Opto-Electronic Engineering
光 电 工 程

Article
2022 年, 第 49 卷, 第 4 期

DOI: 10.12086/oee.2022.210363

基于多尺度特征融合的遥感图像小目标检测

马　梁[1,2,3]，苟于涛[1,2,3]，雷　涛[1,2*]，
靳　雷[1,2]，宋怡萱[1,2,3]

[1]中国科学院光电探测技术研究室，四川 成都 610209；
[2]中国科学院光电技术研究所，四川 成都 610209；
[3]中国科学院大学，北京 100049

摘要：本文提出了一种鲁棒的基于多尺度特征融合的遥感图像小目标检测方法。考虑到常用的特征提取网络参数量庞大，过多的下采样可能导致小目标消失，同时基于自然图像的预训练模型直接应用到遥感图像中可能存在特征鸿沟。因此，根据数据集中所有目标尺寸的分布情况(即：先验知识)，首先提出了一种基于动态选择机制的轻量化特征提取模块，它允许每个神经元依据目标的不同尺度自适应地分配用于检测的感受野大小并快速从头训练模型。其次，不同尺度特征所反应的信息量各不相同且各有侧重，因此提出了基于自适应特征加权融合的 FPN (feature pyramid networks) 模块，它利用分组卷积的方式对所有特征通道分组且组间互不影响，从而增加图像特征表达的准确性。另外，深度学习需要大量数据驱动，由于遥感小目标数据集匮乏，自建了一个遥感飞机小目标数据集，并对 DOTA 数据集中的飞机和小汽车目标做处理，使其尺寸分布满足小目标检测的任务。实验结果表明，与大多数主流检测方法对比，本文方法在 DOTA 和自建数据集上取得了更好的结果。

关键词：多尺度特征；小目标检测；特征融合；场景复杂度
中图分类号：TP751　　**文献标志码：**A

马梁，苟于涛，雷涛，等. 基于多尺度特征融合的遥感图像小目标检测 [J]. 光电工程，2022，**49**(4): 210363
Ma L, Gou Y T, Lei T, et al. Small object detection based on multi-scale feature fusion using remote sensing images[J]. *Opto-Electron Eng*, 2022, **49**(4): 210363

Small object detection based on multi-scale feature fusion using remote sensing images

Ma Liang[1,2,3], Gou Yutao[1,2,3], Lei Tao[1,2*], Jin Lei[1,2], Song Yixuan[1,2,3]

[1]Photoelectric Detection Technology Laboratory, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[2]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[3]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: This paper proposes a robust small object detection method based on multi-scale feature fusion using remote sensing images. When the natural image-based pre-training model is directly applied to the remote sensing images, the large number of parameters and excessive down sampling in widely feature extractions may lead to the disappearances of small objects due to feature gaps. Therefore, based on the distribution of all object sizes in the dataset (i.e., prior knowledge), a lightweight feature extraction module is first integrated via dynamic selection

收稿日期：2021−11−15；收到修改稿日期：2022−01−06
*通信作者：雷涛，taoleiyan@ioe.ac.cn。

mechanism that allows each neuron to adaptively allocate the receptive field size for detection. Meanwhile, the information reflected by various scale features has different amounts and emphasis. To increase the accuracy of image feature expression, the FPN (feature pyramid networks) module based on adaptive feature weighted fusion is applied by using the grouping convolution to group all feature channels without affecting each other. In addition, deep learning needs a large amount of data to drive. Due to the lack of remote sensing small object dataset, we built a remote sensing plane small object dataset, and processed the plane and small-vehicle objects in DOTA dataset to make its distribution of size meet the requirement of small object detection. Experimental results show that compared with most mainstream detection methods, the proposed method achieves better results on DOTA and self-built datasets.

Keywords: multi-scale features; small object detection; feature fusion; scene complexity

1 引 言

近年来，随着遥感光学技术的不断发展，高分辨率遥感图像的大量获取促进了环境监测、动物保护、交通管理、国防军事等领域的建设。在众多的遥感图像视觉任务中，遥感飞机检测对于民用和国防具有重要意义。与大中型目标的检测精度已提升到一个全新的高度相比，小型目标的检测受特征信息少以及目标区域存在复杂背景等影响，使得检测精度不高。因此，本文针对遥感小目标检测展开研究，以期对相关领域的发展起到一定的推动作用。

目前，基于深度学习的目标检测器大致可分为两大类：双阶段检测器 (如：RCNN[1]、Fast R-CNN[2]、Faster R-CNN[3]) 和单阶段检测器 (如：SSD[4]、YOLOv3[5]、RetinaNet[6])。这些方法在大中型目标检测任务中取得了优异的成绩，但小目标检测的效果较差。为了提升小目标的检测能力，Lin[7] 提出的 FPN (feature pyramid networks) 结构通过将神经网络中包含高级语义特征的深层特征图与包含丰富纹理细节特征的浅层特征图相融合。基于此工作，随后研究人员提出了多种特征融合方法[8-17]，检测性能均得到不同程度的提升。Pang[18] 通过引入注意力模型，降低了复杂背景对小目标检测的影响，降低了虚警率。针对遥感目标尺度变化大、遥感图像背景复杂等问题，文献 [19] 基于 RCNN 和 FPN 结构进行改进，设计并融合了全局上下文网络和金字塔局部上下文网络，分别在全局和局部提取上下文信息并引入空间感知注意力模块，引导网络关注信息更丰富的区域并生成更合适的图像特征。最近，Gong[20] 开始对特征融合时的权重进行研究，其通过统计的方法生成一组融合权重引入 FPN 结构，进一步提升了小目标的检测性能。这些检测算法在自然图像中虽有出色的检测能力，但在遥感小目标检测方面的表现与应用均欠佳，主要原因有以下几点：

1) 模型复杂，实时性差。很多基于深度学习的检测方法是通过增加网络深度和模型的复杂度来提升检测性能，庞大的计算量对硬件提出了更高的要求。与之相反，很多遥感检测任务需要在一些算力有限的边缘设备上部署，对模型的实时性有一定要求。因此，很多优秀的检测算法无法应用到其中。

2) 遥感图像背景复杂且目标尺度分布范围较广。小目标自身可用于区分的特征相对较少，因此相似的背景会对小目标检测产生严重干扰。如图 1 所示，与飞机形状过于接近的背景增加了网络训练的难度。同时，不同图像间由于分辨率不同，可能导致目标类内甚至类间巨大的尺度差异，大大增加了目标检测的难度。单一尺度很难覆盖所有的目标，因此，多尺度目标检测成为遥感图像检测的标配方式。

另外，目前虽有一些公开的遥感目标检测数据集如 DOTA[21]、DOAI[22]，但其多针对通用检测任务。对于特定任务 (如：遥感小目标检测) 的数据集极度匮乏。针对这一问题，一些检测或跟踪算法[23-28] 采用的方式是基于大型公开数据集 (如：ImageNet[29]) 上预训练好的模型在遥感数据上微调。而将通用数据集作为某一特定任务的数据支撑，其最终结果很难有保证。还有针对遥感目标方向任意分布的问题，文献 [30-32] 提出了多种基于旋转框的目标检测算法。其中，文献 [32] 的实验结果显示，基于旋转框的检测算法在面对不同目标类别时表现出了不同的性能：在检测飞机、棒球场这些长宽比接近 1:1 的目标时性能下降；在检测网球场、足球场等这些近似矩形目标时性能提升，说明旋转框并不适用于所有遥感目标。

为了解决上述问题，本论文提出了一种鲁棒的基

图 1 遥感图像中的复杂背景

Fig. 1 Complex background in remote sensing images

于多尺度特征融合的遥感图像小目标检测方法，其主要特点是：1) 由于图像输入常用的神经网络 (如：ResNet、VGG-16) 后会进行多次采样和卷积，造成小目标特征严重丢失，影响最终的检测精度。为此，根据数据集中所有目标尺寸的分布情况 (即：先验知识)，我们提出了一种基于动态选择机制的轻量化特征提取模块，它允许每个神经元依据目标的不同尺度自适应地分配用于检测的感受野大小并控制采样次数。2) FPN 虽已被广泛用于解决小目标漏检问题，但是不同尺度特征所反应的信息量通常各不相同且各有侧重，因此提出了基于自适应特征加权融合的 FPN 模块，它利用分组卷积的方式对所有特征通道分组且组间互不影响，从而进一步增加图像特征表达的准确性。3) 针对遥感小目标数据集匮乏的问题，本文自建了一个遥感飞机小目标数据集，并对 DOTA 数据集中的飞机和小汽车目标做处理，使其尺寸分布满足小目标检测的任务。最后，在 DOTA 和自建数据集上的实验结果表明，本文所提方法与主流检测算法相比均是最优结果。

