GCRPNet：用于光学遥感图像中显著目标检测的图增强上下文和区域感知网络

Mengyu Ren†, Yutong Li†, Hua Li*, *Member, IEEE*, Runmin Cong, *Senior Member, IEEE*, Sam Kwong, *Fellow, IEEE*

***Abstract*—光学遥感图像中的显著目标检测面临着众多挑战，包括目标尺度的显著变化以及目标与背景之间的低对比度。基于视觉变换器（ViTs）和卷积神经网络（CNNs）架构的现有方法旨在利用全局和局部特征，但难以有效整合这些异构特征限制了它们的整体性能。为克服这些限制，我们提出了一种图增强的上下文和区域感知网络（GCRPNet），该网络基于 Mamba 架构以同时捕获长距离依赖性和增强区域特征表示。具体而言，我们采用视觉状态空间（VSS）编码器来提取多尺度特征。为了进一步实现对这些特征的深度引导和增强，我们首先设计了一个差异-相似性引导的层次图注意模块（DS-HGAM）。该模块增强了不同尺度特征之间的跨层交互能力，同时增强了模型的结构感知能力，使其能够更有效地区分前景和背景。随后，我们设计了 LEVSS 模块作为 GCRPNet 的解码器。该模块整合了我们提出的自适应扫描策略和多粒度协作注意力增强模块（MCAEM）。它对经过多尺度卷积处理的特征图进行自适应补丁扫描，从而捕捉丰富的局部区域信息，增强了 Mamba 的局部建模能力。大量实验结果表明，所提出的模型达到了最先进的性能，验证了其有效性和优越性。**

***Index Terms*—Optical remote sensing images, salient object detection, visual state space model**

显著目标检测（SOD）是计算机视觉领域的一项基本任务，旨在识别和分割场景中最引人注目的物体或区域。作为一个关键的前置视觉任务，SOD 的研究已广泛应用于物体跟踪、图像分割和质量评估等领域。传统上，SOD 方法主要集中在自然场景图像，并在该领域取得了显著成果。近年来，随着卫星遥感技术的不断进步，光学遥感图像中的显著目标检测越来越受到研究人员的关注。与自然图像相比，光学遥感图像为显著目标检测带来了许多挑战，如低的目标与背景对比度、显著的目标尺度变化以及复杂的物体结构。这些因素使得许多在自然图像上表现良好的 SOD 方法难以直接转移到遥感情景中。因此，研究人员开发了针对遥感图像特征优化的专用模型，旨在有效检测光学遥感图像中的显著目标。

传统的光学遥感图像显著目标检测（ORSI-SOD）方法 [1]–[6] ，通常依赖于颜色先验和人工设计的特征，如位置线索和空间结构来识别显著目标。然而，由于缺乏可学习性，这些方法难以适应遥感场景的复杂和多变特性，从而限制了它们的性能。

基于深度学习的 ORSI-SOD 方法在过去十年中取得了显著进展。卷积神经网络（CNNs）和视觉变换器（ViTs）的引入，极大地提升了 ORSI-SOD 模型的性能。CNNs 具有局部感受野，提供了高计算效率和对尺度变化的强适应性，使其在 ORSI-SOD 任务中得到广泛应用。例如，SFANet [7] 通过渐进流和不确定性感知细化增强了多尺度语义表示。LiteSalNet [8] 通过结合轻量级注意力模块和多流解码器以实现精确的显著性检测，专注于效率。然而，由于 CNNs 固有的局部滑动窗口操作，这些方法往往缺乏对全局语义依赖关系的建模能力。如图中 LiteSalNet [8] 和 SFANet [7] 在图 **??** 中的描述，这种限制可能会导致分割错误和不精确的边界，尤其是在处理包含低对比度和大尺寸物体的遥感图像时。

为了解决这个问题，研究人员将 ViTs 引入到 ORSI-SOD 任务中，以增强模型的全局信息建模能力。ViTs 利用自注意力机制进行全局建模，有效地弥补了 CNN 的感受野限制。它们强大的全局建模能力使其在涉及复杂遥感图像的分割任务中特别具有优势。例如，Gao 等人 [9] 提出了一种新颖的自适应空间分割 Transformer 框架，能够生成高质量的显著性图。同时，HFANet [10] 集成了结合 CNN 和 Transformer 的混合编码器，以获取包含全局语义关系和局部空间细节的特征，从而在复杂背景中实现对显著对象的精准检测。然而，ViTs 在建模空间结构方面缺乏局部归纳偏置 [11], [12] ，使其在表现细粒度目标细节时困难重重，常常导致边界模糊的分割结果。在混合架构中 [10], [13], [14] ，从理论上讲，CNN 的局部细节建模和 ViTs 的全局关系表达应该是互补的。然而，相关研究 [15]–[17] 表明，仅仅将两者结合并不能实现预期的协同效益。这主要是由于它们在特征表示、感受野机制和输出范式方面存在显著差异。如图 Fig. **??** 中的 HFANet [10] 所示，这些差异导致了语义不一致和融合噪声，从而阻碍了有效的集成，最终降低了整体模型的性能。

为了有效解决上述挑战，我们提出了一种新的网络架构，称为图增强的上下文和区域感知网络（GCRPNet），该网络特别设计用于解决 ORSI-SOD 任务中的问题，如低物体背景对比度和物体尺度变化大等问题。GCRPNet 基于传统的编码器-解码器架构，采用 Mamba 模型 [18], [19] 作为其核心建模主干。主要动机在于 Mamba 使用选择性扫描算法来捕捉远程依赖，同时具有良好的稳定性和可扩展性，这使得在全局上下文和局部区域细节之间实现有效的建模平衡成为可能。为了应对 ORSI 中目标边界模糊和显著目标与背景对比度低的挑战，我们提出了一种差异-相似性引导的分层图注意力模块（DS-HGAM）。该模块在跳跃连接处集成多尺度特征，并利用图神经网络 [20] 来建模融合特征之间的空间关系和上下文依赖性。这样，我们的模型可以更好地学习前景和背景之间的结构差异，从而减少由边界模糊引起的分割错误。同时，为了解决多尺度物体变化的问题，并减轻 Mamba 扫描过程对局部区域结构造成的干扰以增强其局部建模能力，我们引入了局部增强视觉状态空间（LEVSS）模块。该模块首先将特征输入到多粒度协同注意力增强模块（MCAEM）中，以提取多尺度

† equal contribution, * corresponding author.

Mengyu Ren and Hua Li are with the School of Computer Science and Technology, Hainan University, Haikou, 570228 China.

Yutong Li is with the School of Information and Communication Engineering, Hainan University, Haikou, 570228 China.

Runmin Cong is with the School of Control Science and Engineering, Shandong University, Jinan, 250061 China.

Sam Kwong is with the School of Data Science, Lingnan University, Hong Kong, China.

局部信息。随后进行局部增强二维选择性扫描（LESS2D）操作。与传统的扫描策略不同，LESS2D 根据空间分辨率自适应地将特征图划分为块，并在每个块中进行方向性扫描，以建模局部像素之间的依赖关系。此方法不仅加强了模型表示多尺度特征的能力，提升了其对 ORSI 中尺度多样目标的适应性，而且通过块划分、方向性扫描和整合的过程，使模型能够增强局部区域特征表示的同时实现全局上下文建模。局部细节建模能力的这一提升进一步提升了模型在复杂场景中的分割准确性。我们在两个广泛使用的基准数据集（即 ORRSD [21] 和 EORRSD [22] ）上进行了广泛的对比实验，结果表明 GCRPNet 在 ORSI-SOD 任务上显著优于现有的最新方法。我们的主要贡献可以总结如下：

早期的 SOD 方法主要依赖于手工制作的低级特征和启发式先验，例如颜色对比、边界连通性和中心-环绕假设。这些方法在简单背景中可以实现合理的性能，但由于缺乏对图像内容的语义理解，它们的有效性在复杂场景或多对象背景中经常受到限制。近年来，基于深度学习的 SOD 方法逐渐成为主流。这些方法通常采用端到端网络架构，从大型标注数据集中自动学习显著性表示，导致检测精度和细粒度性能的显著提升。

