E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP751　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)07-2570-14

论文引用格式:Xiao Z J, Li S B, Qu H C and Li F K. 2025. Remote sensing image detection guided by context information and multi-scale feature sequences. Journal of Image and Graphics, 30(7):2570-2583(肖振久, 李士博, 曲海成, 李富坤. 2025. 上下文信息和多尺度特征序列引导的遥感图像检测. 中国图象图形学报, 30(7):2570-2583)[DOI:10.11834/jig.240361]

上下文信息和多尺度特征序列引导的遥感图像检测

肖振久[1],李士博[1*],曲海成[1],李富坤[2]

1. 辽宁工程技术大学软件学院,葫芦岛　125105; 2. 河南师范大学计算机与信息工程学院,新乡　453007

摘　要:**目的**　针对遥感图像(remote sensing image,RSI)检测中目标尺寸小且密集、尺度变化大,尤其在复杂背景信息下容易出现漏检和误检问题,提出一种上下文信息和多尺度特征序列引导的遥感图像检测方法,以提升遥感图像的检测精度。**方法**　首先,设计自适应大感受野机制(adaptive large receptive field,ALRF)用于特征提取。该机制通过级联不同扩张率的深度卷积进行分层特征提取,并利用通道和空间注意力对提取的特征进行通道加权和空间融合,使模型能够自适应地调整感受野大小,从而实现遥感图像上下文信息的有效利用。其次,为解决颈部网络特征融合过程中小目标语义信息丢失问题,设计多尺度特征序列融合架构(multi-scale feature fusion,MFF)。该架构通过构建多尺度特征序列,并结合浅层语义特征信息,实现复杂背景下多尺度全局信息的有效融合,从而减轻深层网络中特征模糊性对小目标局部细节捕捉的影响。最后,因传统交并比(intersection over union,IoU)对小目标位置偏差过于敏感,引入归一化Wasserstein距离(normalized Wasserstein distance,NWD)。NWD将边界框建模为二维高斯分布,计算这些分布间的Wasserstein距离来衡量边界框的相似性,从而降低小目标位置偏差敏感性。**结果**　在NWPU VHR-10(Northwestern Polytechnical University very high resolution 10)和DIOR(dataset for object detection in aerial images)数据集上与10种方法进行综合比较,结果表明,提出的方法优于对比方法,平均精度(average precision,AP)分别达到93.15%和80.89%,相较于基准模型YOLOv8n(you only look once version 8 nano),提升了5.48%和2.97%,同时参数量下降6.96%。**结论**　提出一种上下文信息和多尺度特征序列引导的遥感图像检测方法,该方法提升目标的定位能力,改善复杂背景下遥感图像检测中的漏检和误检问题。

关键词:遥感图像(RSI);目标检测;感受野(RF);特征融合;归一化Wasserstein距离(NWD)

Remote sensing image detection guided by context information and multi-scale feature sequences

Xiao Zhenjiu[1], Li Shibo[1*], Qu Haicheng[1], Li Fukun[2]

1. School of Software, Liaoning Technical University, Huludao　125105, China;

2. School of Computer and Information Engineering, Henan Normal University, Xinxiang　453007, China

Abstract: **Objective** Remote sensing images (RSIs) have been widely used in environmental monitoring, resource management, and emergency response due to their wide coverage, rich information content, and high temporal and spatial resolution. However, in the actual process of object detection, the objects in RSIs are typically small in size, densely distrib-

收稿日期:2024-07-15;**修回日期**:2024-12-15;**预印本日期**:2024-12-22

*通信作者:李士博　lishibo@stu.htu.edu.cn

基金项目:辽宁省高等学校基本科研项目(LJKMZ20220699);辽宁工程技术大学学科创新团队资助项目(LNTU20TD-23)

Supported by: Scientific Research Foundation of the Higher Education Institutions of Liaoning Province (LJKMZ20220699); Discipline Innovation Team of Liaoning Technical University (LNTU20TD-23)

uted, and exhibit large-scale variations, which can easily lead to missed and false detection problems, particularly in complex backgrounds. To address these challenges, we propose an RSI detection method guided by contextual information and multi-scale feature sequences. **Method** First, we designed an adaptive large receptive field (ALRF) during the feature extraction stage. The design of a reasonable receptive field (RF) is crucial for RSI processing. A large RF cannot only effectively capture the global structure of large targets, but can also use contextual information to capture the details of small targets, improving detection performance. Large kernel convolution is a common method for expanding RF, but its computational cost is high. As the size of the convolution kernel increases, the number of parameters will also increase rapidly, affecting the inference speed and efficiency of a model. To solve this problem, ALRF adopts multilayer cascade dilated convolution to expand RF by gradually increasing the size and dilation rate of the convolution kernel while maintaining a low number of parameters, achieving an RF and context perception capability comparable with that of large kernel convolution. For the hierarchical features extracted via cascade dilated convolution, ALRF performs channel weighing through channel attention and then spatial fusion through spatial attention, adaptively adjusting RF size in accordance with the features of different targets, effectively capturing contextual information in RSIs, and realizing the extraction of multi-scale features. Secondly, we proposed a multi-scale feature fusion (MFF) architecture during the feature fusion stage. MFF is primarily composed of a focal scale fusion engine (FSFE) and a fine-scale feature encoder (FFE), and it fuses the multi-scale feature sequences extracted from the backbone network through a bidirectional feature pyramid network structure. FSFE constructs a multi-scale feature sequence by using Gaussian convolution kernels with increasing standard deviations. By contrast, FFE introduces detailed information from the shallow feature layer of P2 and uses hybrid coding to enhance semantic expression. MFF achieves effective fusion of multi-scale global information under complex background information through the collaborative work of FSFE and FFE, reducing the effect of feature ambiguity in deep networks on capturing small target details. Finally, we introduce the normalized Wasserstein distance (NWD) to replace the traditional intersection over union (IoU). NWD measures the similarity of bounding boxes by modeling them as 2D Gaussian distributions and calculating the Wasserstein distance between these distributions. As a robust similarity metric, NWD can accurately evaluate the alignment between the bounding box of a small object and the ground truth box, reducing sensitivity to positional deviations. **Result** Our method is experimentally verified on two datasets, namely, the Northwestern Polytechnical University very-high resolution 10 (NWPU VHR-10) dataset and the dataset for object detection in aerial images (DIOR), and compared with 10 detection methods from the two-stage, one-stage, and DETR series. Experimental results show that the method exhibits significant advantages in small target detection and complex backgrounds. On the NWPU VHR-10 dataset, the average precision (AP) of the method is improved by 5.48% to 93.15% compared with the baseline model, You Only Look Once v8 Nano (YOLOv8n), while the number of parameters is reduced by 6.96%. In DIOR, the AP of the method is 80.89%, which is 2.97% higher than that of YOLOv8n. The effectiveness of ALRF, MFF, and NWD is verified through ablation experiments. ALRF achieves efficient use of contextual information through cascaded dilated convolutions. MFF enhances the ability to capture small target details by constructing multi-scale feature sequences and introducing shallow features. NWD improves the balance of positive and negative sample distribution by modeling the bounding box as a 2D Gaussian distribution. The simultaneous use of ALRF, MFF, and NWD can achieve the best effect on the model, with precision, recall, and AP increased by 2.76%, 4.61%, and 5.48%, respectively, compared with the baseline model. In addition, the precision–recall curve of this method is closer to the upper right and more stable, indicating that it is stable in the detection of different categories of targets. Simultaneously, this method is effective in reducing false and missed detections and exhibits stronger generalization and robustness compared with other methods. **Conclusion** An RSI detection method guided by context information and multi-scale feature sequences is proposed to solve the problems of dense small targets, large target size variations, and complex background in RSI detection. This method first uses ALRF to effectively utilize context information. Then, MFF is used to reduce the semantic information loss of small targets. Finally, NWD is introduced to replace the traditional IoU to optimize the distribution of positive and negative samples. The experimental results show that the AP of this method on the NWPU VHR-10 dataset and DIOR reaches 93.15% and 80.89%, respectively. Meanwhile, the model parameters are only 2.94 M, which is significantly better than that of the comparison method. In our future work, the proposed method will be extended to target detection under multi-view imaging

conditions, and a lightweight model suitable for real-time detection applications will be developed.