2 方法原理

基于多尺度特征融合的遥感图像小目标检测方法，由基于动态选择机制的轻量化特征提取模块、基于自适应特征加权融合的 FPN 模块、目标分类及位置回归模块组成，其网络框架见图 2。

2.1 网络概述

遥感图像的尺寸一般较大，直接输入网络会导致庞大的计算量，从而引起内存不足。受文献 [33] 工作的启发，本文对大尺寸遥感图像预处理，以图像中任一目标为中心统一裁剪为 600 pixels×600 pixels 大小，超出部分用 0 填充，然后输入网络进行训练。网络结构如图 3 所示。首先，基于动态选择机制的轻量化特征提取模块负责对输入图像进行特征提取，随后通过自适应特征加权融合的 FPN 模块实现多尺度特征间的信息互补与加强，增强后的特征用来进行目标分类和位置回归。整个网络可以端到端从头训练模型。

2.2 基于动态选择机制的轻量化特征提取模块

目前，大多数目标检测方法都采用 VGG16、

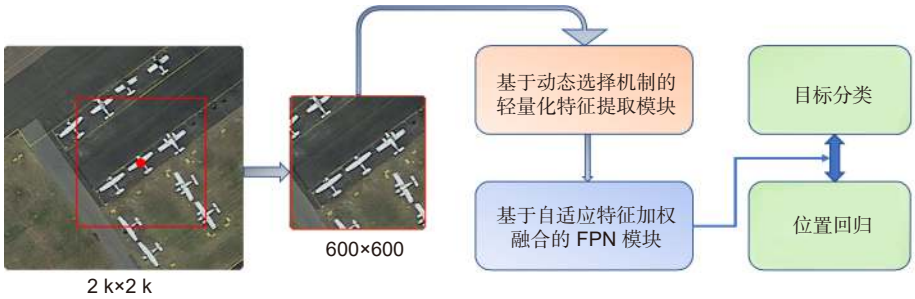


图 2 网络框架

Fig. 2 Network framework

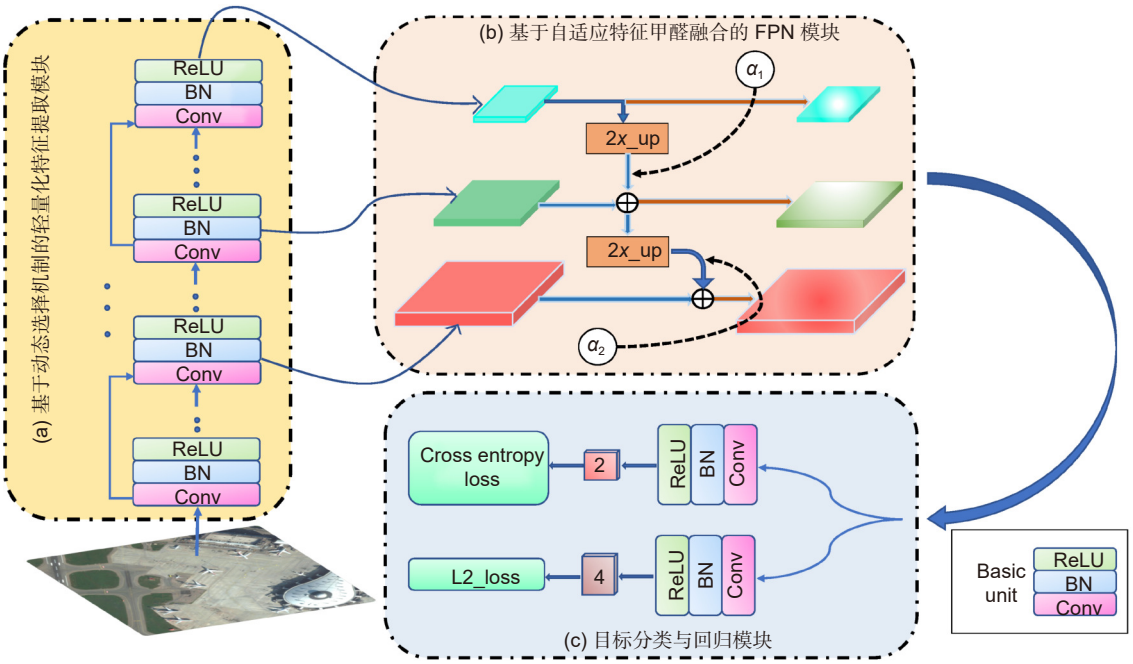


图 3 网络结构图

Fig. 3 Network structure

ResNet50、ResNet101 等网络，训练时基于 ImageNet 上预训练好的模型在自己的数据上进行微调。这种方式虽在一定程度上可以加快模型的收敛速度，但也存在一些弊端：1) 上述提到的特征提取网络参数量庞大，在要求实时性的任务中并不适用，此外当训练数据不足时模型泛化性也较差；2) VGG[34] 和 ResNet[35] 网络原本是针对图像分类任务设计的，过多的下采样操作可能导致深层特征无法有足够的分辨率来支持目标检测任务，这一点对小目标尤为不利；3) 由于自然图像和遥感图像间存在一定的特征鸿沟，将自然图像的预训练模型应用到遥感图像中可能导致次优的训练结果。

为了解决前述问题，本文提出了一种基于动态选择机制的轻量化特征提取模块，其结构如图 3(a) 所示。首先，我们统计数据集目标的整体尺寸分布情况 (作为一种先验知识)，然后依据数量将目标近似均匀地划分为多个尺度，同时根据划分的尺度分别计算用于检测的感受野大小 (小尺度目标自身特征稀少，需要上下文信息辅助判断，因此感受野与目标尺度比值要略大)。随后，模型依据输入目标的尺寸以及对应的感受野大小自主动态选择相应的特征图进行检测。感受野计算表达式如下：

$$R_{\mathrm{F}}(i)=R_{\mathrm{F}}(i-1)+(k-1)\times\sum_{j=1}^{i-1}S_j, \tag{1}$$

式中：$R_{\mathrm{F}}(i)$表示网络第i层的感受野大小，k表示第i层卷积核的大小，S_j表示第j层的卷积步长。根据式 (1) 可知，感受野大小与网络的深度和累积卷积步长成正相关关系。因此需要合理设置卷积层的卷积步长，使网络既能快速达到满足检测所有目标的感受野大小，保证网络轻量化，同时又能使小目标在深层网络中保留尽可能多的特征信息。

通过实验分析，该模块具有以下优势：1) 网络参数量少。如表 1 所示，相比 VGG 和 ResNet，我们的参数量大幅度减少，这使得可以不用依赖任何预训练模型就可以快速从头训练网络，这也规避了不同数据集间因存在特征鸿沟而引起的模型性能下降。2) 网络结构易于修改，普适性高。可根据任务需求快速修改网络结构，同时还便于和其他通用检测模块相结合，如 FPN，进而执行更复杂的检测任务。

2.3 基于自适应特征加权融合的 FPN 模块

区别于图像中的大中型目标，小目标往往呈现外观模糊，不易察觉的特点，有些小目标甚至只有几个像素到十几个像素，这就导致小目标存在特征稀少、高区分度特征提取困难等问题。神经网络浅层特征缺乏足够的语义信息，较小的感受野很难提取到全局语义，因此需要高层语义的融合，辅助浅层网络检测小目标。

表 1　不同网络的参数量

Table 1　Parameters of different networks

模型	参数量M
VGG16	138
ResNet50	25.6
ResNet101	44.6
Ours	0.49

自 FPN 问世后，当前对多尺度特征融合的研究多集中在以下几个方面：

1) 多尺度融合结构的探索。通过构建更加复杂的融合结构和融合策略来提升目标检测的性能，但复杂的结构带来了大量运算和内存开销。另外，不同尺度的特征对融合的贡献值不同。绝大多数融合结构都是将多尺度特征无差别的 1:1 直接融合，这种策略并非在所有任务中都能产生最优的融合效果。一个 10 像素的目标在经过三次两倍下采样操作后，在特征图上仅剩 1 到 2 个像素的信息，几乎没有可用来检测的特征。这种情况下，传统的 FPN 结构可能并非最优方案甚至在某些情况下带来负面影响。

2) 目标尺度匹配问题研究，FSAF[36] 抛开锚框匹配策略，提出了一种目标自适应尺度匹配方法，但过多的计算开销以及不易设计合理的损失函数使其在实际应用中比较困难。

