# 动态特征融合的遥感图像目标检测

谢星星 程 塥 姚艳清 姚西文 韩军伟
（西北工业大学自动化学院 西安 710129）

**摘 要** 目标尺度差异性和类间相似性是遥感图像目标检测面临的两个重要挑战.多尺度特征融合作为一种解决目标尺度差异性大和类间相似度高的方法,受到了广泛关注.然而目前大多数融合方法使用固定权重融合不同尺度特征,使所有的输入图像共享融合方式,忽略输入图像中目标尺度对特征融合的影响.针对上述问题,本文提出了一种动态特征融合网络.该网络由特征门控模块和动态融合模块构成,能够实现多尺度特征的动态融合.其中,特征门控模块旨在对融合前的特征进行选择性抑制或增强,降低背景信息对后续融合的干扰.动态融合模块旨在建立输入目标尺度和特征融合之间的联系,根据输入目标尺度动态学习融合权重.最后,在采用特征金字塔的Faster R-CNN上构建动态特征融合网络,并在大规模的遥感图像目标检测数据集DIOR和DOTA上验证了动态特征融合网络的有效性.

**关键词** 目标检测;遥感图像;动态特征融合;特征门控
**中图法分类号** TP391 **DOI号** 10.11897/SP.J.1016.2022.00735

# Dynamic Feature Fusion for Object Detection in Remote Sensing Images

XIE Xing-Xing CHENG Gong YAO Yan-Qing YAO Xi-Wen HAN Jun-Wei
(*School of Automation, Northwestern Polytechnical University, Xi'an* 710129)

**Abstract** Remote sensing images contain more valuable information, which has opened a door to help us observe and measure the earth's surface. Thanks to the advance of earth observation techniques, remote sensing images with different spectral and spatial resolutions are increasing daily. How to understand these huge volumes of remote sensing images is becoming more and more important. As a fundamental task of remote sensing image understanding, object detection in remote sensing images has been an active research area. The goal of object detection in remote sensing images is to locate ground objects and classify them into different categories. It supports a wide range of real-world applications, including aerial reconnaissance, emergency rescue, and urban management. In recent years, deep learning techniques and large-scale datasets with annotations have provided a major improvement in general object detection, e.g., Fast/Faster R-CNN, RetinaNet, and FCOS. Driven by these improvements, object detection in remote sensing images has achieved significant progress. However, the large variations of object sizes and inter-class similarity are still two big challenges for object detection in remote sensing images. To address these challenges, many works have been introduced. One of the typical methods, termed Feature Pyramid Network (FPN), creates a feature pyramid with strong semantics at all scales by combining low-resolution, semantically strong features with high-resolution,

收稿日期:2021-01-20;在线发布日期:2022-01-12. 本课题得到国家自然科学基金(No. 61772425)、陕西省杰出青年科学基金(2021JC-16)、西北工业大学博士论文创新基金(No. CX2021082)资助. **谢星星**,博士研究生,主要研究领域为计算机视觉、遥感图像目标检测. E-mail: xiexing@mail.nwpu.edu.cn. **程 塥**(通信作者),教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为模式识别、计算机视觉、遥感图像理解. E-mail: gcheng@nwpu.edu.cn. **姚艳清**,博士研究生,主要研究领域为计算机视觉、遥感图像理解. **姚西文**,博士,副研究员,主要研究领域为计算机视觉、遥感图像处理. **韩军伟**,教授,博士生导师,长江学者特聘教授、IEEE Fellow,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、脑图像分析.

semantically weak features. After that, Libra R-CNN fuses the features of different scales with the same weights for enchaining the discriminability of features. PANet enhances the entire feature hierarchy by top-down and bottom-up path augmentation, which shortens the information path between lower features and top ones. These methods greatly improve detection accuracy. However, most of them utilize fixed weights to fuse the features of different scales, in which all input images share the fusion method, ignoring the influence of object scales of input images on feature fusion. On the one hand, the feature fusion approach is static, which is unable to change fusion weights according to the size of objects adaptively, thus preventing the robustness of detection. On the other hand, it can introduce useless features and suppress the feature representation when fusing features. To this end, we design a dynamic feature fusion network for minimizing the influence from the variations and improving the representation of features. The network contains a feature gate module and a dynamic fusion module. The feature gate module aims to selectively attenuate useless features and enhance useful features before dynamic feature fusion, and minimize the interference of background information on subsequent dynamic fusion. We model it by a gate unit, which consists of spatial, channel, and global attention. The dynamic fusion module is to establish the connection between the object scales and the feature fusion weights, thus learning the fusion weights dynamically according to object scales. We achieve this by a lightweight fully-connected network, which takes the multi-scale features as the input. Finally, we propose a dynamic feature fusion network on the Faster R-CNN with FPN, and conduct extensive experiments on two large-scale remote sensing image object detection datasets, named DIOR and DOTA. The experimental results demonstrate the effectiveness of our proposed method.

**Keywords** Object detection; remote sensing images; dynamic feature fusion; feature gate

## 1 引 言

遥感图像目标检测旨在定位和识别遥感图像中感兴趣的目标，是遥感图像智能解译的关键技术.作为视觉目标检测的重要分支，遥感图像目标检测在自然灾害检测、军事侦察、城市规划等领域有着广泛的用途.近些年，随着卷积神经网络(Convolutional Neural Network，CNN)的飞速发展、计算机并行运算能力的提升、以及大规模标注数据集的出现[1,2]，一系列先进的基于卷积神经网络的遥感图像目标检测算法相继被提出.例如Cheng等人[3]针对遥感图像目标方向多变的问题，设计了旋转不变卷积神经网络实现遥感图像目标旋转不变特征的提取，极大地提升了遥感图像目标检测的性能.Zhang等人[4]针对遥感图像目标尺度差异性，设计了尺度自适应的遥感图像目标检测区域生成网络.Cheng等人[5]针对遥感图像目标尺度差异性和类间相似性，在特征金字塔基础上设计了跨尺度的特征融合方法来提升特征的判别性.Ding等人[6]针对遥感图像目标密集排布和方向任意的问题，设计了遥感图像有向目标检测算法.虽然基于卷积神经网络的遥感图像目标检测在精度上有了极大的提升，但是目标尺度差异性和类间相似性(如图1)仍然是遥感图像目标检测面临的两大重要挑战.

融合生成判别性的多尺度特征是解决遥感图像目标尺度差异性大和类间相似性高的有效方法，可以极大地提升遥感图像目标检测的性能.然而，目前大多数融合方法以固定权重融合不同尺度的特征.例如特征金字塔[7](Feature Pyramid Network，FPN)通过自上而下和横向连接的方式依次对相邻尺度的特征进行相加融合.Libra R-CNN[8]使用相同权重融合多尺度特征，实现多尺度特征的有效平衡.PANet[9]以自上而下和自下而上的方式对相邻尺度特征进行相加融合.这种以固定权重进行特征融合的方法会使所有输入图像共享融合方式，忽略输入图像目标尺度对特征融合的影响、制约特征融合的适应性、影响特征融合的效果.