CNNs 在 SOD 领域取得了显著的进展，这归功于其强大的多层次特征学习能力。Li 等人 [23] 通过利用从深度卷积网络中提取的多尺度特征，结合空间一致性增强和多层图像分割融合技术，显著提高了视觉显著性检测性能。Sun 等人 [24] 提出了一种新颖的 U 形网络架构，有效整合了多尺度特征和边缘信息，以解决现有方法中常见的边界模糊和定位不准确等问题。Zhao 等人 [25] 同时建模显著对象和显著边缘信息，使用边缘线索引导显著对象特征的融合，从而实现更精确的定位和边界检测。由于其优越的全局建模能力和捕捉长距离依赖关系的能力，基于 Transformer 的模型被越来越多地应用于 NSI-SOD 任务。Liu 等人 [26] 提出了一种纯粹基于 Transformer 的架构，从序列到序列的角度重新思考 SOD 任务，通过建模全局长距离依赖关系来增强检测性能。Zhang 等人 [27] 介绍了一种基于生成视觉 Transformer 的新框架，结合潜在变量和基于能量的先验分布来生成与人类感知一致的不确定性图，从而显著改善检测性能。Liu 等人 [28] 提出了 TriTransNet，从多尺度特征融合的角度重新考虑 RGB-D SOD 任务。通过建模跨层次特征的长距离依赖关系，并用深度信息增强 RGB 表示，该模型实现了更精确的显著对象定位和边界划分。最近，Mamba 因其在建模长距离依赖关系方面的强大能力而在显著性目标检测（SOD）领域受到关注。Zhang 等人 [29] 引入了一种基于 Mamba 的双流特征提取器和跨模态融合模块，有效整合了 RGB 和深度信息，提高了显著性目标检测性能。He 等人 [30] 提出了一种基于 Mamba 架构的新型统一框架，名为 Samba，在各种 SOD 任务中实现了优异的性能和低计算成本。

虽然 SOD 在自然场景图像（NSIs）上取得了显著进展，但直接将现有方法应用于遥感图像（ORSIs）仍然面临众多挑战。NSI-SOD 方法主要是针对从地面视角的自然场景设计的，通常依赖于明确的物体边界、丰富的纹理和突出的语义上下文。然而，遥感图像中的物体常常表现出小尺寸、低对比度、类间差异小和背景干扰强等特征，使现有的基于 NSI 的方法难以有效地在遥感场景中进行泛化。为了解决这些问题，近年来出现了一系列专门为 ORSIs 量身定制的 SOD 方法，即 ORSI-SOD，旨在提升复杂遥感环境中的检测性能。

得益于其强大的语义表示和局部特征提取能力，卷积神经网络（CNNs）能够在光学遥感图像（ORSIs）中精确捕捉细粒度的局部特征。近年来，许多基于 CNN 的 ORSI-SOD 模型取得了显著进展。LV-Net [21] 利用双流金字塔模块和嵌套编码器-解码器结构来有效处理 ORSIs 中的多尺度和复杂背景挑战。DAFNet [22] 将低层注意力线索传递到深层，以生成精确的高层注意力图，有效抑制背景噪声并突出显著对象。ACCoNet [31] 通过协调相邻特征和捕捉上下文信息来增强显著性检测，从而更全面地激活显著区域。MCP-Net [32] 引入了一个记忆模块，以利用图像间和图像内的上下文信息，显著提高了光学 RSIs 中显著对象的检测性能。同样，基于 Transformer 的方法越来越多地应用于 ORSI-SOD 任务，旨在克服 CNN 在处理复杂背景、多尺度目标和光学遥感图像弱边界方面的局限性。TLCKD-Net [33] 采用基于 Transformer 的编码器来有效捕捉全局特征，解决传统方法在处理复杂背景和不同尺度、形状的显著对象时的不足。GeleNet [34] 采用基于 Transformers 的全局到局部范式，将全局特征提取与局部增强模块相结合，更精确地定位显著对象并突出细节。然而，纯 ViT 架构在建模局部信息方面有限，而光学遥感图像中的目标通常显示出多尺度和复杂形状，这容易导致不精确的边界分割。为了解决这个问题，研究人员提出了混合架构。ASNet [13] 通过双分支编码器整合了 Transformer 和 CNN 技术，捕捉全局依赖和局部细节特征，适应 RSI 的复杂性并提高检测性能。ADSTNet [14] 结合了 CNN 和 Transformers 的优势，有效地补偿全局和局部信息，边界引导模块进一步有助于准确检测和定位。

最近，状态空间模型（SSMs）的出现提供了一种新的全局建模范式，特别是在 Mamba 模型的引入中。Mamba 通过引入选择性扫描机制增强了全局建模能力，并采用并行扫描策略，有效捕捉图像中的全面上下文信息，从而更好地表示长距离依赖关系。基于这一基础，Mamba 在计算机视觉领域迅速发展，产生了多种视觉 Mamba 架构。尽管 Mamba 在视觉应用中表现出令人印象深刻的潜力，它仍然面临局部建模的挑战。在当前的视觉 Mamba 模型中，常见的做法是将二维特征图分割成补丁，并在输入 SSMs 进行处理之前将其扫描成一维序列。然而，图像本质上缺乏序列结构，这种序列化与图像数据的空间局部性相冲突。结果是邻近像素之间的依赖关系被扰乱，导致对细粒度结构（如边缘和纹理）的建模不足。这最终阻碍了 SSMs 生成高质量的视觉特征。

作为视觉领域 Mamba 架构的关键组成部分，高效的扫描机制在提升模型性能方面发挥着重要作用，并已成为当前有关 Mamba 的研究的重点。Vision Mamba 采用双向扫描机制来同时处理前向和后向信息。VMamba [35] 引入了一种交叉扫描机制，通过沿四个不同路径对图像进行扫描，以实现更全面的特征提取，并已得到广泛采用。Hilbert 扫描策略 [36] 结合了多种扫描方法和方向，使其能够从多个角度捕捉局部特征信息。LocalMamba [37] 模型通过将图像划分为多个不重叠的局部窗口并对每个窗口内的像素进行独立扫描，纳入了一种局部扫描策略。

在本节中，我们提出了一种名为 GCRPNet 的新颖架构，其整体结构如图 1 所示。GCRPNet 采用了典型的编码器-解码器框架，并整合了三个创新模块：DS-HGAM、MGAM 和 LESS2D。为了清晰地展示我们的工作，我们首先在第三节 A 中总结和回顾了现有的 SSM 方法。然后，我们介绍 GCRPNet 的整体架构，接着详细描述了三个核心模块，包括它们的设计原则和实现细节。最后，我们解