为了充分发挥多尺度特征间的互补作用，本文提出了一种基于自适应特征加权融合的 FPN 模块，如图 3(b) 所示。具体研究思路如下：我们首先对特征图的通道进行分组操作。然后，通过分组卷积操作使各特征 f 独立取得融合因子 α 并获得加权特征图 (见图 4)，其计算表达式：

$$F_{\text{refined}}(f_{ij},\alpha_{ij}) = f_{ij}*\alpha_{ij}, \tag{2}$$

式中：$1 \leqslant i \leqslant G; 1 \leqslant j \leqslant C/G$，$G$和$C$分别表示分组数目和特征图的通道数，$f_{ij}$表示第$i$组第$j$个通道的特征，$\alpha_{ij}$是作用于$f_{ij}$上的融合因子。在模型训练过程中，它可以根据不同的数据和目标函数，通过梯度反向传播不断学习优化，最终达到最优的融合效果。注意：在组内进行常规卷积运算，组间互不影响。另外，该方法具有压缩计算量的特点。其参数量是相同标准卷积的 $1/G$。在获得加权特征图F_{refined}的基础上，进一步融合其它尺度的特征，获得信息更完整，更有区分度的特征F_{det_i}，从而提升目标检测的精度。多尺度特征融合表达式如下：

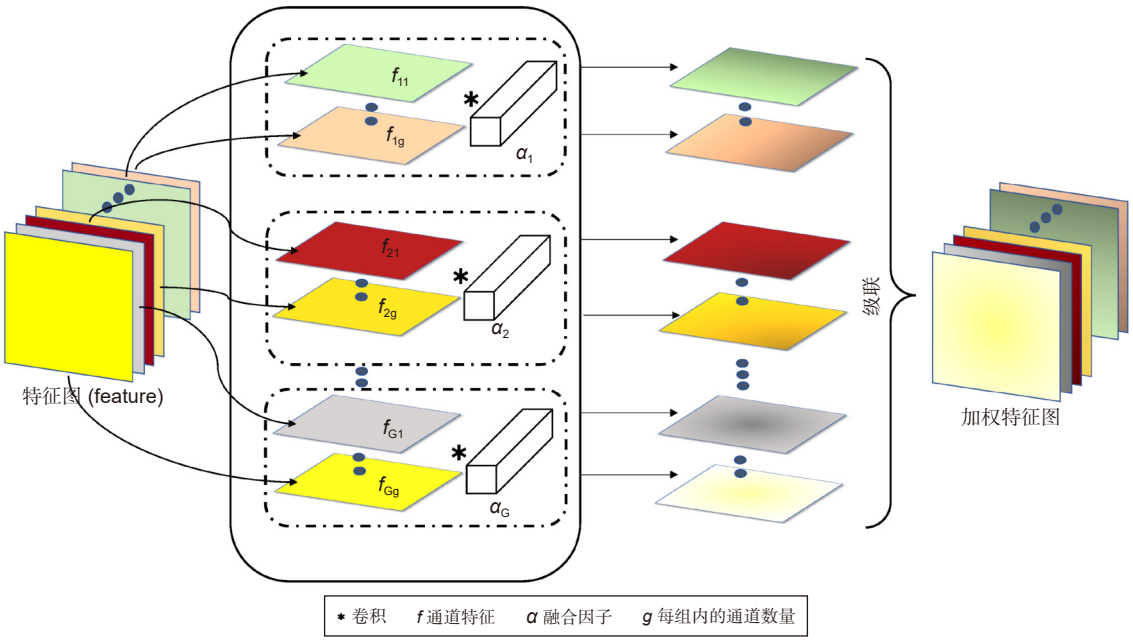


图 4　基于分组卷积的特征加权方法

Fig. 4　Feature weighting method based on grouped convolution

$$F_{\text{det\_}i}=conv(F_{\text{refined}}(Deconv(conv(F_{\text{det\_}i-1},k_1),k_2),\alpha)+conv(f_i,k_3),k_4), \tag{3}$$

式中：$F_{\text{det_}i}$表示用于检测第i个尺度目标的最终特征图，f_i表示第i个尺度的原始特征图，$conv$和$Deconv$分别表示卷积和反卷积运算，k_1、k_2、k_3、k_4分别表示各卷积运算对应的卷积核。

2.4 目标分类及回归模块

Anchor-base 是一种常见的目标分类与回归策略，但这种方式存在一些不足：锚框尺寸、长宽比以及数量都需要根据目标的尺度分布情况而定，这是一个相当繁琐的过程。

人眼能看到的区域称为“视场”，只有进入其中的目标才有可能被人发现，卷积网络某一层的特征图的某个位置的特征向量是从前一层的特定区域计算出来的，即感受野。如图 5(a) 所示，对于 3×3 卷积，底部左下角的 3×3 红色区域就是顶部左下角的感受野。而感受野就是神经网络的“视场”。我们基于感受野对目标进行分类和回归，如图 5(b) 所示，红色虚线框代表某一感受野范围，当目标位于其中时，分为正样本 (绿色标定目标)，反之为负样本 (蓝色标定目标)；若目标同时处于多个感受野内时，则将其忽略 (黄色标定目标)。分类损失采用交叉熵损失，回归损失使用 L2 损失。

$$L_{\text{cls}}=-(y\cdot\log(\hat{y})+(1-y)\cdot\log(1-\hat{y})), \tag{4}$$

$$L_{\text{reg}}=\sum_{i=1}^{n}(y_i-f(x_i))^2, \tag{5}$$

其中：y 和y_i表示目标值，$\hat{y}$和$f(x_i)$表示预测值。

3 实验结果

3.1 数据集

DOTA-v1.5：武汉大学公开发布的航空图像目标检测数据集，共计 2806 张，尺寸 800 pixels×800 pixels~4000 pixels×4000 pixels 不等，包含 16 个类别共计 40 万个实例。数据集的飞机目标尺度分布范围较广，我们所关注的目标主要是 32 pixels 以下的小目标。将数据集中所有的飞机目标选出后进行处理，最终 99.7% 的目标分布在 6 pixels~70 pixels 之间，且 6 pixels~30 pixels 的小目标占比达到 74.8%，其中，训练集 646 张图像，13790 个目标；测试集 160 张图像，3102 个目标。此外，我们还将数据集中的小汽车 (small-vehicle) 目标选出后进行处理，最终 99.8% 的目标分布在 6 pixels~60 pixels 之间，且 6 pixels~25 pixels 的小目标占比达到 74.7%，其中，训练集 9120 张图像，185044 个目标；测试集 2610 张图像，60193 个目标。

自建数据集：本文建立了一个包含 3576 张图像，24853 个目标实例的遥感飞机小目标数据集。其中，99.7% 的目标分布在 6 pixels~50 pixels 之间，且 20 pixels 以下的目标占比达到了 64.2%，其中，训练集有 2835 张图像，19722 个目标；测试集 741 张图像，5131 个目标。

实验数据集目标的尺度分布统计情况见图 6，经过预处理后 DOTA 数据集中飞机与小汽车的训练集和测试集样图见图 7。

3.2 数据预处理

为了得到泛化能力更强的检测模型，在数据输入网络进行训练前，我们对数据进行了如下预处理：

1) 通用数据增强方法：为了提升数据的表达能力，我们对数据进行了翻折、随机亮度调整、随机饱和度调整、随机对比度调整、均值滤波以及高斯滤波。通过这些方法让图像数据体现更多实际中的场景，进而提升训练模型的泛化能力。

2) 随机目标中心裁剪：为了让网络学习更加泛化

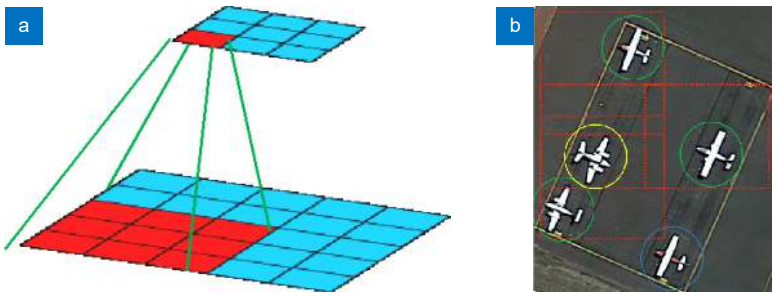


图 5 (a) 卷积网络感受野示意图；(b) 基于感受野的目标分类策略

Fig. 5 (a) Schematic diagram of convolutional network receptive field; (b) Object classification strategy based on receptive field

