基于遥感图像的多模态小目标检测

胡 俊， 顾晶晶， 王秋红

(南京航空航天大学计算机科学与技术学院，江苏 南京 210016)

摘 要：由于遥感图像目标往往较小且容易受光线、天气等因素的影响，所以单一模态下基于深度学习的遥感图像目标检测的准确度较低。然而，不同模态间的图像信息可以相互增强提高目标检测的性能。因此，基于 RGB 和红外图像，提出了一种适用于遥感图像多模态小目标检测的平衡多模态深度模型。相比简单地相加、点乘和拼接的方式融合 2 个模态的特征信息，设计了一种平衡多模态特征的方法增强目标特征，以弥补单一模态信息不足的缺点。首先分别对 RGB 和红外图像进行浅层特征提取；其次，融合 2 个模态的特征信息并进行深层的特征提取；然后，基于 YOLOv4 方法，构建了多模态小目标检测模型。最后，基于 VEDAI 数据集，在遥感图像多模态小目标检测实验结果中验证了该方法的有效性。

关 键 词：遥感图像；平衡多模态深度模型；小目标检测；融合；VEDAI 数据集

中图分类号：TP 753 **DOI**：10.11996/JG.j.2095-302X.2022020197

文献标识码：A **文 章 编 号**：2095-302X(2022)02-0197-08

Multimodal small target detection based on remote sensing image

HU Jun, GU Jing-jing, WANG Qiu-hong

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu 210016, China)

Abstract: Since targets in remote sensing images are relatively small and easily affected by illumination, weather, and other factors, deep-learning based target detection methods from single modality remote sensing images suffer from low accuracy. However, the image information between different modalities can enhance each other to improve the performance of target detection. Therefore, based on RGB and infrared images fusion, we proposed a balanced multimodal depth model (BMDM) for multimodal small target detection from remote sensing images. As opposed to simple element-wise summation, element-wise multiplication, and concatenation to fuse the feature information of the two modalities, we designed a balanced multimodal feature method to enhance target features to make up for the shortcomings of single modal information. We first extracted low-level features from RGB and infrared images, respectively. Secondly, we fused the feature information of the two modalities and extracted deep-level features. Thirdly, we constructed a multimodal small target detection model based on the one-stage method. Finally, the effectiveness of the proposed method was verified by the experimental results of multimodal small target detection performed on the public dataset VEDAI of remote sensing images.

Keywords: remote sensing images; balanced multimodal deep model; small target detection; fusion; VEDAI dataset

收稿日期：2021-08-26；**定稿日期**：2021-11-26

Received：26 August，2021；**Finalized**：26 November，2021

基金项目：国家自然科学基金项目(62072235)

Foundation items：National Natural Science Foundation of China (62072235)

第一作者：胡 俊(1994–)，男，硕士研究生。主要研究方向为数字图像处理与数据挖掘。E-mail：hujunyn@163.com

First author：HU Jun (1994–), master student. His main research interests cover digital image processing and data mining. E-mail：hujunyn@163.com

通信作者：顾晶晶(1986–)，女，教授，博士。主要研究方向为网络数据挖掘、智能系统等。E-mail：gujingjing@nuaa.edu.cn

Corresponding author：GU Jing-jing (1986–), professor, Ph.D. Her main research interests cover data mining, intelligent system, etc. E-mail：gujingjing@nuaa.edu.cn

遥感图像在实时、动态、宏观等特点的基础上，为军事侦察、地质灾害调查与救治等方面提供了一种新的探测手段。近些年，随着卫星遥感技术和深度卷积神经网络(deep convolutional neural network，DCNN)技术的快速发展，遥感图像目标检测在军事、情报、商业、经济、规划等领域有着重要的应用。