针对上述问题，本文提出了一种动态特征融合网络.该网络由特征门控模块和动态融合模块组

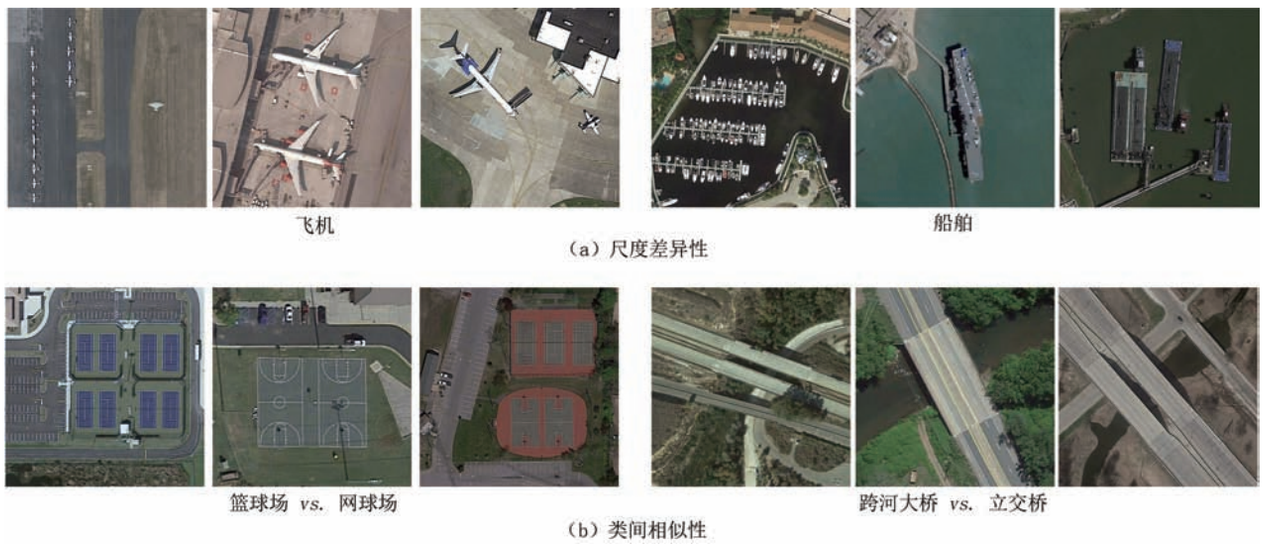


图1　遥感图像目标检测面临的两大挑战：(a)尺度差异性；(b)类间相似性

成，能够对FPN提取的多尺度特征进行动态融合. 其中，特征门模块控旨在对融合前的特征进行选择性增强或抑制，降低背景信息对后续融合结果的干扰. 动态融合模块旨在建立输入图像目标尺度和特征融合之间的联系，依据输入图像目标尺度动态地学习融合权重. 最后，我们在采用FPN的Faster R-CNN[10]上构建了动态特征融合网络，并在大规模的遥感图像目标检测数据集DIOR[1]和DOTA[2]上验证了动态特征融合网络的有效性.

## 2　相关工作

### 2.1　遥感图像目标检测

深度学习的飞速发展和大规模遥感数据集[1, 2]的出现，使得遥感图像目标检测迎来了跨越式的发展[11]. 受R-CNN[12](Region-based Convolution Neural Network，R-CNN)目标检测框架的启发和影响，深度学习在遥感图像目标检测领域逐渐开始发展起来，一系列基于R-CNN[12]的遥感图像目标检测算法被提出. Cheng等人[13]在R-CNN框架的基础上设计了旋转不变层，实现遥感图像目标的旋转不变检测. Long等人[14]为实现更精确的遥感图像目标定位，在R-CNN的基础上设计了基于评分的无监督边界回归. Cheng等人[3]通过对CNN特征设计旋转不变和费舍尔(Fisher)判别约束，显著地提升了CNN特征的旋转不变性和判别性.

2015年，Faster R-CNN的出现为实现更快更精确的遥感图像目标检测提供了基础. Deng等人[15]受Faster R-CNN和区域建议网络(Region Proposal Network，RPN)的启发，设计了针对遥感图像车辆检测的区域生成网络和车辆位置以及属性预测网络. Li等人[16]通过设计旋转不变的RPN来解决遥感图像目标旋转和形变问题. Zhang等人[4]针对遥感图像目标的特点，设计了尺度自适应的遥感图像目标检测区域生成网络. Cheng等人[5]在具有FPN结构的Faster R-CNN框架上引入跨尺度连接和注意力机制，来提升特征的判别性. Li等人[17]通过设计特征注意力和自适应多感受野机制来提升遥感图像目标检测的性能.

考虑到双阶段遥感图像目标检测算法的速度问题，基于单阶段的遥感图像目标检测算法[18-20]开始被研究，Hou等人[18]在单阶段目标检测框架RetinaNet[21]的基础上通过特征融合和检测头级联，实现了高效的遥感图像目标检测. 虽然这些基于锚框的检测算法在检测速度和精度上都取得了长足的进步，但它们涉及的超参数多且锚框设置和数据集中目标尺度的分布存在关联. 这使得无锚框的遥感图像目标检测开始流行[22, 23]. Pan等人[23]在目标检测算法CornerNet[24]的基础上设计了特征选择和自适应的检测头，实现无锚框的遥感图像目标检测.

另外，由于遥感图像采用俯视成像，地物目标会以任意方向出现，使用水平边界框不能紧密地去定位物体. 一些学者陆续开始研究遥感图像有向目标检测. Ding等人[6]针对遥感图像目标密集排布、方向任意的问题，通过水平的RoI(Region of Interesting，RoI)来学习旋转的RoI，实现遥感图像有向目标检测. Xu等人[25]通过滑动水平边界框的顶点来实现遥感图像有向目标检测. Xu等人[26]针对已有遥感图像

目标检测算法没有充分利用多层语义信息的问题，提出了一种多层语义信息指导的遥感图像目标检测算法来同时输出水平和有向边界框.

### 2.2 多尺度特征融合

融合生成更具判别性的多尺度特征是提高目标检测性能的关键.特征金字塔FPN[7]作为简单有效的特征融合网络，通过侧向连接和自顶向下的路径来增强顶层特征的语义信息，极大地提升了目标检测的性能.之后在FPN基础上衍生出许多特征融合网络.Liu等人[9]提出了PANet，在特征金字塔FPN的基础上加入自下而上的路径来融合相邻尺度的特征，以促进特征信息之间的流通.Pang等人[8]在Libra R-CNN中用相等固定的权重融合多个尺度的特征，实现多尺度特征的再平衡.Tan等人[27]设计了双向加权特征金字塔，在训练时设置可学习的特征融合参数，在测试时使用由训练学习得到的权重去融合相邻尺度特征.Liu等人[28]设计了自适应多尺度特征融合网络，为每个尺度位置上的特征学习一个融合系数，实现特征的自适应融合.Zhao等人[29]设计了多层次多尺度特征融合的金字塔网络，将骨干网络不同阶段的多尺度特征进行级联融合，极大地丰富了特征的语义信息.Wang等人[30]提出了尺度均衡特征金字塔用于挖掘不同尺度特征的内在关联性，并以不同权重去融合相邻尺度的特征.Ghiasi等人[31]利用神经架构搜索来搜索最优的特征金字塔结构，极大地提升了特征的判别性，但是神经架构搜索需要巨大的GPU资源，制约了金字塔网络的通用性.

与上述多尺度特征融合网络相比，本文提出的是一种动态特征融合网络，可根据输入目标尺度动态调整特征融合权重.这种动态特征融合网络涉及的学习参数少，同时考虑了输入目标尺度对特征融合的影响，是一种输入目标尺度驱动的特征融合网络.

## 3 方 法

### 3.1 整体概述

目前已有的大多数特征融合网络采用静态的融合方式，所有的输入图像共享融合权重，忽略了输入图像中目标尺度对特征融合的影响、制约了特征融合的适应性、影响融合效果.针对上述问题，我们提出了动态特征融合网络.该网络能够依据输入目标尺度对FPN的多尺度特征进行动态融合，生成判别性的多尺度特征.网络的整体框架如图2所示，包括两个模块：特征门控模块和动态融合模块.动态特征融合网络的大致工作流程如下.

