从光学到SAR:基于多级跨模态对齐的SAR图像舰船检测算法

何佳月[1]，宿南[1]，徐从安[2]，尹璐[3]，廖艳苹[1]，闫奕名[1]

1. 哈尔滨工程大学 信息与通信工程学院, 哈尔滨 150001;
2. 海军航空大学 信息融合研究所, 烟台 264001;
3. 北京市遥感信息研究所, 北京 100192

摘 要：合成孔径雷达（SAR）舰船检测是近年来的研究热点。然而，与光学图像不同，SAR成像的特点会导致不直观的特征表示。此外，由于SAR图像数据量不足，现有的基于大量标记SAR图像的方法可能难以达到较好的检测效果。为了解决这些问题，本文提出了一种基于多级跨模态对齐的SAR图像舰船检测算法MCMA-Net（Multi-level Cross-Modality Alignment Network），通过将光学模态中丰富的知识迁移到SAR模态来增强SAR图像的特征表示。该算法首先设计了一个基于邻域—全局注意力的特征交互网络NGAN（Neighborhood-Global Attention Network），通过对骨干网络的浅层特征采用邻域注意力机制进行局部交互、对深层特征采取全局自注意力机制进行全局上下文交互，在兼顾全局上下文建模能力的同时，提升局部特征的编码能力，使得网络在不同层级更合理的关注相应的信息，从而能够促进后续的多级别模态对齐。其次，本文设计了一个多级模态对齐模块MLMA（Multi-level Modality Alignment），通过从局部级别到全局级别再到实例级别的对两种模态不同隐含空间中的特征进行对齐，促进模型有效地学习模态不变特征，缓解了光学图像和SAR图像之间的模态鸿沟，实现了从光学模态到SAR模态的知识传输。大量的实验证明我们的算法优于现阶段的检测算法，取得了最好的实验结果。

关键词：遥感，SAR，目标检测，跨模态，特征对齐，注意力机制

中图分类号：P2

引用格式：**何佳月,宿南,徐从安,尹璐,廖艳苹,闫奕名.2024.从光学到SAR:基于多级跨模态对齐的SAR图像舰船检测算法.遥感学报,28(7):1789-1801**

He J Y, Su N, Xu C A, Yin L, Liao Y P and Yan Y M. 2024. From optical to SAR: A SAR ship detection algorithm based on multi-level cross-modality alignment. National Remote Sensing Bulletin, 28(7): 1789-1801[DOI: 10.11834/jrs.20243249]

1 引 言

合成孔径雷达SAR（Synthetic Aperture Radar）是一种主动式微波成像传感器，具有全天时、全天候观测地球的能力。近年来，随着SAR图像数据的不断增长，SAR舰船检测作为SAR图像解译的一个重要分支，因其在海洋监测、国防安全等方面的重要价值而备受关注。在民事领域，对民船进行准确的检测，有利于推动渔业安全管理、海面监管、海洋救援等工作的开展，同时可以在一定程度上打击偷渡和违法捕捞等行为。利用SAR图像数据在深度学习领域中开展目标检测工作已经成为一个重要的研究方向。

由于技术限制，早期遥感图像的成像分辨率较低，研究时将舰船等效为点目标，传统方法针对舰船与海平面背景有较强的对比度这一特性，对SAR图像中的舰船目标进行检测（Pappas等，2018）。其中，恒虚警检测算法CFAR（Constant

收稿日期：2023-07-11；**预印本**：2024-03-06

基金项目：国家自然科学基金（编号：62271159，62071136，62002083，61971153）；黑龙江省优秀青年基金（编号：YQ2022F002）；黑龙江省博士后基金（编号：LBH-Q20085，LBH-Z20051）；中央高校基本科研业务费资金资助（编号：3072022QBZ0805，3072021CFT0801，3072022CF0808）；高分专项中俄边境地区国家安全监测及综合服务产业化示范（编号：72-Y50G11-9001-22/23）

第一作者简介：何佳月，研究方向为深度学习、SAR图像目标检测。E-mail：hejiayue@hrbeu.edu.cn

通信作者简介：宿南，研究方向为深度学习、多模态遥感图像处理和应用。E-mail：sunan08@hrbeu.edu.cn

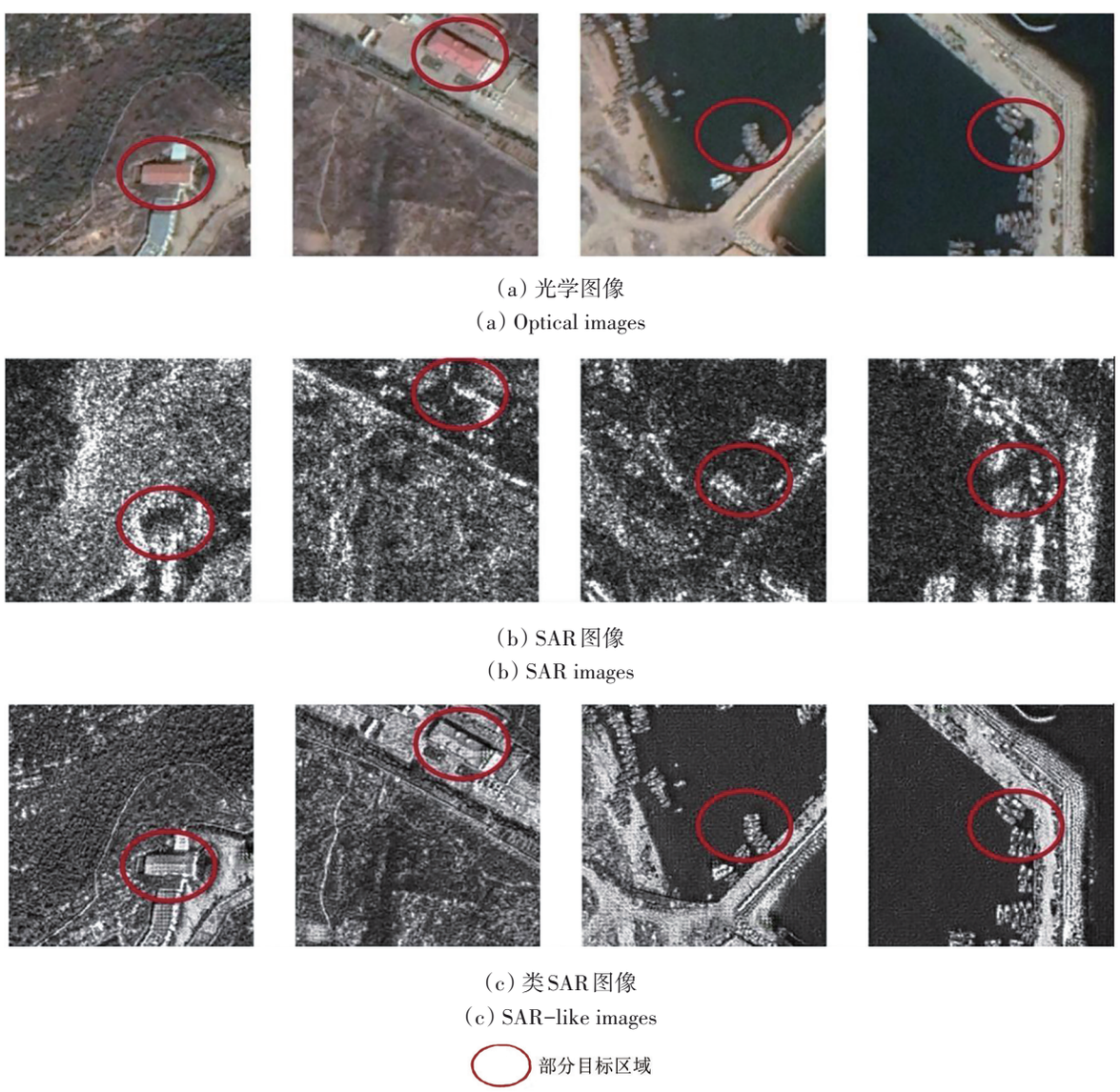
False Alarm Rate）是一种具有代表性的传统算法。然而，CFAR算法对于干扰存在严重的复杂场景会产生较高的虚警率，其高度依赖于背景杂波模型的构建，实际应用效率较低（Zhang等，2018；侯卫和李勇，2023）。此外，CFAR算法使用时需要专业人员有针对性地设置参数，因此其泛化能力较差。因此，这些方法在实际应用中仍然面临着巨大的困难。

