的多尺度特征更具代表性,有效提高了不同尺度目标的检测精度。

1.2 注意力机制

人类视觉通过快速扫描全局图像,获得需要重点关注的目标区域,然后赋予这些区域更多的注意力获得更多感兴趣目标的细节信息。类似地,在神经网络中的注意力机制就是从所有信息中选择出对当前目标任务贡献最大的信息。在注意力机制中,所有信息在被聚合之前会根据学到的注意力权重进行重新加权,相关性最高的信息被赋予的权重较高,相关性较低的信息被赋予的权重较低,以此分离出重要信息,并抑制不重要信息的干扰。根据注意力域的不同,常见的注意力模型可分为空间域注意力机制、通道域注意力机制和混合域注意力机制3类。空间域注意力机制利用训练得到的空间变换模块将原始图像中空间域信息变换到另一个空间中并保留关键信息。将输入的图像信息映射为对应的张量,然后将这个张量输入空间变换模块中,在该模块中,张量信息进入定位网

络, 定位网络会学习一组参数作为网格生成器的参数, 生成一个采样信号, 也即一个变换矩阵, 将这个变换矩阵与原始张量相乘得到变换后的张量(图像特征), 利用这个变换矩阵对输入的关键信息进行识别. 通道域注意力机制就是给特征图上每个通道的信号都增加一个权重, 代表该通道与关键信息的相关性, 这个权重越高代表相关性越高. 具有代表性的是 SENet^[27], 其主要是通过挤压过程使 C 个特征图变成一个 $1 \times 1 \times C$ 的向量, 然后在激活过程中对该向量进行非线性变换, 通过学习得到的参数生成每个特征通道的权重, 增强对关键通道的注意, 最后通过乘法逐通道加权到先前的特征上, 从而增强重要的特征信息, 弱化不重要的特征信息. 经过对空间域注意力机制和通道域注意力机制的介绍可知, 空间域注意力机制忽略

了通道域的信息, 将每个通道的特征做同样的处理, 通道域注意力机制则直接对单个通道内的信息进行全局平均池化, 忽略了通道内的局部信息. 结合 2 种注意力机制的思想, 文献[28]提出混合域注意力机制, 混合域注意力机制是利用掩码机制得到注意力权重, 其中每个通道的特征图中的每个元素都对应一个注意力权重, 这就同时利用到空间域和通道域的信息, 而不会只关注空间域的信息或通道域的信息, 因此称之为混合域注意力机制.

2 FAGNet 检测框架

本文以 Faster R-CNN 检测框架和特征金字塔网络为基线结构, 提出遥感图像多尺度目标检测方法 FAGNet, 其整体框架结构如图 1 所示.

图 1 FAGNet 目标检测框架

FAGNet 依然遵循两阶段方法路线, 整个网络主要包括 3 部分: 特征提取网络、RPN 以及分类/回归网络. 在特征提取网络部分设计了 MAFPN 模块, 其以 ResNet101 为基础构造了密集连接特征金字塔, 融合多个层次的特征图, 增加目标尺度上下文信息, 该模块通过添加空间域注意力机制和通道域注意力机制使融合后的多层级特征图有选择性地聚焦于显著性的部分, 使含有目标的区域在通道维度和空间维度上的特征信息都得到增强. 然后将从 RPN 中得到的候选区域送入分类/回归网络, 在分类/回归网络中使用 GVR 机制预测边框的 9 个参数, 获得旋转边框预测结果, 最后通过旋转

非最大值抑制操作获得任意方向目标的检测结果, 有效提高了遥感图像多尺度任意方向目标的检测性能.

2.1 策略特征提取网络 MAFPN

MAFPN 主要通过多层特征融合以及组合注意力机制使最终的特征图包含更丰富更有效的语义信息, 提高算法对目标尺度的自适应性.