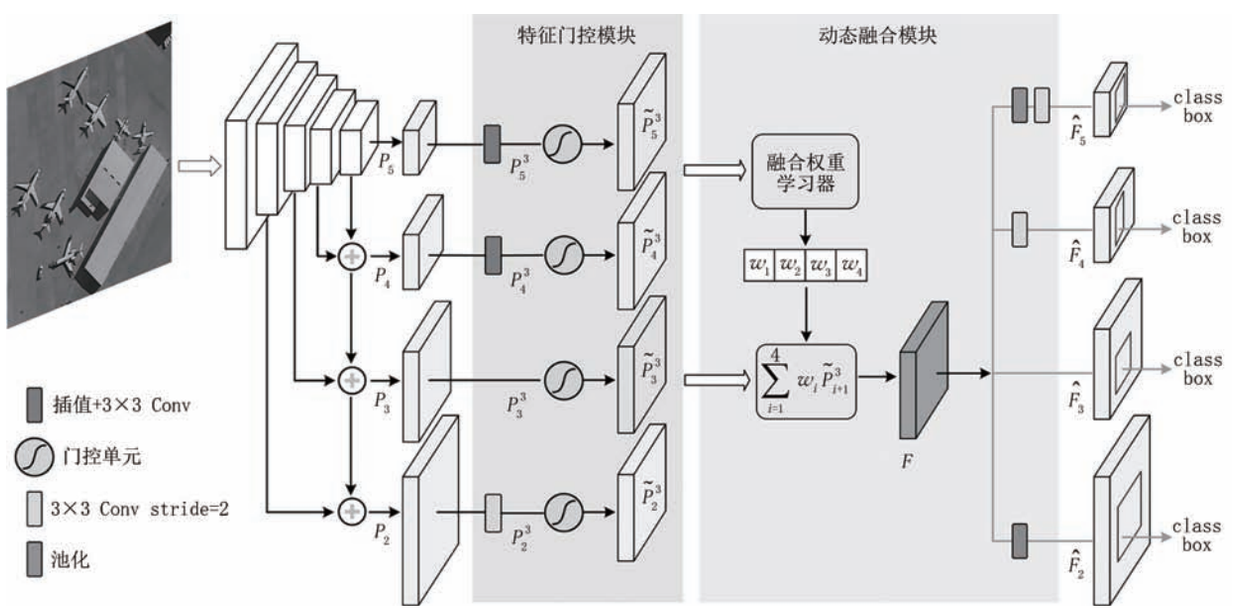


图2 动态特征融合网络的整体框架，包括特征门控和动态融合两个模块

先将FPN提取的多尺度特征$\{\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4,\boldsymbol{P}_5\}$进行尺寸调整，得到和特征$\boldsymbol{P}_3$相同尺寸大小的特征$\{\boldsymbol{P}_2^3,\boldsymbol{P}_3^3,\boldsymbol{P}_4^3,\boldsymbol{P}_5^3\}$.其中，尺寸小于$\boldsymbol{P}_3$的特征先使用双线性插值，再使用步长为1、卷积核大小为3×3的卷积运算进行调整.尺寸大于$\boldsymbol{P}_3$的特征使用步长为2、卷积核大小为3×3的卷积运算进行调整.之后分别使用四个门控单元对同尺寸的特征$\{\boldsymbol{P}_2^3,\boldsymbol{P}_3^3,\boldsymbol{P}_4^3,\boldsymbol{P}_5^3\}$进行选择性地增强或抑制，得到新的同尺寸特征

$\{\tilde{\boldsymbol{P}}_2^3,\tilde{\boldsymbol{P}}_3^3,\tilde{\boldsymbol{P}}_4^3,\tilde{\boldsymbol{P}}_5^3\}$. 将$\{\tilde{\boldsymbol{P}}_2^3,\tilde{\boldsymbol{P}}_3^3,\tilde{\boldsymbol{P}}_4^3,\tilde{\boldsymbol{P}}_5^3\}$作为融合权重学习器的输入，学习得到一组特征融合权重$\{w_1,w_2,w_3,w_4\}$. 根据融合权重$\{w_1,w_2,w_3,w_4\}$对$\{\tilde{\boldsymbol{P}}_2^3,\tilde{\boldsymbol{P}}_3^3,\tilde{\boldsymbol{P}}_4^3,\tilde{\boldsymbol{P}}_5^3\}$进行线性加权融合得到特征$\boldsymbol{F}$. 对$\boldsymbol{F}$重新采样得到多尺度特征$\{\hat{\boldsymbol{F}}_2,\hat{\boldsymbol{F}}_3,\hat{\boldsymbol{F}}_4,\hat{\boldsymbol{F}}_5\}$. 其中$\boldsymbol{F}$到$\hat{\boldsymbol{F}}_3$不经过任何操作. $\hat{\boldsymbol{F}}_5$是先通过池化，再经过步长为2、卷积核大小为3×3的卷积运算获得. $\hat{\boldsymbol{F}}_4$是经过步长为2、卷积核大小为3×3的卷积运算获得. $\hat{\boldsymbol{F}}_2$是先通过插值，再经过步长为1、卷积核大小为3×3的卷积运算获得. 多尺度特征$\{\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4,\boldsymbol{P}_5\}$和$\{\hat{\boldsymbol{F}}_2,\hat{\boldsymbol{F}}_3,\hat{\boldsymbol{F}}_4,\hat{\boldsymbol{F}}_5\}$的尺寸保持一致. 最终多尺度特征$\{\hat{\boldsymbol{F}}_2,\hat{\boldsymbol{F}}_3,\hat{\boldsymbol{F}}_4,\hat{\boldsymbol{F}}_5\}$用于后续的分类和定位任务.

整个动态特征融合网络可以嵌入到具有FPN的Faster R-CNN框架中，实现端到端训练. 训练损失计算与Faster R-CNN保持一致，其中回归损失采用Smooth L1，类别损失采用交叉熵损失(Cross Entropy Loss). 接下来我们分别对特征门控模块和动态融合模块进行介绍.

### 3.2　特征门控模块

由于遥感图像背景复杂，直接融合多尺度特征会将背景信息融入，降低前景和背景的区分度，影响后续的检测. 所以在融合前对多尺度特征进行选择性增强或抑制是非常有必要的操作. 受注意力机制的启发[32]，我们设计了特征门控模块. 该模块先对多尺度特征的尺寸进行调整，再对调整后的同尺寸特征进行选择性增强或抑制. 其中尺寸调整主要通过插值和卷积运算实现，特征的选择性增强或抑制通过门控单元实现. 门控单元如图3所示，采用了通道注意力、全局注意力和残差连接. 通道注意力是在不引入学习参数的条件下建立通道特征间的相关关系，根据相关关系自适应地增强或抑制通道特征. 全局注意力是通过全连接层学习一个全局注意力系数，实现输入特征的整体增强或抑制. 我们设计的门控单元只需要学习一个系数—全局注意力系数，不需要设置任何阈值，是一种输入特征驱动的门控机制. 整体的实现流程如下. 对于输入特征$\boldsymbol{P}_k^3\in\mathbb{R}^{c\times h\times w}$，通过维度变换(R)得到特征$\boldsymbol{A}\in\mathbb{R}^{c\times hw}$. 通过维度变换加转置(RT)获得特征$\boldsymbol{B}\in\mathbb{R}^{hw\times c}$. 将特征$\boldsymbol{A}$和$\boldsymbol{B}$进行矩阵相乘得到特征通道间的关系矩阵$\boldsymbol{X}\in\mathbb{R}^{c\times c}$，对关系矩阵$\boldsymbol{X}$进行softmax操作得到归一化后的关系矩阵$\hat{\boldsymbol{X}}\in\mathbb{R}^{c\times c}$. softmax操作如公式(1)所示：

$$\hat{\boldsymbol{X}}_{ij}=\frac{\exp(\boldsymbol{X}_{ij})}{\sum_{j=1}^{c}\exp(\boldsymbol{X}_{ij})} \tag{1}$$