早期基于传统机器学习方法的检测工作[1-3]，其检测性能十分有限。随着发展，DCNN 在目标检测任务中占据着主导地位，根据网络阶段可分为 one-stage (YOLO[4]，SSD[5])和 two-stage (R-CNN[6]，Fast RCNN[7]，Faster RCNN[8])。2 种方法在一般的目标检测任务中分别具有速度和精度上的优势，但面向由卫星、遥感器、无人机等设备采集的遥感数据时，由于遥感数据目标较小，这些算法容易受到光线、天气等因素影响，性能往往达不到预期，因此和一般的目标检测任务存在一定差异性，给检测任务带来了许多挑战。虽然已经开展了很多的遥感图像目标检测工作[9-17]，但大部分还是仅针对 RGB 图像的目标检测。现今，用于航空目标检测的多模态数据的可用性显著增加，如高光谱、合成孔径雷达(synthetic aperture radar，SAR)和红外(infra-red，IR)图像，其均有自身的优势，为 RGB 图像提供了补充信息。考虑到 RGB 图像通常无法在较差的亮度条件下捕获信息，利用红外模式捕获更长的热波长，以完成不同天气条件下检测物体，有助于补偿特征信息损失来扩展 RGB 的能力。所以也有一小部分工作是关于 RGB 和 IR 图像相结合的目标检测，但只是通过特征提取[18-20]并简单地运用早期融合方式或相加、拼接[13, 21-23]等中期融合方式进行目标检测。

为了解决以上问题，本文结合了 RGB 和 IR 图像信息，提出一种适用于遥感图像多模态小目标检测的平衡多模态深度模型(balanced multimodal deep model，BMDM)。考虑到简单地相加、点乘和拼接的方式融合 2 个模态的特征信息往往达不到理想效果，本文设计了一种平衡多模态特征的方法增强目标特征，弥补了单一模态信息不足的缺点。

1 平衡多模态深度模型

如图 1 所示，首先输入 RGB 和 IR 图像。在第 1 阶段，分别对输入的 2 种图像提取浅层特征(low-level feature)，使用平衡多模态融合方法(balanced multimodal fusion，BMF)对提取的浅层特征进行融合，再经过深层网络提取深层特征(deep-level feature)。第 2 阶段，对第 1 阶段的特征进行空间金字塔池化(spatial pyramid pooling，SPP)和路径聚合网络(path aggregation network，PAN)处理。第 3 阶段，输出目标信息，并与真实数据计算损失来训练模型。

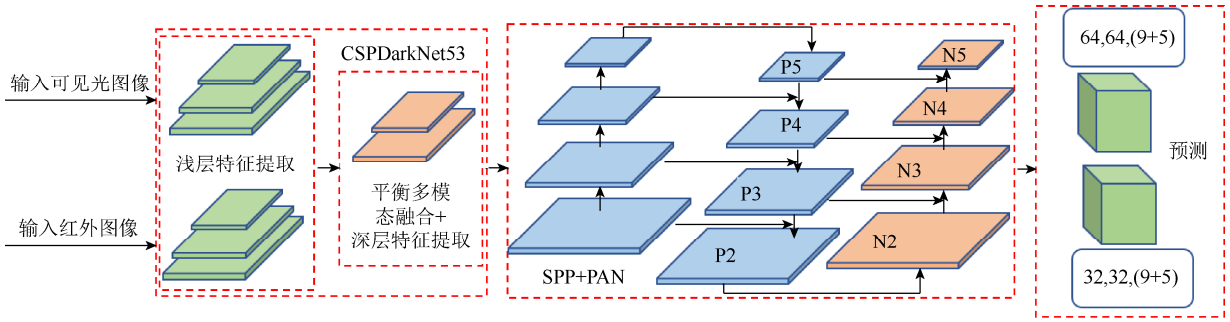


图 1 平衡多模态深度模型框架结构

Fig. 1 The framework of balanced multimodal deep model

1.1 基于 YOLOv4 的遥感小目标网络结构

遥感小目标检测网络结构基于 YOLOv4 实现，整体分为主干、颈部和头部 3 部分。主干用于 RGB 和 IR 图像特征提取和融合。颈部是位于主干和头部的一些网络层，通常用来收集不同阶段的特征图。头部是目标的检测器，包括目标在图像的位置、置信度和目标类别信息。其中主干又包括浅层特征提取层(low-level layer)、中间特征融合层(mid-level fusion layer)和深层特征提取层(deep-level layer)。浅层特征提取层由 1 个卷积层和 1，4，8 个跨阶段局部残差块组成。中间特征融合层由一个全局平均池化层、全连接层和 Softmax 激活函数组成。深层特征提取层由 8 个跨阶段局部残差块组成。主干网络结构如图 2 所示。为了使网络获取更多小目标的特征信息，提高其检测率，BMDM 在 CSPDarknet53 的基础上对网络结构进行改进，将该网络中的第 2 个残差块增加 2 个，同时去除第 5 个残差块。跨阶段局部残差块结构如图 3 所示。