Key words: remote sensing image (RSI); target detection; receptive field (RF); feature fusion; normalized Wasserstein distance (NWD)

0 引 言

基于深度学习的目标检测方法在遥感领域的应用日益广泛,无论在民用还是军事领域,都显现出巨大的应用潜力。这项任务旨在预测图像中目标的边界框和类别。目前主流的检测算法主要分为双阶段检测和单阶段检测两大类。双阶段检测算法通过初步生成大量候选区域,再通过分类和回归步骤对这些区域进行定位和识别。代表方法有 Fast R-CNN (fast region-based convolutional neural network) (Girshick, 2015)、Faster R-CNN (faster region-based convolutional neural network) (Ren 等, 2017) 和 Dynamic R-CNN (dynamic region-based convolutional network) (Zhang 等, 2020),这些算法以高检测精度著称,但处理速度较慢。相比之下,单阶段检测算法从主干网络直接提取特征,并据此预测目标的类别和位置。代表方法有 SSD (single shot multibox detector) (Liu 等, 2016) 和 YOLO (you only look once) 系列算法 (Redmon 等, 2016),它们提升了检测速度,实现了效率与精度的平衡。

然而,由于遥感图像 (remote sensing image, RSI) 独特的成像特性,直接将通用目标检测器应用于遥感图像时,容易出现漏检和误检问题。遥感图像中包含大量小尺寸目标,小目标仅占据少量像素,导致其特征信息不足,为特征提取和表达带来挑战。为应对这一挑战,学者们不断探索和改进遥感图像检测算法。Chen 等人 (2021) 通过引入目标增强和衰减非最大值抑制机制,来提升遥感图像中小目标的检测性能。Ma 等人 (2023) 进一步发展这一领域,提出多级加权深度感知网络 (multi-level weighted depth perception network, MwdpNet)。该网络采用组残差结构和多级特征加权融合策略,并结合深度感知与通道注意引导模块,以优化遥感图像中小目标特征捕捉能力。Zhang 等人 (2024a) 通过生成高分辨率特征图并引入背景退化策略,改善了小目标的细节捕捉。而李朝辉等人 (2023) 通过数据增强和改进正样本采样策略,增加模型训练中正样本的数量,使模型能够充分地学习目标特征,提升小目标的检测精度。

除小目标数量众多问题外,遥感图像中目标多尺度特性和复杂背景信息同样构成挑战。在遥感场景中,不同类别的目标以及同一类别内不同大小的目标共存,导致目标尺度差异显著。尺度差异使小尺寸目标特征难以在深层网络中有效传递,进而影响多尺度目标检测的准确性。为应对这一挑战,Shamsolmoali 等人 (2022) 提出旋转等变特征图像金字塔网络 (rotation equivariant feature image pyramid network, REFIPN),通过特征图金字塔在不同尺度和角度上提取特征,缓解因目标旋转和尺度差异带来的漏检与误检问题。随后,佘浩东和赵良瑾 (2024) 利用椭圆旋转高斯核,通过形状和尺寸的适应性调整,实现与目标形状的紧密贴合,从而进一步提取多尺度目标特征。在此基础上,Liu 等人 (2024) 设计自适应多尺度特征金字塔网络 (adaptive multi-scale feature pyramid metwork, AMFF),通过集成自适应多尺度特征增强与融合模块,动态调整特征图的层次与分辨率,以更好地适应不同尺度目标的检测需求。而为减轻复杂背景信息的干扰,Zhang 等人 (2024b) 提出语义驱动特征增强模块 (semantic-driven feature enhancement module, SFEM)。SFEM 利用语义分割将遥感图像中的目标对象和背景区域在特征图中分离,并在空间域内分别处理这些区域,从而实现目标对象的特征增强和背景特征的抑制。尽管现有方法在遥感图像检测任务中取得了较高的准确率,但在面对小目标密集分布、尺度差异显著以及背景信息复杂情况时,仍存在漏检和误检问题,需要进一步优化和改进。

在遥感图像检测中,感受野 (receptive field, RF) 合理的设计至关重要。较大的感受野不仅能有效捕捉大尺寸目标的全局结构,还能利用上下文信息为小目标提供细节特征,从而提升检测性能。例如,在图 1 中,桥梁的部分结构与高速公路相似,模型必须观察到目标的完整边界和形状,否则会仅检测到目标的一部分。此外,汽车和船只这样的小型目标,由于体积小、特征不明显,仅凭自身的像素信息难以准确识别,此时,模型需要结合目标周围的上下文信

息，如车辆所在的道路、船只所在的水域等。而为了捕获这些信息，感受野必须足够大，以涵盖相关的上下文环境。基于以上分析，本文以 YOLOv8n（Jocher 等，2023）为基准，提出一种上下文信息和多尺度特征序列引导的遥感图像检测方法。首先，为了充分利用上下文信息并扩展目标感知范围，设计自适应大感受野机制（adaptive large receptive field，ALRF）。该机制通过级联不同扩张率的深度卷积以提取多尺度特征，并在通道维度上对这些特征进行加权处理后，在空间维度上进行特征选择与融合。这些权重通过训练和反向传播优化，使模型能自适应调整不同目标在空间中的感受野大小，从而为准确定位目标提供丰富的上下文信息。其次，针对原始网络中小目标语义信息丢失问题，设计多尺度特征序列融合架构（multi-scale feature fusion，MFF）替换特征金字塔网络（feature pyramid network，FPN）（Lin 等，2017）和路径聚合网络（path aggregation network，PAN）（Liu 等，2018）。MFF 通过不同偏差的卷积核构建多尺度特征序列，并引入 P2 层的浅层特征以保留细节信息，从而减轻深层网络中的特征模糊性，增强对小目标细节捕捉能力的表达。最后，因传统 IoU（intersection over union）对位置偏差过于敏感，引入归一化 Wasserstein 距离（normalized Wasserstein distance，NWD）（Xu 等，2022）。NWD 将边界框建模为二维高斯分布，并通过计算这些分布间的 Wasserstein 距离，再将结果通过指数函数归一化为适用于小目标检测的度量。NWD 解决了 IoU 在标签分配中的不准确和正负样本不平衡问题。

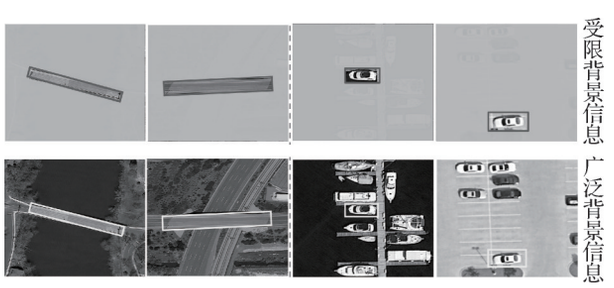


图 1　桥和高速公路、船和汽车的不同上下文信息对比

Fig. 1　Comparison of different contextual information between bridges and roads，ships and cars

1　模型构建

本文方法的整体网络结构如图 2 所示，网络由 3 部分组成：主干网络（backbone）、颈部网络（neck）和检测头（head）。主干网络用于特征提取，分为第 1 至第 5 阶段（P1～P5），分别用于提取不同层次的图像特征，在后 4 个阶段引入 ALRF 机制，并将其置于 CBS（convolution-batchnorm-silu）与 C2f（CSP bottleneck with 2 convolutions）模块之间。颈部网络采用特征融合技术，将主干网络在不同阶段提取的特征进行有效融合。用 MFF 替换 FPN 与 PAN，MFF 的核心部件为序列融合引擎（focal scale fusion engine，FSFE）和精细尺度特征编码器（fine-scale feature encoder，FFE）。FSFE 融合多尺度目标的全局高层语义信息，而 FFE 则整合多尺度的浅层局部细节信息。MFF 通过构建多尺度特征序列并引入浅层 P2 特征，克服了 FPN 和 PAN 在特征融合时的不充分及小目标语义信息丢失问题。MFF 有效缓解深层网络中的特征模糊性，增强对小目标细节的捕捉能力，同时减少模型参数量。最后，检测头承担最终的分类和回归任务，在 YOLOv8 中，损失计算采用 TAL（task alignment learning）（Feng 等，2021）正负样本分配策略。本文通过引入 NWD 与 TAL 结合，解决了传统 IoU 对位置偏差过于敏感所导致的标签分配不准确和正负样本不平衡的问题。