其中，$\hat{\boldsymbol{X}}_{ij}$和$\boldsymbol{X}_{ij}$分别代表关系矩阵$\hat{\boldsymbol{X}}$和$\boldsymbol{X}$第$i$行第$j$列的值. 再将关系矩阵$\hat{\boldsymbol{X}}$与输入特征$\boldsymbol{P}_k^3$相乘得到特征$\boldsymbol{M}\in\mathbb{R}^{c\times h\times w}$，相乘运算如公式(2)所示：

$$\boldsymbol{M}=\hat{\boldsymbol{X}}\times\boldsymbol{P}_k^3 \tag{2}$$

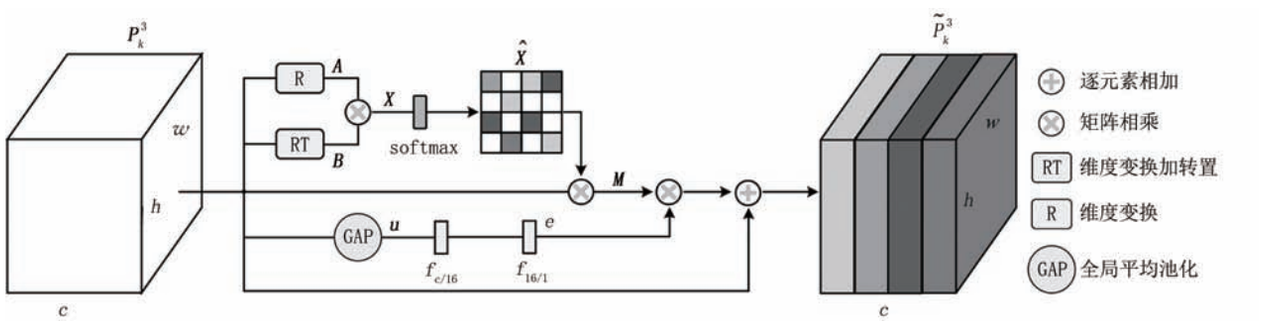


图3　门控单元

另外，为了能自适应的对输入特征进行抑制或增强，我们使用了全局注意力机制. 全局注意力机制主要通过全连接层学习全局特征调节系数$e\in\mathbb{R}$实现. $e$主要通过两个步骤学习得到，第一步对输入特征$\boldsymbol{P}_k^3$进行全局平均池化(GAP)，得到特征向量$\boldsymbol{u}\in\mathbb{R}^c$. GAP的实现如公式(3)所示：

$$\boldsymbol{u}(l)=\frac{1}{w\times h}\sum_{i=1}^{w}\sum_{j=1}^{h}(\boldsymbol{P}_k^3)_l(i,j) \tag{3}$$

其中，$(\boldsymbol{P}_k^3)_l$代表输入特征$\boldsymbol{P}_k^3$的第$l(1\leqslant l\leqslant c)$个通道，$\boldsymbol{u}(l)$代表输入特征$\boldsymbol{P}_k^3$第$l$个通道经过全局平均池化后的值. 第二步将$\boldsymbol{u}$作为输入，使用全连接层$f_{c/16}$和$f_{16/1}$学习全局特征调节系数$e$，如公式(4)所示：

$$e=\sigma\left(f_{16/1}\left(f_{c/16}(\boldsymbol{u})\right)\right) \tag{4}$$

其中，$f_{c/16}$代表第一个全连接层(输入维度为$c$，输出

维度为16).$f_{16/1}$代表第二个全连接层(输入维度为16,输出维度为1),$\sigma$代表sigmoid激活函数.最后将全局特征调节系数$e$作用于特征$\boldsymbol{M}$,并与输入特征进行残差连接,获得最终的输出$\tilde{\boldsymbol{P}}_k^3$,$\tilde{\boldsymbol{P}}_k^3$的计算如公式(5)所示:

$$\tilde{\boldsymbol{P}}_k^3=e\cdot\boldsymbol{M}+\boldsymbol{P}_k^3 \tag{5}$$

### 3.3 动态融合模块

在目标检测中,以固定权重融合多尺度特征会使所有输入图像共享融合方式,忽略输入图像目标尺度对特征融合的影响、制约了特征融合的适应性.所以根据输入目标尺度动态调整融合权重,对提升特征融合的适应性是非常有必要的.基于上述思想,我们设计了动态融合模块.该模块建立输入目标尺度和特征融合之间的联系,依据输入目标尺度去学习融合权重,为多尺度特征分配不同的融合权重.动态融合模块主要通过融合权重学习器实现.

学习器的整体结构如图4所示,它以相同尺寸特征$\left\{\tilde{\boldsymbol{P}}_2^3,\tilde{\boldsymbol{P}}_3^3,\tilde{\boldsymbol{P}}_4^3,\tilde{\boldsymbol{P}}_5^3\right\}$作为输入,输出一组融合权重.权重的学习通过全连接层实现,整体的工作流程如下.首先对相同尺寸的特征$\left\{\tilde{\boldsymbol{P}}_2^3,\tilde{\boldsymbol{P}}_3^3,\tilde{\boldsymbol{P}}_4^3,\tilde{\boldsymbol{P}}_5^3\right\}$进行级联,得到特征$\tilde{\boldsymbol{P}}\in\mathbb{R}^{4c\times h\times w}$,然后将特征$\boldsymbol{V}$进行全局平均池化(GAP)得到特征向量$\boldsymbol{S}\in\mathbb{R}^{4c}$,全局平均池化操作$F_{avg}$如公式(6)所示:

$$\boldsymbol{S}(l)=F_{avg}(\boldsymbol{V}_l)=\frac{1}{w\times h}\sum_{i=1}^{w}\sum_{j=1}^{h}\boldsymbol{V}_l(i,j) \tag{6}$$

其中,$\boldsymbol{V}_l(i,j)$代表特征$\boldsymbol{V}$第$l(1\leqslant l\leqslant 4c)$通道$(i,j)$位置的值,$\boldsymbol{S}(l)$代表特征$\boldsymbol{V}$第$l$通道特征$\boldsymbol{V}_l$经过全局平均池化后的值.接着使用全连接层去学习特征融合权重.考虑到运算成本,本文使用两个全连接层$f_{4c/(c/4)}$和$f_{(c/4)/4}$去学习特征融合权重$\boldsymbol{Z}\in\mathbb{R}^4$,整个学习过程如公式(7)所示:

$$\boldsymbol{Z}=f_{(c/4)/4}\left(f_{4c/(c/4)}(\boldsymbol{S})\right) \tag{7}$$

其中,$f_{4c/(c/4)}$代表第一个全连接层(输入维度为$4c$,输出维度为$c/4$),$f_{(c/4)/4}$代表第二个全连接层(输入维度为$c/4$,输出维度为4).考虑到融合权重学习器训练的稳定性,我们对$\boldsymbol{Z}$进行softmax操作,得到归一化后的融合权重$\left\{w_1,w_2,w_3,w_4\right\}$.

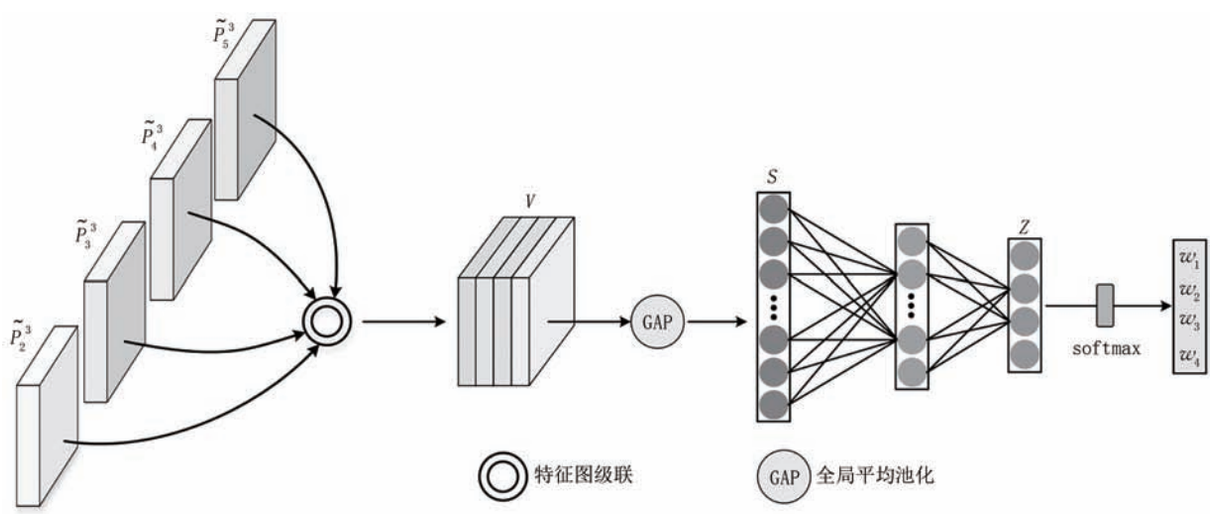


图4 融合权重学习器

softmax操作如公式(8)所示:

$$w_i=\frac{\exp(\boldsymbol{Z}(i))}{\sum_{i=1}^{4}\exp(\boldsymbol{Z}(i))} \tag{8}$$