随着深度学习技术的飞速发展，目标检测领域也受到了广泛的关注。目前基于深度学习的方法大致可以分为双阶段（Two-stage）算法和单阶段（One-stage）算法两大类。双阶段算法首先对输入图像进行特征提取和区域筛选，得到大量的候选框，之后进行分类检测，在检测精度上具有优势。经典的双阶段算法如：Fast R-CNN（Girshick，2015）、Faster R-CNN（Ren等，2015）、Cascade R-CNN（Cai和Vasconcelos，2018）等。单阶段算法简化了整个流程，将其视为一个回归问题，推理速度得到极大的提高。经典的单阶段算法如：SSD（Liu等，2016）、YOLO系列（Redmon等，2016；Redmon和Farhadi，2017，2018）、RetinaNet（Lin等，2017）等。

受到光学目标检测算法（Liu等，2018；Wu等，2020；Lu等，2019；Dai等，2017）的启发，遥感图像目标检测领域发展迅速（Zhou等，2021；Yao等，2021；Yu等，2020）。当前阶段的SAR图像目标检测算法主要集中在网络模型的创新上，以提升SAR图像目标检测的性能。研究人员通过调整网络结构和设计先进的特征提取器来实现这一目标。例如，有研究（Lin等，2019；Zhao等，2020）选择引入新的注意力机制，增强骨干网络的特征提取能力。另一些研究（Wang等，2023a；Zhang等，2022）则采用特征融合的思想，通过有效地利用提取到的不同特征，将它们进行联合建模，以提升模型的鲁棒性和性能。此外，还有一些研究（Miao等，2022）选择设计更轻量化的模型，以提高算法的效率和实时性。但是由于SAR图像的成像机理与光学图像存在显著差异，使得SAR图像具有独特的特征和问题，相比于光学图像，SAR图像的成像机理和特征表示具有一定的复杂性，其特征更加抽象和难以直观理解。这些创新方法虽然为SAR图像目标检测带来了显著的改进，但是大多都是直接参照光学图像算法的改进思路，并不完全适用于SAR图像，尽管在网络模型和特征提取器的创新方面取得了进展，但在SAR图像目标检测的性能提升方面仍存在一定的限制。因此，仅仅将光学图像算法直接应用于SAR图像往往不能取得理想的结果。另一方面，由于SAR图像数据获取和标注较为困难，需要巨大的经济成本。相较于光学数据而言，现阶段SAR图像的数据量比较匮乏，仅仅采用现有的SAR图像数据去训练出一个检测性能好、鲁棒性又高的模型难度较高。然而光学数据量要远远大于SAR图像数据，并且光学图像具有丰富的细节信息和直观的特征表示，因此更易于观察和解译。在遥感图像研究领域，为了解决由单源数据的局限性而导致的模型性能提升遇到瓶颈的问题，一些学者（Cao等，2019；Wang等，2023b，2023c；Li等，2022；Zhang等，2023）也选择采用多源数据融合的方法来进行智能解译，从而提升网络性能。在SAR图像目标检测任务中，光学图像可以作为一种补充数据源，通过大量的光学数据来辅助训练SAR数据，有助于学习出一个知识更为丰富的模型，因此选用什么样的训练方式能够更为合理的将这两种模态的图像进行利用是一个值得探索的问题。

由于SAR模态数据与光学模态数据在成像机理和特征表示上存在显著差异，实现这种跨模态知识的异质迁移具有很大的挑战性。Li等（2019）选择基于预训练模型进行微调的方法，通过在大规模光学图像数据上预训练的卷积神经网络，将预训练模型的权重用于初始化跨模态任务的模型，并在目标任务上进行微调，可以加快模型的收敛速度和提高性能。但是这种方法过度依赖于训练的数据规模和多样性。如果预训练数据集较小或不够多样化，预训练模型的特征表示可能无法充分适应跨模态任务的需求。Bao等（2021）选择使用配对的光学—SAR数据进行预训练，使得预训练模型更贴合目标任务。然而，目前可用于训练数据的一一配对的光学图像和SAR图像较少且难以获得，因此该方案可能是次优的。Shi等（2022）选择基于风格迁移的方法，这类方法通过风格技术，将光学图像转换为类似于SAR图像的风格和外观，以减少模态差异带来的影响。常见的风格迁移方法包括基于风格的生成器架构StyleGAN（Karras等，2019）、循环一致性生成对

抗网络CycleGAN（Zhu等，2017）等。然而，这些方法仅将光学图像转换为SAR图像，没有涉及SAR图像的物理特性。这些类SAR图像并没有表现出与真实SAR图像相同的散射特性，尤其是对于目标而言（图1）。类SAR图像中存在大量的低质量数据，会影响检测的最终结果。Guo等（2021）选用基于领域自适应的方法：这类方法旨在通过领域自适应技术，将源域（光学数据）和目标域（SAR数据）的特征分布进行对齐，以实现跨模态的知识迁移。然而，SAR和光学图像的特征空间之间存在广泛的潜在异构性，域适应方法可能难以弥合如此大的域间隙。但是不可否认的是，基于域自适应的算法为使用光学图像去辅助SAR图像进行检测提供了一个很好的思路，尤其是对于不同场景下的图像而言。



（a）光学图像
（a）Optical images

（b）SAR图像
（b）SAR images

（c）类SAR图像
（c）SAR-like images

部分目标区域

图1 同场景下光学图像、SAR图像、以及经过风格迁移之后的类SAR图像的对比图
Fig. 1 In the same scene, the comparison of optical images, SAR images, and SAR-like images after style transfer

综上所述，本文提出了一种基于多级跨模态对齐的SAR图像舰船检测算法MCMA-Net，该算法包括两个模块：一个基于邻域—全局注意力的特征交互网络NGAN以及一个多级模态对齐模块MLMA。通过对不同级别的模态特征信息采取不一样的关注策略以及从多个级别对齐不同模态间的特征，实现了利用光学模态中丰富的信息去辅助SAR图像数据进行训练。首先基于邻域—全局注意力的特征交互网络来进行特征提取，对于局部信息更加丰富的浅层特征而言，我们采用邻域注

意力机制，直接将注意力操作范围限制到了每个像素的邻域，提升了网络对局部信息的提取能力。对于全局信息比较丰富的深层特征，采用全局自注意力机制，通过对深层特征建立一种全局级别的关联，能够获取更丰富的语义信息表示。接着，为了借助光学图像中的细节信息丰富SAR图像的特征表示，提出了多级模态对齐模块，通过从局部级别到全局级别再到实例级别的对两种模态的特征进行对齐，逐步减小光学模态图像和SAR模态图像的差异性，通过在对齐过程中探索更多的模态共享特征，实现跨模态的知识传输。最后，与现阶段的先进算法在SSDD（Li等，2017）数据集和HRSID（Wei等，2020）数据集的实验结果进行对比，证明了我们的模型具有一定的鲁棒性，且取得了较优越的性能。

