Computer Science and Application 计算机科学与应用, 2025, 15(4), 275-286
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154100

Hans汉斯

基于多特征融合的图像显著性检测方法研究

潘　磊，李子龙*，秦培鑫，周文婧，时晶晶，李　辉

徐州工程学院信息工程学院(大数据学院)，江苏　徐州

收稿日期：2025年3月16日；录用日期：2025年4月16日；发布日期：2025年4月24日

摘　要

图像显著性检测是计算机视觉的关键分支，它试图模仿人类视觉选择性注意的特性，在图像中精确识别并突出最受关注的目标或区域，影响环境中亮点检测精度的因素包括光照变化、复杂背景、不同的目标尺寸以及低对比度，近年来，以卷积神经网络(CNN)为典型的深度学习技术，虽然大幅提高了检测性能，但在面对复杂背景干扰和低对比度目标时，单一模型的泛化性能仍受到限制。鉴于这些问题，一种依靠多种特征组合来检测图像较大性的新方法，该方法充分融合了多尺度特征提取、跨模态信息融合以及多层次注意力机制，有效提高了模型在复杂背景情况和低对比度环境下的鲁棒性与精确性，实验结果显示，所提出的多特征融合方法在S、F和MAE等主要指标方面的性能有明显提升，准确性和稳定性也有所提高。实验结果基于大量公共标准数据集，并与主流模型的性能进行了对比，本研究还探讨了不同特征与融合策略所起的作用，为复杂场景中的较大性检测研究给予了新的思考方向。

关键词

图像显著性检测，多种特征的融合，卷积神经网络技术，注意力相关机制，跨模态的融合手段

Research on Image Saliency Detection Method Based on Multi Feature Fusion

Lei Pan, Zilong Li*, Peixin Qin, Wenjing Zhou, Jingjing Shi, Hui Li

College of Information Engineering (Big Data College), Xuzhou University of Technology, Xuzhou Jiangsu

Received: Mar. 16th, 2025; accepted: Apr. 16th, 2025; published: Apr. 24th, 2025

Abstract

Image saliency detection is a key branch of computer vision. It attempts to imitate the characteristics of human visual selective attention, accurately identify and highlight the most concerned target

*通讯作者。

文章引用: 潘磊, 李子龙, 秦培鑫, 周文婧, 时晶晶, 李辉. 基于多特征融合的图像显著性检测方法研究[J]. 计算机科学与应用, 2025, 15(4): 275-286. DOI: 10.12677/csa.2025.154100

or area in the image. The factors affecting the accuracy of highlight detection in the environment include illumination change, complex background, different target sizes and low contrast. In recent years, convolutional neural network (CNN) as a typical deep learning technology has greatly improved the detection performance, but the generalization performance of a single model is still limited in the face of complex background interference and low contrast targets. In view of these problems, a new method based on a variety of feature combinations to detect the large image is proposed. This method fully integrates multi-scale feature extraction, cross modal information fusion and multi-level attention mechanism, and effectively improves the robustness and accuracy of the model in complex background and low contrast environment. The experimental results show that the performance of the proposed multi feature fusion method has been significantly improved in terms of S, F and Mae, and the accuracy and stability have also been improved. The experimental results are based on a large number of public standard data sets, and compared with the performance of mainstream models. This study also discusses the role of different features and fusion strategies, which provides a new direction for the research of large scale detection in complex scenes.

Keywords

Image Saliency Detection, Multiple Features Fusion, Convolutional Neural Network Technology, Attention Related Mechanisms, Cross Modal Fusion Means

1. 引言

图像显著性检测作为计算机视觉的关键分支，致力于仿照人类视觉的选择性注意机制，在图像里精准地辨别并凸显出最为引人注目的目标或者区域。较高的像素值代表吸引人类注意力的概率较高。显著性检测在目标识别与分割、场景理解、图像检索、人机交互等众多领域中得到了广泛运用。很明显，通过去除前景和背景之间的显著区域，高级视觉任务的性能得到了提高。

早期的显著性检测手段大多依靠人工设计的低层次视觉特性以及启发式的规则。这与大多数人所使用的方法的根本区别在于，它不是依赖于颜色、亮度、纹理、边缘等属性，而是基于显著性的属性，例如中心－周围对比度，来计算所谓的显著图。该模型在显著性检测研究领域起到了开创性的作用。例如，Achanta 及其同事利用对比度来获取显著图，并提出了一种结合光谱差异和颜色显著性的方法[1]，该方法直接将显著性定义为背景和前景之间的差异。后来，已经提出了很多改进的模型。