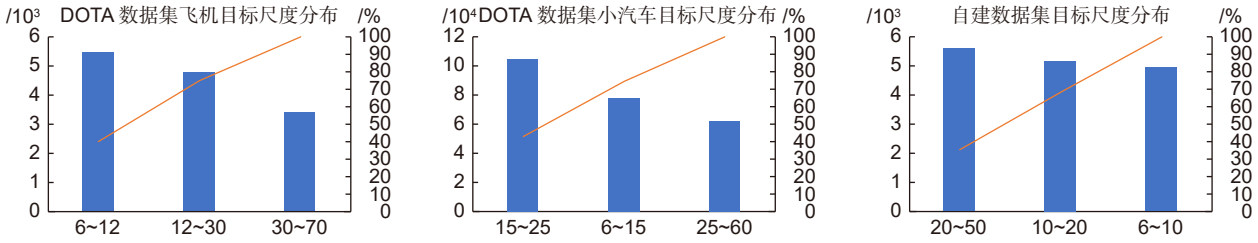


图 6 数据集目标尺度分布

Fig. 6 Object scale distribution of the dataset

图 7 DOTA 数据集中的飞机与小汽车图像样例。(a) 训练集，(b) 测试集

Fig. 7 Sample of plane and small-vehicle image of DOTA dataset used in the experiment. (a) Training set; (b) Testing set

的特征，每次随机以图像中的某个目标为中心，裁剪一张 600 pixels×600 pixels 的“新图像”，然后送入网络进行训练。这样处理可以进一步提高训练数据的多样性，即使同一张图像，在不同 mini-batch 的训练中，也可以生成完全不同的训练数据，同时随机裁剪又可以避免网络学习一些无关紧要的位置信息，迫使其学习更加鲁棒的目标特征。

3) 小目标增强：目标越小越难以捕捉和学习其特征，我们使用的数据集中有大量 6 pixels~15 pixels 的极小目标，为了保证网络对这些小目标的学习能力，在每个 mini-batch 的数据中，至少保证有一个目标处于该尺度，否则重新选择一批数据送入网络训练。这样处理可以使网络在每次训练的过程中都有这些极小目标的特征去学习，从而避免网络过多学习大尺寸目标的特征而导致对小目标检测能力的下降。

4) 目标剪切粘贴：DOTA 数据集处理完成后，统计发现 15 pixels 以下的飞机目标数量较少，而该尺度内的小目标最难检测，需要保证有足够数量的目标位于该尺度范围内。因此采用随机裁剪粘贴的方法，裁剪目标一共尝试了两种方案：沿物体边缘将目标抠出和连带目标周围小部分背景按矩形扣出。接下来，对抠出的目标随机进行数据增强，包括翻折，旋转，对比度、饱和度变换等，使目标更好地模拟实际遥感图像中可能出现的不同情况，并且能够进一步增加数据量。最后，将目标随机粘贴到背景中的某一区域。经过实验发现，按第一种裁剪目标方案抠出的目标会导致网络不收敛。主要原因是按边缘裁剪会破坏目标自身的边缘特征，而目标的边缘特征是网络执行目标回归重要的信息支撑，这些被破坏的边缘特征分布与其他真实目标的边缘特征分布不一致，模型难以在这样的数据下拟合。为了避免边缘的影响，最终采用了第二种目标裁剪方案，该方案不会对目标的边缘信息造成破坏。该方法流程如图 8 所示。

5) 目标标注优化：首先，在处理 DOTA 数据集时，我们经过了降采样处理，处理之后的某些小目标已经无法辨认，如果将这些目标继续作为正样本参与训练会干扰模型的拟合效果，因此将这些目标剔除作为负样本，剔除方法包括真值框去除和背景覆盖两种形式。其次，对剪切粘贴后的扩充目标重新标记。最后，对一些密集排布的小目标标注框重新标注校正，确保标签可靠准确。

3.3 评估指标

目标检测常用的评价指标有准确率 (precision)、召回率 (recall)、AP (average precision) 和 mAP (mean

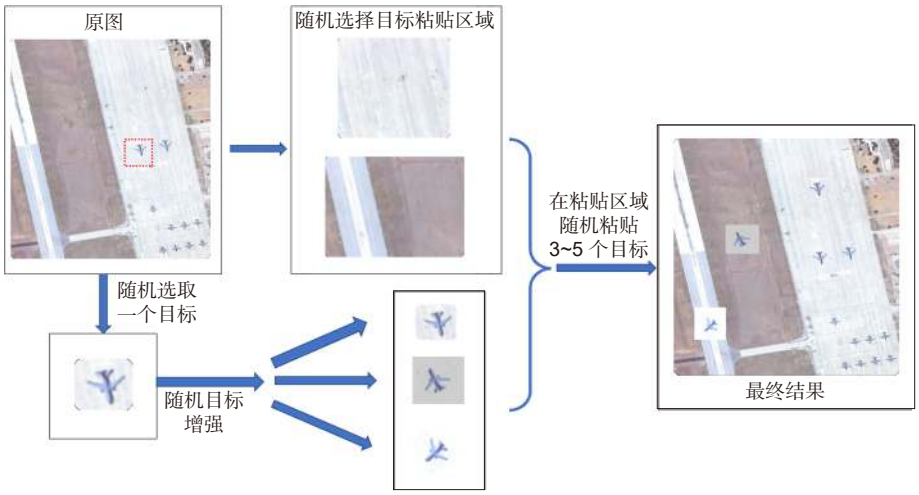


图 8 目标剪切粘贴流程示意图

Fig. 8 Objects cut and copy flow diagram

average precision)。单纯依靠准确率或召回率来进行算法优劣的评价是不严谨的，因为同一种算法可以通过调整置信度阈值来改变准确率或者召回率。而 AP 是 P-R 曲线与坐标轴所围区域的面积，其值可以综合反映算法在准确率和召回率上的优劣，越接近 1 算法性能越好。mAP 是所有类别 AP 值的均值。由于我们的检测目标只有飞机一个类别，因此，对于本文的任务来说，mAP 和 AP 指标一致。

3.4 参数设置

我们使用 SGD 优化器训练网络，动量设为 0.9，权重衰减为 0.00001。mini-batch 为 8，初始学习率为 0.1，共训练 50 万次，其中在 30 万和 45 万次后学习率下调 10 倍。NMS 阈值设为 0.4。

3.5 对比实验与数据分析

3.5.1 不同金字塔层数的融合对比实验

根据自建数据集目标尺度的分布情况，将目标分为多个尺度，我们实验了两种方案：划分为 2 个尺度和 3 个尺度，并分别计算其对应的感受野大小，具体参数见表 2。实验结果见表 3 (B_*x* 表示特征提取网络有 *x* 个 Basic unit)。其中，双尺度方案使 mAP 提升 0.4%，三尺度方案使 mAP 提升 1.1%，同时，准确率和召回率相比前者也均有明显提升。这是由于前者在单一尺度上划分比较广，导致用较大感受野检测小目标，同时 4 倍下采样使小目标信息丢失严重，对小目标检测造成较大干扰。作为对比，后者在尺度划分和感受野大小匹配上更加精细，在提供小目标检测所需

表 2 特征图感受野与对应目标尺寸参数

Table 2 Receptive field of feature map and corresponding object size parameters

金字塔层数		检测目标尺寸	下采样倍数	感受野	感受野步长	感受野/目标尺寸
两层	1	6~25	4	55	4	3.5
	2	25~50	8	95	8	2.5
三层	1	6~10	2	23	2	2.9
	2	10~20	4	47	4	3.1
	3	20~50	8	79	8	2.3

表 3 不同特征融合方案的检测结果

Table 3 Detection results of different feature fusion schemes

Basic unit		FPN	mAP	Precision	Recall
两层	B_11	-	86.8	76.8	88.9
	B_11	√	87.2	82.6	88.8
三层	B_10	-	87.4	47.4	**91.8**
	B_10	√	**88.5**	**83.8**	90.4

上下文信息的同时，保留了更多小目标自身的特征信息。

之后的所有实验都采用 3 个尺度的划分方案。虽然此实验是在自建的数据集上进行的，但根据统计信息，处理后的 DOTA 数据集目标尺度分布情况与其非常接近，因此，训练检测 DOTA 数据集也沿用该方案。

3.5.2 基于自适应特征加权融合的 FPN 模块的有效性

我们的融合因子取自 1×1 卷积核，随后通过分组卷积实现融合因子与特征图的加权操作。为了验证融合因子的有效性，在 DOTA 和自建数据集上分别进行了多组对比实验，实验结果如表 4、表 5 和表 6。可以看出，不同方式下的加权融合相比传统 FPN 均带来不同程度的性能提升。同时如图 9 和图 10 所示，模型收敛速度进一步加快。具体检测效果如图 11、图 12 所示。