这里学习到的融合权重代表多尺度特征在融合时的重要性.最后依据学习的融合权重$\left\{w_1,w_2,w_3,w_4\right\}$对输入特征$\left\{\tilde{\boldsymbol{P}}_2^3,\tilde{\boldsymbol{P}}_3^3,\tilde{\boldsymbol{P}}_4^3,\tilde{\boldsymbol{P}}_5^3\right\}$进行线性加权融合得到特征$\boldsymbol{F}$,线性融合的计算如公式(9)所示:

$$\boldsymbol{F}=\sum_{i=1}^{4}w_i\cdot\tilde{\boldsymbol{P}}_{i+1}^3 \tag{9}$$

为了与FPN的多尺度特征在尺寸上保持一致,后续将融合特征$\boldsymbol{F}$进行重新调整得到新的多尺度特征$\left\{\hat{\boldsymbol{F}}_2,\hat{\boldsymbol{F}}_3,\hat{\boldsymbol{F}}_4,\hat{\boldsymbol{F}}_5\right\}$.

## 4 实 验

### 4.1 数据集和评价指标

为了验证本文方法的有效性,我们在两大公开的遥感图像目标检测数据集DIOR[1]和DOTA[2]上

进行了实验. DIOR[1]数据集包括20个目标类别，分别为飞机、机场、棒球场、篮球场、跨河大桥、烟囱、水坝、高速公路服务区、高速公路收费站、高尔夫球场、田径场、码头、立交桥、船舶、体育馆、储油罐、网球场、火车站、车辆、风车（表1）. 总共23 463幅图像，192 472个目标实例. 其中5862幅图像用于训练，5863幅图像作为验证，11 738幅图像作为测试. 图像的大小为800×800，图像的空间分辨率为0.5米～30米. 实验中我们将训练和验证图像合并在一起用于训练. 实验采用准确率（Average Precision，AP）和平均准确率（mean Average Precision，mAP）作为检测评价指标. AP和mAP的计算方式采用PASCAL VOC2007[33]标准.

DOTA[2]数据集包括15个类别，分别为飞机、棒球场、跨河大桥、田径场、小型车辆、大型车辆、船舶、网球场、篮球场、储油罐、足球场、交通环岛、码头、游泳池、直升机（表2）. DOTA[2]数据集总共2806幅图像，图像尺寸从800×800到4000×4000变化，图像的空间分辨率为0.1 m～4.5 m. 其中训练集包含1411幅图像，验证集包含458幅图像，测试集包含937幅图像. 实验中将训练集和验证集合并在一起进行训练，测试结果提交至DOTA[2]测试服务器进行测评. 对于训练集和验证集中的原始图像以1024×1024的滑窗进行裁图，滑窗步长为824. 对于测试集中的原始图像以1024×1024的滑窗进行裁图，滑窗步长为512.

**表1　DIOR[1]数据集目标类别索引表**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 飞机 | 机场 | 棒球场 | 篮球场 | 跨河大桥 | 烟囱 | 水坝 | 高速公路服务区 | 高速公路收费站 | 高尔夫球场 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 田径场 | 码头 | 立交桥 | 船舶 | 体育馆 | 储油罐 | 网球场 | 火车站 | 车辆 | 风车 |

**表2　DOTA[2]数据集目标类别索引表**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 飞机 | 棒球场 | 跨河大桥 | 田径场 | 小型车辆 | 大型车辆 | 船舶 | 网球场 | 篮球场 | 储油罐 | 足球场 | 交通环岛 | 码头 | 游泳池 | 直升机 |

### 4.2　实验设置

实验使用两块显存为11 GB的GeForce RTX 2080Ti GPU进行训练和测试. 两个数据集训练时的批次大小都设置为4，共迭代12个epoch，初始学习率为0.005，动量系数为0.9，权重衰减系数为0.000 1. 在第9个epoch和11个epoch之间学习率为0.000 5，第12个epoch的学习率为0.000 05. 训练时50%的训练图像进行随机翻转. 其中DIOR[1]数据集训练时输入图像尺寸为800×800. DOTA[2]数据集训练时输入图像尺寸为1024×1024. 无特殊说明外，实验所使用的基础骨干网络都为ResNet101[34]. 所有实验在目标检测开源工具mmdetection[35]上运行. 实验中所采用的其它参数均为mmdetection[35]的默认参数.

### 4.3　消融实验

#### 4.3.1　门控单元消融实验

我们以Faster R-CNN加FPN为基准方法. 在DIOR[1]数据集上进行了4组消融实验，分别验证了通道注意力、全局注意力和残差连接的效果，结果如表3所示. 基准方法Faster R-CNN加FPN在DIOR[1]数据集的平均准确率（mAP）为70.3%. 当门控单元只融合通道注意力特征时，mAP从70.3%提升至70.7%. 当门控单元只融合全局注意力特征时，mAP从70.3%提升至70.6%. 同时融合通道注意力特征和全局注意力特征，mAP可从70.3%提升至70.9%，提升0.6%. 在融入通道注意力特征和全局注意力特征的基础上，引入残差连接，mAP从70.9提升至71.0%. 由此可见在门控单元中，通道注意力带来的增益大于全局注意力和残差连接.

**表3　门控单元消融实验结果**

| | 通道注意力 | 全局注意力 | 残差连接 | mAP |
|---|---|---|---|---|
| 基准方法 | | | | 70.3 |
| 基准方法＋ | √ | | | 70.7 |
| | | √ | | 70.6 |
| | √ | √ | | 70.9 |
| | √ | √ | √ | 71.0 |

#### 4.3.2　特征门控和动态融合消融实验

为了验证特征门控和动态融合的有效性，本文以Faster R-CNN加FPN的目标检测算法为基准方法，在DIOR[1]数据集上进行了3组消融实验结果如表4所示. 基准方法Faster R-CNN加FPN在DIOR[1]数

据集的平均准确率(mAP)为70.3%.在FPN多尺度特征的基础上,加入特征门控,mAP从70.3%提升至71.0%,提升0.7%.这里加特征门控是指分别对FPN的每个尺度添加特征门控,没有特征调整操作.为了验证动态融合带来的增益,在不加特征门控的条件下,在基准方法中加入动态融合能将mAP从70.3%提升到71.7%,提升了1.4%.这里不加特征门控,主要是先将FPN的多尺度特征统一到$\boldsymbol{P}_3$尺度,然后进行动态融合得到特征$\boldsymbol{F}$.最后通过对$\boldsymbol{F}$重新采样获得多尺度特征.可见动态特征融合带来的增益大于特征门控.在基准方法中同时加入特征门控和动态特征融合,将mAP从70.3%提升到72.3%,提升2.0%.

图5展示了将不同尺度目标作为输入,动态特征融合网络学习到的特征融合权重,其中红色代表最大融合权重,蓝色代表次最大融合权重.当输入图像目标尺度较小时,底层特征$\boldsymbol{P}_2$和$\boldsymbol{P}_3$中包含了目标的细节信息有利于小目标检测,在特征融合时底层特征$\boldsymbol{P}_2$和$\boldsymbol{P}_3$会采用较大的权重,如图5(a),当输入图像目标尺度较大时,顶层特征$\boldsymbol{P}_4$和$\boldsymbol{P}_5$提取的语义信息有利于大目标检测,在特征融合时顶层特征$\boldsymbol{P}_4$和$\boldsymbol{P}_5$会采用较大的权重,如图5(b).