2 模型方法

针对由于SAR图像特殊的成像机理导致的特征不明显，以及SAR图像数据获取和标注困难导致训练样本不足等问题，提出了一种基于多级跨模态对齐的SAR图像舰船检测算法MCMA-Net，算法具体框图如图2所示。首先对于输入的图像进行特征提取，之后针对骨干网络不同级别的特征所独有的优势，对其采用不同的注意力机制。通过对浅层特征和深层特征采用不同的关注策略，提升骨干网络对不同模态不同层级特征的提取能力，挖掘更多有用信息，促进后续更好的实现跨模态对齐。接着通过采用多级模态对齐网络，尽可能的降低两种模态之间的差异性，分别对骨干网络浅层的局部特征，深层的全局特征，以及实例级的特征采用不同的对齐方式进行对齐。最后，通过上述步骤获取的信息将分别传输到边框回归以及分类子网络中进行定位和判别任务。接下来，我们将对MCMA-Net进行详细的描述。

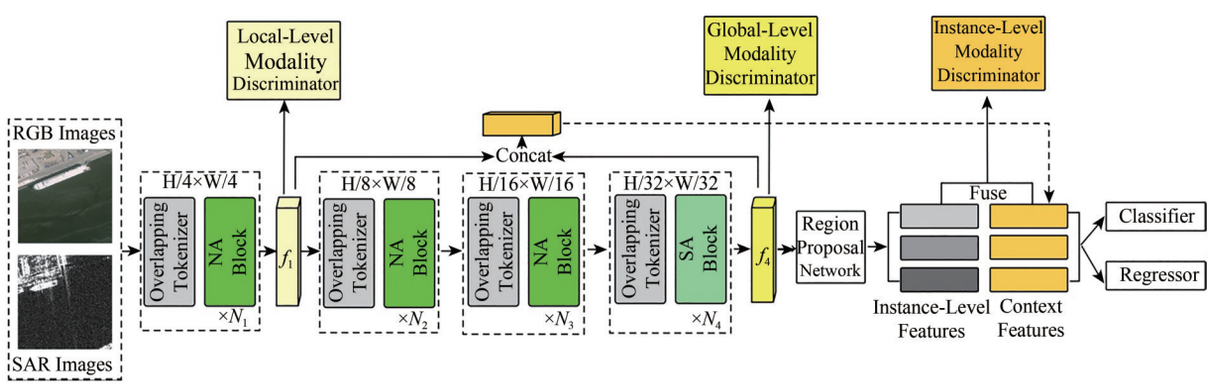


图2 MCMA-Net算法的整体流程图

Fig. 2 The overall framework of MCMA-Net algorithm

2.1 基于邻域—全局注意力的特征交互网络NGAN

现阶段的骨干网络对于浅层特征的处理还有深层特征的处理大多使用同一种方法，但是由于浅层特征和深层特征具有不同的语义信息和表达能力，这种处理方式可能并不是最优的。具体而言，在浅层网络层中，卷积和池化操作的感受野较小，只能捕捉到输入图像的局部细节和纹理等低级特征。因此，浅层特征更接近输入图像的原始信息，通常包含更多的局部信息。相比之下，深层特征具有更大的感受野，对于检测物体的整体结构、关系等全局信息更敏感，能够提供更丰富的语义表达能力。如果采用相同的处理方法来处理这些特征，可能无法充分发掘和利用它们的不同表达能力，从而限制了网络的表示能力。

针对这个问题，本文提出了一种基于邻域—全局注意力的特征交互网络NGAN，通过对浅层特征、深层特征分别采用邻域注意力机制和全局自注意力机制，能够在兼顾全局上下文建模能力的同时，提升局部特征的编码能力，同时也能充分挖掘不同模态之间的局部信息和全局信息，便于后续模态对齐。

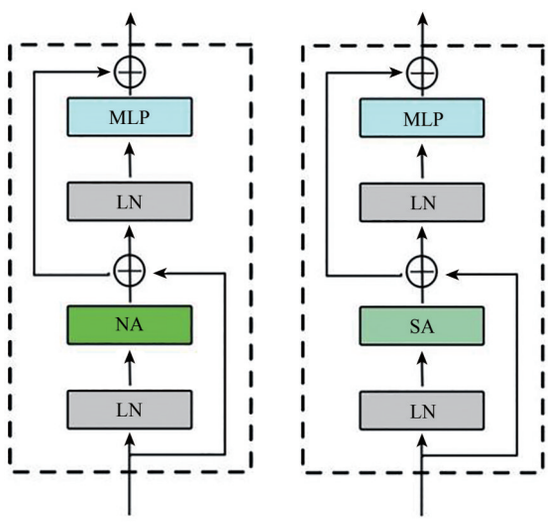
对于输入图像，首先采用两个卷积核为3、步长为2的卷积进行嵌入输入，骨干网络的总体由4个

级别组成，值得注意的是，前3个级别由多个邻域自注意力NA（Neighborhood Attention）块组成（Hassani等，2023），后一个级别由多个全局自注意力SA（Self Attention）块组成（Dosovitskiy等，2021），NA块和SA块的结构如图3所示。这些块类似于Transformer编码器层。每个层由邻域自注意力（NA）或者是全局自注意力（SA）、多层感知机MLP（Multi-layered Perceptron）、每个模块前的层归一化LN（Layer Norm）以及跳跃连接组成。每个级别后面都连接一个卷积核为3、步长为2的下采样器，除了最后一个级别的特征。通过使用下采样器，特征的空间大小减少了为原来的一半，而通道数量增加了一倍。

具体而言，对于包含更多局部信息的浅层特征（骨干网络前3个阶段），选择采用邻域注意力机制，如图4（a）所示，网络可以学习到每个像素与其邻域像素之间的依赖关系。这有助于模型更好地理解和建模像素之间的空间关系，有助于网络能更好的利用局部信息。我们令$\rho(i,\ j)$代表来$(i,\ j)$处的一个像素的相邻区域，对于$L\times L$的邻域，$\left\|\rho(i,\ j)\right\|=L^2$。因此，单个像素的邻域注意力为

$$NA\left(X_{ij}\right)=\mathrm{softmax}\left(\frac{Q_{ij}K^{\mathrm{T}}_{\rho(ij)}+B_{ij}}{\mathrm{scale}}\right)\boldsymbol{V}_{\rho(ij)} \tag{1}$$

式中，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$为变量$\boldsymbol{X}$的线性投影，$B_{i,\ j}$为相对位置偏差。将其加入每个注意力权重，依照其相对位置。最后，扩展到所有的像素$(i,\ j)$中，构成了邻域注意力。

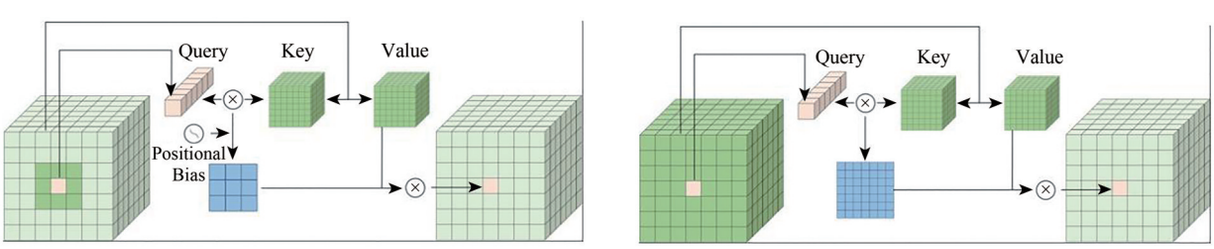


（a）NA块
（a）NA Block

（b）SA块
（b）SA Block

图3 NA块与SA块的结构示意图

Fig. 3 The structure of NA block and SA block



（a）邻域注意力
（a）Neighborhood Attention

（b）全局自注意力
（b）Global Self-attention

图4 邻域注意力和全局自注意力（对于单个像元）的query-key-value结构示意图

Fig. 4 Query-key-value structure of neighborhood attention and global self-attention (For a single pixel)