2.1.1 MAFPN——特征融合

在卷积神经网络中, 低层特征图包含更多小目标的位置信息, 高层特征图感受野更大, 包含更多有益于分类的语义信息. 随着网络层数的加深, 经过多次的卷积和池化操作, 目标的特征信息逐

渐丢失, 无法提供必要的分辨率信息, 小目标的位置信息和细节信息都被弱化, 导致在高层级特征图中丢失了小目标大部分的特征信息. 在特征金字塔这种网络结构中, 通过自底向上、自顶向下以及横向连接的方式融合高层和底层的特征图. 自底向上的路径是网络的前馈计算, 通过不同比例的特征映射得到不同的特征层级, 自顶向下的路径通过对空间上更抽象但语义信息更丰富的高层特征图上进行上采样, 得到与其相邻的上一个层级特征图相同尺度的特征图, 原始的上一个层级的特征图经过一个 1×1 的卷积减少特征图的通道数, 从而使相同层级的 2 个特征图的通道数一致, 便于后续的特征分类. 接着将二者进行逐像素相加融合, 为了降低上采样所带来的混叠效应, 最后再经过一个 3×3 的卷积生成最终的融合特征图, 使其同时具有不同尺度的特征信息和分辨率信息. 由众多研究可知, 使用特征金字塔结构可以在一定程度上改善多尺度目标的检测效果, 但针对尺度差异较大的目标, 如田径场和小汽车等, 检测效果仍需进一步提高, 因此本文在 MAFPN 中的特征融合部分设计多级特征金字塔网络结构(multi-level feature pyramid network, MLFPN), 从骨干网络 ResNet101 中选取卷积层 $\{C_2, C_3, C_4, C_5\}$ 作为特征金字塔的基础结构, 在卷积层 C_5 的后面添加 2 个卷积层 C_6 和 C_7 , 来获得更精细的语义信息. 计算 C_6 和 C_7 的方式为

$$C_6 = \text{Conv2D}(k = 2048, s = (3,3))(C_5) \quad (1)$$

$$C_7 = \text{Conv2D}(k = 2048, s = (3,3))(C_6) \quad (2)$$

其中, Conv2D 表示卷积操作; k 表示输出通道数; s 表示卷积核大小. 因此, 最后生成自底向上的路径为 $\{C_2, C_3, C_4, C_5, C_6, C_7\}$, 构建得到的特征金字塔网络为 $\{C_2, C_3, C_4, C_5, C_6\}$, 然后分别将 C_3, C_4, C_5, C_6, C_7 上采样得到与 C_2, C_3, C_4, C_5, C_6 相同大小, 然后自顶向下逐层进行融合得到融合特征图 $C'_2, C'_3, C'_4, C'_5, C'_6$, 具体的融合网络结构如图 2 所示.

2.1.2 MAFPN——组合注意力模块

通过将特征图可视化发现, 将不同特征图融合, 虽然同时兼顾了大/小尺度的目标特征信息, 但会彼此弱化, 并引入了大量的噪声, 使目标边界变得模糊, 导致漏检和误检增加. 为解决不同尺度特征图融合导致的特征信息弱化以及噪声的问题, 在融合后的特征图上引入空间域注意力机制和通道域注意力机制, 突出增强特征图上的有效信息

并抑制无关信息, 减少噪声. 通道域注意力机制可以使网络更加关注有效通道的特征图, 空间域注意力机制可以分配更多注意力给有效的位置信息.

图2 特征融合网络示意图

为了更好地提取遥感图像中目标的特征信息, MAFPN使用通道域注意力和空间域注意力顺序组合方式, 图3展示了组合注意力模块的具体结构. 将特征融合中得到的融合特征图 C'_4 输入注意力模块, 首先通过全局平均池化(global average pooling, GAP)把每个通道的所有特征值相加再取平均, 把二维特征图变成一个实数, 得到一个 $1 \times 1 \times C$ 的向量, C 表示通道数. 由文献[29]可知全局最大池化(global max pooling, GMP)也有益于有效特征信息的筛选, 因此沿着通道维度同时使用 GAP 和 GMP; 并分别将其送入 2 个全连接层训练学习, 得到 2 个一维的通道权重序列; 将 2 组权重序列相加再经过 Sigmoid 激活函数将这些数值映射到 $[0, 1]$, 最终得到 1 组权重序列; 将其与 C'_4 进行特征加权得到中间特征图 C''_4 . 文献[27,29]指出, 不同的注意力模块组合所达到的信息筛选效果也不同, 经过实验证明, 通道域注意力模块连接空间域注意力模块组合性能最佳. 因此, 本文方法使用这种组合方式, 在通道域注意力模块后面接一个空间域注意力模块. 与通道域注意力关注通道信息不同的是, 空间域注意力主要是关注位置信息, 将得到的特征图 C''_4 分别经过 GAP 和 GMP 得到 2 个单通道的特征图, 将二者按照通道维度连接并进行一个卷积操作得到空间域注意力特征图, 然后连接