1.1　ALRF

在遥感图像检测中，感受野的大小直接影响模型对不同目标的捕捉能力。大核卷积是一种常见的方法，它通过较大的卷积核一次性覆盖更广的区域，从而提升模型对上下文信息的感知能力。大核卷积在检测较大目标或依赖远距离上下文信息的任务中尤其有效。然而，大核卷积的计算代价高，随着卷积核尺寸的增加，参数量会急剧增长，从而影响模型的推理速度和效率。为此，本文采用不同扩张率的深度卷积进行多层级联，以实现与大核卷积相当的感受野和上下文感知能力，同时降低模型参数。受 SKNet（selective kernel network）（Li 等，2019）和 LSKNet（large selective kernel network）（Li 等，2023）启发，设计了 ALRF 机制，该机制由级联扩张卷积、通道注意力（channel attention，CA）和空间注意力（spatial attention，SA）组成，网络结构如图 3 所示。ALRF 能够自适应地调整空间中不同目标的感受野大小，从而有效捕捉和利用上下文信息。ALRF 融合 SKNet 和 LSKNet 的优势，SKNet 在通道维度上自适应聚焦大核卷积信息，而 LSKNet 通过对大核卷积

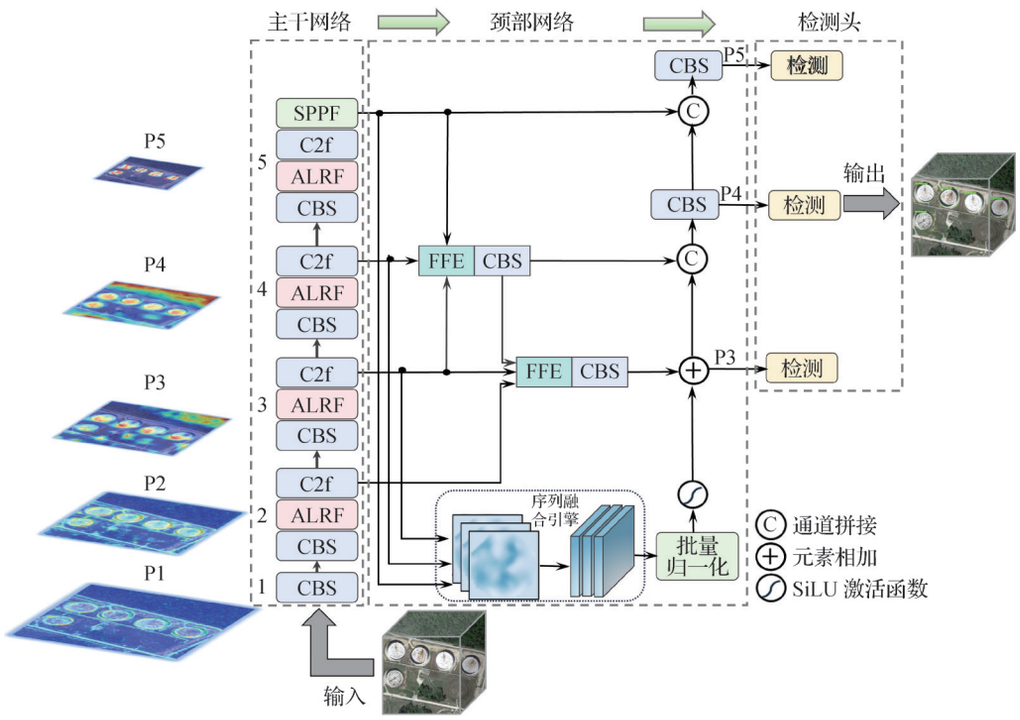


图2　本文方法整体网络结构

Fig. 2　Diagram of the overall network structure of the proposed method

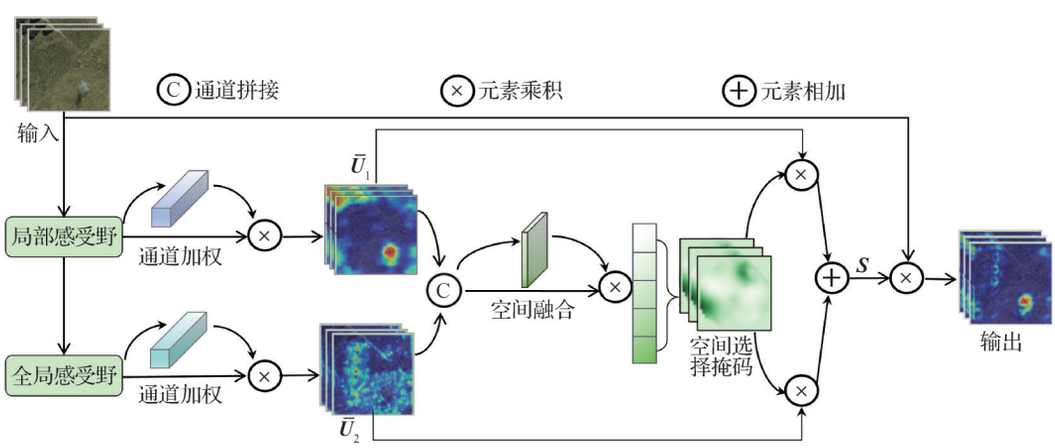


图3　ALRF机制结构

Fig. 3　Structure of the ALRF mechanism

解耦，在空间维度上进行融合。本文方法进一步发展，采用多层级联扩张卷积来提取不同感受野范围的特征，并在通道维度上对这些特征加权处理后，在空间维度上融合，实现对上下文信息的有效利用，同时保持模型的计算效率。

1.1.1　级联扩张卷积

根据先验知识，小感受野通常用于提取小目标特征。然而，为了有效利用遥感图像中的上下文信息，本文将含有扩张率的深度卷积进行级联组合，并逐步增大卷积核的大小和扩张率，实现与大核卷积相近的感受野与上下文感知能力，具体为

$$k_{i-1} \leqslant k_i, d_1 = 1, d_{i-1} < d_i \leqslant RF_{i-1} \tag{1}$$

$$RF_1 = k_1, RF_i = d_i(k_i - 1) + RF_{i-1} \tag{2}$$

式中，k_i为第i次深度卷积核大小、d_i为扩张率，RF_i为感受野大小。

通过增大卷积核大小和扩张率可确保目标有足够大的感受野，从而实现上下文信息的利用。例如，使用2～3个扩张卷积的级联组合，即可近似模拟一个大核卷积的效果。如表1所示，在通道数为32的情况下，对两个具有代表性的示例进行理论效率比

表 1　通道数为 32 时级联扩张卷积和大核卷积参数对比

Table 1　Comparison of parameters between cascaded dilated convolution and large kernel convolution with 32 channels

感受野大小	(k, d)序列或k	参数量/K
19	19	369.70
	(19, 1)	11.58
	(3, 1)→(5, 5)	1.15
23	23	541.73
	(23, 1)	16.96
	(3, 1)→(5, 2)→(5, 3)	2.06

较，其中(k, d)序列表示使用扩张卷积，k表示常规卷积。

级联后有以下两个优点，首先，能够生成多个不同感受野的特征图，实现分层特征提取；其次，与直接应用单一大核卷积相比，级联组合在保持相同感受野的同时，显著减少参数的数量，级联过程为

$$\boldsymbol{U}_0 = \boldsymbol{X}, \boldsymbol{U}_{i+1} = DW_i(\boldsymbol{U}_i) \tag{3}$$

式中，$\boldsymbol{X}$表示输入特征图，$\boldsymbol{U}_0$为初始特征图，DW_i为第i次深度卷积操作。初始特征图等于输入特征图，每个后续的特征图$\boldsymbol{U}_{i+1}$是通过在前一层特征图$\boldsymbol{U}_i$上应用深度卷积DW_i得到的。

1.1.2　通道加权和空间融合

为了使网络动态地聚焦于与被检测目标最相关的上下文信息，本文先采用通道注意力进行通道加权，使每个分支在通道方向上学习“关注什么”，其结构如图 4 所示，然后使用空间注意力进行空间选择和融合，即在空间方向上学习“关注哪里”，其结构如图 5 所示。