而对于包含更多全局信息的深层网络特征而言（骨干网络最后一个阶段），我们通过采用全局自注意力机制，如图4（b）所示，在深层特征中引入全局上下文信息，弥补局部信息的不足，使网络可以学习到不同区域之间的长距离依赖关系，使得特征能够更好地理解和编码目标的全局结构和语义。此时，函数ρ将每个像素映射到所有像素，即$\rho(i,\ j)$包含全部可能的像素。除此之外，这时候的$\boldsymbol{K}_{\rho(i,\ j)}=\boldsymbol{K}$，$\boldsymbol{V}_{\rho(i,\ j)}=\boldsymbol{V}$，真正实现了全局自注意力机制，通过去除偏置项，全局自注意力机制可以表示为

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

2.2 多级模态对齐模块MLMA

由于SAR图像数据获取困难以及人工标注困难导致现阶段SAR数据量要远远低于光学数据，除此之外SAR图像特殊的成像机理导致的其特征

表示不直观。因此利用少量的SAR图像数据训练出一个性能较优越的检测模型存在一定的困难。与现阶段的SAR图像检测算法不同，我们选择借助包含更多细节丰富特征信息以及数据量更加庞大的光学数据，希望利用光学特征对SAR图像的模态表示进行补偿，从而建立稳健的SAR模态特征。

受域自适应算法的启发（Saito等，2019），本文采用了一种多级模态对齐模块，通过模态对齐的方式来实现这种知识传输。但是由于两种模态之间巨大的差异性，仅仅进行一次模态对齐很难达到好的效果。因此为了能提取到更多的模态不变特征以及缩小这种模态之间的差异性最终选择通过分阶段的从3个级别对两种模态的特征进行对齐，使光学图像特征和SAR图像特征在特征空间中的分布尽可能相似。如图2所示，我们分别从局部级别、全局级别以及实例级别进行模态对齐。通过对不同级别的特征采取更适合自身特点的对齐方式，能够更合理的利用光学特征去辅助SAR图像特征，减少误差。

模态对齐操作主要是通过特征提取器和模态分类器来实现的。其中模态分类器的主要目的是通过分析输入特征的模态信息，判断特征是来自光学模态还是SAR模态。而特征提取器的任务是它通过从输入数据中学习提取模态不变的特征表示，从而利用提取有用的特征来欺骗模态判别器，使得不同模态的特征在特征空间中更加接近。在训练过程中，特征提取器和模态判别器进行博弈，特征提取器通过最小化模态判别器对特征的模态判断误差来学习模态不变的信息。同时，通过最大化对特征模态判断的准确性来使模态判别器区分不同模态的特征。通过这种训练方式，特征提取器能够生成具有高度相似性的特征表示，从而使模态判别器更难区分特征的来源模态，使得不同模态的特征在共享的特征空间中趋于对齐，以更好的实现跨模态的信息传输。

具体而言，对于浅层特征，特征往往具有较小的感受野，浅层特征上的局部信息比较丰富，因此我们对具有局部性和通用性的低级特征（骨干网络第一阶段输出的特征）采取局部对齐的方式进行处理。在局部级别，通过利用模态判别器D_{local}用来区分这些浅层的特征都来自那个模态，而我们的特征提取器就不断的提取特征来对D_{local}进行欺骗，通过这样可以使模态间的差异性降低。D_{local}是一个卷积核等于1的全卷积网络，且输入模态的特征预测图与输出模态的特征的维度是相匹配的。在训练局部级别的模态判别器的时候，我们采取的是最小二乘损失，具体表示为

$$L_{\text{local}_{\text{Opt}}}=\sum_{i,w,h} D_{\text{local}}\left(F_1^{\text{Opt}}\left(\boldsymbol{x}_i^{\text{Opt}}\right)\right)_{\text{wh}}^2 \tag{3}$$

$$L_{\text{local}_{\text{SAR}}}=\sum_{i,w,h}\left(1-D_{\text{local}}\left(F_1^{\text{SAR}}\left(x_i^{\text{SAR}}\right)\right)_{\text{wh}}\right)^2 \tag{4}$$

$$L_{\text{local}}=L_{\text{local}_{\text{Opt}}}+L_{\text{local}_{\text{SAR}}} \tag{5}$$

式中，$\boldsymbol{x}_i^{\text{T}}$代表的是输入图像，$F_1^{\text{T}}\left(x_i^{\text{T}}\right)$代表的是局部特征，也就是骨干网络第一个阶段输出的特征，$D_{\text{local}}\left(F_1^{\text{T}}\left(x_i^{\text{T}}\right)\right)_{\text{wh}}^2$代表着局部级别的模态判别器的输出，$T$代表着输入图像的模态，$w$代表宽度，$h$代表着高度。随着网络的加深，感受野的范围扩大。同时，深层的特征图中的全局信息变得更加丰富。为了更好的处理利用这些具有全局性和区分度的高层特征（骨干网络的第3阶段输出），我们在全局层面对这些特征进行对齐操作。为了减少具有特殊性的深层特征之间的差异，我们引入了一个模态判别器D_{global}，其主要任务同样也是学习区分输入特征是来自于哪个模态。通过这样的训练，模态判别器能够学习到不同模态之间的差异，并为特征对齐提供指导。D_{global}的分类损失可以表示为

$$L_{\text{global}_{\text{opt}}}=-\sum_i\left(1-D_{\text{global}}\left(F_4^{\text{Opt}}\left(\boldsymbol{x}_i^{\text{Opt}}\right)\right)\right)\log\left(D_{\text{global}}\left(F_4^{\text{Opt}}\left(\boldsymbol{x}_i^{\text{Opt}}\right)\right)\right) \tag{6}$$

$$L_{\text{global}_{\text{SAR}}}=-\sum_i D_{\text{global}}\left(F_4^{\text{SAR}}\left(\boldsymbol{x}_i^{\text{SAR}}\right)\right)\log\left(1-D_{\text{global}}\left(F_4^{\text{SAR}}\left(\boldsymbol{x}_i^{\text{SAR}}\right)\right)\right) \tag{7}$$

$$L_{\text{global}}=L_{\text{global}_{\text{opt}}}+L_{\text{global}_{\text{SAR}}} \tag{8}$$

式中，$F_4^{\text{T}}\left(x_i^{\text{T}}\right)$代表着骨干网络第3个阶段输出的特征，$D_{\text{global}}\left(F_4^{\text{T}}\left(x_i^{\text{T}}\right)\right)$代表着全局级别的模态判别器的输出。此外，由于实例级特征之间仍然存在着模态间的差异，例如外观、尺度、视角等，忽略这些差异可能会对检测结果产生不利的影响。为了解决这个问题，本文进一步进行实例级对齐，以帮助模型更准确理解光学和SAR模态之间的关系，并学习更广泛的规则，从而增强模型的泛化能力。由于实例级的特征独立地表示局部目标，缺乏对上下文整体信息的感知，有效地利用整体上下文信息可以准确地诱导实例级对齐，对后续的检测

任务也至关重要。因此，我们我们首先将上下文特征（f_1和f_4）与实例级特征（f_{ins}）进行级联融合，之后对融合的特征（f_{fus}）采用了实例级判别器D_{ins}，通过不断的通过损失对其进行优化，从而实现实例级别的严格对齐。其损失函数表示为

$$L_{\text{ins}_{\text{Opt}}} = -\sum_j \log\left(D_{\text{ins}}\left(f_{\text{fus}_j}^{\text{Opt}}\right)\right) \tag{9}$$

$$L_{\text{ins}_{\text{SAR}}} = -\sum_j \log\left(1 - D_{\text{ins}}\left(f_{\text{fus}_j}^{\text{SAR}}\right)\right) \tag{10}$$

$$L_{\text{ins}} = L_{\text{ins}_{\text{Opt}}} + L_{\text{ins}_{\text{SAR}}} \tag{11}$$