特征图的每个通道都包含了一组特定的特征信息，它们彼此之间的关联性有强有弱，有效地利用它们之间的关联性，可以进一步提升融合特征图的信息表达能力。具体的融合方式有两种：1) 利用分组卷积组间计算互不影响的特点，对特征通道进行不同数量的分组 (例如分为 3 组等)，然后通过实验结果可以在某种程度上间接分析这些通道特征之间的关联性。但这种方式缺乏理论支撑，只能盲目地通过实验不断尝试。2) 考虑到每一个通道特征都有决定最终融合效果的能力。直接将分组卷积的分组数量设为与对应特征图通道 (channel) 数量相等，这样即考虑了每一个通道特征对于融合的贡献值，又使每个通道特征都独立获得一个属于自己的融合权重，加之通过网络的不断学习，最终得到一组最佳的融合权值。我们进行了一系列对比实验 (见表 4、表 5 和表 6)，结果证明了所提方法的有效性。同时分组卷积相比常规卷积，可以大大降低计算量，分组数量越多，计算量越小。因此，该方法在提升目标检测性能的同时，还最大程度兼顾了算法的实时性。

与可自适应学习的融合因子一样，常数因子也可

表 4　网络不同配置下的 DOTA 飞机数据集测试结果

Table 4　DOTA plane dataset test results under different network configurations

B_13	FPN	分组数量(3)	特征图通道(channel)	常数融合因子[0.71,0.87]	mAP	Precision	Recall
√	-	-	-	-	80.5	63.4	82.8
√	√	-	-	-	82.0	81.4	85.1
√	√	√	-	-	82.3	85.1	84.5
√	√	-	√	-	**83.6**	**85.5**	**87.0**
√	√	-	-	√	82.5	82.3	85.6

表 5　网络不同配置下的 DOTA 小汽车数据集测试结果

Table 5　DOTA small-vehicle dataset test results under different network configurations

B_12	FPN	分组数量(3)	特征图通道(channel)	常数融合因子[0.63,1.28]	mAP	Precision	Recall
√	-	-	-	-	63.7	56.8	**73.9**
√	√	-	-	-	65.9	86.1	68.5
√	√	√	-	-	66.3	83.3	68.9
√	√	-	√	-	**68.7**	**86.4**	71.7
√	√	-	-	√	64.4	84.0	67.3

表 6　网络不同配置下的自建数据集测试结果

Table 6　Test results of our dataset under different network configurations

B_10	FPN	分组数量(3)	特征图通道(channel)	常数融合因子[1.08,1.05]	mAP	Precision	Recall
√	-	-	-	-	89.9	44.2	**93.7**
√	√	-	-	-	90.2	83.6	91.4
√	√	√	-	-	90.6	84.8	92.0
√	√	-	√	-	**91.0**	**87.7**	92.4
√	√	-	-	√	未收敛	未收敛	未收敛

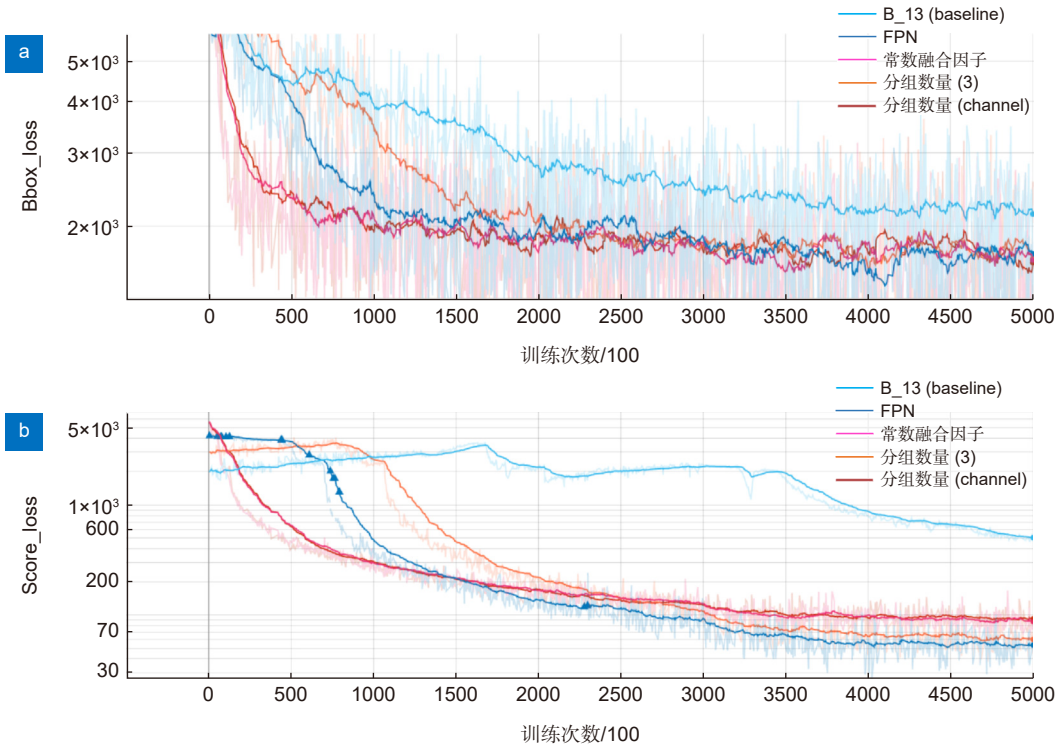


图 9　网络在 DOTA 飞机训练集上训练的 loss 曲线

Fig. 9　The loss curve of the network trained on the DOTA plane training set

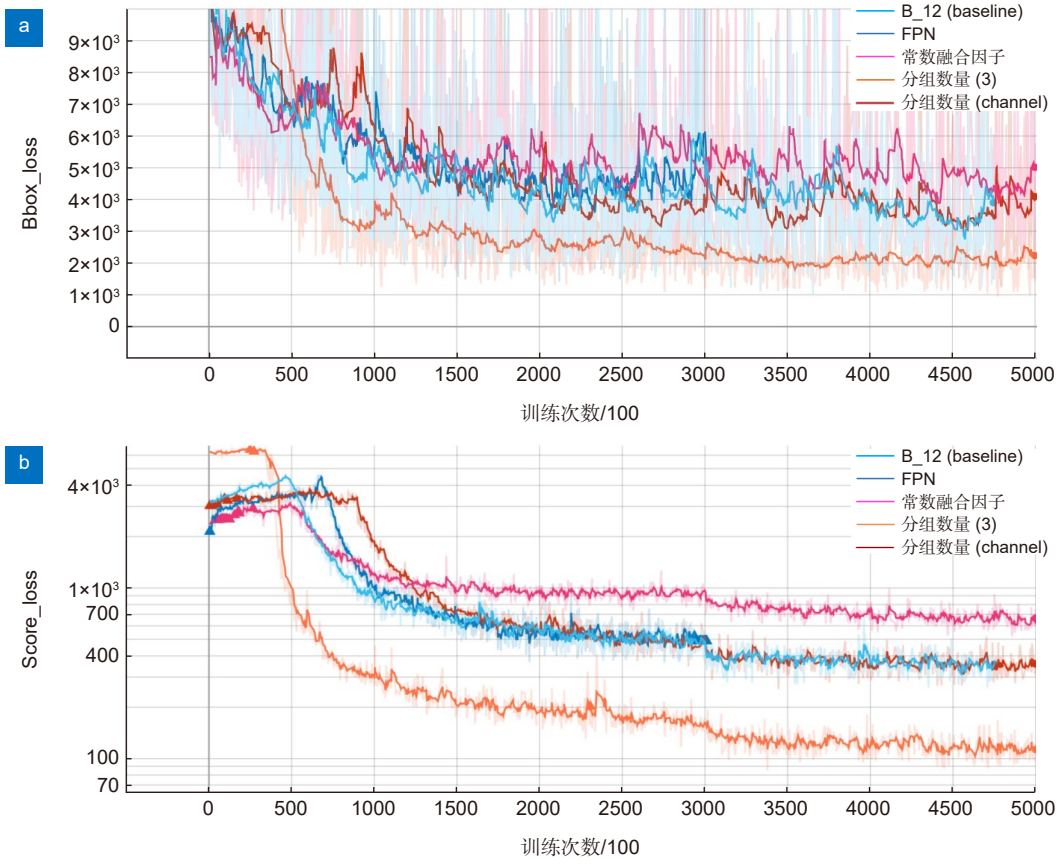


图 10　网络在 DOTA 小汽车训练集上训练的 loss 曲线

Fig. 10　The loss curve of the network trained on the DOTA small-vehicle training set