一个 Sigmoid 激活函数将这些数值映射到 $[0, 1]$, 就得到空间域注意力权值, 将其与中间特征图 C_4'' 相乘, 即可得到最终的特征图 P_4 . 将这种组合注意力模块嵌入到网络中, 有效增强了候选区域的相关特征信息, 达到特征优选的目的.

图 3 组合注意力模块结构

2.2 分类/回归网络

2.2.1 旋转边框回归

MAFPN 主要通过多层特征融合以及注意力机制使最终的特征图包含更丰富的有效语义信息. 在得到最终的融合特征图后, 将其输入 RPN 中获得最终的候选区域, 然后将候选区域送入到分类/回归网络进行分类/回归. 首先通过感兴趣区域对齐(region of interest align, ROI Align)^[4]操作将候选区域映射为固定尺度的特征图块, 然后将其送入 2 个全连接层进行分类/回归. 在回归的阶段本文借鉴文献[30]中的回归方法, 除了表示水平边框的中心点 (x, y) 、宽和高 (w, h) 4 个值外, 还得到表示旋转边框所需要的 4 个滑动顶点的比例参数 $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ 以及用来表示该边框偏离水平方向程度的旋转因子 r , 通过预测得到的 $(x, y, w, h, \alpha_1, \alpha_2, \alpha_3, \alpha_4, r)$ 9 个值, 将水平边框的 4 个顶点依据 4 个滑动比例参数滑动一定的距离得到旋转边框, 具体的原理如图 4 所示.

图 4 中黑色矩形框为水平矩形框, 可以用 (c'_1, c'_2, c'_3, c'_4) 和 (x, y, w, h) 来表示; 红色四边形可以用 (x, y, w, h) 以及 4 个比例系数 $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ 表示, 其中, $\alpha_1 = d_1/w, \alpha_2 = d_2/h, \alpha_3 = d_3/w, \alpha_4 = d_4/h, d_1, d_2, d_3, d_4$ 分别表示水平矩形框 4 个顶点要滑动的

图 4 旋转边框回归原理

距离; 水平矩形框的 4 个顶点 (c'_1, c'_2, c'_3, c'_4) 分别与红色 4 边形的 4 个滑动顶点 (c_1, c_2, c_3, c_4) 一一对应, 其中 c_1 是 c'_1 滑动距离 d_1 得到的点, 其他类似. 为了得到红色四边形的旋转信息引入旋转因子 r , 计算公式为 $r = S_{\text{black}} / S_{\text{red}}$, 其中 S_{black} 表黑色水平矩形框的面积, S_{red} 表示红色四边形的面积. r 越小, 红色四边形越偏离水平方向; r 越大, 红色四边形越接近水平方向. 设置阈值为 0.9, 当 $r > 0.9$ 时, 则将其视为水平目标; 若 $r < 0.9$, 则利用 $x, y, w, h, \alpha_1, \alpha_2, \alpha_3, \alpha_4$ 计算新的旋转边框.