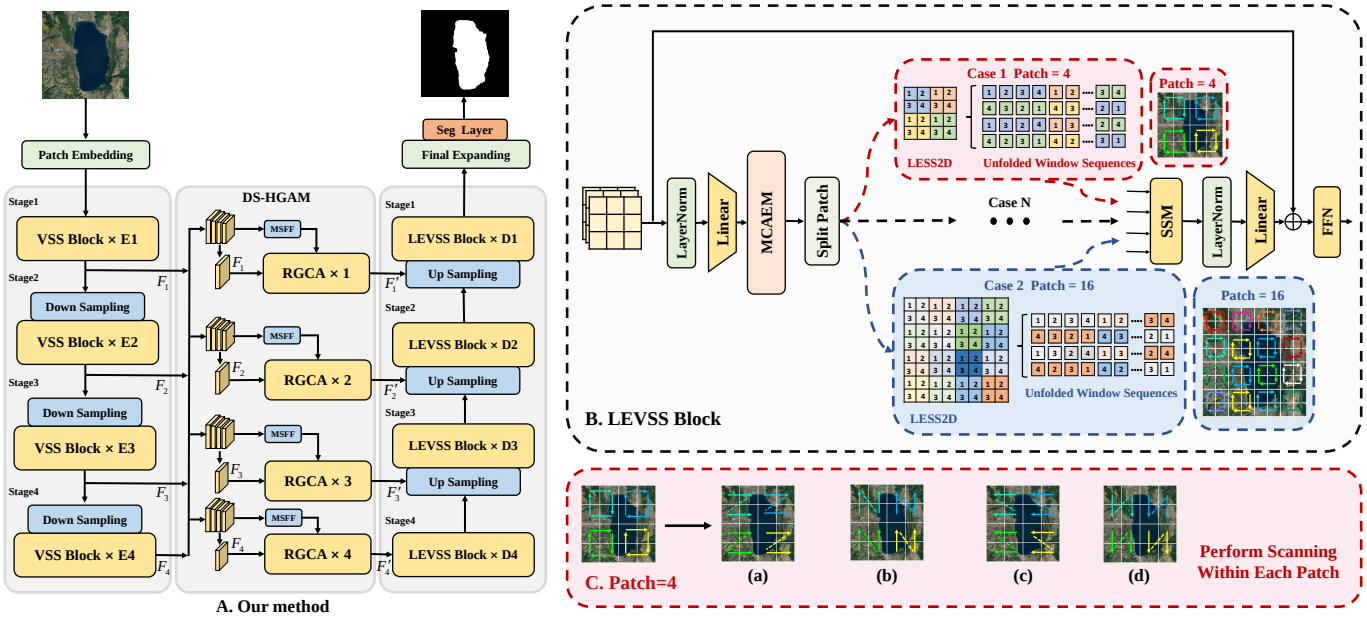

Fig. 1: 所提方法的流程。A：GCRPNet 采用 U 形的编码器-解码器架构。编码器建立在 VMamba 的视觉状态空间（VSS）上。在跳跃连接中，使用了一种 DS-HGAM 模块，通过特征融合和图注意力机制来增强模型的表示能力。随后，在解码阶段，经过 DS-HGAM 处理的特征被输入到 LEVSS 块中进行解码。B：所提的 LEVSS 块通过 MCAEM 和 LESS2D 增强了模型捕捉局部信息的能力，以实现更精确的边界表示。C：LESS2D 在每个图像块内执行四方向扫描。

释了我们方法中使用的损失函数。

A. 预备知识

状态空间模型：状态空间模型（SSM）[38], [39]，如 S4 [40] 和 Mamba [18]，是受循环神经网络（RNNs）和卷积神经网络（CNNs）启发而构建的结构化序列架构，具有随序列长度线性或近线性扩展的能力。这些模型来源于连续系统，将输入序列 $x(t) \in \mathbb{R}^L$ 映射到输出序列 $y(t) \in \mathbb{R}^L$ 通过一个隐藏状态 $h(t) \in \mathbb{R}^N$。更具体地说，SSM 通过以下连续时间常微分方程（ODEs）来描述：

$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t) \tag{1}$$
$$y(t) = \boldsymbol{C}h(t) \tag{2}$$

其中 $h(t)$ 是当前隐藏状态，$h'(t)$ 是更新后的隐藏状态，$x(t)$ 是当前输入，$y(t)$ 是输出。$\boldsymbol{A} \in \mathbb{R}^{N\times N}$ 是 SSM 的演化矩阵，$\boldsymbol{B} \in \mathbb{R}^{N\times 1}$ 和 $\boldsymbol{C} \in \mathbb{R}^{1\times N}$ 分别是输入和输出投影矩阵。

离散状态空间模型：为了将连续时间状态空间模型（SSM）整合到深度模型中，它们需要被离散化。这涉及将连续时间函数转换为离散时间的序列到序列映射。例如，这在像 S4 [40] 和 Mamba [18] 这样的模型中可以看到。一种常用的方法是零阶保持（ZOH）。它引入了一个时间尺度参数 Δ，将连续参数 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 转换为其离散版本 $\overline{\boldsymbol{A}}$，$\overline{\boldsymbol{B}}$。

$$\overline{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A}) \tag{3}$$
$$\overline{\boldsymbol{B}} = (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}, \tag{4}$$
$$h_t = \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t \tag{5}$$
$$y_t = \boldsymbol{C}h_t \tag{6}$$

B. 总体结构

GCRPNet 的整体结构如图 1（A）所示，其总体架构采用 U 形设计，由三个主要组件组成：（1）一个基于 VMamba 骨架构建的四阶段 VSS 编码器，负责多尺度特征提取；（2）嵌入跳跃连接中的四个差异-相似性引导的层次化图注意模块（DS-HGAM），这些模块融合来自编码器的多尺度特征并应用图注意机制对其进行建模；（3）一个四阶段局部增强 VSS（LEVSS）解码器，旨在增强局部建模并逐步解码特征以获得细粒度分割地图。在编码器阶段，输入图像 $I \in \mathbb{R}^{H\times W\times 3}$ 首先通过由两个卷积层组成的补丁嵌入模块，该模块执行逐块嵌入。每个编码器阶段 $s \in \{1,2,3,4\}$ 包含一个 2× 降采样层，后跟多个堆叠的 VSS 块以提取多尺度特征 F_i，其中 $i \in \{1,2,3,4\}$。在编码器之后，来自四个阶段的特征以级联方式输入到差异-相似性引导的层次化图注意模块（DS-HGAM）。这些模块处理多尺度特征并将其整合到各自的特征表示中。通过图注意机制，获得的特征 F_i^g 能够在融合的特征图中捕捉像素之间的长距离关系，从而使模型能够在严重的背景干扰下识别显著对象，并增强其对显著区域的辨别能力。在解码器中，网络包含四个解码阶段，每个阶段 D_s，其中 $s \in \{4,3,2,1\}$ 包含多个堆叠的 LEVSS 模块，以增强局部特征建模能力。第一个解码阶段 D_4 直接采用编码器中与跳跃连接的特征 F_4^g 作为输入。其余三个阶段 $s \in \{3,2,1\}$ 将高级跳跃连接特征 F_s^g 与前一阶段的上采样中间特征沿通道维度进行拼接，以形成融合输入特征。传统 U 形网络架构中的跳跃连接由于其简单且直接的特征拼接策略，缺乏空间上下文和细粒度细节的建模，通常无法在低对比度情况下有效区分前景和背景，从而导致边界模糊和定位不准确。然而，尽管 ORSIs 在复杂背景和低对比度时很困难，显著对象通常在各区域中展现出类似的视觉模式的聚集分布。为了更好地利用这一特性，我们提出了 DS-HGAM。DS-HGAM 不是简单地将编码器和解码器的特征进行拼

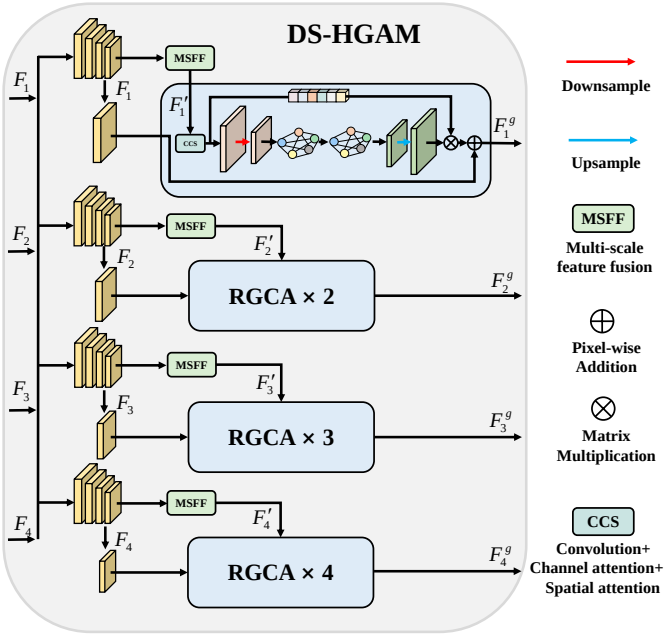

Fig. 2: DS-HGAM 示意图

接，而是进行跳跃连接特征的多尺度融合，并采用图神经网络来跨层次级别建模空间关系和上下文依赖。通过计算区域和尺度间的注意力，DS-HGAM 有效捕获不同分辨率下视觉相似片段之间的互补信息。这不仅增强了模型的结构感知能力，还加强了细粒度细节的表达，使得在低对比度和杂乱场景下能进行更准确的显著性预测。