近些年来，深度学习的蓬勃发展给显著性检测带来了具有变革意义的进步。然后，卷积神经网络的特征会自动学习，这比人工描述的特征更加灵活，显著检测的性能也大大提高了。自从诸如 Fully Convolutional Network (FCN) [2]等端到端的深度模型现身以来，以 CNN 为基础的显著性检测逐步替代了传统的方法，掀起了第三波研究的热潮。近年来，深度学习在图像识别领域引发了一场革命。随着深度学习的发展，诸如视觉转换器之类的新深度学习模型能够突破卷积神经网络的局限。卷积神经网络(CNN)可以自行学习多层次的视觉特征表达，其表达能力远超手工特征，让显著性检测的性能提升到了前所未有的水平。自从端到端深度学习出现以来，例如全卷积网络(FCNs)，已经使用了具有多层深度的分层网络。

基于 CNN 的较大性检测持续取代传统方法，引发了又一波研究热潮，部分学者把视觉 Transformer

引入较大目标检测，以便更全面地观察场景，视觉 Transformer 的自注意力机制突破了卷积神经网络的感受野，有全局和局部关系的优势，生成对抗网络(GAN)也应用于较大图生成，凭借生成器与判别器之间的博弈训练，提高较大图的对比度与细节真实性。GAN 是卷积神经网络和生成神经网络的结合，利用判别器引导较大图预测。

引入这些深度学习方法后，较大性检测模型的鲁棒性和准确率有了提高，不断刷新各类基准指标，即便如此，现有方法在复杂场景中仍面临诸多难题，当前景物体与背景相似(比如是伪装的)，或者场景中有几个大小不同的物体时，一般难以用单一特征或单一模态区分较大物体与非较大背景。这促使研究者探索融合多种互补特征信息，以提高较大性检测的鲁棒性和精度，多特征融合方法是凭借结合颜色、形状、纹理、深度等各种特征获取更全面信息，确定场景中最较大像素。如图 1 展示了显著性检测的基本效果示意。在原始图像中，模型能够准确地定位出最引人注目的目标区域(如人物、动物、车辆等)，并将其从背景中清晰分离出来。这种前景与背景之间的显著性差异正是后续视觉任务(如目标分割和识别)的基础。对于 RGB-D 这类多模态数据，引入深度图像能起到辅助检测作用，如将颜色上与背景接近的目标中的前景区分出来，特征和模态结合的方法已被证实是提高较大目标检测性能的有效途径。

尽管取得了上述进展，但现有的方法在复杂场景中仍然面临着诸多挑战，若场景中存在几个大小不同的感兴趣对象，那么单一特征或单一感官的信息便不足以可靠地区分前景和背景，这促使研究者探索融合多种互补的特征信息，以提高较大性检测的鲁棒性与精度，多特征融合可整合来自不同维度的线索，比如颜色、形状、纹理和深度，并对像素的关键性做出更为准确的判断。多特征和多模式信息是提升较大性检测性能的有效方式，基于上述背景，一种借助多特征融合进行图像较大性检测的全新方法，结合注意力机制，该方法可在复杂环境中生成准确的较大图，接下来将阐述相关的理论基础、所提方法的详细设计、实验安排以及结果分析。

Figure 1. Example of significance detection
图 1. 显著性检测示例

2. 相关理论基础

2.1. 视觉显著性检测概述

视觉显著性(Visual Saliency)是指在图像里最能够吸引人类关注的区域所具备的特征。“信号检测”

任务可分为两个子问题：1) “尖锐”物体的检测；2) 人眼注视点的预测。本文所聚焦的显著物体检测一般会被形容成二元分割的任务，也就是要在图像中检测并分割出最为显著的前景物体。首先，有必要检测图片中最重要的对象，其次，获取对象的确切轮廓。此任务包括两个阶段。在实际运用的时候，这两个阶段往往是一同完成的，也就是直接给出前景物体的二值掩膜。检测能力(尽可能多地发现真正显著的区域，且不将背景误标记为显著)、空间精度(显著图应定位准确并显示细节)和计算速度(为了使其成为视觉任务的预处理，显著图应尽可能快地创建)等等。