2.2.2 损失函数设计

由于增加了需要预测的边框回归有关参数的数目, 所以损失函数也要做出相应改变, 在损失函数中 RPN 的分类损失和回归损失都与水平框目标检测中一样. 由于在回归阶段增加了 5 个参数, 所以在 RCNN 的回归损失中除了水平框的损失之外增加了 $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ 以及旋转因子 r 对应的损失. 损失函数分别为

$$L = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}} + \frac{1}{N_{\text{reg}}} \sum_i p_i^* \times L_{\text{reg}} \quad (3)$$

$$L_{\text{cls}}(p_i, p_i^*) = -\sum_{j=1}^K p_i^* \log(p_i) \quad (4)$$

$$L_{\text{reg}} = \lambda_1 \times L_h + \lambda_2 \times L_\alpha + \lambda_3 \times L_r \quad (5)$$

$$L_h = s_{L_1}(t_i - t_i^*) \quad (6)$$

$$L_\alpha = \sum_i s_{L_1}(\alpha_i - \alpha_i^*) \quad (7)$$

$$L_r = s_{L_1}(r - r^*) \quad (8)$$

$$s_{L_1}(x) = \begin{cases} 0.5x^2, & |x| \leq 1 \\ |x| - 0.5, & \text{其他} \end{cases} \quad (9)$$

通过式(3)计算网络的损失, N_{cls} 和 N_{reg} 分别表示批次的大小以及在同一批次里 RPN 提议的框的个数, i 表示提议的框的索引. 若第 i 个框为正

样本, 则 $p_i^* = 1$; 否则 $p_i^* = 0$. 分类损失由式(4)交叉熵损失函数得到, K 表示目标类别总数, p_i 表示提议的框被预测为目标 j 类的概率. 回归损失由式(5)计算得到, 其中包括 3 个部分: 水平框的损失 L_h , 4 个滑动因子 $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ 的损失 L_α 以及旋转因子 r 的损失 L_r , $\lambda_1, \lambda_2, \lambda_3$ 是平衡损失的超参数. 这 3 部分损失均使用式(9)所示的平滑 L_1 损失函数 s_{L_1} 进行计算. 式(6)中 $t_i = (t_x, t_y, t_w, t_h)$ 表示预测的水平边框的 4 个参数化坐标, $t_i^* = (t_x^*, t_y^*, t_w^*, t_h^*)$ 表示与预测的正锚框对应的真值边框的参数化坐标. 式(7)中 α 和 α_i 分别表示预测边框的滑动比例参数与对应的真值边框的滑动比例参数. 式(8)中 r 和 r^* 分别表示预测边框的旋转因子与对应的真值边框的旋转因子.

3 实验

3.1 数据与参数设置

本文在 DOTA 数据集上进行实验, 该数据集包括从不同传感器和平台收集的 2806 幅遥感图像, 图像尺寸从 800×800 到 4000×4000 不等. 图像中的目标分别被标记为足球场、直升机、游泳池、大型车辆、小型车辆、桥梁、港口、篮球场、网球场、棒球场、油罐、舰船、飞机、田径场和环形路线等 15 个类别, 其中每类目标都涵盖多种不同的形状、方向和尺度. 为使训练数据和测试数据中目标种类和比例分配均衡, 使用官方划分的训练集和测试集进行实验. 由于遥感图像尺寸较大, 不适合直接用于训练, 因此需对数据集进行多尺度裁剪, 分别对原始图像进行 1.0 倍和 0.5 倍的下采样, 1.0 倍即不做下采样处理直接裁剪, 0.5 倍下采样是先对原始图像进行 0.5 倍下采样然后进行裁剪, 这一点主要是考虑到数据集中如田径场和足球场等大尺度目标很可能在裁剪中被截断, 所以对原始图像进行 0.5 倍的下采样可以在一定程度上增加完整大尺度目标的比例, 改善大尺度目标被截断的情况. 裁剪窗口大小设置为 1024×1024 , 裁剪步长为 200. 对于尺寸不足 1024×1024 的图像, 通过在图像周围填充 0 的方法将其填充为 1024×1024 大小, 此时刚好与所设置裁剪窗口大小一致, 无需再进行裁剪. 通过这种方式保证裁剪后所有图像切片大小均为 1024×1024 , 输入网络的切片维度为 $1024 \times 1024 \times 3$. 据统计, 裁剪扩充后数据集中目标尺度的纵横比分布如图 5 所示.