具体而言，DS-HGAM（如图 2 所示）首先针对跳接特征 F_i 的四组，其中 $i \in \{1,2,3,4\}$ 来自不同的编码层，并为每组引入多尺度特征融合（MSFF）模块。这些模块在多尺度特征上执行空间对齐，将它们融合为相应分辨率的特征。特征融合过程如下：

$$F_i' = Conv_{1\times1}(Concat(\phi_1(F_1),\phi_2(F_2),\phi_3(F_3),\phi_4(F_4)) \tag{7}$$

其中，$\phi_i(\cdot)$ 表示通过上采样或下采样将特征调整到 i 层分辨率的操作；$Concat$ 表示逐通道的连接；$Conv_{1\times1}$ 表示连接后用于压缩通道数量的卷积操作。

随后，RGCA 模块将融合特征 F_i' 作为输入，其中 $i \in \{1,2,3,4\}$。特征 F_i' 首先被输入到 CCS 模块中。CCS 模块结合了通道注意力和空间注意力，以增强对目标内显著区域的关注，然后进行卷积运算以处理融合特征并提高其表现力。

$$F_i^c = CCS(F_i') \tag{8}$$

同时，为了减少后续图注意网络操作处理的参数数量，RGCA 模块对融合特征进行下采样，特征 $F_i^c \in \mathbb{R}^{h\times w\times c}$ 被输入到一个核大小为 $s\times s$、步长为 s 的卷积层中。然后，输出特征 $F_i^{c'} \in \mathbb{R}^{h/s\times w/s\times c}$ 的每一行像素被视为一个图节点 $f_i \in \mathbb{R}^d$。由图像块组成的图可以表示为：$F_{in} = \{\vec{f_1},\vec{f_2},...,\vec{f_n}\}, n = h/s \times w/s$。为了将输入特征转换为更高层次的表示并增强网络的表达能力，我们应用一个共享的可学习权重矩阵 $W \in \mathbb{R}^{d'\times d}$ 对每个节点进行线性变换。然后，使用一个共享的自注意力权重矩阵 $l \in \mathbb{R}^{2d'\times1}$ 计算节点对之间的注意力系数。我们仅为节点 j 计算 a_{ij}，其中 $\mathcal{N}_i$ 表示节点 i 的一级邻居节点集。为了更好地比较不同节点之间的注意力系数 a，我们应用 softmax 函数来规范化所有被选中的 j。因此，a_{ij} 的公式为：

其中，T 表示转置操作，$||$ 表示拼接操作，σ 是 LeakyReLU 激活函数。假设 δ 是 ELU 激活函数，最终输出特征可以表示为：

$$\vec{f_i'} = \delta(\sum_{n\in\mathcal{N}_i} a_{ij} W \vec{f_j}) \tag{9}$$

最终输出特征与输入特征具有相同的维度，记作 $F_{out} = \{\vec{f_1'},\vec{f_2'},...,\vec{f_n'}\}, \vec{f_n'} \in \mathbb{R}^d$。然后将此输出特征 F_{out} 通过反卷积和上采样操作，并与获得的通道注意权重 w 相乘，形成 F_i 的残差信息流，从而补偿下采样过程中丢失的信息。

整体操作过程如下：

$$F_i^{gat} = GAT(F_i^c) \otimes CA(F_i^c) \tag{10}$$

$$F_i^g = F_i^{gat} + F_i \tag{11}$$

最后，通过残差连接将原始特征 F_i 与增强特征 F_i^{gat} 融合，以生成优化的特征表示 F_i^g。

C. 局部增强视觉状态空间

在遥感图像中的显著物体通常表现出显著的尺度变化和复杂的结构特征。这要求模型在局部区域和多尺度表示方面具备强大的能力，以实现准确的显著性检测。然而，传统的 Mamba 扫描策略将二维图像展平成一维序列，破坏了局部区域空间的依赖性，使得模型难以捕捉微小的细节。如何使 Mamba 模型有效地建模局部区域仍然是一个挑战。因此，我们提出了 LEVSS 模块，该模块是在原始 Mamba 模块基础上设计的，以增强局部区域信息的表示，并显著提高模型捕捉局部区域依赖的能力。LEVSS 的结构如图 1 (B) 所示，主要由两个组件组成：MCAEM 和 LESS2D。

1) 多粒度协同注意力增强模块： 在 ORSI-SOD 任务中，有效地适应多尺度物体的变化对于稳健的显著性检测至关重要。为此，我们提出了一种 MCAEM，它有效地结合了多尺度特征提取与通道和空间注意机制，以增强特征表示（如图 3 所示）。该模块集成到 LEVSS 块中，并在序列扫描之前应用，旨在加强多尺度特征表示，并补偿序列建模结构局部区域感知能力的不足。

具体来说，MCAEM 使用深度可分离卷积 [41] 和三个不同的核大小 $(3\times3, 5\times5, 7\times7)$ 在多个感受野中提取特征，从而增强跨尺度的结构信息表示。同时，通道注意力和空间注意力机制被应用于提取的特征以实现自适应增强。通道注意力从全局的角度调整不同语义类型（*e.g.*，边缘，纹理，颜色）的响应，这有助于提高检测精度。与此同时，空间注意力专注于定位图像中的重要区域并强调判别性区域，而并行设计有助于缓解冗余并减少特征通道或空间位置之间的偏差依赖性。

给定输入特征图 $F_i \in \mathbb{R}^{h\times w\times c}$，MCAEM 的整体处理可以总结为以下步骤：

$$F_k' = DWConv_{k\times k}(F), k \in \{3,5,7\} \tag{12}$$

$$F_k^{att} = CA(F_k') + SA(F_k') \tag{13}$$

$$F_{out} = F + Conv_{1\times1}(Concat(F_3^{att}, F_5^{att}, F_7^{att})) \tag{14}$$

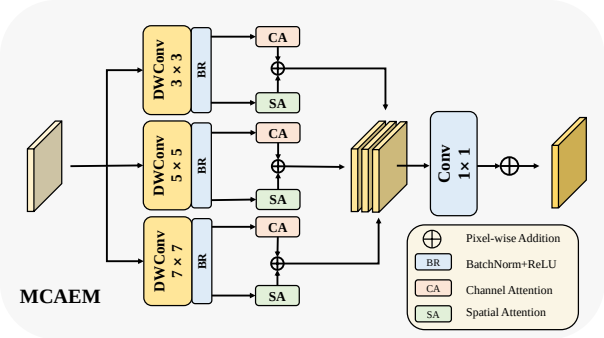

Fig. 3: 所提出的 MCAEM 的示例

这里，$CA(\cdot)$ 和 $SA(\cdot)$ 分别指代通道和空间注意力模块，$Conv_{1\times1}$ 表示一个 1×1 卷积，$Concat$ 表示在通道维度上的连接。

在复杂物体的显著性目标检测中，精确地建模局部细节对于提高检测精度起着至关重要的作用。为此，我们提出了一个 LESS2D 模块，旨在解决传统 SS2D 在捕捉局部特征方面的局限性，同时在增强对局部细节敏感性的同时，平衡 Mamba 模型的全局上下文建模能力。