浅层特征提取

	type	filters	size	output
	convolutional	32	3×3	512×512
	convolutional	64	3×3/2	256×256
1×	convolutional	32	1×1	
	convolutional	64	3×3	
	residual			256×256
	convolutional	128	3×3/2	128×128
4×	convolutional	64	1×1	
	convolutional	64	3×3	
	residual			128×128
	convolutional	256	3×3/2	64×64
8×	convolutional	128	1×1	
	convolutional	128	3×3	
	residual			64×64

中间层特征融合

	BMF			32×32

深层特征提取

	convolutional	512	3×3/2	32×32
8×	convolutional	256	1×1	
	convolutional	256	3×3	
	residual			32×32

图 2　主干网络结构

Fig. 2　Structures of backbone network

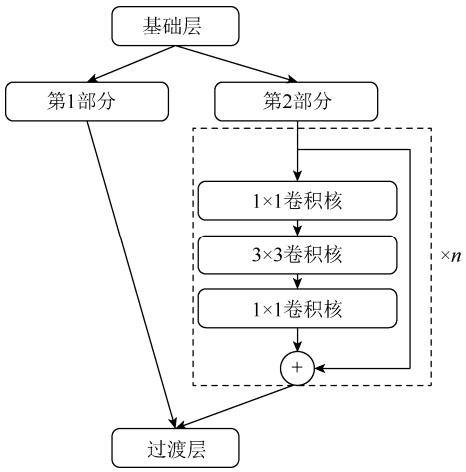


图 3　跨阶段局部残差块结构

Fig. 3　Structures of cross stage partial residual block

(1) 浅层特征提取层。低层特征映射可保留小目标的位置信息，而高层特征映射可包含高层语义线索。网络输出 2 个 3D 张量，分别表示 RGB 和 IR 图像的低层特征。

(2) 中间特征融合层。为了平衡融合 2 个模态的局部特征信息，提出的 BMF 方法。

(3) 高层特征提取层。为了更好的丰富目标特征信息，将 2 个模态的特征信息融合后，再对其特征进行高层特征的提取。

(4) 颈部层。为了使输入的头部信心更丰富，将自底而上的数据流信息进行聚合，以使低层信息能更好地使用。颈部层使用了 SPP 池化层和 PAN 层。

(5) 头部层。输出预测目标的信息，与真实标注数据计算损失来优化网络。输出 2 个 3D 的张量，每个张量均包含检测出目标的位置、置信度和目标类别。

1.2　候选框的选取

候选框是一组宽高固定的初始框，对其选择会直接影响网络对目标的检测精度与速度。K-means 聚类算法具有可解释性强、聚类效果较优、收敛速度快等优势。本文利用 K-means 聚类算法对遥感图像车辆检测数据集中的目标框进行聚类。对数据集进行聚类分析，度量方法采用均值交并比方法 (average intersection over union，avg IOU)，其目标函数为

$$f=\operatorname{argmax}\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_k} I_{IOU}(B,C)}{n} \tag{1}$$

其中，i 为样本号；j 为聚类中心的序号；B 为样本；C 为簇的中心；n_k 为聚类中心样本的第 k 个数；n 为样本的总个数；$I_{IOU}(B,C)$为簇的中心框和聚类框的交并比；聚类数分别选取 m=2,3,⋯,10 进行聚类，得到 m 与均值交并比的关系图(图 4)。随着 m 值的增大，目标函数也在变化，而 m=6 后的曲线逐渐趋于平稳，因此选取候选框的数量为 6，以加快损失函数的收敛，并消除候选框带来的误差。由于本文方法采用 2 个尺度对目标进行检测，所以尺度 1 和 2 分别对应的候选框为(22，10)，(11，22)，(20，19)和(22，40)，(40，17)，(47，43)。