图 5 DOTA 数据集中目标尺度的纵横比分布信息统计

由图 5 可知在扩充后数据集中的目标尺度纵横比大都分布在 $[0, 5]$, 因此本文实验中将锚框尺度纵横比分别设置为可以覆盖大部分目标纵横比的 $[1, 1/2, 2, 1/5, 5]$. 使用平均精度(average precision, AP)和平均精度均值(mean average precision, mAP)作为评价指标, 其中预测结果与真值框的交并比(intersection over union, IoU)值大于 0.5 时视为检测正确, 如果在同一位置同时有多个预测框的 IoU 值大于 0.5, 则取其中 IoU 最大的预测框为正确检测结果. 本文实验所采用的硬件配置为 Intel i9-CPU, NVIDIA GeForce RTX 2080 Ti 显卡, CUDA10.0, CUDNN7.4, 实验中使用随机梯度下降法进行训练, 初始学习率设置为 0.0025, 动量设置为 0.9, 正则化系数设置为 0.0001, 批量处理大小为 2, 损失函数中的超参数 $\lambda_1, \lambda_2, \lambda_3$ 分别设置为 1, 1, 16.

3.2 实验结果及分析

为使得 MAFPN 中注意力模块对于特征信息的筛选效果达到最佳, 本文分别将通道域注意力模块和不同注意力组合模块嵌入网络进行训练测试, 以 mAP 作为评价指标, 实验结果如表 1 所示.

表 1 不同注意力模块对网络检测性能的影响 %

网络结构	mAP
Faster RCNN+MLFPN+GVR(FMG)	71.530
FMG +通道注意力	72.216
FMG +空间注意力+通道注意力	72.879
FMG +空间注意力与通道注意力并行	72.637
FMG +通道注意力+空间注意力(本文)	73.354

可以看出, 在网络中加入不同注意力模块可以在不同程度上提高网络的检测精度, 使用组合注意力模块的性能优于仅使用通道域注意力模块的性能, 并且按照顺序组合注意力模块比并行组合注意力模块的性能更好, 对比之下性能最佳的

是本文方法所使用的通道域注意力模块和空间域注意力模块的组合顺序,检测精度达到 73.354%。为了进一步证明 MAFPN 可以更有效地提取和融合特征信息,本文分别对 ResNet101 骨干网络输出的原始特征图、原始特征图经过 MAFPN 中特征融合模块和组合注意力模块的特征图进行可视化,可视化结果如图 6 所示。

对比图 6 可知,在逐次经过 MAFPN 的特征融合和注意力模块后,特征图中对应目标存在的区域(红框标出)逐渐变得高亮,这也就表示在 MAFPN 中的特征融合和组合注意力模块的作用

下,目标存在区域逐渐被赋予了更高的权重,存在目标的区域将对目标检测任务的贡献更大,也从另一方面证明了 MAFPN 模块的有效性。

为了证明本文方法的有效性,与已有的一些经典目标检测算法进行对比,如使用旋转边框训练的 Faster RCNN(faster R-CNN trained on oriented bounding boxes, FR-O)^[22], RoI 转换器(RoI transformer, RoI Trans)^[31], R2CNN^[32]、旋转区域建议网络(rotation region proposal networks, RRPN)^[33], CAD-Net^[34], R2CNN++^[35]等。表 2 和图 7 所示为不同算法关于各类目标的 AP 值及其均值 mAP。

图 6 特征图可视化示例图

由表 2 可知,FR-O 的 mAP 为 54.13%,该算法将 Faster RCNN 中的水平边框回归改进为旋转边框回归。由于检测精度受感受野尺度的影响较大,而 FR-O 只在最后一层特征图上进行预测,其感受野尺度单一,并且经过网络的前向传播目标信息损失严重,因此其对于多尺度目标的检测效果相对较差。与之相比,本文的 FAGNet 的 mAP 值提升了 19.22%,其中每类目标的 AP 值都要高于 FR-O。