具体来说，该模块的核心动机在于观察到随着特征图分辨率的增加，其空间几何结构变得更加复杂，包含的局部细节更为丰富。将二维图像展平成一维序列不仅破坏了原有的空间邻接关系，还导致局部区域特征在序列中变得分散，从而削弱了模型捕捉局部依赖关系的能力。为了解决这个问题，我们提出了 LESS2D（如图 1 (B) 所示）并将其集成到一个四阶段解码器架构中。对于给定的输入图像 $I\in\mathbb{R}^{h\times w\times c}$，我们对提取的多尺度特征图执行分块操作，将其分为多个图像块，以便于后续局部特征建模和处理。在我们的方法中，对于低分辨率特征图 ($e.g.$，$\frac{1}{16}h\times\frac{1}{16}w$)，由于其空间建模范围有限且几何关系集中，LESS2D 采用全局四方向扫描策略以充分保留长距离依赖建模能力。相反，对于中高分辨率特征图 ($e.g.$, $\frac{1}{8}h\times\frac{1}{8}w, \frac{1}{4}h\times\frac{1}{4}w, \frac{1}{2}h\times\frac{1}{2}w$)，图像被划分为多个不重叠的局部块（例如，分别为 4，16 和 64 块），并在每个块内独立执行四方向扫描。来自所有块的扫描序列随后按顺序连接以形成一个完整的序列，然后输入下游模块进行进一步处理。

以大小为 $\frac{1}{4}\boldsymbol{h}\times\frac{1}{4}\boldsymbol{w}$ 的特征图 F 为例，我们将其划分为 16 个不重叠的子区域，每个子区域记为 F_i，其中 $i\in\{1,2,...,16\}$。整体过程如下进行：

$$\begin{aligned}Scan_4(F_i) = [&Scan^{\rightarrow}(F_i),\ Scan^{\downarrow}(F_i),\\ &Scan^{\leftarrow}(F_i),\ Scan^{\uparrow}(F_i)]\end{aligned} \tag{15}$$

$$\vec{S} = Concat(Scan^{\rightarrow}(F_1), Scan^{\rightarrow}(F_2), ..., Scan^{\rightarrow}(F_{16})) \tag{16}$$

每个块贡献四个方向扫描，分别记为 $Scan^{\rightarrow}$、$Scan^{\downarrow}$、$Scan^{\leftarrow}$ 和 $Scan^{\uparrow}$（如图 1 (C) 所示），它们分别表示从左到右、从上到下、从右到左和从下到上的扫描。方程 16 展示了从左到右的扫描过程，其中 $\vec{S}$ 表示完整的图像序列。

我们定义了一个综合损失函数来约束整个训练过程。我们的损失函数由两个组成部分构成：二元交叉熵（BCE）损失和交并比（IoU）损失。具体而言，我们计算四个不同尺度上的显著性预测损失：P_1，P_2，P_3 和 P_4，其中 P_1 是最终的显著性图，而 P_2，P_3 和 P_4 是用于辅助训练的中间尺度显著性图。总损失函数定义为：

$$\mathcal{L}_{total} = \sum_{i=1}^{4}\lambda_i(\mathcal{L}_{BCE}(P_i,G_i) + \mathcal{L}_{IoU}(P_i,G_i)) \tag{17}$$

其中 λ_i 是权重系数，P_i 代表显著性图，G_i 是相应的真实图。在本文中，我们设置 $\lambda_1=\lambda_2=\lambda_3=\lambda_4$ 以确保所有尺度的特征图对最终损失函数的贡献相等。

BCE 损失用于确保逐像素的分类精度。对于每个预测的显著性图 P_i 及其对应的真实图 G_i，二元交叉熵损失被定义为：

$$\mathcal{L}_{BCE}(P_i,G_i) = -\frac{1}{N}\sum_{j=1}^{N}[G_i^j log(P_i^j) + (1-G_i^j)log(1-P_i^j)] \tag{18}$$

其中 N 表示像素总数，P_i^j 和 G_i^j 分别表示预测图和真实图在第 j 个像素处的值。IoU 损失用于最大化预测区域和真实区域之间的重叠。IoU 损失定义为：

$$\mathcal{L}_{IoU}(P_i,G_i) = 1 - \frac{\sum(P_i\odot G_i)+\epsilon}{\sum(P_i+G_i-P_i\odot G_i)+\epsilon} \tag{19}$$

其中 $\odot$ 表示逐元素乘法，而 ϵ 是用于防止除零错误的平滑项。

I. 实验

A. 实验设置

1) 数据集： 为了评估该方法的有效性，我们在两个公共的 ORSI-SOD 基准数据集上对其进行了训练和测试：ORSSD [21] 和 EORSSD [22]。

ORSSD 数据集是 ORIS-SOD 领域中首个公开发布的数据集。该数据集包含总共 800 张高质量的可见光遥感图像，其中包括八种类型的遥感物体，例如岛屿、船只和车辆。其中，600 张图像被分配用于算法训练和优化的训练集，而剩余的 200 张图像则被指定为测试集，以评估模型的性能和泛化能力。

EORSSD 数据集是 ORSSD 数据集的增强版本，在数据量和多样性方面有了显著改进。与原始的 ORSSD 相比，EORSSD 的数据量增加了一倍多，提供了更丰富的训练和测试样本。该数据集包括 2,000 张高分辨率图像及其对应的像素级真值，涵盖了高分辨率卫星图像和低分辨率航空图像。它包含九种不同的遥感对象：船只、飞机、汽车、建筑物、水面、岛屿、道路、无以及其他。其中，训练集包含 1,400 张图像，而测试集由 600 张图像构成。在本研究中，我们使用 PyTorch 框架实现了所提方法，并在 RTX 4090Ti GPU 上进行了所有实验。我们将学习率设置为 1e-4，批大小为 4，并采用 AdamW 优化策略进行训练。为确保输入数据的一致性，在输入模型前将每张图像调整为 384×384 像素。此外，为了在训练期间有效减轻过拟合，我们引入了一系列数据增强技术，包括随机缩放和裁剪、随机图像变换和归一化。最后，我们在 ORSSD 和 EORSSD 数据集上训练了 100 个周期。

2) 评估指标： 为了全面评估所提出方法的性能，我们采用八个广泛使用的评价指标来评估模型的性能。这些指标包括平均绝对误差（MAE）、S-测量（S_α）、F-测量分数（F_β）、E-测量（E_ξ）、自适应 E-测量（E_ξ^{adp}）、平均 E-测量（E_ξ^{mean}）、最大 E-测量（E_ξ^{max}）、自适应 F-测量（F_β^{adp}）、平均 F-测量（F_β^{mean}）和最大 F-测量（F_β^{max}

TABLE I: 基于 F_β^{max} 、F_β^{mean} 、F_β^{adp} 、E_ξ^{max} 、E_ξ^{mean} 、E_ξ^{adp} 、S_α 和 MAE 评分的定量比较结果在两个基准数据集上进行了展示。在表格中，↑ 和 ↓ 分别表示偏好较高和较低的数值。每列中的前三个结果分别用红色、蓝色和绿色标出。