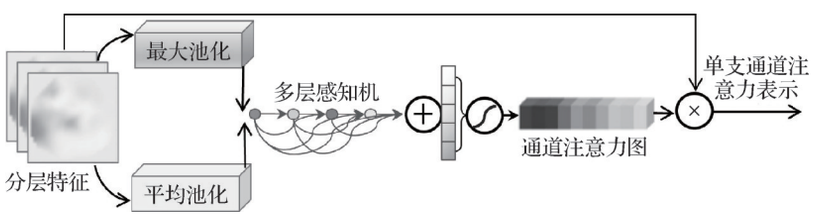


图 4　通道加权结构

Fig. 4　Channel weighted structure

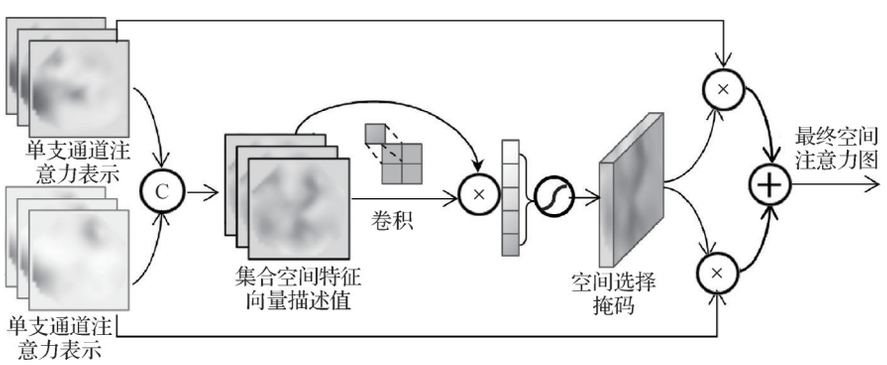


图 5　空间融合结构

Fig. 5　Spatial fusion structure

对卷积级联后提取的分层特征U_i进行通道加权以提取不同重要层次的特征，其过程为

$$\overline{\boldsymbol{U}}_i = \sigma(MLP(P_{\text{avg}}(\boldsymbol{U}_i)) + MLP(P_{\max}(\boldsymbol{U}_i))) \times \boldsymbol{U}_i = \sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{U}_i^{\text{avg}})) + \boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{U}_i^{\max}))) \times \boldsymbol{U}_i \tag{4}$$

$$\overline{\boldsymbol{U}} = C[\overline{\boldsymbol{U}}_1; \cdots; \overline{\boldsymbol{U}}_i] \tag{5}$$

式中，P_{avg}和$P_{\max}$为平均池化和最大池化，C为通道拼接操作，MLP为多层感知机，$\boldsymbol{W}_0$和$\boldsymbol{W}_1$是权重矩阵，σ为 sigmoid 激活函数，$\overline{\boldsymbol{U}}_i$为单支通道注意力表示，$\overline{\boldsymbol{U}}$为最终通道注意力表示。

将拼接得到的最终通道注意力特征表示$\overline{\boldsymbol{U}}$送入空间注意力，从含有权重信息的特征中进行空间融合。首先通过对$\overline{\boldsymbol{U}}$应用基于通道的平均池化和最大池化提取空间关系。与通道注意力不同的是，空间注意力关注捕捉不同空间位置的通道信息，以便生成空间注意力图，表示为

$$\boldsymbol{SA}_{\text{avg}} = P_{\text{avg}}(\overline{\boldsymbol{U}}), \boldsymbol{SA}_{\max} = P_{\max}(\overline{\boldsymbol{U}}) \tag{6}$$

式中，SA_{avg}和SA_{max}分别为最终通道注意力特征经平均池化和最大池化得到的集合空间特征向量描述值。

为了使不同空间描述值之间产生信息交互，使用卷积操作将空间特征集合进行合并，将合并后的空间特征与$\overline{U}$相乘，并通过激活函数σ将合并后的征转换成空间选择掩码。之后，将式(4)得到的含有不同权重信息的特征$\overline{U}_i$与空间选择掩码加权融合得到最终空间注意力图，具体过程为

$$\widehat{SA} = \sigma(Conv^{2\to 1}([SA_{avg}; SA_{max}]) \times \overline{U}) \tag{7}$$

$$S = \sum_{i=1}^{N}(\widehat{SA} \times \overline{U}_i) \tag{8}$$

$$Y = X \times S \tag{9}$$

式中，$Conv^{2\to 1}$为卷积操作，$\widehat{SA}$为空间选择掩码，S为最终空间注意力图。ALRF的最终输出结果Y是输入特征X与S之间的元素乘积。

图6展示了使用ALRF后与基准模型在第2阶段到第5阶段的感受野变化及对比情况。

1.2 MFF

FPN和PAN通过逐层拼接或加和操作进行特征融合。然而，简单的线性融合忽略了不同特征层间的语义和尺度差异，无差异化的处理会导致特征融合不充分。此外，FPN和PAN未能充分利用浅层特征中的丰富语义信息，这对小目标细节的捕捉至关重要。为克服这些局限，设计一种新的特征融合结构MFF，其由序列融合引擎(FSFE)模块和精细尺度特征编码器(FFE)组成，并以双向特征金字塔网络(bidirectional feature pyramid network, BiFPN)(Tan等，2020)结构融合从骨干网中提取的多尺度特征。FSFE模块通过不同偏差的高斯卷积核构建多尺度特征序列，实现差异化处理来有效融合不同粒度的语义信息。而FFE引入P2浅层特征中的细节信息，并通过混合编码增强语义表达，从而减少特征融合中小目标语义信息的丢失。

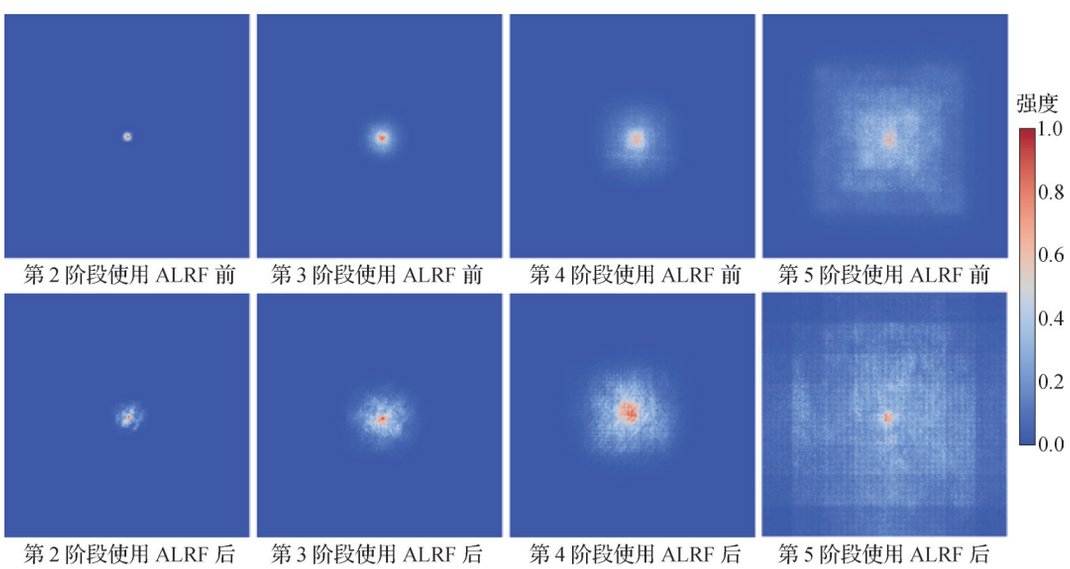


图6 使用ALRF前后感受野变化

Fig. 6 Changes in receptive field before and after using ALRF

1.2.1 FSFE

FSFE模块旨在有效融合P3、P4和P5层的特征，以捕获不同空间尺度、不同大小和形状的遥感图像信息，其结构如图7所示。

首先，为进一步构建从骨干网生成的多尺度特征层(P3、P4和P5)的序列表示，以捕捉不同尺度的特征信息，不同尺度特征层会与一系列标准偏差不断增大的高斯核进行卷积操作，具体为

$$G_{\sigma_k}(x,y) = \frac{1}{2\pi\sigma_k^2} e^{-\frac{x^2+y^2}{2\sigma_k^2}} \tag{10}$$