**表4 特征门控和动态融合消融实验结果**

| | 特征门控 | 动态融合 | mAP |
|---|---|---|---|
| 基准方法 | | | 70.3 |
| 基准方法 + | √ | | 71.0 |
| | | √ | 71.7 |
| | √ | √ | 72.3 |

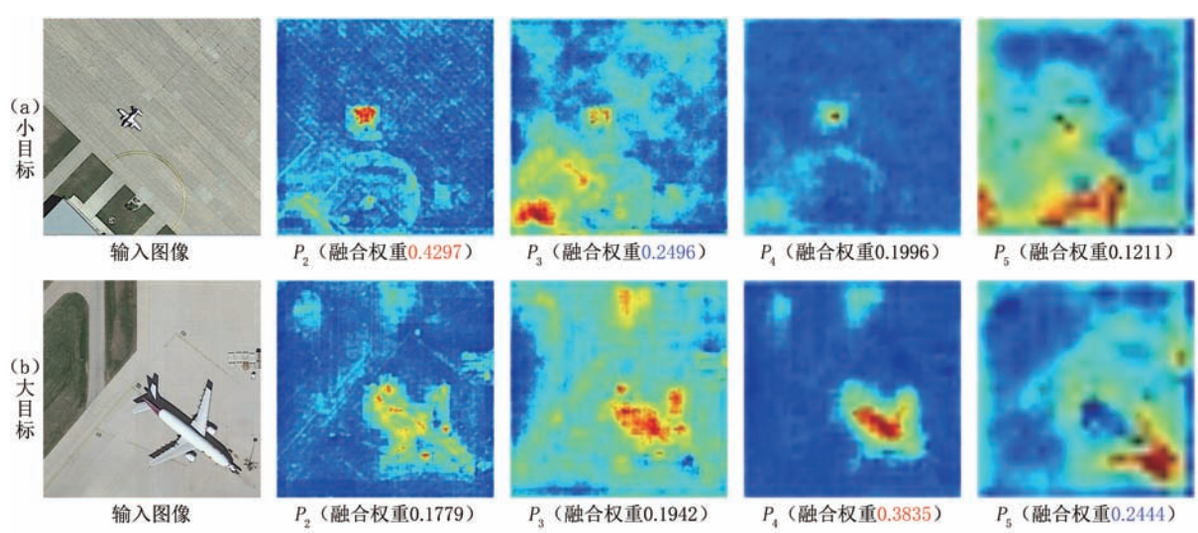


图5 不同尺度目标输入对应的特征融合权重,红色代表最大融合权重,蓝色代表次最大融合权重

另外我们在图6中呈现了动态融合前后特征的可视化结果,其中融合前的特征取自FPN的$\boldsymbol{P}_3$,融合后的特征取自$\hat{\boldsymbol{F}}_3$.从特征可视化的结果可以看出融合后的特征包含了更少的干扰信息,特征更聚焦于目标的位置,如图6(b)和6(d)所示.特征门控可以减少干扰信息的融入,提升融合特征中前景和背景的区分度,如图6(c)和(d)所示.

### 4.4 与其它算法比较

为了进一步证明本文方法的有效性,我们在DIOR和DOTA数据集上,将本文方法和其它的遥感图像目标检测方法进行了对比,对比结果分别如表5和表6所示.在DIOR[1]数据集上我们比较了Faster R-CNN[10]、Libra R-CNN[8]、Mask R-CNN[37]、RetinaNet[21]、PANet[9]、CornerNet[24]、YOLOv3[36]、CSFF[5]等8种方法.其中Faster R-CNN、Libra R-CNN、Mask R-CNN、RetinaNet、PANet、CSFF的基础骨干网络为ResNet101,检测头采用FPN结构.CornerNet的基础骨干网络为Hourglass-104,YOLOv3的基础骨干网络为Darknet53[36].从表5中可以看出本文方法的平均检测精度(72.3% mAP)优于其它两阶段目标检测方法.相比于基准方法,本文方法在飞机、船舶、篮球场、网球场、立交桥等目标类别上的检测精度都有提升.

图7给出了本文方法在DIOR[1]数据集上的一些检测结果,其中绿色框代表正确检测结果、蓝色框代表虚警、红色框代表漏检,绿色框左上角的数字分别代表不同的目标类别.从可视化结果可以看出,本文方法在目标尺度差异大、类间相似度高、密集排

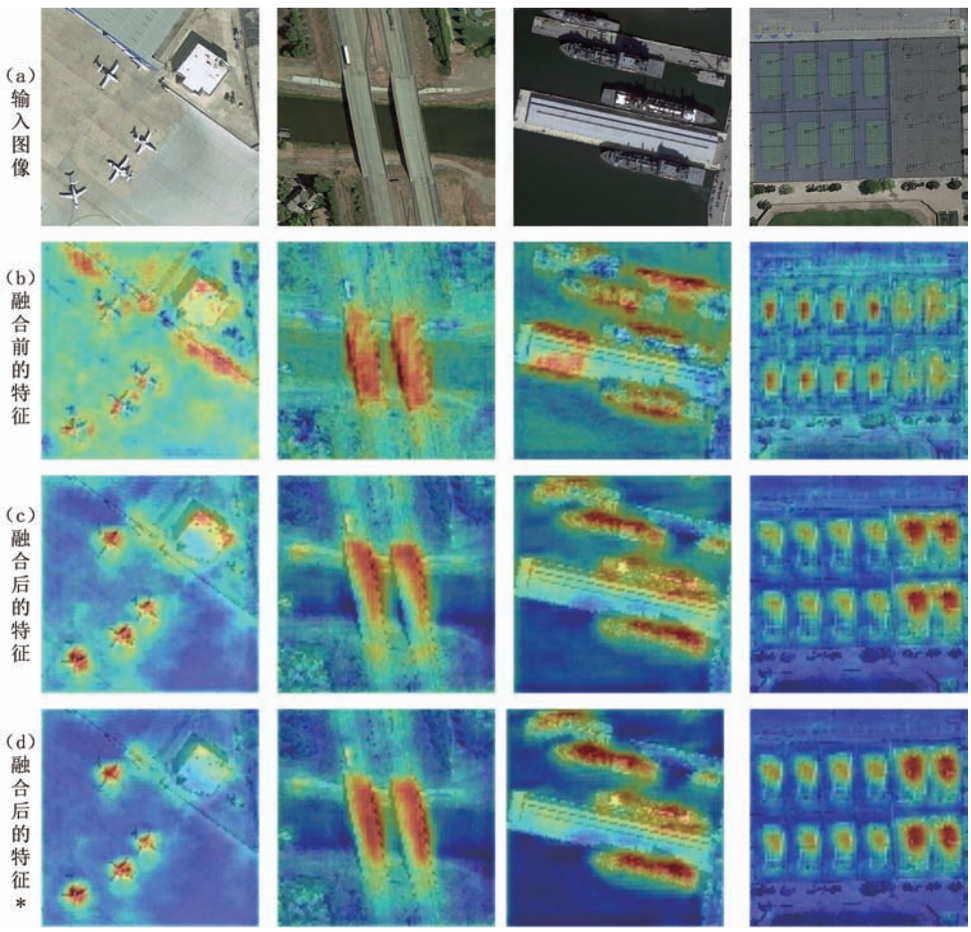


图6　动态融合前后特征的可视化结果：(a)为输入的原始图像；(b)为融合前特征的可视化结果；(c)为未加特征门控动态融合后特征的可视化结果；(d)为加特征门控动态融合后特征的可视化结果. *代表加特征门控