Methods	Pubs	EORSSD								ORSSD							
		$F_\beta^{max}\uparrow$	$F_\beta^{mean}\uparrow$	$F_\beta^{adp}\uparrow$	$E_\xi^{max}\uparrow$	$E_\xi^{mean}\uparrow$	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$F_\beta^{mean}\uparrow$	$F_\beta^{adp}\uparrow$	$E_\xi^{max}\uparrow$	$E_\xi^{mean}\uparrow$	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$
EMFINet	TGRS'21	0.8721	0.8488	0.7987	0.9712	0.9603	0.9502	0.9292	0.0084	0.9002	0.8856	0.8617	0.9737	0.9671	0.9663	0.9366	0.0109
DAFNet	TIP'21	0.8735	0.7982	0.6525	0.9816	0.9382	0.8216	0.9186	0.0053	0.8999	0.8442	0.7565	0.9821	0.9536	0.8928	0.9188	0.0106
ERPNet	TCYB'22	0.8744	0.8272	0.7174	0.9665	0.9364	0.8995	0.9254	0.0082	0.9036	0.8799	0.8392	0.9738	0.9604	0.9531	0.9352	0.0114
SBA-Net	ISPRS'22	0.8164	0.7984	0.7104	0.9382	0.9235	0.8704	0.8612	0.0142	0.8373	0.8246	0.8159	0.9394	0.9205	0.9286	0.8618	0.0264
HFANet	TGRS'22	0.8950	0.8805	0.8651	0.9682	0.9608	0.9461	0.9298	0.0127	0.9059	0.8917	0.8697	0.9727	0.9662	0.9467	0.9355	0.0116
ESGNet	TCSVT'23	0.8907	0.8754	0.8634	0.9781	0.9732	0.9462	0.9374	0.0051	0.9220	0.9113	0.9076	0.9858	0.9814	0.9661	0.9484	0.0065
ACCoNet	TCYB'23	0.8822	0.8464	0.7787	0.9759	0.9659	0.9381	0.9303	0.0067	0.9149	0.8952	0.8830	0.9819	0.9775	0.9770	0.9428	0.0087
BSCGNet	TGRS'23	0.8809	0.8666	0.8506	0.9693	0.9650	0.9644	0.9296	0.0079	0.9101	0.8977	0.8897	0.9784	0.9744	0.9773	0.9411	0.0093
SFANet	TGRS'24	0.8834	0.8682	0.8495	0.9770	0.9725	0.9682	0.9351	0.0058	0.9192	0.9063	0.8984	0.9830	0.9789	0.9800	0.9453	0.0077
RAGRNet	TGRS'24	0.8853	0.8771	0.8646	0.9785	0.9736	0.9761	0.9364	0.0057	0.9242	0.9157	0.9104	0.9861	0.9817	0.9835	0.9507	0.0066
UDCNet	TGRS'24	0.8821	0.8517	0.8214	0.9775	0.9689	0.9551	0.9311	0.0056	0.9239	0.9068	0.8932	0.9850	0.9803	0.9770	0.9497	0.0068
SOLNet	TGRS'24	0.8609	0.8514	0.8394	0.9624	0.9576	0.9607	0.9171	0.0078	0.9012	0.8946	0.8925	0.9734	0.9688	0.9728	0.9284	0.0111
LiteSalNet	TGRS'25	0.8878	0.8810	0.8798	0.9743	0.9633	0.9729	0.9276	0.0063	0.9124	0.9061	0.9056	0.9789	0.9699	0.9752	0.9372	0.0090
LGIPNet	TGRS'25	0.8837	0.8702	0.8563	0.9761	0.9711	0.9701	0.9339	0.0065	0.9149	0.9027	0.8924	0.9806	0.9758	0.9755	0.9450	0.0082
DPU-Former	IJCAI'25	0.8931	0.8758	0.8464	0.9816	0.9749	0.9665	0.9402	0.0056	0.9263	0.9142	0.9024	0.9868	0.9827	0.9838	0.9412	0.0062
Ours	—	0.9032	0.8876	0.8719	0.9801	0.9748	0.9457	0.9487	0.0049	0.9356	0.9243	0.9196	0.9861	0.9819	0.9650	0.9590	0.0056

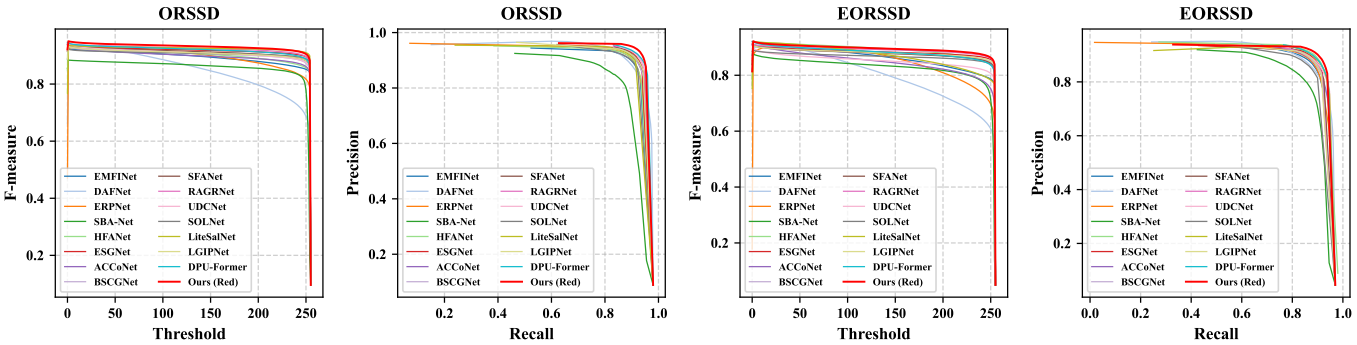

Fig. 4: 我们的方法与其他 ORSI-SOD 方法在 ORSSD [21] 和 EORSSD [22] 数据集上的定量比较如下。

)。具体来说，S_α 量化预测显著图和真实情况之间的结构相似性。F_β 是精度和召回率的加权谐调平均数，其中 β^2 设为 0.3，以便在召回率之上更强调精度。MAE 计算预测与真实情况之间的平均像素误差，提供预测偏差的直接度量。此外，E_ξ^{adp} 、E_ξ^{mean} 和 E_ξ^{max} 分别通过自适应阈值、图像级平均值和最大值来评估 E-测量。这些指标从多个角度捕捉结构和分布的一致性。类似地，F_β^{adp} 、F_β^{mean} 和 F_β^{max} 在相同的三种阈值策略下评估 F-测量，强调精度与召回率之间的权衡。

B. 与现有技术的比较

我们将我们的方法与专门为 ORSI-SOD 任务设计的 15 种最新方法进行比较，包括 EMFINet [42] 、DAFNet [22] 、ERPNet [43] 、SBA-Net [44] 、HFANet [10] 、ESGNet [45] 、ACCoNet [31] 、BSCGNet [46] 、SFANet [7] 、RAGRNet [47] 、UDCNet [48] 、SOLNet [49] 、LiteSalNet [8] 、LGIPNet [50] 和 DPU-Former [17] 。为了确保公平评估，所有实验均在 ORSI-SOD 和 EORSSD 数据集上的相同条件下进行。

1) 定量比较： 我们提出的 GCPRNet 与其他方法的定量比较结果总结在 Tab I 中。显然，我们的模型实现了 SOTA 性能。在 EORSSD 数据集上，我们的方法在八个评估指标上的定量比较结果呈现在表格底部。可以观察到，我们的方法在几个主要指标上实现了最佳性能，包括 F_β^{max} 、F_β^{mean} 、S_α 和 MAE ，同时在指标 F_β^{adp} 、E_ξ^{max} 和 E_ξ^{mean} 上取得了第二好的结果。具体而言，与第二好的方法相比，我们的方法在 F_β^{max} 上实现了 0.82 % 的性能提升，在 F_β^{mean} 上提升了 0.66 %，在 S_α 上提升了 0.85 %，进一步凸显了其在显著目标检测任务中的优势。在 ORSSD 数据集上，我们的方法同样实现了卓越的性能。具体来说，它在五个关键指标上达到了最佳结果：F_β^{max} 、F_β^{mean} 、F_β^{adp} 、S_α 和 MAE ，同时在指标 E_ξ^{max} 和 E_ξ^{mean} 上取得第二好的结果。与第二好的方法相比，我们的方法在指标 F_β^{max} 上提升了 0.93 %，在指标 F_β^{mean} 上提升了 0.86 %，在指标 S_α 上提升了 0.83 %，同时还将 MAE 减少了 0.06 %，进一步验证了我们方法的有效性和鲁棒性。