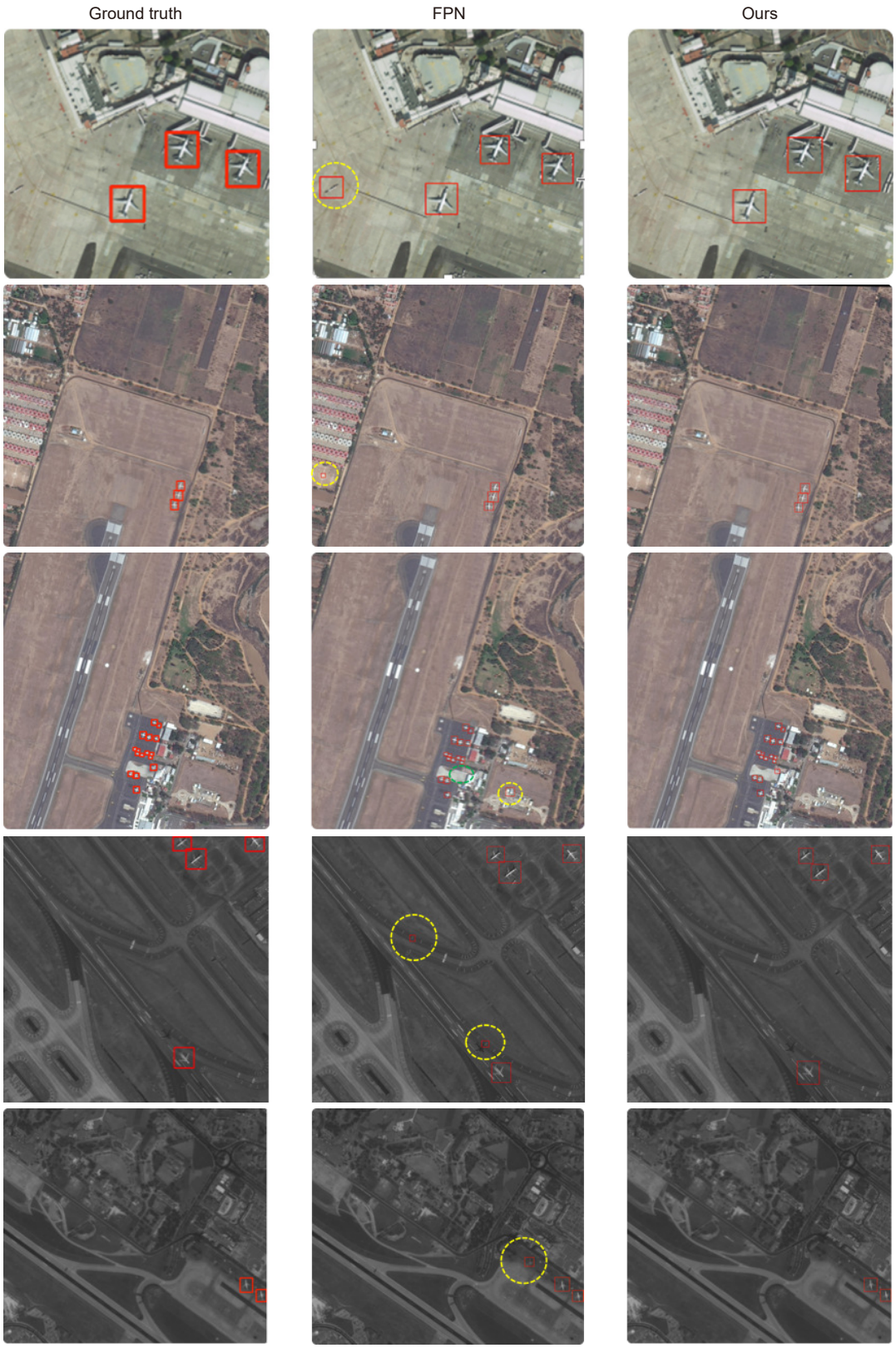


图 11　部分飞机检测结果。

黄色圆圈代表虚警，绿色圆圈代表漏检。

Fig. 11　Partial plane test results.

Yellow circles represent false alarms and green circles represent missed detection.

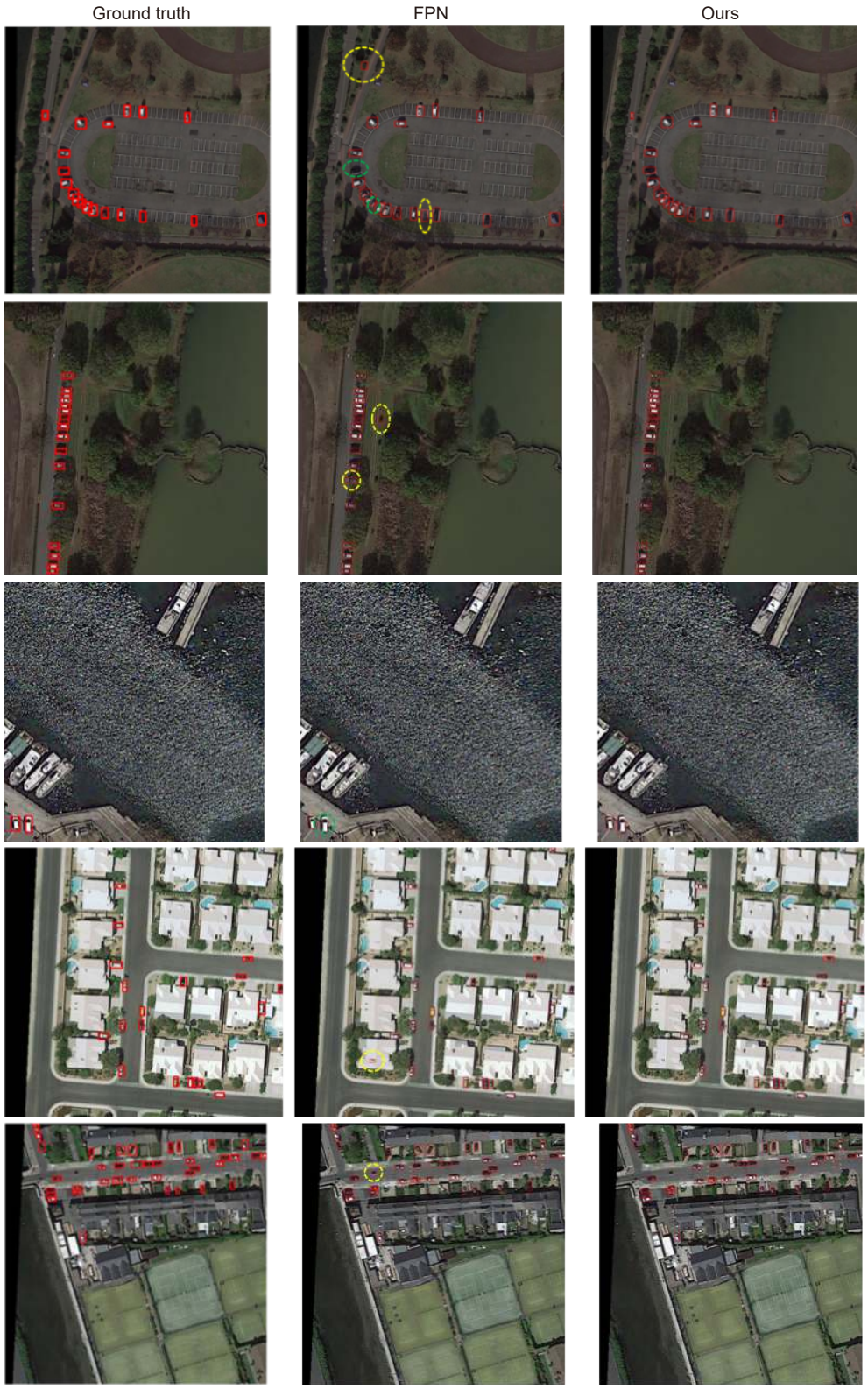


图 12　部分小汽车检测结果。

黄色圆圈代表虚警，绿色圆圈代表漏检。

Fig. 12　Partial small-vehicle test results.

Yellow circles represent false alarms and green circles represent missed detection.