传统的显著性检测手段大多建立在自下而上的视觉注意模型以及底层特征的融合基础之上。Itti 等人提出了一种显著度检测架构[3]，该架构采用了多通道和多尺度的特征融合，从而模拟了早期视觉系统中的中心－周边对比度过程。该模型于颜色、亮度以及方向这三个通道中提取多尺度特征图，同时通过计算中心区域和周边区域的差别来获取像素的显著值。在接下来的几年里，基于这个定义的第一轮大规模研究活动结束了，出现了许多改进的模型。

除了局部对比与频域分析之外，其他的经典方式还涵盖：基于随机游走以及图论的图结构办法(像 Graph-Based Visual Saliency 模型借助图节点来传播显著值[4])基于朴素贝叶斯的显著性概率模型、借助“物体性”(objectness)先验进行的显著性检测，还有结合少量标注开展训练的传统机器学习方式等等。这个原理直观，计算容易，但它们对复杂背景缺乏鲁棒性。大多数经典方法使用人工设计的显著性。比如，在前景和背景的对比度不高或者背景干扰较为强烈的情况下，单独的低层特征通常难以精确辨别出显著目标。

2.2. 多特征融合方法

为增强复杂状况中的显著性检测能力，研究者引入了多特征融合这一理念，也就是将多种相互补充的特征线索加以整合以判定显著性。多尺度特征融合是一种将不同尺度的图像信息结合起来的常用方法，它可以同时考虑目标的全局轮廓和局部细节。举例来说，Klein 与 Frintrop 借助信息论的方式来度量多通道特征的显著差别[5]：他们针对亮度、颜色、方向等众多特征通道构建了具有伸缩性的特征表述，并通过计算中心区域和周边区域特征分布的 Kullback-Leibler 散度来判定显著性。然后，不同的特征在感知过程中被赋予不同的作用：颜色和亮度与背景形成对比，纹理和运动提供边缘，形状和运动提供有关物体的结构和动态信息。合理地整合这些不同质的特征，有利于突破单一特征的限制，增强检测的全面性与准确性。除了把手工设计的特征给予融合之外，对多模态数据展开融合也是极为关键的发展趋向，深度与颜色相互结合所有的优势十分明显：彩色图像蕴含着丰富的纹理以及颜色信息，然而当场景中的前景与背景颜色相近时，其效果大多时候不尽如人意，深度图像包含着关于场景的几何信息，可用来把前景与复杂的背景区分开来。

2.3. 深度学习在显著性检测中的应用

CNN 的出现为显著性检测的进步提供了巨大助力。端到端学习的优点在于特征是从数据中学习得到的，无需事先设计重要信息，这使得模型更具灵活性。典型的 CNN 显著性检测模型运用的是编码器－解码器的结构：编码器(常常是经过预训练的分类网络，像 VGG、ResNet 之类)对多层次的特征图予以提取，解码器则逐层将这些特征进行融合并实施上采样操作，从而形成与输入具有相同尺寸的显著图。Hou 等人提出的 DSS 模型就是一个例子[6]。它将每个卷积层的特征图作为输入，并通过层融合和迭代细化来提高边缘信息的准确性。

Transformer 作为近些年来新兴的另外一种深度模型，于显著性检测方面同样呈现出发展潜力。一种最初用于自然语言处理，后来用于视觉领域的神经网络形式。“转换器”是一种使用“自注意力”机制对

输入序列的全局相关性进行建模的机制。不过，纯 Transformer 模型于显著性检测里有着因过度平滑从而造成细节丢失的情况。神经网络和变形器的混合方法现在是解决此问题的最佳方法。这些将 CNN 与 Transformer 相融合的模型在显著性检测基准方面有着出众的表现，据悉在多项评价指标上都超越了当下最新的二十多种 Liu 等人使用纯“Transformer”架构来设计显著度检测方法[7]，该方法具有长距离的序列间依赖关系。其令牌融合和上采样策略实现了高分辨率的输出图。

总的来说，显著性检测的发展呈现出从传统方式向深度学习方式转变的历程。当前的显著图模型基于卷积神经网络的多层次表示和转换器的全局依赖建模的结合。与以前的模型相比，当前模型的性能指标有了显著提高。基于这一理论，在下一节当中，将会阐述我们所提出的基于多特征融合的显著性检测模型以及其实现的具体细节。