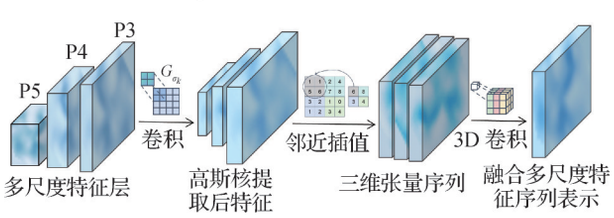


图7 FSFE结构图

Fig. 7 FSFE structure diagram

$$F_{\sigma_k}(i,j) = \sum_u \sum_v f(i-u, j-v) \times G_{\sigma_k}(u,v) \tag{11}$$

式中，G_{σ_k}为二维高斯函数，σ_k为高斯核的标准偏差，

$f(i-u,j-v)$为输入特征层经过高斯核中心点(i,j)平移后(u,v)的值，$\boldsymbol{F}_{\sigma_k}$为经高斯核卷积后的结果。为捕捉不同尺度的特征信息，不仅仅使用单一的高斯核，而是采用一组标准差逐渐增大的高斯核，每个σ_k对应一个特定高斯函数，其决定了函数影响范围的宽度。

为了使这些多尺度特征层$\boldsymbol{F}_{\sigma_k}$可以协同工作，首先使用近邻插值方法将其调整到相同的尺寸。然后，将调整后的特征层在深度方向堆叠，构建出一个三维张量。在此基础上，进一步应用3D卷积、3D批量归一化以及SiLU(sigmoid linear unit)激活函数，以生成融合多个序列的特征表示，具体过程为

$$\boldsymbol{F}'_{\sigma_k}=f_{\text{Upsample}}(\boldsymbol{F}_{\sigma_k}) \tag{12}$$

$$\boldsymbol{F}_{\text{stack}}=C[\boldsymbol{F}'_{\sigma_1};\boldsymbol{F}'_{\sigma_2};\boldsymbol{F}'_{\sigma_3}] \tag{13}$$

$$\boldsymbol{F}_{\text{seq}}=f_{\text{Conv3D}}(\boldsymbol{F}_{\text{stack}}) \tag{14}$$

式中，Upsample为邻近插值法，$\boldsymbol{F}'_{\sigma_k}$为调整到相同尺寸的特征层，$\boldsymbol{F}_{\text{stack}}$为三维张量序列，$\boldsymbol{F}_{\text{seq}}$为融合多个序列的特征表示。

通过上述步骤，FSFE利用不同偏差的高斯核捕捉不同尺度目标的特征信息，并将其统一到相同尺寸后利用3D卷积进行融合。该方法通过对不同尺度特征差异化处理，克服特征融合不充分问题。

1.2.2 FFE

为了检测遥感图像中的小目标，可以使用浅层大尺寸特征图，而在FPN和PAN融合机制中，通常对小尺寸特征进行上采样，并简单地与上层特征相加，忽略了大尺寸特征层中丰富细节信息。为解决这一问题，设计FFE，其结构如图8所示，C和S分别表示特征图的通道数和空间尺寸大小。

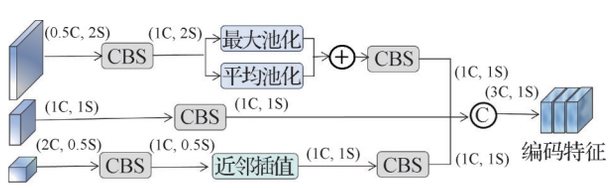


图8 FFE结构图

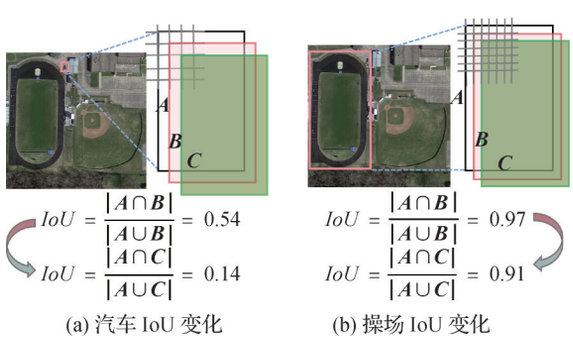
Fig. 8 FFE structure diagram

FFE处理过程的输入为主干网络提取的多尺度特征层，分别对应于大、中、小尺度的特征表示。首先，多尺度特征层经过一系列卷积操作，调整其通道数，与主尺度(中等尺寸)的特征层保持一致。接着，对于大尺寸特征层，FFE通过混合池化结构减少空间维度，增强网络对输入特征平移变化的鲁棒性。而小尺寸的特征图，FFE使用近邻插值方法进行上采样，以保留局部特征，避免小物体特征信息的失真。将大尺寸和小尺寸特征层编码为与中等尺寸特征层相同的空间尺寸后在通道维度上拼接，形成FFE的输出。

FFE通过利用浅层特征信息并对不同特征层进行编码，克服特征融合中语义信息丢失问题，提升模型在小目标检测任务中的精度。

1.3 归一化Wasserstein距离

传统IoU度量及其衍生版对微小物体的位置偏差极为敏感，如图9所示，以5×8像素的小物体为例，轻微的位置偏差，IoU值便从0.54急剧下降至0.14，从而影响标签分配的准确性。相比之下，尺寸为200×320像素的常规物体，在相同的位置偏差下，IoU值仅从0.97轻微下降至0.91。小物体位置偏差敏感性严重影响了检测器的标签分配质量。因此，本文引入归一化Wasserstein距离来替换IoU，以解决小物体对位置偏差过于敏感而导致的标签分配不准确和正负样本不平衡问题。



(a) 汽车IoU变化 (b) 操场IoU变化

图9 不同尺寸物体IoU变化

Fig. 9 Variation in IoU for objects of different sizes

((a)IoU variation of vehicle;(b)IoU variation of playground)

遥感图像中微小物体很多不是严格的矩形，其边界框内混杂着背景像素。小物体边界框中，前景像素一般集中在中心区域，而背景像素则多位于边缘。为了更好地描述边界框内不同位置像素的重要性，NWD采用二维高斯分布$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$对边界框$\boldsymbol{B}=(cx,cy,w,h)$进行建模，具体为

$$\boldsymbol{\mu}=\begin{bmatrix}cx\\cy\end{bmatrix},\boldsymbol{\Sigma}=\frac{1}{4}\begin{bmatrix}w^2&0\\0&h^2\end{bmatrix} \tag{15}$$

式中，(cx,cy)是边界框的中心坐标，(w,h)表示框的宽高，$\boldsymbol{\mu}$是均值矩阵，$\boldsymbol{\Sigma}$是协方差矩阵。

对于两个边界框的高斯分布$\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$和$\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$,其Wasserstein距离计算为

$$W(\mathcal{N}_1, \mathcal{N}_2) = \left\| \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \right\|_2^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\left(\boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{1/2} \right)^{1/2} \right) \tag{16}$$

式中,$\| \cdot \|_2$表示欧氏距离,tr表示矩阵的迹。为了将Wasserstein距离的值域从$[0, +\infty)$映射到$(0, 1]$以适应损失函数,选择一个指数非线性变换函数将Wasserstein距离重新映射到另一个空间,定义NWD为

$$NWD(\mathcal{N}_1, \mathcal{N}_2) = \exp\left(-\frac{\sqrt{W(\mathcal{N}_1, \mathcal{N}_2)}}{C} \right) \tag{17}$$

式中,C是用于归一化的超参数。

最后,将NWD与TAL相结合,实现动态样本选择,在训练过程中根据预测质量和对齐度动态调整正负样本的分配。NWD作为更鲁棒的相似性度量,能更准确评估微小物体边界框与真实框的对齐程度。

2 实验结果与分析

2.1 实验环境和数据集

实验环境为Ubuntu 20.04操作系统,CPU为E5-2680v4,显卡为NVIDIA GeForce GTX 3090,显存为24 G,在PyTorch 1.13.1、CUDA11.6框架下运行。训练过程中采用SGD(stochastic gradient descent)优化器,初始学习率为0.010,动量大小为0.937,batchsize设为8,训练轮次为300。