R2CNN 和 RRPN 是用于任意方向文本检测的算法,同样可以应用到遥感图像中,其整体架构和

Faster RCNN 类似,RRPN 是将 Faster R-CNN 中的 RPN 替换为 RRPN,生成的是倾斜的候选区域,在生成锚框的时,除了尺度和纵横比 2 个参数之外,还加入一个方向变量,所以生成的锚框是具有倾斜角度的。R2CNN 在分类/回归阶段中除了分类和水平框回归 2 个分支外又增加一个倾斜框回归分支。与 FR-O 算法相比,R2CNN 和 RRPN 这 2 种方法在检测性能上有一定程度的改善,但由于它们仍是使用骨干网络末端特征图进行预测,在经过多次的卷积和池化操作之后,特征信息损失严重,

表 2 不同算法针对每一类目标的 AP 值及其均值 mAP

%

目标类别	FR-O ^[22]	RoI Trans ^[31]	RRPN ^[33]	R2CNN ^[32]	CADNet ^[34]	R2CNN++ ^[35]	本文
舰船	37.16	83.59	57.25	55.81	76.60	66.83	85.68
油罐	52.98	81.46	67.38	72.39	73.30	86.15	86.39
港口	47.89	62.83	53.08	55.14	62.00	65.34	66.79
桥梁	17.70	43.44	31.66	35.34	49.40	45.50	45.52
飞机	79.42	88.64	88.52	80.94	87.80	89.66	87.99
直升机	46.30	47.67	53.58	48.22	62.20	62.05	58.51
游泳池	47.40	58.93	51.94	53.35	67.00	68.01	68.02
田径场	64.05	75.92	59.30	67.44	73.50	75.10	71.65
棒球场	77.13	78.52	71.20	65.67	82.40	81.22	82.57
足球场	50.30	58.39	56.69	55.06	48.40	64.05	62.79
篮球场	69.64	77.27	72.84	66.92	79.20	80.69	79.90
大型车辆	38.02	73.68	56.19	50.91	63.50	60.17	74.15
小型车辆	35.30	68.81	51.85	59.92	71.10	68.27	72.29
网球场	89.41	90.74	90.81	90.67	90.90	90.90	90.66
环形路线	52.91	53.54	52.84	52.23	60.90	63.48	67.34
mAP	54.13	69.56	61.01	60.67	69.90	71.16	73.35

注. 粗体字为最高 AP 值.

图 7 不同算法检测性能对比

对于多尺度目标的特征表达能力严重不足, 因此其检测精度仍然较低, FAGNet 与之相比 mAP 分别提升了 12.49% 和 13.43%.

RoI Trans, CADNet 以及 R2CNN++ 相比于前 3 者的 mAP 有了很大的提升. RoI Trans 旨在通过有监督的旋转感兴趣区域学习和基于位置敏感对齐的特征提取来改善目标检测性能, 其仍使用单个尺度的特征图进行预测. CADNet 则是在多层特征图融合的基础上使用了 SENet 选择有效特征信息, 并在分类/回归阶段引入上下文信息指导目标的分类/回归. 从表 2 来看, RoI Trans 和 CADNet 的 mAP 差别不大, FAGNet 与二者相比 mAP 分别提升

了 3.79% 和 3.45%. R2CNN++ 相比于前面几种算法检测性能最好, 基本每个类别的 AP 值都是最高, 这是由于其融合了 2 个不同尺度的特征图, 得到的最终的融合后的特征图上的信息表征能力更强, 在此基础上又使用通道域注意力和像素级注意力来捕捉对检测最有益的特征信息, 并且其在生成锚框的时设置 15 个纵横比, 使生成的锚框有更大的概率可以刚好覆盖目标, 其 mAP 达到 71.16%.