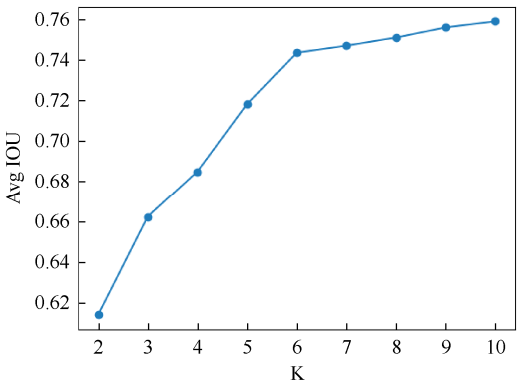


图 4　K-means 聚类分析结果图

Fig. 4　K-means clustering analysis result

1.3　平衡多模态融合

假设 $X=\{X_1,X_2,\cdots,X_n\}$ 是 n 个训练集浅层特征，其中 X_i 包含了 2 个模态的特征信息，$x_i\in\Re^d$ 为第 i 幅 RGB 图像的浅层特征，$y_i\in\Re^d$ 为第 i 幅 IR 图像的浅层特征。一般来说，将其简单拼接、相加或点乘在一起均不可靠，因为这 2 种模态之间存在某种

关系，且相互依存。为了更好地利用 2 种模态的特征信息，本文在 2 种模态特征信息中找到一个平衡点，在 RGB 图像局部特征较好时更多地使用其信息，相反地，在 IR 图像局部特征较好时更多地使用其信息。因此，本文找出一个平衡矩阵 $\boldsymbol{a}\in\Re^d$ 来平衡 2 种模态，即

$$z_i=\boldsymbol{a}x_i+\boldsymbol{a}y_i+\boldsymbol{\beta} \tag{2}$$

$$\boldsymbol{a}=\delta(\omega_i*GAP(x_i+y_i)+b_i) \tag{3}$$

其中，z_i 为融合 2 种浅层特征后的特征信息；$\boldsymbol{\beta}\in\Re^d$ 为一个正则项矩阵；δ为 Softmax 激活函数；ω_i 和 b_i 为全连接层的参数；*为元素相乘；$GAP(\cdot)$为全局平均池化。正则项矩阵$\boldsymbol{\beta}$的加入是为了对参数进行约束，防止出现过拟合情况，从而使得模型的泛化能力更强。矩阵初始值赋为 0，最终由网络训练出平衡矩阵和正则项矩阵。

1.4 数据增强

鉴于遥感图像数据受环境和拍摄器等因素影响，使得图像不可避免的存在各种噪声，尤其是遥感图像中相对一般的检测目标要小得多，受这些噪声的影响更是严重。为了更好地训练模型，采用数据增强来减小噪声并且使得数据集“更强”，从而使目标检测器获得更好的精度且不增加推理成本，同时使模型在不同环境获得的图像具有更高的鲁棒性。数据增强包括：

(1) 几何变换，包括旋转、翻转、裁剪。

(2) 图像增强，包括高斯噪声、模糊处理、擦除、填充和颜色扰动。

(3) 混合削减，将图像的部分区域剪掉并填充上训练集中其他数据的区域值像素值。

(4) 马赛克数据增强，将 4 张训练图像组合成 1 张图像用于训练，使模型能够训练到小的目标。如图 5 所示，图 5(a)为原图，经数据增强后得到图 5(b)增强图。

1.5 网络输出

本文方法将整个图像分割成一个网格，且根据 2.2 节选择 2 个尺度，每个尺度有 3 个候选框。网络整体输出是 2 个 3D 张量(64×64×42，32×32×42)，张量包含了位置、置信度和类别信息，2 个张量分别对应 32×32 和 16×16 的网格。

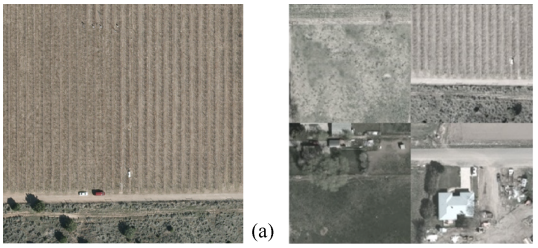
对于第 i 个网格，预测输出 3×14 维的向量对应 3 个候选框，每个候选框的输出为

$$\boldsymbol{V}_i=[x^i,y^i,w^i,h^i,conf^i,c^1,c^2,\cdots,c^9] \tag{4}$$

其中，(x^i,y^i)为预测框的中心点，相对于该网格的左上角；$conf^i$ 为该预测框里目标的置信度，取值为 0 到 1，越高说明置信度越大；c^j, j=1,2,⋯,9 为预测框里目标的类别，类别对应于汽车、卡车、拖拉机、露营车、客货车、其他车、皮卡、船、飞机。同时，输出的评估方式采用平均正确率均值，即

$$mAP=\frac{\sum_{i=1}^{k}AP_i}{k} \tag{5}$$

$$AP=\sum_{i=1}^{n-1}(r_{i+1}-r_i)p_{\text{inter}}(r_i+1) \tag{6}$$



(a) (b)