本文实验数据选自NWPU VHR-10(Northwestern Polytechnical University very high resolution 10)数据集(Cheng和Han,2016)和DIOR(dataset for object detection in aerial images)数据集(Li等,2020)。NWPU VHR-10数据集是由西北工业大学发布,由650幅有标注和150幅无标注的遥感图像组成。数据集涵盖10个类别的地理空间对象,本文将650幅有标记图以8∶2随机划分为训练集和测试集。DIOR数据集是一个大规模光学遥感图像目标检测数据集,同样由西北工业大学提出。该数据集包含23 463幅遥感图像和190 288个目标实例,覆盖20个目标类别,其中,训练集包含11 725幅图像,测试集包含11 738幅图像。

2.2 评价指标

用精度(precision,P)、召回率(recall,R)、平均精确度(average precision,AP)及参数量为综合评价指标,评估遥感图像检测效果,其中mAP50为IoU阈值为0.5时的AP值。

2.3 消融实验

为了探究ALRF、MFF和NWD对模型的影响,在NWPU VHR-10数据集上进行多组实验。根据式(1)和式(2),本文采用带扩张率的深度卷积进行两层级联,并设置步幅为4来逐步扩大感受野,以探究感受野大小对ALRF的影响,结果见表2。结果表明,当感受野大小为19时,ALRF效果最佳。

表2 NWPU VHR-10数据集上感受野大小对ALRF影响
Table 2 The impact of receptive field size on ALRF on the NWPU VHR-10 dataset

感受野大小	(k, d)序列	mAP50/%
11	(3, 1)→(5, 2)	90.06
15	(3, 1)→(5, 3)	91.17
19	(3, 1)→(9, 2)	91.98
23	(5, 1)→(7, 3)	91.64

接着,为探究通道注意力和空间注意力对ALRF的影响,将通道加权和空间融合所在的位置分别记为位置1和位置2。同时,在不同位置又分别选取坐标注意力(coordinate attention,CoA)、局部注意力(local attention,LA)和全局注意力(global attention,GA)进行实验,结果如表3所示。实验表明,在采用不同扩张率的深度卷积进行分层特征提取后,先进行通道加权,再进行空间融合时,ALRF的表现最优。

将ALRF的感受野大小设定为19,并调整注意力的顺序为“先进行通道加权,再进行空间融合”后,对模型中各模块的检测效果进行探究,结果见表4。结果显示,ALRF对遥感图像检测影响最为显著,其中精度、召回率和mAP50分别提高1.86%、3.40%和4.31%,这得益于大感受野的应用,能够有效地利用目标的上下文信息。MFF在提高模型检测精度的同时显著降低模型参数量,从基准的3.16 M降低至2.83 M,减少10.44%。这得益于MFF在融合多尺度特征时,采用不同卷积核和混合编码技术,实现精准的特征放大与融合,避免对所有特征无差别处理,

表 3　NWPU VHR-10 数据集上不同注意力在不同位置对 ALRF 的影响

Table 3　The impact of different attention positions on ALRF on the NWPU VHR-10 dataset

/%

位置 1	位置 2	精度	召回率	mAP50
CoA	LA	92.85	85.32	90.81
CoA	GA	92.97	85.55	91.87
LA	GA	92.11	84.23	90.89
LA	CoA	92.74	84.86	91.68
GA	CoA	93.13	**86.08**	91.90
GA	LA	93.07	85.47	91.56
SA	CA	92.35	84.20	90.75
CA	SA	**93.18**	85.61	**91.98**

注：加粗字体表示各列最优结果。

从而减少冗余计算。此外，NWD 的引入进一步优化小目标的检测性能，使正负样本分配更为均衡。将 3 种方法结合可使模型效果达到最优，其中精度、召回率和 mAP50 分别提高 2.76%、4.61% 和 5.48%，同时参数量减少 6.96%。

2.4　对比实验

实验选取 10 组方法对比，其中双阶段检测方法包括 Faster R-CNN、SABL（Huang 等，2022）和 Dynamic R-CNN；单阶段检测方法包括 TOOD（task-aligned one-stage object detection）（Feng 等，2021）、RTMDet（real-time models for object detection）（Lyu 等，2022）、YOLOv8n 和 YOLOv10n（Wang 等，2024）；DETR（end-to-end object detection with transformers）系列检测方法包括 DETR（Carion 等，2020）、DAB-DETR（Liu 等，2022）和 DDQ（dense distinct query for end-to-end object detection）（Zhang 等，2023）。不同目标检测器在 NWPU VHR-10 数据集上的检测效果见表 5。

表 4　NWPU VHR-10 数据集上消融实验结果

Table 4　Ablation results on the NWPU VHR-10 dataset

ALRF	MFF	NWD	精度/%	召回率/%	mAP50/%	参数量/M
–	–	–	91.32	82.21	87.67	3.16
√	–	–	93.18	85.61	91.98	3.37
–	√	–	92.56	84.38	90.25	**2.83**
–	–	√	91.41	82.43	87.93	3.16
√	√	√	**94.08**	**86.82**	**93.15**	2.94

注：“√”和“–”分别表示采用和未采用，加粗字体表示各列最优结果。

表 5　NWPU VHR-10 数据集上不同方法的 mAP50 对比

Table 5　Comparison of mAP50 with different methods on the NWPU VHR-10 dataset

/%

类别	方法	飞机	储罐	棒球场	网球场	篮球场	田径场	汽车	桥	港口	船只	平均 mAP50
双阶段检测器	Faster R-CNN	84.02	60.00	86.61	77.91	78.20	99.40	54.81	59.52	52.11	78.60	73.12
	SABL	93.31	75.91	77.50	75.80	74.48	96.13	55.31	43.22	68.99	84.09	77.47
	Dynamic R-CNN	97.84	59.35	97.85	91.35	94.16	99.52	89.46	80.35	83.04	95.74	88.87
单阶段检测器	TOOD	98.23	65.43	99.51	**98.33**	96.54	99.51	95.33	85.04	**89.54**	97.65	92.51
	RTMDet	98.70	**86.69**	99.50	93.71	**98.33**	99.49	88.50	80.59	86.10	94.59	92.62
	YOLOv8n	97.37	75.40	99.29	82.39	83.10	99.49	84.09	81.40	81.48	92.70	87.67
	YOLOv10n	99.13	57.45	99.52	78.72	79.94	98.80	84.76	74.48	78.68	91.83	84.33
	本文	**99.48**	81.11	98.26	91.84	91.36	**99.54**	**96.65**	**88.55**	86.84	**97.85**	**93.15**
DETR 系列检测器	DETR	97.94	50.67	97.46	85.64	92.45	99.47	74.25	85.54	77.94	95.34	85.67
	DAB-DETR	97.65	52.04	99.04	89.14	91.54	99.51	84.35	85.04	86.24	97.36	88.19
	DDQ	98.28	63.65	**99.53**	94.78	98.18	99.49	96.48	87.68	86.18	97.81	92.21

注：加粗和下划线字体分别表示各列最优和次优结果。

在双阶段模型中，各模型在不同类别上的表现存在较大差异。例如，Dynamic R-CNN 在多数类别上表现良好，但在储罐类别上的检测精度较低，仅为 59.35%。在单间段检测器中，本文方法的 mAP50 最

高，RTMDet次之，两者在多个类别上表现突出。特别是在汽车和船只小目标检测任务中，本文方法的mAP50分别达到96.65%和97.85%，展现出优异的检测能力。DETR系列检测器在储罐类别上的检测精度普遍较低，但在棒球场这一类别上，DDQ表现出色。整体来看，本文方法在所对比的10种检测器中表现优异，mAP50达到93.15%。其中，与基准模型YOLOv8n相比，在汽车和船只小目标的检测任务中AP50分别提升12.56%和5.15%，实现了显著的性能提升。

为了更详细地进行对比，图10展示了各检测器的PR(precision-recall)曲线。PR曲线揭示了精度与召回率之间的关系，而曲线下方与坐标轴围成的面积，即为模型的平均精度(AP)值，是衡量检测效果的重要指标。本文方法的PR曲线更靠近右上方且更为平稳，表明其效果优于其他对比方法。