图 4 展示了在 ORSSD 和 EORSSD 数据集上的 PR 和 F-衡量曲线。更高的曲线表明更好的性能。与其他方法相

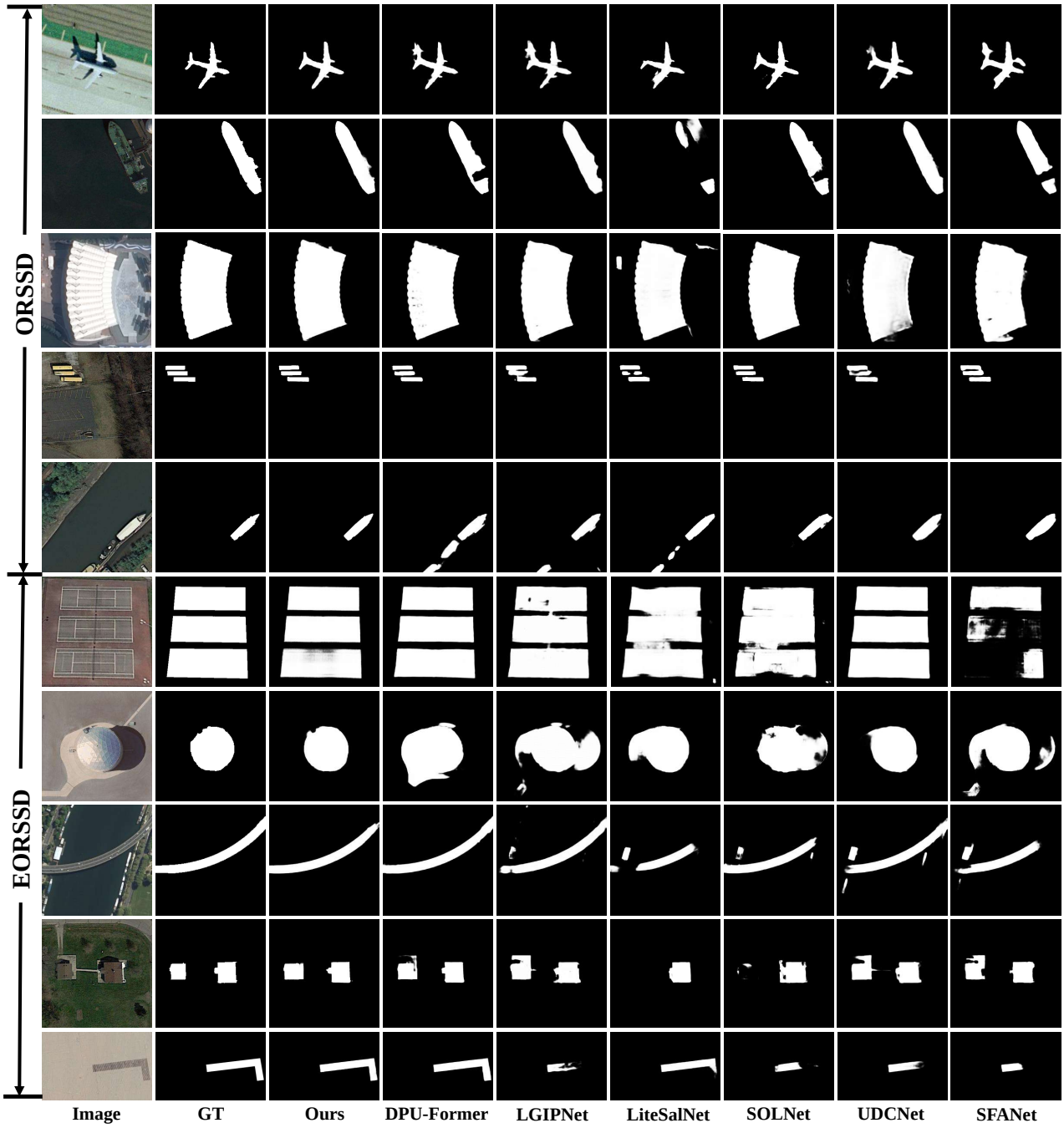

Fig. 5: 我们的方法与其他六种方法在各种具有挑战性的场景中的定性比较。

比，我们的方法始终实现更高的精度和召回率。具体而言，我们的 PR 曲线更接近右上角，显示了在不同阈值下的卓越性能。同样地，我们的 F-衡量曲线优于所有竞争方法。这些结果共同突出了我们的方法在各种条件下的有效性和鲁棒性。

2) 定量比较： 图 fig 5 展示了我们模型与近年来六种最新的先进方法在 ORSSD 和 EORSSD 数据集上的定性比较。我们选择了多个具有代表性的遥感图像场景，包括大型物体、小型物体、细长结构和低对比度物体。与图中的其他方法相比，我们的模型在这些复杂条件下表现出色的显著性检测性能。例如，在第七行的低对比场景中，只有我们的模型能够精确分割出圆形建筑的完整轮廓，而其他方法存在不同程度的过度分割，表明它们对这些条件的适应性有限。这凸显了我们模型在捕捉细微局部变化和精确边界定位方面的强大能力。同样，在第四行涉及小物体检测中，我们的模型正确分割出所有三个黄色物体，并且边界清晰且不重叠。相比之下，其他方法生成模糊且不完整的结果，特别是对于中间物体。这些结果展示了我们模型在不同物体尺度上出色的分割能力，反映出其强大的多尺度特征表示能力。此外，在最后一行的 L 形建筑场景中，只有我们的方法和 DPU-Former 能够完整识别并分割出

TABLE II: GCRPNet 中每个组件的有效性。在表格中，↑ 表示偏好较高的值，而 ↓ 表示偏好较低的值。

Methods	EORSSD				ORSSD			
	$F_\beta^{max}\uparrow$	$E_\xi^{max}\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta^{max}\uparrow$	$E_\xi^{max}\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$
w/o DS-HGAM	0.8907	0.9748	0.9389	0.0061	0.9135	0.9758	0.9416	0.0096
w/o MCAEM	0.8924	0.9752	0.9411	0.0055	0.9214	0.9835	0.9562	0.0059
w/o LESS2D	0.8870	0.9725	0.9394	0.0055	0.9170	0.9774	0.9450	0.0081
Ours	0.9032	0.9801	0.9487	0.0049	0.9356	0.9861	0.9590	0.0056

完整的 L 形结构，而所有其他方法都未能分割出完整的建筑形状，出现明显的分割错误。总之，图 fig 5 中的定性结果全面展示了我们所提出模型在各种复杂遥感图像场景中的显著性物体检测的鲁棒性和准确性。我们的方法不仅确保了有效的全局上下文建模，还在提取关键局部特征方面表现出色，使得模型能够更有效地学习显著性信息。

在本节中，我们对 EORSSD 和 ORSSD 数据集进行全面的消融研究，以展示我们模型中每个组件的有效性。具体结果如表 II 所示。

在表格中，w/o DS-HGAM 表示去除了差异相似性引导的层次图注意模块，意味着编码器特征是通过跳跃连接直接传递到解码器的。w/o MCAEM 表示去除了多粒度协作注意增强模块。w/o LESS2D 表示在解码器阶段没有使用所提出的局部增强 2D 选择扫描，而是采用了传统的 VMamba 扫描策略。

我们的消融研究表明，完整的模型在两个数据集上均取得了卓越的性能。具体来说，它在 EORSSD 数据集的四个关键指标 F_β^{max} 、E_ξ^{max} 、S_α 和 MAE 上分别获得了 0.9032、0.9810、0.9487 和 0.0049 的得分，而在 ORSSD 数据集上则分别获得了 0.9356、0.9861、0.9590 和 0.0056 的得分。相比之下，其他变体在这些指标上表现出不同程度的性能下降。这些结果清楚地证明了我们提出的模块的有效性。w/o DS-HGAM 的实验结果表明，通过引入 DS-HGAM 模块，所有指标均有改善。在 EORSSD 数据集上，指标 F_β^{max} 、E_ξ^{max} 和 S_α 分别增加了 1.25 %、0.53 % 和 0.98 %，而指标 MAE 则减少了 0.12 %。同样地，在 ORSSD 数据集上，指标 F_β^{max} 、E_ξ^{max} 和 S_α 分别提高了 2.21 %、1.03 % 和 1.74 %，而指标 MAE 下降了 0.4 %。这些结果表明，将 DS-HGAM 模块引入跳跃连接有助于模型捕捉跨空间和跨尺度的特征关系，从而实现更有效的特征整合，并提高显著性图的质量。在没有 MCAEM 的设置中，结果表明通过引入 MCAEM 模块，EORSSD 数据集上的指标 F_β^{max} 从 0.8924 增至 0.9032，实现了 1.08 % 的性能提升；指标 E_ξ^{max} 从 0.9752 提升到 0.9801，增加了 0.49 %，这表明多尺度卷积和注意机制能有效提取多尺度局部特征，从而增强模型对显著区域的表示能力，显著改善掩码预测的质量。没有 LESS2D 的实验结果表明，通过使用 LESS2D 模块，EORSSD 数据集上的指标 F_β^{max} 提高了 1.62 %，指标 E_ξ^{max} 提高了 0.76 %，指标 S_α 提高了 0.93 %。在 ORSSD 数据集上，指标 F_β^{max} 提高了 1.86 %，指标 E_ξ^{max} 提高了 0.87 %，指标 S_α 提高了 1.4 %，而指标 MAE 下降了 0.25 %。LESS2D 采用的逐块扫描策略增强了局部像素间的依赖关系，促进了精细局部特征表示的学习，并充分利用了 Mamba 的序列建模能力。这使得模型能够更有效地整合全局上下文与局部细节，生成更加精确的显著性图。