**表5　DIOR[1]数据集的比较结果，*代表基准方法**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| YOLOv3[36] | **72.2** | 29.2 | 74.0 | 78.6 | 31.2 | 69.7 | 26.9 | 48.6 | 54.4 | 31.1 | 61.1 | 44.9 | 49.7 | **87.4** | 70.6 | **68.7** | **87.3** | 29.4 | **48.3** | 78.7 | 57.1 |
| Faster R-CNN [10] | 54.0 | 74.5 | 63.3 | 80.7 | 44.8 | 72.5 | 60.0 | 75.6 | 62.3 | 76.0 | 76.8 | 46.4 | 57.2 | 71.8 | 68.3 | 53.8 | 81.1 | 59.5 | 43.1 | 81.2 | 65.1 |
| Mask R-CNN [37] | 53.9 | 76.6 | 63.2 | 80.9 | 40.2 | 72.5 | 60.4 | 76.3 | 62.5 | 76.0 | 75.9 | 46.5 | 57.4 | 71.8 | 68.3 | 53.7 | 81.0 | 62.3 | 43.0 | 81.0 | 65.2 |
| RetinaNet[21] | 53.3 | 77.0 | 69.3 | 85.0 | 44.1 | 73.2 | 62.4 | 78.6 | 62.8 | 78.6 | 76.6 | 49.9 | 59.6 | 71.1 | 68.4 | 45.8 | 81.3 | 55.2 | 44.4 | 85.5 | 66.1 |
| PANet[9] | 60.2 | 72.0 | 70.6 | 80.5 | 43.6 | 72.3 | 61.4 | 72.1 | 66.7 | 72.0 | 73.4 | 45.3 | 56.9 | 71.7 | 70.4 | 62.0 | 80.9 | 57.0 | 47.2 | 84.5 | 66.1 |
| CornerNet[24] | 58.8 | **84.2** | 72.0 | 80.8 | 46.4 | 75.3 | 64.3 | 81.6 | 76.3 | 79.5 | 79.5 | 26.1 | 60.6 | 37.6 | 70.7 | 45.2 | 84.0 | 57.1 | 43.0 | 75.9 | 64.9 |
| Libra R-CNN[8] | 54.1 | 81.7 | 71.6 | 81.4 | 46.4 | **79.7** | 66.1 | 83.8 | 70.2 | 76.4 | 82.2 | 50.3 | 58.8 | 71.1 | 68.4 | 53.7 | 81.3 | 63.9 | 42.9 | 81.3 | 68.3 |
| CSFF[5] | 57.2 | 79.6 | 70.1 | 87.4 | 46.1 | 76.6 | 62.7 | 82.6 | 73.2 | 78.2 | 81.6 | 50.7 | 59.5 | 73.3 | 63.4 | 58.5 | 85.9 | 61.9 | 42.9 | 86.9 | 68.0 |
| Faster R-CNN[10]* | 54.1 | 81.7 | 71.5 | 86.3 | **51.3** | **79.7** | 67.8 | 84.9 | 70.6 | 80.8 | 82.5 | 52.7 | 61.3 | 72.0 | **73.1** | 62.2 | 81.4 | 66.8 | 43.7 | 81.3 | 70.3 |
| 本文算法 | 57.2 | **84.2** | **74.4** | **88.7** | 50.8 | 79.2 | **70.8** | **87.7** | **77.0** | **82.4** | **85.1** | **53.7** | **62.7** | 74.7 | 69.4 | 60.2 | 87.0 | **67.8** | 45.5 | **87.2** | **72.3** |

**表6　DOTA[2]数据集的比较结果，*代表基准方法**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| YOLOv2[38] | 76.90 | 33.87 | 22.73 | 34.88 | 38.73 | 32.02 | 52.37 | 61.65 | 48.54 | 33.91 | 29.27 | 36.83 | 36.44 | 38.26 | 11.61 | 39.20 |
| R-FCN[39] | 80.01 | 58.96 | 31.64 | 58.97 | 49.77 | 45.04 | 49.29 | 68.99 | 52.07 | 67.42 | 41.83 | 51.44 | 45.15 | 53.30 | 33.89 | 52.58 |
| ICN[40] | **90.00** | 77.70 | 53.40 | **73.30** | 73.50 | 65.00 | 78.20 | 90.80 | 79.10 | 84.80 | **57.20** | 62.10 | 73.50 | 70.20 | 58.10 | 72.50 |
| Adaptive R-CNN[41] | 88.62 | **80.22** | 53.18 | 66.97 | 76.30 | **72.59** | **84.07** | 90.66 | 80.95 | 76.24 | 57.12 | **66.65** | **84.08** | 66.36 | 56.85 | 72.72 |
| Faster R-CNN[10]* | 89.04 | 78.84 | 51.60 | 60.45 | **78.23** | 66.78 | 78.56 | 90.63 | **80.99** | 81.80 | 48.76 | 63.83 | 72.68 | 70.83 | 56.30 | 71.28 |
| 本文算法 | 89.77 | 76.00 | **54.13** | 71.05 | 74.05 | 67.57 | 78.90 | **90.88** | 80.04 | **84.86** | 57.12 | 63.15 | 76.61 | **73.63** | **58.74** | **73.10** |

图7 DIOR[1]数据集目标检测结果，其中，绿色框代表正确检测结果，红色框代表漏检，蓝色框代表虚警

布时，可以获取不错的检测结果.但是当目标存在遮挡和严重的背景干扰时，检测结果会存在少量虚警和漏检.例如丛林中的车辆和水坝、公路斑马线旁停放的车辆.此外我们给出了本文方法和基准方法的检测结果对比，如图8所示.从对比结果来看，本文方法能够准确检测出基准方法漏检的目标，减少基准方法中的虚警，例如图8(b)第一列中漏检的田径场和错检的水坝.

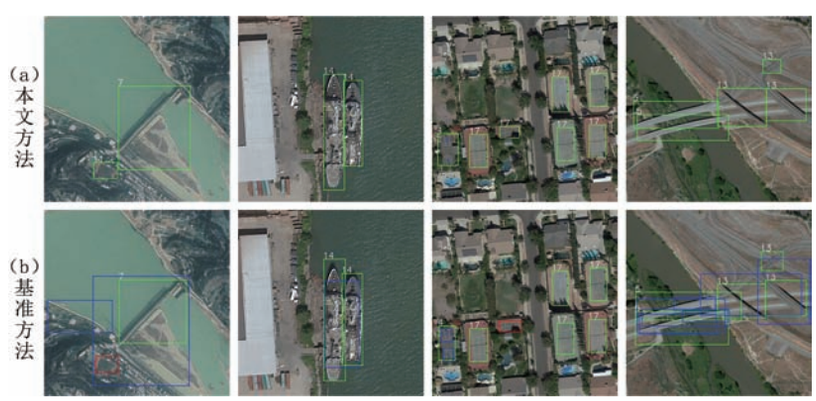


图8 (a)本文方法和(b)基准方法在DIOR[1]数据集上的检测结果对比，其中，绿色框代表正确检测结果，红色框代表漏检，蓝色框代表虚警

在DOTA[2]数据集上，我们主要比较了5种方法，分别为ICN[40]、Adaptive R-CNN[41]、R-FCN[39]、Faster R-CNN[10]和YOLOv2[38].其中ICN[40]、R-FCN[39]、Adaptive R-CNN[41]和Faster R-CNN[10]4种算法所用

的骨干网络都为ResNet101[34]，YOLOv2[38]骨干网络为Darknet-19[38]. ICN[40]、Adaptive R-CNN[41]、Faster R-CNN[40]检测头都采用FPN结构，Faster R-CNN[40]加FPN为本文基准方法.从表6可以得出本文方法相比于基准方法，mAP有1.82%的提升，整体结果优于其它检测方法.相比于基准方法，本文方法在飞机、网球场、跨河大桥等目标类别上的检测精度都有提升.

## 5 结 论

针对遥感图像目标尺度差异性和类间相似性，本文提出了一种动态特征融合网络.该网络包括特征门控模块和动态融合模块.特征门控模块对融合前的多尺度特征进行选择性增强或抑制，减少背景信息对后续特征融合的干扰.动态融合模块建立输入图像目标尺度和特征融合之间的联系，依据输入图像目标尺度动态学习融合权重，使得特征融合能够依据输入图像目标尺度动态调整.最后我们在具有FPN结构的Faster R-CNN上构建了动态特征融合网络，并在大规模遥感图像目标检测数据集DIOR和DOTA上验证了动态特征融合网络的有效性.