3. 方法与模型

3.1. 研究思路与整体框架

本文构建了一种基于多特征融合的图像显著性检测模型，其整体运用了“编码器－融合模块－解码器”这样的结构。其概念是，在通过多层次和多特征编码获得丰富特征后，通过特殊的融合程序对这些特征进行组合和融合，最后逐渐提高解码的分辨率，以输出高质量的空间信息图。如图 2 呈现出了模型的整体框架概貌。在编码阶段，每种输入的数据类型都被单独处理。为了得到多尺度信息，主干编码器于不同的尺度(分辨率)当中对输入予以处理：高分辨率支路能够留存细致的细节，低分辨率支路能够获取全局语义。具体来讲，我们构建了三种尺寸各异的图像金字塔输入(像是原尺寸、二分之一尺寸、四分之一尺寸)，然后将它们分别输入到结构相同但参数彼此独立的卷积编码器里面。

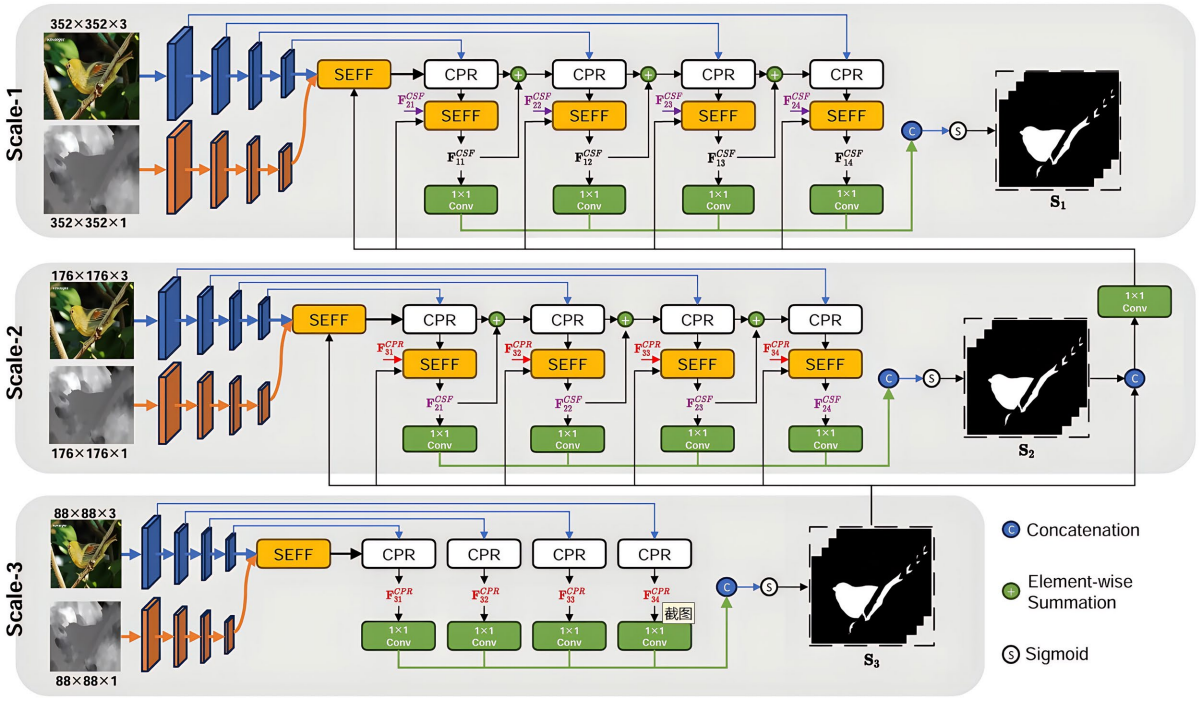

Figure 2. The overall architecture of the proposed multi feature fusion saliency detection model
图 2. 所提多特征融合显著性检测模型的整体架构

在完成特征提取工作后，我们构建了具有显著增强效果的特征融合模块，用于融合处理多模态、多

尺度的编码特征。在这一步中，对来自先前组织规模的特征和来自先前组织规模的显著图进行门控和增强，以提高融合特征的辨别力。具体的操作方式为：把上一尺度已经预测出来的显著图当作引导，针对当前尺度的 RGB 特征以及深度特征各自实施加权，强化显著区域的特征、削弱非显著区域的特征，随后再把二者加以融合(比如逐像素相加或者级联卷积)。这种由显著图引导的融合能够保障融合流程着重凸显出真正关键的区域特性，降低无关背景特性造成的干扰。我们的融合输出既是两种输入的混合，但仍处于合理的维度，这使得分解和利用变得容易得多。融合模块所输出的特征不仅整合了不同模态的信息，而且还融合了不同尺度的上下文。解码器获取融合特征，并通过连续的上采样和卷积操作生成与输入大小相同的概率图。为增强解码阶段的细节还原精准度，我们于解码流程里导入跳跃连接，把编码器浅层的高分辨率特性跟解码特性加以融合，助力复原目标的边缘与细节。