对比可知, 在多种算法中 FAGNet 的 mAP 值最高, 达到 73.35%, 比次优 mAP 值提升 2.19%, 并且在舰船、大型车辆、小型车辆、港口、游泳池、棒球场、油罐以及环形路线等大部分目标类别的 AP 值均为最高, 尤其是舰船、大车辆、小车辆以及油罐等尺度较小并且大部分场景下排布密集的目标的检测效果较为友好, 分别比次优的 AP 值提高了 2.09%, 0.47%, 1.19%, 0.24%. 对于港口这种纵横比较大并且排列较为密集的目标, FAGNet 的 AP 值比次优 AP 值提升 1.45%. 对于环形路线, FAGNet 比次优 AP 值提升 3.48%. 综合分析, FAGNet 在特征提取部分有效地利用多个尺度的特征, 提高了网络对于目标尺度的自适应能力, 同时还利用组合注意力模块进行特征信息增强, 大大提高有效信息的利用率, 在后期利用 GVR 机制更好地学习边框的旋转信息, 从而得到更为准确的旋转边框. 但是对于田径场和足球场等尺度较大目标的检测精度仍需要进一步提高, 分析有以下 2

点原因: (1) 虽然多尺度裁剪可以在一定程度上增加完整的大尺度目标的个数, 但是仍不能完全避免由于目标尺度大于裁剪的滑动步长而出现的目标被截断的情况, 正如第 3.1 节中的分析, 非完整目标的存在会干扰网络对于大尺度目标的学习; (2) 数据集存在类别不平衡现象, 其中足球场、田径场、直升机、桥梁等几类目标的实例个数相对其他类别的目标较少, 算法无法充分学习这些目标

的特征信息. 对于桥梁这种纵横比较大且类内差异较大的目标, 其检测性能仍需进一步提高.

由图 7 可以看出, FAGNet 的检测性能整体上具有优势. 在同一幅图像中同时存在尺度差异较大的足球场、小型车辆等都可以被检测出来, 在不同分辨率的图像中的同一类目标绝大部分都能被准确检测到, 尤其是对小尺度目标以及排布密集的目标效果较好, 具体检测结果示例如图 8 所示.

图 8 FAGNet 在 DOTA 测试集上检测结果示例

4 结 语

针对遥感图像目标检测领域中多尺度多方向目标的检测问题, 本文提出的基于多尺度注意力特征金字塔网络和滑动顶点回归机制目标检测方法 FAGNet, 其优势有 2 个方面: (1) 构造特征提取网络 MAFPN, 利用深度特征金字塔结构进行特征

融合, 在多个尺度的融合特征图上使用组合注意力模块提高有效特征信息的强度, 增强网络对于目标多尺度特征的代表能力; (2) 在回归阶段引入 GVR 机制, 加强网络对于不同方向目标的边框位置学习能力. 经实验证明 FAGNet 能够更加准确地检测遥感图像中不同尺度不同方向的目标, 尤其是对尺度较小并且密集分布的目标检测效果较好.

在未来的工作中,将针对如何进一步提高遥感图像中尺度范围更大、类内差异较大的目标的检测精度,如何更好地降低误检和漏检开展研究,期望在遥感图像中实现更好的目标检测性能。

参考文献(References):