## 参 考 文 献


[1] Li K, Wan G, Cheng G, Meng L, and Han J. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307

[2] Xia G-S, Bai X, Ding J, Zhu Z, Belongie S, Luo J, et al. DOTA: A large-scale dataset for object detection in aerial images// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3974-3983

[3] Cheng G, Han J, Zhou P, and Xu D. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE Transactions on Image Processing, 2018, 28(1): 265-278

[4] Zhang S, He G, Chen H, Jing N, and Wang Q. Scale Adaptive Proposal Network for Object Detection in Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 2019, 16(6): 864-868

[5] Cheng G, Si Y, Hong H, Yao X, and Guo L. Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 2020: 1-5

[6] Ding J, Xue N, Long Y, Xia G-S, and Lu Q. Learning RoI transformer for oriented object detection in aerial images// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2849-2858

[7] Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, and Belongie S. Feature pyramid networks for object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117-2125

[8] Pang J, Chen K, Shi J, Feng H, Ouyang W, and Lin D. Libra r-cnn: Towards balanced learning for object detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 821-830

[9] Liu S, Qi L, Qin H, Shi J, and Jia J. Path aggregation network for instance segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8759-8768

[10] Ren S, He K, Girshick R, and Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[11] Liu Xiao-Bo, Liu Peng, Cai Zhi-Hua, Qiao Yu-Lin, Wang Min. Research Progress of Optical Remote Sensing Image Object Detection Based on Deep Learning. Acta Automatica Sinica, 2019, 47(9): 2078-2089 (in Chinese)
(刘小波，刘鹏，蔡之华，乔禹霖，王凌，and 汪敏. 基于深度学习的光学遥感图像目标检测研究进展. 自动化学报，2019，47(9)：2078-2089)

[12] Girshick R, Donahue J, Darrell T, and Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 580-587

[13] Cheng G, Zhou P, and Han J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7405-7415

[14] Long Y, Gong Y, Xiao Z, and Liu Q. Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2486-2498

[15] Deng Z, Sun H, Zhou S, Zhao J, and Zou H. Toward Fast and Accurate Vehicle Detection in Aerial Images Using Coupled Region-Based Convolutional Neural Networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(8): 3652-3664

[16] Li K, Cheng G, Bu S, and You X. Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2017, 56(4): 2337-2348

[17] Li C, Xu C, Cui Z, Wang D, Zhang T, and Yang J. Feature-Attentioned Object Detection in Remote Sensing Imagery// Proceedings of the IEEE International Conference on Image Processing. Taipei, China, 2019: 3886-3890

[18] Hou L, Lu K, Xue J, and Hao L. Cascade Detector With Feature Fusion For Arbitrary-Oriented Objects In Remote Sensing Images//Proceedings of the IEEE International Conference on Multimedia and Expo. London, UK, 2020: 1-6

[19] Wang C, Bai X, Wang S, Zhou J, and Ren P. Multiscale visual attention networks for object detection in VHR remote sensing

images. IEEE Geoscience and Remote Sensing Letters, 2018, 16(2): 310-314

[20] Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, et al. Scrdet: Towards more robust detection for small, cluttered and rotated objects//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 8232-8241

[21] Lin T-Y, Goyal P, Girshick R, He K, and Dollár P. Focal loss for dense object detection//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2980-2988

[22] Yi J, Wu P, Liu B, Huang Q, Qu H, and Metaxas D. Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021: 2150-2195

[23] Pan X, Ren Y, Sheng K, Dong W, Yuan H, Guo X, et al. Dynamic Refinement Network for Oriented and Densely Packed Object Detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11207-11216

[24] Law H and Deng J. Cornernet: Detecting objects as paired keypoints//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 734-750

[25] Xu Y, Fu M, Wang Q, Wang Y, Chen K, Xia G-S, et al. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 34(4): 1452-1459

[26] Xu C, Li C, Cui Z, Zhang T, and Yang J. Hierarchical Semantic Propagation for Object Detection in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(6): 4353-4364

[27] Tan M, Pang R, and Le Q V. Efficientdet: Scalable and efficient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10781-10790

[28] Liu S, Huang D, and Wang Y. Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516, 2019

[29] Zhao Q, Sheng T, Wang Y, Tang Z, Chen Y, Cai L, et al. M2det: A single-shot object detector based on multi-level feature pyramid network//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 9259-9266

[30] Wang X, Zhang S, Yu Z, Feng L, and Zhang W. Scale-Equalizing Pyramid Convolution for Object Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13359-13368

[31] Ghiasi G, Lin T-Y, and Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7036-7045

[32] Hu J, Shen L, and Sun G. Squeeze-and-excitation networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132-7141

[33] Everingham M, Van Gool L, Williams C K, Winn J, and Zisserman A. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010, 88(2): 303-338

[34] He K, Zhang X, Ren S, and Sun J. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[35] Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, et al. Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019

[36] Redmon J and Farhadi A. Yolov3: An incremental improvement. // Proceedings of the Computer Vision and Pattern Recognition. Heidelberg, Germany, 2018:1804-2767

[37] He K, Gkioxari G, Dollár P, and Girshick R. Mask r-cnn// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961-2969

[38] Redmon J and Farhadi A. YOLO9000: better, faster, stronger// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7263-7271

[39] Dai J, Li Y, He K, and Sun J. R-fcn: Object detection via region-based fully convolutional networks// Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 379-389

[40] Azimi S M, Vig E, Bahmanyar R, Körner M, and Reinartz P. Towards multi-class object detection in unconstrained remote sensing imagery//Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018: 150-165

[41] Yan J, Wang H, Yan M, Diao W, Sun X, and Li H. IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery. Remote Sensing, 2019, 11(3): 286.

**XIE Xing-xing**, Ph. D. candidate. His research interests include computer vision and object detection for remote sensing images.

**CHENG Gong**, professor. His main research interests are computer vision, pattern recognition, and remote sensing image understanding.

**YAO Yan-qing**, Ph. D. candidate. Her research interests are computer vision and remote sensing image understanding.

**YAO Xi-wen**, associate professor. His research interests include computer vision and remote sensing image processing.

**HAN Jun-wei**, professor. His research interests include computer vision and brain-imaging analysis.

**Background**

With the advance of earth observation techniques, it has become easier to access massive amounts of remote sensing data. How to interpret these data has been a particular problem to be solved. As a key yet challenging task for remote sensing image interpretation, object detection in remote sensing images has achieved more attention, as well as presented wide application prospects. However, conventional object detection methods of remote sensing images (e. g. , handcrafted feature-based methods) do not obtain satisfied detection results, because of the highly complex backgrounds and variant appearances of ground objects. Fortunately, the powerful feature representation of convolutional neural networks (CNNs) provides a chances of breaking the deadlocks. Originally, CNN-based detectors have mainly prevailed and achieved the surprised performance in natural scene images. Inspired by the wave of the success of natural scene image object detection, CNN-based object detection in remote sensing images have begun to rise. Compared with natural scene images, remote sensing images show some difference, such as imaging perspective, dense object distribution, and more objects with small size. Thus, it is difficult to obtain promising results by directly transferring natural image object detection methods to remote sensing image object detection.

Benefiting from advanced detectors , such as Faster R-CNN, many researchers have proposed remote sensing image object detection approaches on the basis of Faster R-CNN. Some important works have been published on the top journals or proceedings over the past few years. Promising progress on object detection in remote sensing images have been witnessed. However, the large variations of object sizes and inter-class similarity are two leading challenges for object detection in remote sensing images. This has degenerated the accuracy of object detection in remote sensing images. Feature fusion is an effective approach to address these challenges and so has received wide attention. At present, most existing methods of feature fusion mainly build on Feature Pyramid Network (FPN) and utilize fixed weights to fuse the features of different scales. In these fusion methods, all input images share the fusion method. The static fusion approaches ignore the influence of object scales of input images on feature fusion. To this end, we design a dynamic feature fusion network, which contains two modules: a feature gate module and a dynamic fusion module. The feature gate module is to attenuate useless features and enhance useful features, avoiding fusing background information on the stage of dynamic feature fusion. The dynamic fusion module aims to build the bridge between the scales of input objects and fusion weights, and learn fusion weight based on the scales of input objects. The dynamic feature fusion network could minimize the influence from the variations of object sizes on feature fusion, thus improving the adaptiveness of feature fusion.

This paper is supported by the National Natural Science Foundation of China (No. 61772425) , the Shaanxi Science Foundation for Distinguished Young Scholars (2021JC-16), and the Doctor Dissertation of Northwestern Polytechnical University (No. CX2021082).