式中，$f_{\text{fus}_j}^{\text{T}}$代表着经过表示通过ROI-Pooling获取的目标建议特征，$D_{\text{ins}}\left(f_{\text{fus}_j}^{\text{T}}\right)$代表着实例级的模态判别器的输出。最后，我们方法的检测任务损失可以写为$F_{\text{det}}^{\text{T}}$。因此，MCMA-Net的整体损失表示如下：

$$L = L_{\text{det}}^{\text{Opt}} + L_{\text{det}}^{\text{SAR}} + \lambda(L_{\text{local}} + L_{\text{global}} + L_{\text{ins}}) \tag{12}$$

式中，λ表示权重因子，用于平衡检测任务和域判别器的损失。λ的默认值为0.1。

3 实验与分析

3.1 实验数据

本文实验在训练的过程中采用的是光学图像数据和SAR图像数据，网络在两个模态中是完全共享的，在测试的过程中采用的是SAR图像数据。其中，采用的光学图像数据集为HRSC2016（Liu等，2017）数据集。HRSC2016数据集由1061张光学航空影像组成，图像的尺寸从300×300到1500×900不等。采用的SAR图像数据集为SSDD（Li等，2017）数据集、HRSID（Wei等，2020）数据集、以及自制数据集SSD3。其中SSDD数据集包含了1160幅大小约为500×500的SAR图像，这些SAR图像切片中一共包括2540艘舰船目标。将训练集和测试集的数量按照8∶2进行划分，按照原数据集设定的安排，将图像名称的最后一个数字为1或9的图像指定为测试集，其余图像用于训练。HRSID数据集由5604幅SAR图像组成，分辨率分别为0.5 m、1 m、3 m。这些SAR图像切片中一共包含16951艘舰船，每幅图像具有800×800像素。对于HRSID数据集，按照原始的数据集设定，65%的图像用于训练，35%的图像用于测试。SSD3数据集由910张SAR图像组成，分辨率为1 m。这些SAR图像切片中一共包含1730艘舰船，每幅图像的大小为256×256。我们按照8∶2的比例随机划分训练集和测试集。

3.2 实验环境

所有实验均在相同的硬件平台上进行，包括GPU（GTX-3090）、CPU（Intel 4210R）和32 G内存。实验环境为PyTorch 1.10.0，CUDA 11.1和cuDNN 11.1，Python 3.7。我们在Faster R-CNN上实现了MCMA-Net，设置λ（在总体损失函数中）为0.1。为了保证比较的公平性，包括我们在内的所有船舶检测器都在MMDetection工具箱下实现，所有模块的参数均参照MMDetection工具箱的原始设置，均采用随机梯度下降SGD（Stochastic Gradient Descent）作为优化器，采用0.01的学习率、0.9的动量、0.0001的权重衰减和0.5的IoU（Intersection over Union）阈值。

3.3 评价指标

为了有效的评估本文方法的检测性能，实验主要使用了5个评估指标，即：检出率d（detection probability）、虚警率f（False alarm probability）、准确率p（precision）、召回率r（recall）和平均精度mAP（mean Average Precision）。

$$d = \frac{\text{TP}}{\text{GT}} \tag{13}$$

$$f = \frac{\text{FP}}{\text{TP} + \text{FP}} \tag{14}$$

$$p = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$r = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

式中，TP是将目标正确预测的数量，FP是将负样本错误预测为目标的数量，FN是将目标错误预测为负样本的数量，GT是正样本的数量。mAP是一种被广泛采用的评价目标检测模型有效性的性能指标。它是一种兼顾准确率和召回率的综合度量，提供了对模型准确检测物体能力的全局评价。因此，在目标检测领域，mAP常被作为首要的评价标准：

$$\text{mAP} = \int_0^1 p(r)\mathrm{d}r \tag{17}$$

3.4 实验结果分析

为了证明本文方法在跨模态特征传输上的有效性，本文在SSDD数据集上将本文实验结果与现阶段的6种检测算法：Faster R-CNN（Ren等，2015），PANET（Liu等，2018），Cascade R-CNN

(Cai 和 Vasconcelos，2018)，Double-Head R-CNN (Wu等，2020)，Grid R-CNN (Lu等，2019)，DCN (Dai等，2017) 进行对比，如表1所示。实验结果表明，本文方法 MCMA-Net 取得了优于其他几种算法的最好的实验结果：96.6% mAP。相较于第二高的 DCN 算法，本文算法能够在检测精度上高出4.4%，这表明本文算法的设计是合理的。为了进一步证明本文算法的鲁棒性及优越性，本文还在HRSID数据集上进行了实验，结果如表2所示。可以看出，与现阶段较为先进的算法相比，本文算法 MCMA-Net 仍然取得了最好的检测精度，精度可以达到87.4%。与精度第二高的检测算法 DCN 相比，我们的精度提升了5.4%。

表1 不同的算法在SSDD数据集上的实验结果

Table 1 Experimental results of different algorithms on SSDD dataset

方法	d	f	p	r	mAP/%
Faster R-CNN	0.904	0.130	0.870	0.904	89.7
PANET	0.919	0.132	0.868	0.919	91.1
Cascade R-CNN	0.908	0.059	0.941	0.908	90.5
Double-Head R-CNN	0.919	0.131	0.869	0.919	91.1
Grid R-CNN	0.897	0.123	0.877	0.897	88.9
DCN	0.930	0.138	0.862	0.930	92.2
MCMA-Net	0.979	0.153	0.847	0.979	96.6

表2 不同的算法在HRSID数据集上的实验结果

Table 2 Experimental results of different algorithms on HRSID dataset

方法	d	f	p	r	mAP/%
Faster R-CNN	0.819	0.186	0.814	0.819	80.6
PANET	0.829	0.188	0.812	0.829	81.6
Cascade R-CNN	0.822	0.129	0.871	0.822	81.2
Double-Head R-CNN	0.833	0.183	0.817	0.833	82.0
Grid R-CNN	0.809	0.172	0.828	0.809	79.4
DCN	0.834	0.186	0.814	0.834	82.0
MCMA-Net	0.889	0.196	0.804	0.889	87.4

与此同时，我们还比较了本文算法与其他算法的检出率与虚警率，通过表1和表2可以看出，虽然我们的算法的虚警率并不是最低的，但是与其他算法相较而言，差距并不明显。其中，在SSDD数据集上，本文算法的虚警率仅比基线模型 Faster R-CNN 高2.3%，仅比检测精度第二的 DCN 高1.5%。在HRSID数据集上，本文算法的虚警率仅比基线模型 Faster R-CNN 高1%，仅比检测精度第二的 DCN 高1%。但是与其他所有算法相比，MCMA-Net 的检出率有大幅的提升。其中，在SSDD数据集上，本文算法的检出率比基线模型 Faster R-CNN 高7.5%，比检测精度第二的 DCN 高4.9%。在HRSID数据集上，我们的算法的检出率比基线模型 Faster R-CNN 高7.0%，比检测精度第二的 DCN 高5.5%。因为检出率的提升必然会带来误检导致虚警率增大，所以在虚警率差别不大的同时能大幅的提升检出率，证明本文算法的性能更优越。图5和图6是基于所有算法在SSDD数据集和HRSID数据集上的检出率和虚警率绘制得到的ROC曲线，可以看出本文算法 MCMA-Net 对应的粉色曲线在相同虚警率的情况下检出率最高，在虚警率极大值相差不大的情况下，检出率极大值点均远大于其他曲线，具有明显的优势。