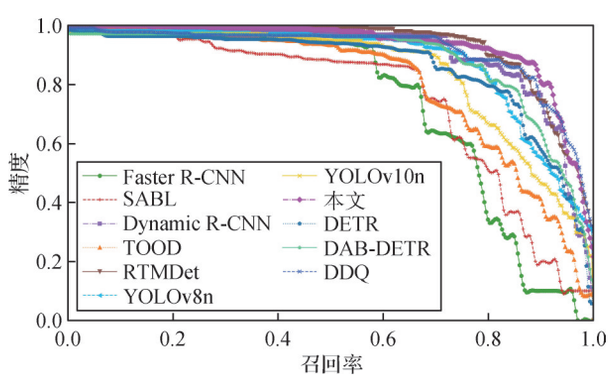


图10　NWPU VHR-10数据集上不同方法PR曲线对比

Fig. 10　Comparison of PR curves of different methods on the NWPU VHR-10 dataset

进一步，为对比模型性能，图11展示了各模型的平均精度和参数量间的散点图。图中对参数量坐标轴进行翻转，散点靠近右上方，模型参数量更低，同时平均精度更高，即模型性能更好。结果表明，在平衡检测精度与模型参数量间的关系后，本文方法性能最优，RTMDet次优，与表4结果一致。虽然YOLOv10n在对比模型中参数量最低，为2.33 M，但mAP50仅为84.33%。相比之下，本文方法的mAP50达到93.15%，同时参数量也控制在2.94 M，表明本文方法在实现高检测精度的同时，还保持较低的参数量。

为验证本文方法的有效性和泛化性，在DIOR数据集上进一步对各模型进行对比，结果如表6所

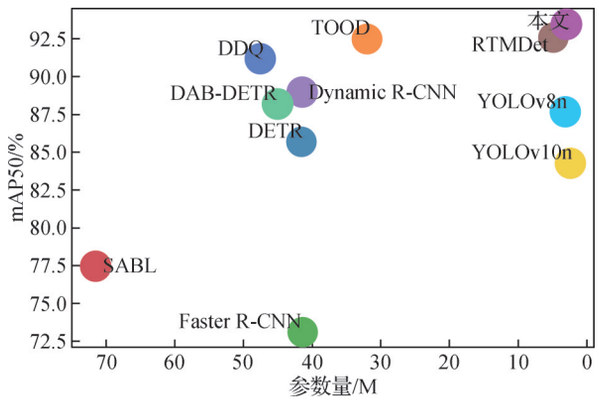


图11　NWPU VHR-10数据集上不同方法性能对比

Fig. 11　Performance comparison of different methods on the NWPU VHR-10 dataset

示。双阶段检测模型在DIOR数据集上的效果不佳，整体平均精度较低，因为它们在应对尺度变化和背景干扰方面的鲁棒性不足。在单阶段模型中，整体效果表现优异，不仅平均精度高，参数量也相对较低。其中，本文方法在小目标检测中的效果尤为突出，相比YOLOv8n，本文方法对汽车、储罐和船只的mAP50分别提升3.61%、1.68%和2.39%，达到所对比模型的最佳效果。此外，在DIOR数据集中最有挑战性的桥这一目标中，本文方法在所有对比模型中的mAP50也达到最高，为56.96%。本文方法在小目标检测和跨度大的桥梁检测中表现优异，得益于其采用了自适应大感受野机制(ALRF)。ALRF不仅能帮助小目标有效利用上下文信息，还能为跨度大的目标提供足够大的感受野。在DETR系列模型中，DDQ表现最好，特别是在复杂和大型目标检测任务(如公路收费站和机场)上表现出色，对于小物体(如汽车和储罐)，DDQ的检测效果不如本文方法。

最后，为了更直观地展示不同方法在小目标检测任务上的表现，选取5种方法在NWPU VHR-10和DIOR数据集上进行可视化对比。结果如图12所示，其中绿色为正确检测，蓝色和红色为漏检和误检。尽管所有方法在一定程度上都能检测到目标，但Dynamic R-CNN在检测过程中出现较多的漏检和误检。RTMDet和YOLOv8n虽然漏检较少，但仍存在较多的误检。DDQ在NWPU VHR-10数据集的汽车检测中表现良好，没有漏检；在DIOR数据集的船只检测中，虽然漏检较少，但误检有所增加。相比之下，本文方法在两个数据集上均表现出色，没有漏检

表 6　DIOR 数据集上不同方法的 mAP50 对比

Table 6　Comparison of mAP50 with different methods on the DIOR dataset

类别	方法	高尔夫球场/%	公路收费站/%	汽车/%	火车站/%	烟囱/%	储罐/%	船只/%	港口/%	飞机/%	田径场/%	网球场/%	水坝/%
双阶段检测器	Faster R-CNN	76.80	63.92	40.01	51.85	74.49	56.71	71.80	45.89	55.36	81.07	81.93	49.38
	SABL	76.99	55.49	41.81	51.44	75.81	63.20	73.28	55.75	62.51	79.73	84.89	58.66
	Dynamic R-CNN	82.72	70.25	41.19	62.21	78.23	59.14	72.47	55.06	61.64	83.27	87.34	69.19
单阶段检测器	TOOD	84.06	73.88	52.90	61.39	**85.24**	71.86	76.65	60.50	72.38	83.54	89.08	69.79
	RTMDet	83.47	69.93	50.66	**69.45**	85.11	71.67	76.49	69.64	86.50	**87.98**	93.07	**77.17**
	YOLOv8n	75.34	69.27	59.48	57.97	77.09	90.16	93.18	67.45	96.38	84.51	93.88	60.62
	YOLOv10n	**84.47**	70.14	54.28	68.16	77.33	78.11	89.72	65.38	80.27	80.69	91.93	70.78
	本文	78.41	75.57	**63.09**	62.99	79.44	**91.84**	**95.57**	**70.23**	**96.88**	86.43	**95.52**	63.81
DETR 系列检测器	DETR	77.15	60.64	34.46	57.36	79.13	33.92	29.92	39.98	52.61	74.01	80.41	64.37
	DAB-DETR	74.88	57.11	38.89	61.52	79.82	46.76	56.06	46.94	70.98	79.08	83.66	67.70
	DDQ	82.58	**85.37**	59.99	67.07	84.76	80.10	91.25	65.13	78.03	86.58	88.61	75.88

类别	方法	篮球场/%	公路服务区/%	体育场/%	机场/%	棒球场/%	桥/%	风车/%	天桥/%	平均 mAP50/%	参数量/M
双阶段检测器	Faster R-CNN	84.84	73.68	66.68	72.52	67.16	36.50	81.32	50.97	64.14	41.39
	SABL	85.40	76.39	64.16	79.21	75.26	39.88	79.03	56.05	66.75	71.50
	Dynamic R-CNN	88.05	85.35	72.57	84.53	73.31	49.30	83.82	67.13	71.34	41.44
单阶段检测器	TOOD	89.09	86.87	74.63	86.95	78.12	51.00	87.38	64.27	74.98	32.04
	RTMDet	**91.99**	**92.02**	81.65	89.38	85.83	47.99	84.45	64.55	77.95	4.89
	YOLOv8n	90.82	82.37	90.33	74.67	92.48	49.40	89.33	63.62	77.92	3.16
	YOLOv10n	89.56	89.45	77.19	85.78	84.36	47.72	85.81	63.27	76.72	**2.33**
	本文	91.82	86.84	**92.48**	79.69	**93.20**	**56.96**	90.69	66.32	**80.89**	2.94
DETR 系列检测器	DETR	84.49	77.89	60.52	80.85	69.14	38.40	81.89	52.98	61.51	41.50
	DAB-DETR	85.14	76.00	67.32	84.57	77.92	37.84	79.51	55.73	66.37	45.02
	DDQ	89.07	91.27	82.78	**92.35**	83.94	56.23	**91.04**	**69.17**	80.06	48.34