图 5 原图和数据增强结果图((a)原图；(b)增强图)

Fig. 5 Figure of original and data enhancement results ((a) The original image; (b) Strengthen figure)

1.6 损失函数

方法的损失函数类似于 YOLOv4，分为类别、置信度和位置损失，损失函数为

$$L=L_{\text{class}}+L_{\text{conf}}+L_{\text{pos}} \tag{7}$$

(1) 类别损失。预测框中存在目标时才进行类别损失计算，根据候选框与实际框的交并比判断是否存在目标。本文采用 2 个特征尺度，分别对应于 8 倍和 16 倍下采样。类别损失计算采用交叉熵损失计算，对每个类别计算交叉熵损失并进行求和运算，类别损失函数为

$$L_{\text{class}}=\lambda_{\text{class}}\sum_{i=0}^{S^2}l_i^{\text{obj}}\sum_{c=1}^{9}[-p_i(c)\log_2(\tilde{p}_i(c))] \tag{8}$$

其中，λ_{class}为一个惩罚项；s^2 为图像划分的网格数；l_i^{obj} 为在 i 处框有目标，其值为 1，否则为 0。

(2) 置信度损失。区分有目标和无目标的置信度损失，采用交叉熵损失，置信度损失函数为

$$L_{\text{conf}}=\lambda_{\text{obj}}\sum_{i=0}^{S^2}\sum_{j=1}^{3}l_{i,j}^{\text{obj}}[-C_i^{\eta}\log_2(\tilde{C}_i)]+\lambda_{\text{noobj}}\sum_{i=0}^{S^2}\sum_{j=1}^{3}l_{i,j}^{\text{noobj}}[-C_i^{\eta}\log_2(\tilde{C}_i)] \tag{9}$$

其中，λ_{obj}和λ_{noobj}为惩罚项；$l_{i,j}^{\text{obj}}$与$l_{i,j}^{\text{noobj}}$相反。η为 focal 损失参数。

(3) 位置损失。采用完全交并比损失(complete intersection over union，CIOU)[24]，即

$$L_{\text{pos}} = 1 - IOU + \frac{\rho^2(p,g)}{c^2} + \alpha v + l_\theta \tag{10}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^g}{h^g} - \arctan\frac{w^p}{h^p}\right)^2 \tag{11}$$

$$\alpha = \frac{v}{(1-IOU)+v} \tag{12}$$

$$IOU = \frac{P \cap G}{P \cup G} \tag{13}$$

$$l_\theta = (\theta^g - \theta)^2 \tag{14}$$

其中，P，G 分别为预测框和真实框；p，g 分别为 2 个框的中心点；c 为 2 个框的最小包围矩形框的对角线长度；ρ为预测框和真实框的中心端距离；θ^g 和θ分别为真实角度值和预测角度值。

2　实　验

将本文方法在遥感图像车辆检测库(vehicle detection in aerial imagery，VEDAI)中进行测试。实验条件：操作系统为 Red Hat 4.8.5，深度学习框架为 PyTorch，CPU 为 i7-5930 K，内存为 64 G，GPU 为 GeForce RTX 3090，GPU 内存为 16 G。

2.1　实验数据集和设置

VEDAI (https://downloads.greyc.fr/vedai/)是一个用于遥感图像中车辆检测的数据集，是在无约束环境中对自动目标识别算法进行基准测试的工具。数据库中包含的车辆除了体积小之外，还表现出不同的问题，如多个方向、照明阴影变化、镜面反射或遮挡。此外，每幅图像均有多个光谱波段和分辨率。该数据集包含 1 272 张 1024×1024 的遥感图像，对应于 1 272 张 512×512 遥感图像，空间分辨率为 12.5 cm。所有图像均是从同一高度拍摄的，每幅图像有 2 种模态：RGB 和 IR 图像。数据集划分为汽车、卡车、拖拉机、露营车、客货车、其他车、皮卡、船和飞机 9 个类别。本文使用的是 512×512 的图像，所有类别均包括在内，见表 1。