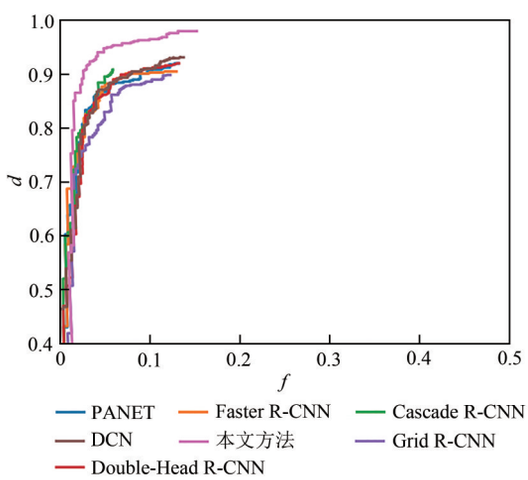


图5 不同算法在SSDD数据集上的ROC曲线

Fig. 5 ROC curves of different algorithms on SSDD dataset

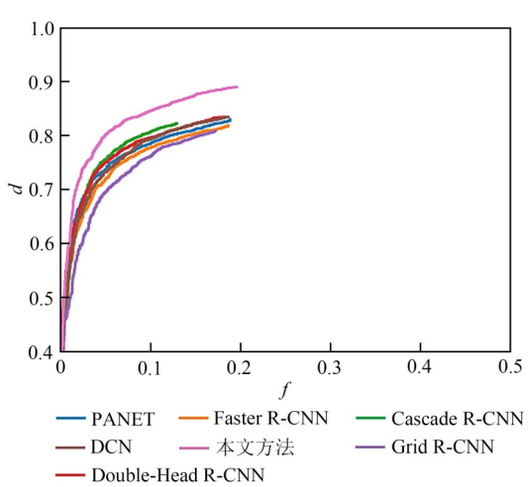


图6 不同算法在HRSID数据集上的ROC曲线

Fig. 6 ROC curves of different algorithms on HRSID dataset

为了进一步验证我们提出的算法的实际应用价值，也在自制数据集SSD3上进行了定量实验，实验结果如表3所示。可以看出，我们的算法MCMA-Net取得了最优的检测性能，检测精度达到了89.2%，比基线模型Faster R-CNN高10.9%。与精度第二高的算法PANET相比，我们的精度提升了9.3%。更值得注意的是，在SSD3数据集上，我们的算法MCMA-Net同时拥有最高的检出率和最低的虚警率，远远优于其他所有算法，证明了我们的算法具有不错的鲁棒性。

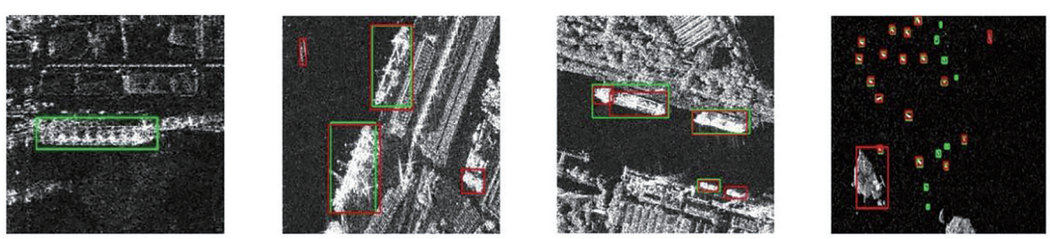
表3 不同的算法在SSD3数据集上的实验结果

Table 3 Experimental results of different algorithms on SSD3 dataset

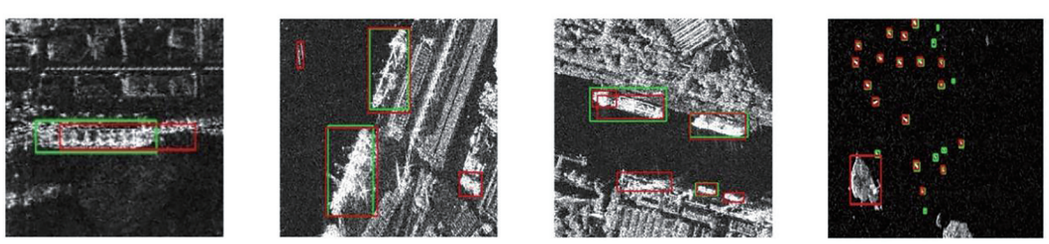
方法	*d*	*f*	*p*	*r*	mAP/%
Faster R-CNN	0.837	0.473	0.527	0.837	78.3
PANET	0.853	0.416	0.584	0.853	79.9
Cascade R-CNN	0.803	0.285	0.715	0.803	78.4
Double-Head R-CNN	0.837	0.357	0.643	0.837	79.6
Grid R-CNN	0.825	0.505	0.495	0.825	76.0
DCN	0.850	0.419	0.581	0.850	78.9
MCMA-Net	0.895	0.153	0.847	0.895	89.2

由这些实验结果可以得出结论，本文设计的基于邻域—全局注意力的特征交互网络能够更好的挖掘不同级别的特征信息以取得更好的检测结果。除此之外，还说明了我们设计的多级模态对齐模块实现了利用光学信息去辅助SAR图像进行检测，并有益于检测效果的提升。

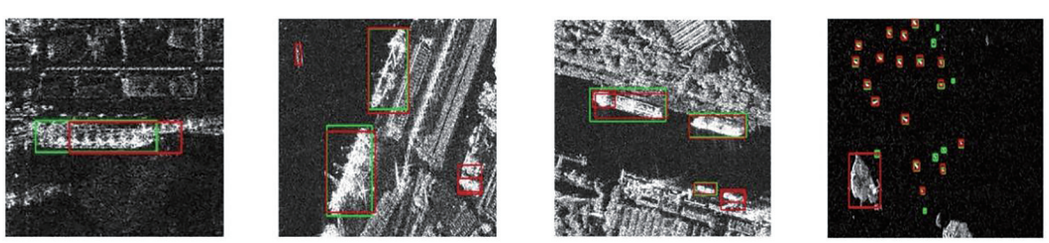
除此之外，为了更直观的看出本文检测算法的效果，如图7所示，我们将Faster R-CNN算法、Double-Head R-CNN、DCN算法与我们的算法MCMA-Net的测试结果进行了可视化。可以看出，无论是大尺度目标还是小尺度目标，对于其他算法检测效果较差的那些，我们的算法几乎都能精准的将其检测出来，这得益于我们设计的基于邻域—全局注意力的特征交互网络能够更有效的提取不同阶段的特征信息。除此之外，在复杂场景下，本文方法对目标的定位也更加准确，且误检也大大降低，这表明通过迁移光学信息去增强SAR图像的特征表示，有助于网络更好的理解SAR模态的特征，学习出一个更强大的模型。



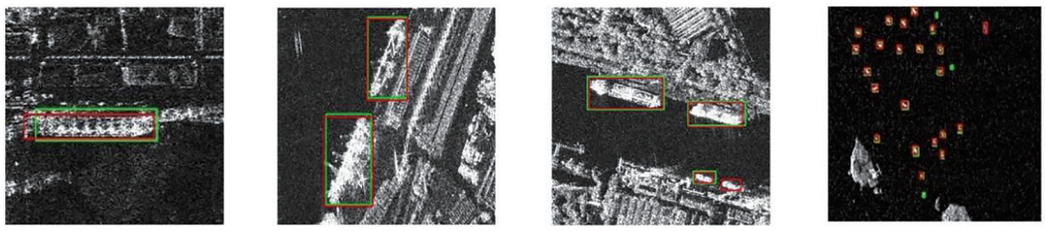
(a) Faster R-CNN



(b) Double-Head R-CNN



(c) DCN



(d) MCMA-Net

目标的真实位置 检测结果

图7 本文算法与其他3种算法的可视化结果图

Fig. 7 Visual detection results of our method and some comparison methods on SSDD