在本文中，我们提出了一种基于 Mamba 的 ORSI-SOD 模型，名为 GCRPNet。具体来说，我们设计了一种 DS-HGAM 模型，用于在跳过连接中执行多尺度特征的交互融合。通过利用图神经网络，DS-HGAM 能建模跨空间关系以整合来自不同区域的特征，从而生成更高质量的显著性图，并增强模型的鲁棒性和泛化能力。同时，为解决现有 Mamba 模型在局部信息建模中的局限性，我们进一步提出了一种新颖的 LEVSS 模块，并将其整合到网络的解码阶段。该模块整合了两个子模块：MCAEM 和 LESS2D。MCAEM 模块通过使用具有多种内核大小和并行注意力机制的深度可分离卷积扩大了感受野，使模型能够更好地适应不同尺度和外观的目标。此外，为进一步提高局部建模能力，LESS2D 模块在多尺度特征图上执行块状扫描，加强局部特征表示，并有效整合全局上下文建模与局部细节学习。广泛的实验结果表明，所提出的模型在 ORSI-SOD 任务上达到了最先进的性能，充分验证了其有效性和优越性。

References

[1] Q. Zhang, L. Zhang, W. Shi, and Y. Liu, "Airport extraction via complementary saliency analysis and saliency-oriented active contour model," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 7, pp. 1085–1089, 2018.

[2] L. Zhang, Y. Liu, and J. Zhang, "Saliency detection based on self-adaptive multiple feature fusion for remote sensing images," *International Journal of Remote Sensing*, vol. 40, no. 22, pp. 8270–8297, 2019.

[3] Z. Liu, D. Zhao, Z. Shi, and Z. Jiang, "Unsupervised saliency model with color markov chain for oil tank detection," *Remote Sensing*, vol. 11, no. 9, p. 1089, 2019.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[5] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2814–2821.

[6] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[7] Y. Quan, H. Xu, R. Wang, Q. Guan, and J. Zheng, "Orsi salient object detection via progressive semantic flow and uncertainty-aware refinement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[8] Z. Ai, H. Luo, and J. Wang, "A lightweight multi-stream framework for salient object detection in optical remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[9] L. Gao, B. Liu, P. Fu, and M. Xu, "Adaptive spatial tokenization transformer for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[10] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[11] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022.

[12] Y. Sun, H. Xuan, J. Yang, and L. Luo, "Glconet: Learning multisource perception representation for camouflaged object detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[13] R. Yan, L. Yan, G. Geng, Y. Cao, P. Zhou, and Y. Meng, "Asnet: Adaptive semantic network based on transformer–cnn for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[14] J. Zhao, Y. Jia, L. Ma, and L. Yu, "Adaptive dual-stream sparse transformer network for salient object detection in optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 5173–5192, 2024.

[15] J. Zhu, Y. Luo, X. Zheng, H. Wang, and L. Wang, "A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 720–11 730.

[16] H. Huang, Y. Huang, S. Xie, L. Lin, R. Tong, Y.-W. Chen, Y. Li, and Y. Zheng, "Combinatorial cnn-transformer learning with manifold constraints for semi-supervised medical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2330–2338.

[17] Y. Sun, J. Yan, J. Qian, C. Xu, J. Yang, and L. Luo, "Dual-perspective united transformer for object segmentation in optical remote sensing images," *International Joint Conference on Artificial Intelligence*, 2025.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[19] Y. Shi, B. Xia, X. Jin, X. Wang, T. Zhao, X. Xia, X. Xiao, and W. Yang, "Vmambair: Visual state space model for image restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018.

[21] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9156–9166, 2019.

[22] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Transactions on Image Processing*, vol. 30, pp. 1305–1317, 2020.

[23] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.

[24] H. Sun, Y. Bian, N. Liu, and H. Zhou, "Multi-scale edge-based u-shape network for salient object detection," in *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18*. Springer, 2021, pp. 501–514.

[25] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8779–8788.

[26] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4722–4732.

[27] J. Zhang, J. Xie, N. Barnes, and P. Li, "Learning generative vision transformer with energy-based latent space for saliency prediction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 448–15 463, 2021.

[28] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4481–4490.

[29] Y. Zhan, Z. Zeng, H. Liu, X. Tan, and Y. Tian, "Mambasod: Dual mamba-driven cross-modal fusion network for rgb-d salient object detection," *Neurocomputing*, vol. 631, p. 129718, 2025.

[30] J. He, K. Fu, X. Liu, and Q. Zhao, "Samba: A unified mamba-based framework for general salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025, pp. 25 314–25 324.

[31] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, vol. 53, no. 1, pp. 526–538, 2022.

[32] K. Huang, N. Li, J. Huang, and C. Tian, "Exploiting memory-based cross-image contexts for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[33] P. Dong, B. Wang, R. Cong, H.-H. Sun, and C. Li, "Transformer with large convolution kernel decoder network for salient object detection in optical remote sensing images," *Computer Vision and Image Understanding*, vol. 240, p. 103917, 2024.

[34] G. Li, Z. Bai, Z. Liu, X. Zhang, and H. Ling, "Salient object detection in optical remote sensing images driven by transformer," *IEEE Transactions on Image Processing*, vol. 32, pp. 5257–5269, 2023.

[35] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 103 031–103 063, 2024.

[36] H. He, Y. Bai, J. Zhang, Q. He, H. Chen, Z. Gan, C. Wang, X. Li, G. Tian, and L. Xie, "Mambaad: Exploring state space models for multi-class unsupervised anomaly detection," *arXiv preprint arXiv:2404.06564*, 2024.

[37] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "Localmamba: Visual state space model with windowed selective scan," in *European Conference on Computer Vision*, 2024, pp. 12–22.

[38] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 572–585, 2021.

[39] J. T. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2022.

[40] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[42] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[43] X. Zhou, K. Shen, L. Weng, R. Cong, B. Zheng, J. Zhang, and C. Yan, "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, vol. 53, no. 1, pp. 539–552, 2023.

[44] Z. Huang, T.-Z. Xiang, H.-X. Chen, and H. Dai, "Scribble-based boundary-aware network for weakly supervised salient object detection in remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 290–301, 2022.

[45] A. Gong, J. Nie, C. Niu, Y. Yu, J. Li, and L. Guo, "Edge and skeleton guidance network for salient object detection in optical remote sensing images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7109–7120, 2023.

[46] D. Feng, H. Chen, S. Liu, Z. Liao, X. Shen, Y. Xie, and J. Zhu, "Boundary-semantic collaborative guidance network with dual-stream feedback mechanism for salient object detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[47] J. Zhao, Y. Jia, L. Ma, and L. Yu, "Recurrent adaptive graph reasoning network with region and boundary interaction for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[48] Y. Sun, J. Yang, and L. Luo, "United domain cognition network for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[49] Z. Li, Y. Miao, X. Li, W. Li, J. Cao, Q. Hao, D. Li, and Y. Sheng, "Speed-oriented lightweight salient object detection in optical

remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[50] L. Sun, H. Liu, X. Wang, Y. Zheng, Q. Chen, Z. Wu, and L. Fu, "Local–global information perception network for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–18, 2025.