训练集和测试集的划分比例为 9∶1，即 1 146 张图像用于训练集，126 张图像用于测试集。模型总共训练 300 次迭代，每次训练的最小批量为 4 张图像，梯度累计间隔为 4 次最小批量迭代。使用 Adam 优化器进行训练，初始学习率为 0.001，权重衰减系数为 0.000 5，动量参数为 0.93。在训练模型中，每一次训练迭代均计算了测试集的平均正确率均值。

表 1　VEDAI 数据集的类别数量

Table 1　Number of categories of VEDAI data

类别	实例
汽车	1 371
卡车	307
拖拉机	190
露营车	396
客货车	101
其他车	204
皮卡	951
船	170
飞机	47

2.2　实验结果

在该数据集中，正负样本十分不均衡，负样本数量太大，占总损失的大部分，且多是容易分类的，因此使得模型的优化达不到理想，导致检测准确度不理想。因此在损失函数中采用了 focal 损失函数，该函数可以通过减少易分类样本的权重，使得模型在训练时更专注于难分类的样本来提升检测的精确度。在实验中，本文分别设置 focal 损失参数为 η=0，1，2，3，由此选取更适合的 focal 损失参数。实验结果如图 6 所示，同时本文还对比了 YOLOrs 模型不同 focal 损失参数的训练结果。根据实验结果可知 η=1 和 η=2 时，测试集的平均正确率均值相差不大，但是在接近 300 次迭代时 η=2 的平均正确率均值有下降趋势，因此本文选取了 η=1 作为 focal 的参数。

在对比试验中，为了不让模型受 focal 损失参数的影响，本文统一设置 η=1。本文对比了 YOLOrs (4 通道)[13]、改进的 YOLOv4 (RGB)、改进的 YOLOv4 (IR)、改进的 YOLOv4 (点乘融合)、SSD[5] (4 通道)、RetinaNet (4 通道)(注：RGB 是 3 通道图像，IR 是 1 通道图像，4 通道表示将 2 个模态的图像合并为 4 通道作为模型的输入；RGB 表示只输入 RGB 图像；IR 表示只输入 IR 图像；点乘融合表示使用 BMF 网络，融合方式采用点乘方式)。YOLOrs 是专为多模态遥感图像实时目标检测而设计的，采用中期拼接方式融合。SSD 为经典的一阶段目标检测算法，使用 focal 损失函数。RetinaNet 使用 ResNet+FPN 作为主干，是一种使用 focal 损失参数的一阶段目标检测模型。实验结果见表 2。实验结果表明融合方式十分重要，这是因为 2 种模态图像均包含着丰富的特征信息，模型能够利用不同模态中有用的信息作为补充信息。表 3 给出了本文模型对 9 个类别的准确率和召回率。

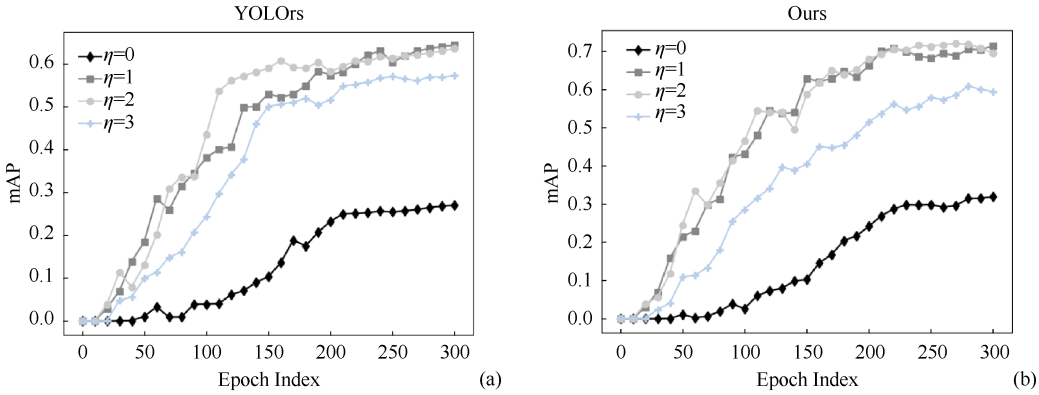


图 6 Focal 参数对比结果图((a) YOLOrs 的不同 focal 参数结果图；(b)本文方法的不同 focal 参数结果图)

Fig. 6 Focal parameter comparison results ((a) Results of different focal parameters of YOLOrs method; (b) Results of different focal parameters of ours)