3.5 消融实验

为了进一步证明我们设计的模块的有效性，我们对提出的MCMA-Net算法进行了消融实验，结果如表4所示，首先，我们评估了基于邻域—全局注意力的特征交互网络NGAN的性能，可以看出，相较于基础模型而言，在加入NGAN模块之后，总体检测精度提升了1.7%，更直接的证明了NGAN模块能够更有效的挖掘和利用SAR图像特征信息。与此同时，相较于基础模型而言，在只加入MLMA模块的时候，总体检测精度提升了0.9%，这证明了MLMA很好的实现了将光学模态特征迁移到SAR模态中，并有助于提升网络的性能。最后，当两个模块一起使用的时候发现并没有出现互斥的效果，也就是整体的检测结果仍然是提升的，相较于基础模型而言，提升了2.7%的检测精度，从而有效的证明了我们的算法的合理性。

表4 MCMA-Net的消融实验结果

Table 4 Ablation experimental results of MCMA-Net

NGAN	MLMA	*d*	*f*	*p*	*r*	mAP/%
×	×	0.953	0.155	0.845	0.953	93.9
√	×	0.968	0.179	0.821	0.968	95.6
×	√	0.961	0.172	0.828	0.961	94.8
√	√	0.979	0.153	0.847	0.979	96.6

注：“√”表示使用相应模块；“×”表示没有使用相应模块。

除此之外，为了证明浅层特征与深层特征之间存在互补关系，我们也进行了一组消融实验，即只对浅层特征进行跨模态学习、只对深层特征进行跨模态学习、以及同时对浅层特征和深层特征进行跨模态学习，如表5所示。实验结果表明在没有采用NGAN模块（基于邻域—全局注意力的特征交互网络）的情况下，仅对浅层特征进行跨模态学习的mAP为94.3%，仅对深层特征进行跨模态学习的mAP为94.2%，同时对浅层特征和深层特征进行跨模态学习的mAP为94.6%，检测精度高于前两种情况。由此可见对深浅层特征同时进行跨模态对齐的效果最好，即浅层特征和深层特征存在互补性，更进一步的证明了我们设计的跨模态算法MCMA-Net的合理性。

表5 对深、浅层特征进行跨模态学习的实验结果

Table 5 Experimental results of cross-modality learning for high-level and low-level features

浅层	深层	*d*	*f*	*p*	*r*	mAP/%
×	×	0.953	0.155	0.845	0.953	93.9
×	√	0.955	0.166	0.834	0.955	94.2
√	×	0.959	0.167	0.833	0.959	94.3
√	√	0.961	0.167	0.833	0.961	94.6

注：“√”表示使用相应模块；“×”表示没有使用相应模块。

4 结 论

本文提出了一种基于多级模态对齐的SAR图像舰船检测算法MCMA-Net，通过将光学模态中更为丰富的知识迁移到SAR模态，有效的解决了由于SAR图像数据量少且特征表示不直观带来的问题。该算法首先采用基于邻域—全局注意力的特征交互网络NGAN，对骨干网络的特征表现不同的浅层特征和深层特征采取不同的注意力机制，提升了骨干网络对不同模态特征的提取性能，充分挖掘最具代表性的模态特征，有助于后续不同模态的对齐效果。接着采取多级模态对齐模块，通过逐步探索光学模态与SAR模态之间的模态不变表示，学习到更加丰富的特征表示，使得我们能够更好地利用光学模态的丰富特征来弥补SAR图像的特征表示的不足。最终进行实验验证所提出算法的有效性，实验表明，与现阶段算法相比，本文提出的算法能达到最佳的检测性能，更具有

优越性。在未来研究中，将致力于提升本文算法对复杂场景下小目标的检测精度，在更具有挑战性的数据集上探索所提出方法的性能。

参考文献（References）

Bao W, Huang M Y, Zhang Y Q, Xu Y, Liu X J and Xiang X S. 2021. Boosting ship detection in SAR images with complementary pretraining techniques. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 8941-8954 [DOI: 10.1109/JSTARS.2021.3109002]

Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: delving into high quality object detection//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6154-6162 [DOI: 10.1109/CVPR.2018.00644]

Cao Q, Ma A L, Zhong Y F, Zhao J, Zhao B and Zhang L P. 2019. Urban classification by multi-feature fusion of hyperspectral image and LiDAR data. Journal of Remote Sensing, 23(5): 892-903 (曹琼, 马爱龙, 钟燕飞, 赵济, 赵贝, 张良培. 2019. 高光谱-LiDAR多级融合城区地表覆盖分类. 遥感学报, 23(5): 892-903) [DOI: 10.11834/jrs.20197512]

Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16x16 words: transformers for image recognition at scale//9th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

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

Guo Y C, Du L and Lyu G X. 2021. SAR target detection based on domain adaptive faster R-CNN with small training data size. Remote Sensing, 13(21): 4202 [DOI: 10.3390/rs13214202]

Hassani A, Walton S, Li J C, Li S and Shi H. 2023. Neighborhood attention transformer//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 6185-6194 [DOI: 10.1109/CVPR52729.2023.00599]

Hou W and Li Y. 2023. Multi-resolution CFAR target detection algorithm and accuracy analysis based on SAR image data. Beijing Survey and Mapping, 37(1): 104-109 (侯卫, 李勇. 2023. 基于SAR影像数据的多分辨率CFAR目标检测算法及精度分析. 北京测绘, 37(1): 104-109) [DOI: 10.19580/j.cnki.1007-3000.2023.01.019]

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]

Li J W, Qu C W and Shao J Q. 2017. Ship detection in SAR images based on an improved faster R-CNN//2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA). Beijing, China: IEEE: 1-6 [DOI: 10.1109/BIGSARDATA.2017.8124934]

Li W, Wang J J, Gao Y H, Zhang M M, Tao R and Zhang B. 2022. Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 5526914 [DOI: 10.1109/TGRS.2022.3166252]

Li Y, Ding Z G, Zhang C, Wang Y and Chen J. 2019. SAR ship detection based on resnet and transfer learning//IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama, Japan: IEEE: 1188-1191 [DOI: 10.1109/IGARSS.2019.8900290]

Lin T Y, Goyal P, Girshick R, He K M and Dollar P. 2017. Focal loss for dense object detection//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Lin Z, Ji K F, Leng X G and Kuang G Y. 2019. Squeeze and excitation rank faster R-CNN for ship detection in SAR images. IEEE Geoscience and Remote Sensing Letters, 16(5): 751-755 [DOI: 10.1109/LGRS.2018.2882551]

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

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

Liu Z K, Yuan L, Weng L B and Yang Y P. 2017. A high resolution optical satellite image dataset for ship recognition and some new baselines//Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2017. Porto, Portugal: SciTePress: 324-331

Lu X, Li B Y, Yue Y X, Li Q Q and Yan J J. 2019. Grid R-CNN//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7355-7364 [DOI: 10.1109/CVPR.2019.00754]

Miao T, Zeng H C, Yang W, Chu B C, Zou F, Ren W J and Chen J. 2022. An improved lightweight retinaNet for ship detection in SAR images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 4667-4679 [DOI: 10.1109/JSTARS.2022.3180159]

Pappas O, Achim A and Bull D. 2018. Superpixel-level CFAR detectors for ship detection in SAR imagery. IEEE Geoscience and Remote Sensing Letters, 15(9): 1397-1401 [DOI: 10.1109/LGRS.2018.2838263]

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

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement. arXiv: 1804.02767

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 91-99

Saito K, Ushiku Y, Harada T and Saenko K. 2019. Strong-weak distribution alignment for adaptive object detection//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6949-6958 [DOI: 10.1109/CVPR.2019.00712]