- [1] Xu Chengji, Wang Xiaofeng, Yang Yadong. Attention-YOLO: YOLO detection algorithm that introduces attention mechanism[J]. Computer Engineering and Applications, 2019, 55(6): 13-23(in Chinese)
(徐诚极, 王晓峰, 杨亚东. Attention-YOLO: 引入注意力机制的 YOLO 检测算法[J]. 计算机工程与应用, 2019, 55(6): 13-23)
- [2] Ren S Q, He K M, Girshick R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks[C] //Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2015: 91-99
- [3] Dai J F, Li Y, He K M, *et al.* R-FCN: object detection via region-based fully convolutional networks[C] //Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2016: 379-387
- [4] He K M, Gkioxari G, Dollár P, *et al.* Mask R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2961-2969
- [5] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6154-6162
- [6] Han X B, Zhong Y F, Zhang L P. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery[J]. Remote Sensing, 2017, 9(7): 666
- [7] Deng Zhipeng, Sun Hao, Lei Lin, *et al.* Object detection in remote sensing imagery with multi-scale deformable convolutional networks[J]. Acta Geodaetica et Cartographica Sinica, 2018, 47(9): 1216-1227(in Chinese)
(邓志鹏, 孙浩, 雷琳, 等. 基于多尺度形变特征卷积网络的高分辨率遥感影像目标检测[J]. 测绘学报, 2018, 47(9): 1216-1227)
- [8] Li K, Cheng G, Bu S H, *et al.* Rotation-insensitive and context-augmented object detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 56(4): 2337-2348
- [9] Yao Qunli, Hu Xian, Lei Hong. Object detection in remote sensing images using multiscale convolutional neural networks[J]. Acta Optica Sinica, 2019, 39(11): Article No.1128002 (in Chinese)
(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): Article No.1128002)
- [10] Tang T Y, Zhou S L, Deng Z P, *et al.* Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining[J]. Sensors, 2017, 17(2): Article No.336
- [11] Yang X, Sun H, Fu K, *et al.* Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks[J]. Remote Sensing, 2018, 10(1): Article No.132
- [12] Zhu M M, Xu Y L, Ma S P, *et al.* Airport detection method with improved region-based convolutional neural network[J]. Acta Optica Sinica, 2018, 38(7): Article No.0728001
- [13] Liu L, Pan Z X, Lei B. Learning a rotation invariant detector with rotatable bounding box[OL]. [2020-08-26]. <https://arxiv.org/pdf/1711.09405v1.pdf>
- [14] Redmon J, Divvala S, Girshick R, *et al.* You only look once: unified, real-time object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 779-788
- [15] Liu W, Anguelov D, Erhan D, *et al.* SSD: single shot multibox detector[C] //Proceedings of the European Conference on Computer Vision. Heidelberg, Springer, 2016: 21-37
- [16] Lin T Y, Goyal P, Girshick R, *et al.* Focal loss for dense object detection[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2999-3007
- [17] Law H, Deng J. CornerNet: detecting objects as paired keypoints[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 734-750
- [18] Tang T Y, Zhou S L, Deng Z P, *et al.* Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks[J]. Remote Sensing, 2017, 9(11): Article No.1170
- [19] Liu W C, Ma L, Chen H. Arbitrary-oriented ship detection framework in optical remote-sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(6): 937-941
- [20] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6517-6525
- [21] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [22] Xia G S, Bai X, Ding J, *et al.* DOTA: a large-scale dataset for object detection in aerial images[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3974-3983
- [23] Lin T Y, Dollár P, Girshick R, *et al.* Feature pyramid networks for object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2117-2125
- [24] Singh B, Najibi M, Davis L S. Sniper: efficient multi-scale training[C] //Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2018: 9310-9320
- [25] Li Y H, Chen Y T, Wang N Y, *et al.* Scale-aware trident networks for object detection[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 6054-6063
- [26] Zhao Q J, Sheng T, Wang Y T, *et al.* M2Det: a single-shot ob-

- ject detector based on multi-level feature pyramid network[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33(1): 9259-9266
- [27] Hu J, Shen L, Albanie S, *et al.* Squeeze-and-excitation networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141
- [28] Wang F, Jiang M Q, Qian C, *et al.* Residual attention network for image classification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3156-3164
- [29] Woo S, Park J, Lee J Y, *et al.* CBAM: convolutional block attention module[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19
- [30] Xu Y C, Fu M T, Wang Q M, *et al.* Gliding vertex on the horizontal bounding box for multi-oriented object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1452-1459
- [31] Ding J, Xue N, Long Y, *et al.* Learning RoI transformer for detecting oriented objects in aerial images[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 2844-2853
- [32] Jiang Y Y, Zhu X Y, Wang X B, *et al.* R2CNN: rotational region CNN for orientation robust scene text detection[OL]. [2020-08-26]. <https://arxiv.org/pdf/1706.09579.pdf>
- [33] Ma J Q, Shao W Y, Ye H, *et al.* Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122
- [34] Zhang G J, Lu S J, Zhang W. CAD-Net: a context-aware detection network for objects in remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(12): 10015-10024
- [35] Yang X, Fu K, Sun H, *et al.* R2CNN++: multi-dimensional attention based rotation invariant detector with robust anchor strategy[OL]. [2020-08-26]. <https://arxiv.org/pdf/1811.07126v1.pdf>