表 2 测试集上精度对比

Table 2 Comparison of precision on test set

类别	BMDM	YOLOrs (4 channels)	Improved YOLOv4 (RGB)	Improved YOLOv4 (IR)	Improved YOLOv4 (multiplication)	SSD (4 channels)	RetinaNet (4 channels)
汽车	0.823	**0.841 5**	0.659	0.474 0	0.599	0.376	0.726 0
卡车	**0.760**	0.526 0	0.452	0.477 0	0.593	0.418	0.583 3
拖拉机	0.519	**0.704 8**	0.496	0.180 0	0.493	0.370	0.627 0
露营车	**0.774**	0.688 1	0.557	0.528 0	0.679	0.443	0.631 6
客货车	**0.681**	0.579 1	0.065	0.065 0	0.065	0.155	0.486 1
其他车	**0.655**	**0.457 5**	0.434	0.062 5	0.356	0.433	0.272 6
皮卡	0.763	**0.782 7**	0.579	0.541 0	0.644	0.356	0.615 6
船	**0.465**	0.214 7	0.110	0.070 0	0.173	0.083	0.185 4
飞机	0.995	0.995 0	0.995	0.995 0	0.995	0.975	0.987 0
平均	**0.715**	0.643 0	0.483	0.377 0	0.511	0.401	0.568 3

注：加粗数据为最优值

表 3 平衡多模态方法的准确率和召回率(%)

Table 3 The precision and recall of balanced multimodal methods (%)

类别	准确率	召回率
汽车	0.765	0.860
卡车	0.556	0.789
拖拉机	0.429	0.545
露营车	0.515	0.897
客货车	0.750	0.750
其他车	0.727	0.727
皮卡	0.679	0.889
船	0.583	0.500
飞机	0.667	1.000
平均	0.630	0.764

(1) 定量对比。本文方法在平均正确率均值结果中有着不俗的表现，达到 71.5%，排名第 1，相比第 2 名提高了 11.0%。是因为在模态融合时 2 种模态的信息互补，增强了模型对目标的检测。其中卡车、露营车、客货车、其他车、船类别均优于其他模型。

(2) 定性对比。图 7 第 1~3 列分别为本文、YOLOrs 和 Improved YOLOv4(multiplication)方法的检测结果，数字 1,2,…,9 对应 9 个类别，其中所标出的红色框为多检和错检结果。YOLOrs 模型容易出现多检的情况，Improved YOLOv4 (multiplication)模型容易出现多检和错检的情况，而本文模型能够精准地检测出目标。

2.3 消融实验

本文设计了一系列的消融实验以分析平衡多模态方法和其每一部分的优势，并对比了使用该网络的单模态 RGB 实验。

(1) 数据增强。对于遥感图像的目标检测任务是十分重要的。实验中本文对比了使用和不使用数据增强的单模态和多模态方法，从实验结果可知数据增强会大幅度提高检测的精确度。

(2) 平衡多模态。相比单模态 RGB 图像检测，平衡多模态目标检测受环境等因素的影响更低，由此泛化性也更高。表 4 展示了该方法在遥感图像小目标检测的精度上更高。

图 7　可视化检测结果图((a)本文方法；(b) YOLOrs 方法；(c)改进的 YOLOv4 (点乘)方法)

Fig. 7　Visual detection results ((a) Ours; (b) YOLOrs method; (c) Improved YOLOv4 (multiplication) method)

表 4　消融实验的平均正确率均值

Table 4　The mAP of ablation experimental

RGB	数据增强	平衡多模态	正则项	平均正确率均值
√	-	-	-	0.217
√	√	-	-	0.483
-	-	√	-	0.493
-	√	√	-	0.685
-	√	√	√	**0.715**

注：加粗数据为最优值；√为使用本模块；- 为未使用本模块

(3) 正则项矩阵。其为了防止过拟合，从而增强模型的泛化能力。实验结果也表明了正则项矩阵的优势。

3　总结与展望

为提升目标检测中小目标的检测精度，以解决光线弱、能见度低等环境下目标检测效果不理想的问题，本文联合挖掘了 RGB 和 IR 图像 2 种模态数据之间的相关性及可实现互补增强，并基于改进的 CSP-DarkNet53 网络提出了基于 YOLO 的平衡多模态多类检测网络，以实时检测遥感图像中的小目标。该方法不仅对遥感图像小目标更敏感，且通过 BMF 方式利用 2 种模态信息互补增强，进一步提升网络的小目标检测精度和鲁棒性。同时，算法采用数据增强减弱噪声的影响，进一步优化了训练数据集。在公开的 VEDAI 数据集上进行验证，相比其他方法，本文方法在多个类别的 mAP 均处于领先，总体上也实现了最好的性能表现。