Shi Y, Du L, Guo Y C and Du Y. 2022. Unsupervised domain adaptation based on progressive transfer for ship detection: from optical to SAR images. IEEE Transactions on Geoscience and Remote Sensing, 60: 5230317 [DOI: 10.1109/TGRS.2022.3185298]

Wang J J, Li W, Gao Y H, Zhang M M, Tao R and Du Q. 2023b. Hyperspectral and SAR Image Classification via Multiscale Interactive Fusion Network. IEEE Transactions on Neural Networks and Learning Systems, 34(12): 10823-10837 [DOI: 10.1109/TNNLS.2022.3171572]

Wang J J, Li W, Wang Y J, Tao R and Du Q. 2023c. Representation-enhanced status replay network for multisource remote-sensing image classification. IEEE Transactions on Neural Networks and Learning Systems, 1-13 [DOI: 10.1109/TNNLS.2023.3286422]

Wang S Y, Cai Z C and Yuan J Y. 2023a. Automatic SAR ship detection based on multifeature fusion network in spatial and frequency domain. IEEE Transactions on Geoscience and Remote Sensing, 61: 4102111 [DOI: 10.1109/TGRS.2023.3267495]

Wei S J, Zeng X F, Qu Q Z, Wang M, Su H and Shi J. 2020. HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation. IEEE Access, 8: 120234-120254 [DOI: 10.1109/ACCESS.2020.3005861]

Wu Y, Chen Y P, Yuan L, Liu Z C, Wang L J, Li H Z and Fu Y. 2020. Rethinking classification and localization for object detection// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10183-10192 [DOI: 10.1109/CVPR42600.2020.01020]

Yao Y Q, Cheng G, Xie X X and Han J W. 2021. Optical remote sensing image object detection based on multi-resolution feature fusion. Journal of Remote Sensing, 25(5): 1124-1137 (姚艳清, 程塨, 谢星星, 韩军伟 . 2021. 多分辨率特征融合的光学遥感图像目标检测 . 遥感学报, 25(5): 1124-1137) [DOI: 10.11834/jrs.20210505]

Yu Y, Ai H, He X J, Yu S H, Zhong X and Zhu R F. 2020. Attention based feature pyramid networks for ship detection of optical remote sensing image. Journal of Remote Sensing, 24(2): 107-115 (于野, 艾华, 贺小军, 于树海, 钟兴, 朱瑞飞 . 2020. A-FPN 算法及其在遥感图像船舶检测中的应用 . 遥感学报, 24(2): 107-115) [DOI: 10.11834/jrs.20208264]

Zhang F, Lu S T, Xiang D L and Yuan X Z. 2023. An improved superpixel-based CFAR method for high-resolution SAR image ship target detection. Journal of Radars, 12(1): 120-139 (张帆, 陆圣涛, 项德良, 袁新哲 . 2023. 一种改进的高分辨率SAR图像超像素 CFAR 舰船检测算法 . 雷达学报, 12(1): 120-139) [DOI: 10.12000/JR22067]

Zhang M M, Li W, Zhang Y X, Tao R and Du Q. 2023. Hyperspectral and LiDAR data classification based on structural optimization transmission. IEEE Transactions on Cybernetics, 53(5): 3153-3164 [DOI: 10.1109/TCYB.2022.3169773]

Zhang Y, Wang X Q, Jiang Z Z, Li G and He Y. 2022. An efficient center-based method with multilevel auxiliary supervision for multiscale SAR ship detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 7065-7075 [DOI: 10.1109/JSTARS.2022.3197210]

Zhao Y, Zhao L J, Xiong B L and Kuang G Y. 2020. Attention receptive pyramid network for ship detection in SAR images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 2738-2756 [DOI: 10.1109/JSTARS.2020.2997081]

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhou P C, Cheng G, Yao X W and Han J W. 2021. Machine learning paradigms in high-resolution remote sensing image interpretation. Journal of Remote Sensing, 25(1): 182-197 (周培诚, 程塨, 姚西文, 韩军伟 . 2021. 高分辨率遥感影像解译中的机器学习范式 . 遥感学报, 25(1): 182-197) [DOI: 10.11834/jrs.20210164]

From optical to SAR: A SAR ship detection algorithm based on multi-level cross-modality alignment

HE Jiayue[1], SU Nan[1], XU Cong'an[2], YIN Lu[3], LIAO Yanping[1], YAN Yiming[1]

1. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China;
2. Research Institute of Information Fusion, Naval Aviation University, Yantai 264001, China;
3. Beijing Institute of Remote Sensing Information, Beijing 100192, China

Abstract: In recent years, interest in Synthetic Aperture Radar (SAR) ship detection has considerably grown. Its distinctive strengths position it as a pivotal player in numerous fields of research. However, the inherent characteristics of SAR images have presented a range of challenges. For instance, in contrast to optical images, SAR images have counterintuitive feature representation. Additionally, owing to the

constrained number of SAR image data, achieving satisfactory results with existing methods that depend on a substantial number of annotated SAR images might be challenging.

How to effectively train a high-performance SAR ship detection network with a limited quantity of SAR images should be investigated. Given that single-modality SAR detection algorithms have inherent limitations, other effective modalities that can assist the SAR modality in completing tasks are needed. For instance, in SAR image target detection, optical images can serve as supplementary data sources. A knowledge-rich model can be developed by utilizing a large volume of optical data in training with SAR data. Hence, reasonable training approaches for effectively utilizing images from SAR and optical modalities should be explored.

To address these challenges, a SAR ship detection algorithm called MCMA-Net, which is based on multilevel cross-modality alignment, is proposed in this paper. The MCMA-Net enriches SAR feature representation by incorporating valuable knowledge from optical modality. First, we propose a neighborhood–global attention-based feature interaction network (NGAN), which employs a neighborhood attention mechanism that enables the local interaction of low-level features and a global self-attention mechanism that captures global context from high-level features. When the ability of global context modeling is considered, the encoding ability of local features improves, NGAN enables the network to focus on corresponding information at different levels and can promote the subsequent multilevel modality alignment. Second, we propose a multilevel modality alignment module (MLMA), which aligns features in the different hidden spaces of the two modalities from three levels. MLMA facilitates the model to acquire modality-invariant features, bridging the modality gap and realizing optical knowledge transmission. Valuable information from the optical modality can compensate for certain deficiencies in SAR images. With the aid of these two modules, we have incorporated optical superiority information by leveraging SAR's inherent advantages, achieving an enhancement in the performance of SAR detection tasks.

Our algorithm is superior to current detection algorithms. Notably, whether on public SAR image datasets or our own SAR image dataset, the MCMA-Net consistently achieves optimal detection results, which indicates the model's stable performance and robustness. The visualization results indicate that the MCMA-Net achieves excellent detection capabilities in complex scenarios. The ablation experiments demonstrate that compared with the baseline model, our algorithm achieved a 2.7% increase in mAP on the SSDD dataset. Various experimental results have consistently validated the rationality of the MCMA-Net.

Key words: remote sensing, SAR, target detection, cross-modality, feature alignment, attention mechanism

Supported by National Natural Science Foundation of China (No. 62271159, 62071136, 62002083, 61971153); Heilongjiang Outstanding Youth Foundation (No. YQ2022F002); Heilongjiang Postdoctoral Foundation (No. LBH-Q20085, LBH-Z20051); Fundamental Research Funds for the Central Universities Grant (No. 3072022QBZ0805, 3072021CFT0801, 3072022CF0808); Industrial Demonstration of National Security Monitoring and Integrated Services in the High-Resolution Special Project for Sino-Russian Border Areas (No. 72-Y50G11-9001-22/23)