用于融合加权，例如 FPN 的融合权重为常数 1。考虑到某一尺度的目标越多，其产生的损失往往也越大，这可能导致网络偏向于学习该尺度的目标特征，降低模型的性能。因此，我们稍作改变：根据目标在相邻尺度区间内分布数量的比值作为常数融合因子 (见表 7)。实验发现，其性能在不同数据集上会产生不一致的结果。主要原因是：神经网络是将网络产生的损失值通过梯度反向传播的方式来不断优化网络的拟合效果。FPN 的结构决定了每个尺度的损失都会受其他尺度损失的影响，融合因子可以调节各尺度损失在其他尺度损失中的占比，从而使各尺度更有效地学习各自所需的特征。由于训练时不同 mini-batch 中的目标尺度分布情况不一定与总体分布一致，因此使用基于数据集整体目标尺度分布统计得到的固定融合因子去训练网络可能会得到次优甚至不收敛的情况。

除以上探讨的因素外，自适应融合因子的初始值也是决定特征融合结果的重要因素。因此，在 DOTA 数据集 (飞机类目标) 上对比了融合因子不同初始化方法对检测结果的影响。实验结果见表 8。随机初始化融合因子会导致检测性能骤降。主要是因为随机初始化权重值偏小，甚至接近于 0。过小的融合权重在一开始就稀释了太多的特征信息，致使后续检测无法获取足够的特征导致模型性能衰退。初值为 1 可以保证特征信息在一开始不会受到损失，随后网络通过学习不断优化融合权重，最终得到一组最优解。同时，图 13 展示了融合因子在不同初始值下的模型收敛过

表 7　数据集各尺度目标分布数量统计

Table 7　Statistics of the distribution of each scale objects number

数据集	尺度	目标数量	常数融合因子$\left(\frac{S_{i+1}}{S_i}\right)$
DOTA飞机训练集	S_1[6-12]	5503	0.87
	S_2[12-30]	4807	
	S_3[30-70]	3428	0.71
DOTA小汽车训练集	S_1[6-15]	59875	1.28
	S_2[15-25]	76615	
	S_3[25-60]	48203	0.63
自建数据训练集	S_1[6-10]	4963	1.05
	S_2[10-20]	5196	
	S_3[20-50]	5625	1.08

表 8　融合因子初始值对检测性能的影响

Table 8　Influence of initial value of fusion factor on detection performance

融合因子初始值	mAP
1	**83.6**
随机初始化	80.7

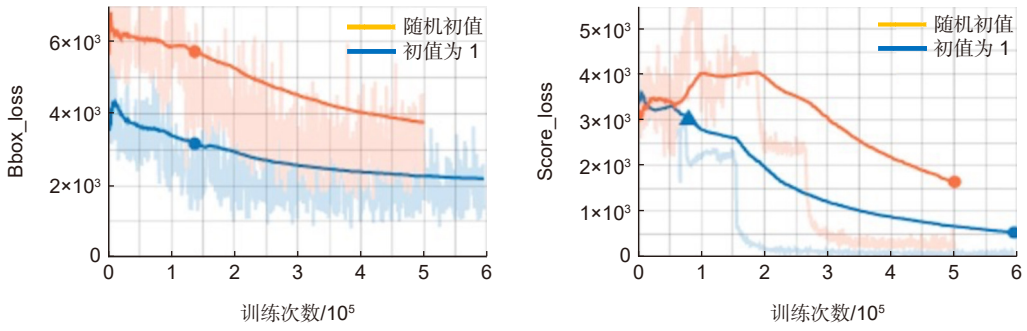


图 13　融合因子不同初始值下的模型收敛情况

Fig. 13　Model convergence under different initial values of fusion factors

程。可以看出，初值为 1 情况下的模型收敛速度不仅更快，而且更趋向于收敛到最优解。

3.5.3 与不同方法的对比

为了让网络更多地关注小目标本身的语义特征，避免背景信息的干扰，尝试引入 CBAM[37] 注意力机制方法，其主要包含两个模块：通道注意力和空间注意力。通道注意力关注物体类别的判定，通过均值池化和最大池化聚合每一个通道的空间信息，通过多层感知机判断不同通道对类别判断的重要性，生成通道注意力权重。空间注意力则关注物体空间位置的判定，在通道维实现最大池化和均值池化，强化空间特征，利用卷积生成空间注意力权重。

CBAM 的通道注意力思想与本文提出的多尺度特征自适应加权融合有相似之处，都是通过生成一组权重来衡量通道间特征的一个重要程度。但前者只是在单一特征图上进行这一操作，并未增加特征的信息量；后者在分析单一特征图各通道权重的同时进一步融合了多个尺度特征间的信息，增强了特征图信息的表达能力。同时生成权值采用的方式也不一样。

对两种方法的实验对比结果如表 9 所示。本文的方法在检测性能以及模型推理速度方面要全面优于 CBAM 注意力机制。主要原因是 CBAM 并没有增加特征图的信息量，对后续目标检测任务的支持能力有限；其次，CBAM 模块引入了大量计算，直接导致推理速度下降三分之一左右，而本文采用分组卷积的方法，在引入自适应融合因子的同时带来了极小的额外计算开销。

为了进一步验证本文方法的有效性，与多种目标检测算法进行比较，实验结果见表 10。本文的方法在三个数据集上均是最优的。同时，DOTA 数据集是彩色图像，我们的数据集是灰度图像，这证明本文的算法不受彩色图或灰度图的限制。此外，在 DOTA 数据集上进一步测试了基于自适应特征加权融合的 FPN 模块在双阶段检测器上的效果，实验见表 11。

表 9 CBAM 与自适应融合模块对检测性能的影响

Table 9 Influence of CBAM and adaptive fusion module on detection performance

模型+数据集	mAP	Precision	Recall	推理速度/(s/张)
B_10+FPN+CBAM(自建数据集)	90.5	83.8	90.6	0.036
B_10+FPN+自适应融合模块(自建数据集)	**91.0**	**87.7**	**92.4**	**0.027**
B_13+FPN+CBAM(DOTA飞机数据集)	83.0	82.6	85.8	0.048
B_13+FPN+自适应融合模块(DOTA飞机数据集)	**83.6**	**85.5**	**87.0**	**0.037**
B_12+FPN+CBAM(DOTA小汽车数据集)	67.6	83.0	71.1	0.043
B_12+FPN+自适应融合模块(DOTA小汽车数据集)	**68.7**	**83.3**	**71.7**	**0.034**

表 10 不同方法检测性能对比

Table 10 Comparison of detection performance of different methods

方法	DOTA飞机数据集(mAP)	DOTA小汽车数据集(mAP)	自建数据集(mAP)
SSD	63.4	43.3	64.4
RetinaNet	55.2	45.1	62.7
Yolov3-tiny	70.8	58.3	74.3
Faster R-CNN	73.0	59.0	88.6
Ours	**83.6**	**68.7**	**91.0**

表 11 基于自适应特征加权融合的 FPN 模块在 Faster R-CNN 上的性能

Table 11 Performance of FPN module based on adaptive feature weighted fusion on Faster R-CNN

Backbone+数据集	mAP
ResNet50+FPN(自建数据集)	88.6
ResNet50+自适应融合模块(自建数据集)	**89.7**
ResNet50+FPN(DOTA飞机数据集)	73.0
ResNet50+自适应融合模块(DOTA飞机数据集)	**73.8**
ResNet50+FPN(DOTA小汽车数据集)	59.0
ResNet50+自适应融合模块(DOTA小汽车数据集)	**63.2**

结果表明该模块在双阶段检测器上依然适用。本文的算法是鲁棒的。

4 结 论

针对遥感小目标易受复杂背景干扰、通用检测算法表现不佳以及相关数据集匮乏的问题，本文提出了解决方法：首先提出了一种基于动态选择机制的轻量化特征提取模块，它允许每个神经元依据目标的不同尺度自适应地分配用于检测的感受野大小，降低复杂背景对小目标检测的影响。其次，不同尺度特征所反应的信息量各不相同且各有侧重，提出了基于自适应特征加权融合的 FPN 模块，它利用分组卷积的方式对所有特征通道分组且组间互不影响，从而进一步增加图像特征表达的准确性。另外，深度学习需要大量数据驱动，本文自建了一个遥感飞机小目标数据集，并对 DOTA 数据集中的飞机和小汽车目标做处理，使其尺寸分布满足小目标检测的任务。最后，在 DOTA 飞机目标、DOTA 小汽车目标和自建数据集上的实验结果显示，所采用的方法分别达到了 83.6%、68.7% 和 91% 的 mAP，相比传统 FPN 带来了 1.6%、2.8% 和 0.8% 的 mAP 提升。同时也验证了本文所提出的自适应融合模块在双阶段检测器上同样适用，我们的方法是鲁棒的。但是，该网络目前也存在不足之处，如：对于密集排布的小目标存在漏检问题，后续的工作将进一步展开研究。

参考文献

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580–587.

[2] Girshick R. Fast R-Cnn[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, 2015: 1440–1448.

[3] 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 Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, 2015, **28**: 91–99.

[4] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 21–37.

[5] Redmon J, Farhadi A. YOLOV3: an incremental improvement[Z]. arXiv: 1804.02767, 2018. https://doi.org/10.48550/arXiv.1804.02767.

[6] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 2999–3007.

[7] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936–944.

[8] Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector[Z]. arXiv: 1701.06659, 2017. https://arxiv.org/abs/1701.06659.