本模型的独特之处在于对多源信息予以充分运用：不但融合了空间尺度信息，还融合了多模态信息，并且借助显著性引导的手段让融合过程能够更具针对性地突出关键特征。尽管从单一来源获取的数据可能不足以对目标进行准确识别，但多特征融合使模型能够在背景变化、尺度变化和跨模态的复杂情况下区分重要对象。

3.2. 多尺度特征提取

在自然场景里，显著目标的大小存在差别：有的占据了整幅图像，有的则只在很小的区域内。在同一个模型中，必须检测大小不同的物体，因此引入了多尺度特征提取机制。也就是说，编码器不但在原始分辨率中进行特征提取，而且还会在其下采样版本上同步开展特征提取工作。这些分支中的每一个都独立地适应其规模。这些高级分支注重细节，而低级分支提取抽象和全局语义。在融合模块与解码器里，不同尺度的特征逐步进行整合，一同对最终的显著预测结果产生作用。

多尺度处理会使模型的计算量有较大提升，我们尝试借助把计算分配到更多权重上，并且降低网络的深度来减小复杂性，以此分散权重或者减少冗余计算，同时引入融合模块在尺度之间传递信息，使低尺度分支可参考高尺度已算出的较大图，对其特征提取给予引导，避免每个尺度都从头开始学习较大性判别。借助这样的方式，该模型可整合多尺度信息，同时将计算成本控制在合理范围之内，多尺度效应不再成为问题。

3.3. 多模态信息融合

除了多尺度这一方面，本模型的另一个关键之处在于 RGB-D 多模态信息的融合。将颜色和深度信息相结合以在 RGB-D 图像中创建深度信息的技术是本文的主题。深度图所给予的空间几何方面的线索在区分前景与背景时极为有效，然而深度数据自身或许存在噪声或者不完整的情况。我们采用多层次和多阶段的融合策略，在编码器和解码器的不同阶段融合 RGB 和深度特征，以便充分利用这两种模态的互补优势。

首先，于编码阶段的初期融合而言：我们在每个尺度的编码器里，在恰当的层融入跨模态注意力或者特征变换单元，促使 RGB 特征与深度特征在提取的时候便开始相互作用。在中间层，我们引入了一个双流交叉注意力网络，其中深度特征的显著度在 RGB 特征的相应位置得到增强，反之亦然。这种处于编码阶段的融合能够被视为针对特征的预融合强化。

随后在特征融合模块中开展主融合操作，这里未采用特征的简单融合方式，优先级是依据特征的较大度来确定的，把上一尺度或者上一级迭代所预测出的较大图当作引导信号，对 RGB 特征以及深度特征进行加权调制，随后将这二者合并。事实上，因为这个原则的合理性，在把两者结合起来时，对观看者真正关键的场景特征会更为突出，而场景中不关键的区域则不太容易被关注到。融合结束后，借助卷积实现通道的压缩投影，这能让两模态信息实现融合，还可将特征维度控制在合理范围，可后续的解码操

作，最终在解码阶段的后续融合环节：当逐层进行上采样以恢复空间分辨率时，促使 RGB 和深度的解码器不断进行信息交互，最终通过融合两种模态的注意力机制来构建。

3.4. 特征交互与注意力机制

在特征融合的流程里，我们运用了多样的注意力机制以推动不同特征的相互作用，增强显著性判别的成效。首先是通道注意力机制。鉴于不同通道的特征于显著性判定里的贡献存在差异，我们在融合特征的前后均引入通道注意力模块(像 SE 模块或者 CBAM 模块)，借由自适应地调整各个通道的权重，以凸显与显著目标有关的通道特征，并且对背景通道特征进行抑制。然后对特征进行归一化处理。这致使例如对于用以呈现目标颜色或者形状的通道，倘若其在当下场景里更具分辨能力，就会被给予更高的权重进行输出。