注：加粗和下划线字体分别表示各列最优和次优结果。

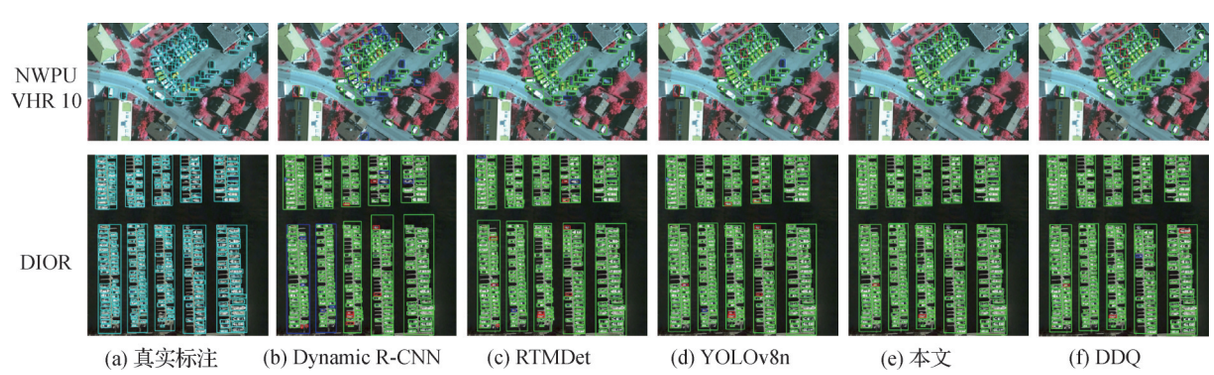


图 12　NWPU VHR 10 和 DIOR 数据集上不同方法可视化对比

Fig. 12　Comparison of different method visualizations on NWPU VHR 10 and DIOR datasets

((a) true labeling; (b) Dynamic R-CNN; (c) RTMDet; (d) YOLOv8n; (e) ours; (f) DDQ)

现象，仅出现少数误检。上述结果表明，本文方法在小目标检测任务中有着较高的准确性和鲁棒性。

3 结 论

针对遥感图像检测中目标尺寸小、尺度变化大和背景复杂等挑战，本文提出一种上下文信息和多尺度特征序列引导的遥感图像检测方法。该方法首先通过自适应大感受野机制（ALRF），根据不同目标的特征自适应调整感受野大小，实现多尺度特征的提取。在特征融合阶段，设计多尺度特征序列融合架构（MFF），通过不同偏差的高斯核构建多尺度序列表示并引入浅层特征，从而解决小目标语义信息丢失问题。最后，引入NWD，并与动态标签分配策略结合，缓解了基于IoU的样本分配策略对小目标位置偏差过于敏感而导致的正负样本不平衡问题。实验结果表明，在NWPU VHR-10和DIOR数据集上，与10种方法相比，本文方法在mAP50指标上均优于对比方法，在参数量指标上仅次于YOLOv10n，展现出优越的性能和实际应用潜力。然而，本文方法主要针对正下视角的遥感图像设计，在多视角或旋转目标检测任务中，可能会面临性能下降的问题。未来工作将拓展至多视角成像条件下的旋转目标检测任务，探索将不同视角下的目标检测问题转化为域适应问题。同时，为满足机载平台实时目标检测的要求，可结合知识蒸馏技术，开发适用于多视角光学遥感图像的轻量化目标检测模型。

参考文献（References）

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chen H B, Jiang S, He G H, Zhang B Y and Yu H. 2021. TEANS: a target enhancement and attenuated nonmaximum suppression object detector for remote sensing images. IEEE Geoscience and Remote Sensing Letters, 18(4): 632-636 [DOI: 10.1109/LGRS.2020.2983070]

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

Feng C J, Zhong Y J, Gao Y, Scott M R and Huang W L. 2021. TOOD: task-aligned one-stage object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3490-3499 [DOI: 10.1109/ICCV48922.2021.00349]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Huang Z L, Wei Y C, Wang X G, Liu W Y, Huang T S and Shi H. 2022. AlignSeg: feature-aligned segmentation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1): 550-557 [DOI: 10.1109/TPAMI.2021.3062772]

Jocher G, Chaurasia A and Qiu J. 2023. YOLO by Ultralytics [EB/OL]. [2024-07-11]. https://github.com/ultralytics/ultralytics

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

Li X, Wang W H, Hu X K and Yang J. 2019. Selective kernel networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 510-519 [DOI: 10.1109/CVPR.2019.00060]

Li Y X, Hou Q B, Zheng Z H, Cheng M M, Yang J and Li X. 2023. Large selective kernel network for remote sensing object detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 16748-16759 [DOI: 10.1109/ICCV51070.2023.01540]

Li Z H, An J T, Jia H Y and Fang Y. 2023. Lightweight object detection model in remote sensing image by combining rotation box and attention mechanism. Journal of Image and Graphics, 28(9): 2706-2718 (李朝辉, 安金堂, 贾红雨, 方艳. 2023. 结合旋转框和注意力机制的轻量遥感图像检测模型. 中国图象图形学报, 28(9): 2706-2718) [DOI: 10.11834/jig.220839]

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

Liu C, Zhang S X, Hu M J and Song Q. 2024. Object detection in remote sensing images based on adaptive multi-scale feature fusion method. Remote Sensing, 16(5): #907 [DOI: 10.3390/rs16050907]

Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018. Path aggregation network for instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8759-8768 [DOI: 10.1109/CVPR.2018.00913]

Liu S L, Li F, Zhang H, Yang X, Qi X B, Su H, Zhu J and Zhang L. 2022. DAB-DETR: dynamic anchor boxes are better queries for DETR//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

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

Lyu C Q, Zhang W W, Huang H A, Zhou Y, Wang Y D, Liu Y Y, Zhang S L and Chen K. 2022. RTMDet: an empirical study of designing real-time object detectors [EB/OL]. [2024-07-11]. https://arxiv.org/pdf/2212.07784.pdf

Ma D L, Liu B Z, Huang Q J and Zhang Q. 2023. MwdpNet: towards improving the recognition accuracy of tiny targets in high-resolution remote sensing image. Scientific Reports, 13(1): #13890 [DOI: 10.1038/s41598-023-41021-8]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Shamsolmoali P, Zareapoor M, Chanussot J, Zhou H Y and Yang J. 2022. Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 60: #5608614 [DOI: 10.1109/TGRS.2021.3112481]

She H D and Zhao L J. 2024. Rotating target detection network that combines key points and guide vectors. Journal of Image and Graphics, 29(2): 533-544 (佘浩东, 赵良瑾. 2024. 结合关键点与引导向量的旋转目标检测网络. 中国图象图形学报, 29(2): 533-544) [DOI: 10.11834/jig.230207]

Tan M X, Pang R M and Le Q V. 2020. EfficientDet: scalable and efficient object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10778-10787 [DOI: 10.1109/CVPR42600.2020.01079]

Wang A, Chen H, Liu L H, Chen K, Lin Z J, Han J G and Ding G G. 2024. YOLOv10: real-time end-to-end object detection [EB/OL]. [2024-07-11]. https://arxiv.org/pdf/2405.14458.pdf

Xu C, Wang J W, Yang W, Yu H, Yu L and Xia G S. 2022. Detecting tiny objects in aerial images: a normalized Wasserstein distance and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 190: 79-93 [DOI: 10.1016/j. isprsjprs. 2022. 06.002]

Zhang H K, Chang H, Ma B P, Wang N Y and Cheng X L. 2020. Dynamic R-CNN: towards high quality object detection via dynamic training//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 260-275 [DOI: 10.1007/978-3-030-58555-6_16]

Zhang H P, Wen S Z, Wei Z X and Chen Z Y. 2024a. High-resolution feature generator for small-ship detection in optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62: #5617011 [DOI: 10.1109/TGRS.2024.3377999]

Zhang S L, Wang X J, Wang J Q, Pang J M, Lyu C Q, Zhang W W, Luo P and Chen K. 2023. Dense distinct query for end-to-end object detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7329-7338 [DOI: 10.1109/CVPR52729.2023.00708]

Zhang Z H, Wang C H, Zhang H Y, Qi D C, Liu Q Y, Wang Y F and Ding W R. 2024b. SREDet: semantic-driven rotational feature enhancement for oriented object detection in remote sensing images. Remote Sensing, 16(13): #2317 [DOI: 10.3390/rs16132317]

作者简介

肖振久，男，副教授，主要研究方向为图像与视觉信息计算。E-mail：xiaozhenjiu@lntu.edu.cn

李士博，通信作者，男，硕士研究生，主要研究方向为遥感大模型。E-mail：lishibo@stu.htu.edu.cn

曲海成，男，副教授，主要研究方向为遥感影像高性能计算。E-mail：quhaicheng@lntu.edu.cn

李富坤，男，硕士研究生，主要研究方向为大模型并行计算。E-mail：lifukun2022@163.com