[9] Li Z X, Zhou F Q. FSSD: feature fusion single shot multibox detector[Z]. arXiv: 1712.00960, 2017. https://doi.org/10.48550/arXiv.1712.00960.

[10] Cui L S, Ma R, Lv P, et al. MDSSD: multi-scale deconvolutional single shot detector for small objects[Z]. arXiv: 1805.07009, 2018. https://doi.org/10.48550/arXiv.1805.07009.

[11] Liang Z W, Shao J, Zhang D Y, et al. Small object detection using deep feature pyramid networks[C]//*Proceedings of the 19th Pacific Rim Conference on Multimedia*, 2018: 554–564.

[12] Cao G M, Xie X M, Yang W Z, et al. Feature-fused SSD: fast detection for small objects[J]. *Proc SPIE*, 2018, **10615**: 106151E.

[13] Zhang S F, Wen L Y, Bian X, et al. Single-shot refinement neural network for object detection[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 4203–4212.

[14] Zhao Q J, Sheng T, Wang Y T, et al. M2Det: a single-shot object detector based on multi-level feature pyramid network[J]. *Proc AAAI Conf Artif Intell*, 2019, **33**(1): 9259−9266.

[15] Xu A L, Du D, Wang H H, et al. Optical ship target detection method combining hierarchical search and visual residual network[J]. *Opto-Electron Eng*, 2021, **48**(4): 200249.
徐安林, 杜丹, 王海红, 等. 结合层次化搜索与视觉残差网络的光学舰船目标检测方法[J]. 光电工程, 2021, **48**(4): 200249.

[16] Zhao C M, Chen Z B, Zhang J L. Research on target tracking based on convolutional networks[J]. *Opto-Electron Eng*, 2020, **47**(1): 180668.
赵春梅, 陈忠碧, 张建林. 基于卷积网络的目标跟踪应用研究[J]. 光电工程, 2020, **47**(1): 180668.

[17] Jin Y, Zhang R, Yin D. Object detection for small pixel in urban roads videos[J]. *Opto-Electron Eng*, 2019, **46**(9): 190053.
金瑶, 张锐, 尹东. 城市道路视频中小像素目标检测[J]. 光电工程, 2019, **46**(9): 190053.

[18] Pang J M, Li C, Shi J P, et al. R^2-CNN: fast tiny object detection in large-scale remote sensing images[J]. *IEEE Trans Geosci Remote Sens*, 2019, **57**(8): 5512−5524.

[19] Zhang G J, Lu S J, Zhang W. CAD-Net: a context-aware detection network for objects in remote sensing imagery[J]. *IEEE Trans Geosci Remote Sens*, 2019, **57**(12): 10015−10024.

[20] Gong Y Q, Yu X H, Ding Y, et al. Effective fusion factor in FPN for tiny object detection[C]//*Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision*, 2021: 1159–1167.

[21] Xia G S, Bai X, Ding J, et al. DOTA: a large-scale dataset for object detection in aerial images[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 3974–3983.

[22] Ding J, Xue N, Xia G S, et al. Object detection in aerial images: a large-scale benchmark and challenges[Z]. arXiv: 2102.12219, 2021. https://doi.org/10.48550/arXiv.2102.12219.

[23] Han J W, Zhang D W, Cheng G, et al. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning[J]. *IEEE Trans Geosci Remote Sens*, 2015, **53**(6): 3325−3337.

[24] Long Y, Gong Y P, Xiao Z F, et al. Accurate object localization in remote sensing images based on convolutional neural networks[J]. *IEEE Trans Geosci Remote Sens*, 2017, **55**(5): 2486−2498.

[25] Hu F, Xia G S, Hu J W, et al. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery[J]. *Remote Sens*, 2015, **7**(11): 14680−14707.

[26] Ševo I, Avramović A. Convolutional neural network based automatic object detection on aerial images[J]. *IEEE Geosci Remote Sens Lett*, 2016, **13**(5): 740−744.

[27] Cheng G, Zhou P C, Han J W. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. *IEEE Trans Geosci Remote Sens*, 2016, **54**(12): 7405−7415.

[28] Zhao C M, Chen Z B, Zhang J L. Application of aircraft target tracking based on deep learning[J]. *Opto-Electron Eng*, 2019, **46**(9): 180261.
赵春梅, 陈忠碧, 张建林. 基于深度学习的飞机目标跟踪应用研究[J]. 光电工程, 2019, **46**(9): 180261.

[29] Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database[C]//*Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 248–255.

[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 Trans Pattern Anal Mach Intell*, 2021, **43**(4): 1452−1459.

[31] Yang X, Yang J R, Yan J C, et al. SCRDet: towards more robust detection for small, cluttered and rotated objects[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 8231–8240.

[32] Azimi S M, Vig E, Bahmanyar R, et al. Towards multi-class object detection in unconstrained remote sensing imagery[C]//*Proceedings of the 14th Asian Conference on Computer Vision*, 2018: 150–165.

[33] He Y H, Xu D Z, Wu L F, et al. LFFD: a light and fast face detector for edge devices[Z]. arXiv: 1904.10633, 2019. https://doi.org/10.48550/arXiv.1904.10633.

[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[Z]. arXiv: 1409.1556, 2014. https://doi.org/10.48550/arXiv.1409.1556.

[35] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[36] Zhu C C, He Y H, Savvides M. Feature selective anchor-free module for single-shot object detection[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 840–849.

[37] Woo S, Park J, Lee J Y, et al. Cbam: convolutional block attention module[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 3–19.

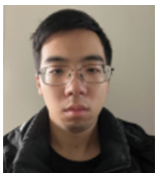
作者简介

马梁 (1997-)，男，硕士，主要从事基于深度学习的目标检测的研究。

E-mail: ml3318276387@163.com

【通信作者】雷涛 (1981-)，男，博士，研究员，主要从事基于传统方法及深度学习技术的图像处理与分析、复杂场景下目标检测识别与跟踪等方面的研究。

E-mail: taoleiyan@ioe.ac.cn



苟于涛 (1997-)，男，硕士，主要从事基于深度学习的目标检测和多模图像融合识别的研究。

E-mail: gouyutao19@163.com

Small object detection based on multi-scale feature fusion using remote sensing images

Ma Liang[1,2,3], Gou Yutao[1,2,3], Lei Tao[1,2*], Jin Lei[1,2], Song Yixuan[1,2,3]

The detection results of the proposed method

Overview: In recent years, with the continuous development of remote sensing optical technology, the acquisition of a large number of high-resolution remote sensing images has promoted the construction of environmental monitoring, animal protection, national defense and military. In numerous remote sensing image visual tasks, remote sensing aircraft detection is of great significance for civil and national defense. Research of the remote sensing small object detection technology is important. Currently, the object detection method based on deep learning has achieved excellent results in large and medium object testing tasks, but the performance and application of remote sensing small object detection are poor. The main reasons are the following: 1) the model is huge, and the real-time is poor; 2) remote sensing image is complicated and the object scale distribution is wide; 3) remote sensing small object detection dataset is extremely lacking.

To solve the above problems, this paper proposes a robust small object detection method based on multi-scale feature fusion using remote sensing images. The main work as follows. First, as the image will be sampled and convolved for many times after being input into common neural networks (such as ResNet and VGG-16), the features of small objects will be seriously lost and the final detection accuracy will be affected. To this end, according to the distribution of all object sizes in the dataset (i.e., prior knowledge), we propose a lightweight feature extraction module based on dynamic selection mechanism, which allows each neuron to adaptively allocate the receptive field size for detection and control the sampling times based on different scale of the objects. Second, although FPN is widely used to solve the problem of small object undetected, the information reflected by various scale features usually has different amounts and emphasis. Therefore, the FPN module based on adaptive feature weighted fusion is proposed, which uses the method of grouping convolution to group all feature channels without affecting each other, so as to further improve the accuracy of image feature expression. Third, for the issue of lack of remote sensing small object dataset, this paper built a remote sensing small object dataset of plane, and processed the plane and small-vehicle objects in DOTA-1.5 dataset to make its distribution of size meet the requirement of small object detection. Finally, experimental results on DOTA and self-built datasets show that our method posseses the best results compared with mainstream detection methods.

Ma L, Gou Y T, Lei T, et al. Small object detection based on multi-scale feature fusion using remote sensing images[J]. *Opto-Electron Eng*, 2022, **49**(4): 210363; DOI: 10.12086/oee.2022.210363

[1]Photoelectric Detection Technology Laboratory, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China; [2]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China; [3]University of Chinese Academy of Sciences, Beijing 100049, China

* E-mail: taoleiyan@ioe.ac.cn