综上所述，本文提出的 BMDM 方法通过融合图像的多模态信息而实现对遥感图像中小目标的精确检测，有效提升了小目标的检测性能，并为后续其他融合方法的选取与尝试提供了参考。虽然目前本文的 BMDM 方法在小目标检测精度上有了较明显地提升，但由于网络需要对 2 个模态的图像进行特征提取并融合，致使网络计算速度上受到限制，且融合方式有待进一步挖掘。因此，如何进一步加速计算、改进融合方法、提高精度，是下一个阶段需要探索的目标。

参考文献 (References)

[1] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[2] VIOLA P, JONES M J. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2005: 886-893.

[4] BOCHKOVSKIY A, WANG C Y, LIAO H Y. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. [2021-07-26]. https://xueshu.baidu.com/usercenter/paper/show?paperid=1q0h0p70e95d0ej0sj1202x0em679337.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[M]//Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[7] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[8] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[9] ETTEN A V. You only look twice: rapid multi-scale object detection in satellite imagery[EB/OL]. [2021-07-26]. https://xueshu.baidu.com/usercenter/paper/show?paperid=196ddb2c129916b9f930a718f09e6348&site=xueshu_se.

[10] YANG X, YANG J R, YAN J C, et al. SCRDet: towards more robust detection for small, cluttered and rotated objects[C]//2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 8231-8240.

[11] 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.

[12] LONG H, CHUNG Y, LIU Z B, et al. Object detection in aerial images using feature fusion deep networks[J]. IEEE Access, 2019, 7: 30980-30990.

[13] SHARMA M, DHANARAJ M, KARNAM S, et al. YOLOrs: object detection in multimodal remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 1497-1508.

[14] KOESTER E, SAHIN C S. A comparison of super-resolution and nearest neighbors interpolation applied to object detection on satellite data[EB/OL].[2021-07-26]. https://xueshu.baidu.com/usercenter/paper/show?paperid=1c520t20gw6f0m60e80u0g106f040545&site=xueshu_se&hitarticle=1.

[15] XIA G S, BAI X, DING J, et al. DOTA: a large-scale dataset for object detection in aerial images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 3974-3983.

[16] LU X C, JI J, XING Z Q, et al. Attention and feature fusion SSD for remote sensing object detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-9.

[17] YANG X, LIU Q Q, YAN J C, et al. R3Det: refined single-stage detector with feature refinement for rotating object[EB/OL]. [2021-07-26]. https://xueshu.baidu.com/usercenter/paper/show?paperid=133q0vw0wg7w04w0pq150pm0kn019941&site=xueshu_se.

[18] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651.

[19] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[EB/OL]. [2021-07-26]. https://xueshu.baidu.com/usercenter/paper/show?paperid=e405c047319275f1026702182776bfdc&site=xueshu_se.

[20] IANDOLA F, MOSKEWICZ M, KARAYEV S, et al. DenseNet: implementing efficient ConvNet descriptor Pyramids[EB/OL]. [2021-07-26]. https://xueshu.baidu.com/usercenter/paper/show?paperid=db44736c4000d1544d02905c43dbf413&site=xueshu_se&hitarticle=1.

[21] 邢素霞, 肖洪兵, 陈天华, 等. 基于目标提取与 NSCT 的图像融合技术研究[J]. 光电子 激光, 2013, 24(3): 583-588.
XING S X, XIAO H B, CHEN T H, et al. Study of image fusion technology based on object extraction and NSCT[J]. Journal of Optoelectronics Laser, 2013, 24(3): 583-588 (in Chinese).

[22] 王春华, 马国超, 马苗. 基于目标提取的红外与可见光图像融合算法[J]. 计算机工程, 2010, 36(2): 197-200.
WANG C H, MA G C, MA M. Fusion algorithm for infrared and visible light image based on object extraction[J]. Computer Engineering, 2010, 36(2): 197-200 (in Chinese).

[23] YANG D F, LIU X, HE H, et al. Air-to-ground multimodal object detection algorithm based on feature association learning[J]. International Journal of Advanced Robotic Systems, 2019, 16(3): 1-9.

[24] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.