其次存在空间方面的注意力。融合阶段是两个模块相结合的地方。显著图是一种特殊的空间注意力，这意味着显著图提供了像素处于前景的概率，高概率区域可用于引导模型对其加以关注。另外，在解码的进程里，我们还能够算出融合特征的空间注意力图(像是依据局部对比或者梯度)，以此来优化输出的显著图，增进前景内部的均一性以及边缘的明晰度。

在更高层面上引入 Transformer 的自注意力理念，以实现长程依赖的特征交互，在此情形下，借助“点积”技术获取每个像素特征与图像其他特征间的相关性，这在处理多个关键目标或消除背景相似带来的干扰方面优势明显，因为模型能“洞察”任意两个区域的关联性实验证明。这种设计可提升复杂场景中较大图的质量。

3.5. 网络训练与优化策略

在模型的训练进程里，我们运用了为显著性检测任务专门定制的训练策略以及损失函数的设计方案，以此保障模型能够有效收敛且具备出色的泛化性能。

在损失函数领域，将像素级二元交叉熵损失(Binary Cross-Entropy, BCE)作为主要优化目标，凭借对预测较大图与真值掩膜逐像素对比来计算误差，为鼓励模型生成结构更正确的结果，添加了结构相似性损失函数，将较大图划分为区域和对象两个层次并与真值对比，在训练过程中提高前景内部区域与整体轮廓的匹配度。在一些实验中，尝试利用生成对抗训练思路，借助添加简单判别器区分预测较大图与真实值，使用对抗损失让较大图分布更接近真实标注图，但在具体实现中，对抗损失的权重需谨慎设置，避免训练不稳定。然后在验证集上实验，找到产生最佳结果的权重，最终总损失是上述各项的加权总和。

在优化算法方面，运用自适应矩估计优化器(Adam)训练模型，初始学习率设为 1e-4，训练过程采用多阶段学习率调度策略：当验证集性能停滞时，将学习率降至原来的 1/10，使模型细化。为防止过拟合，采用提前停止法：若测试集上的损失连续几轮未改善，训练中断，模型参数设为取得最佳结果时的值，训练所有模型时，运用数据增广手段，如随机裁剪、水平翻转、颜色扰动等，提升模型在不同场景下的适应能力。

其次，鉴于我们的模型涵盖了多尺度和多模块，我们施行了分阶段训练的策略。然后为注意力和融合单元赋予权重，并对模型进行进一步训练以学习多特征融合的细节。这种“两步走”的策略能够防止在一开始训练时因模型过于复杂而出现不收敛的情况。我们还为多尺度分支设计了逐步监督：我们不仅计算最终输出的损失，还计算中间尺度的预测图和下采样的真实图的损失，并反馈结果，以便能够监督分支的每个尺度的损失，并有助于多尺度特征学习。

4. 实验与分析

4.1. 实验环境与数据集介绍

我们针对本文所提出的模型在多个公开的数据集中展开了评测工作。实验中使用的硬件和软件为：

英特尔 i7-12700H CPU、16GB 内存和 NVIDIA RTX 3060 GPU；使用的操作系统是 Windows 11，tensorflow 包是 python 3.8，神经网络包是 pytorch 1.1。上述环境被用于模型的训练与测试工作。

我们介绍了 NJU2K 数据集的 RGB-D 图像的公开数据，该数据集由 2000 对室内和室外场景的图像组成。NLPR 数据集由 1000 对日常场景的图像组成。每一个数据集均给出了像素层级的显著目标真实标注(ground truth mask)。训练集由大约 1500 对图像组成，这些图像来自 NLU2K 和 NLPR 数据集，并且在训练过程中这些图像是混合的。使用之前工作中的常见划分，对训练图像进行了以下注释：值得指出的是，鉴于不同数据集的深度图分布存在显著差异，我们在训练过程中对深度通道实施了归一化操作并且进行了数据增广，以此来提升模型针对不同深度质量的稳健性。

在对模型展开训练时，我们仅仅借助上述的 RGB-D 训练集来实施有监督训练(原因在于模型架构对深度通道有所需求)。由于在纯 RGB 图像上进行测试的输入图像没有深度信息，我们使用了空白深度图或从 RGB 图像中获取的深度图，并且没有设置模型的参数，以便公平地评估模型在跨模式条件下的性能。

4.2. 评价指标

为了对显著性检测性能展开定量评估，我们运用了一系列常见的指标，像是 Precision-Recall 曲线、最大 F-measure、S-measure 以及平均绝对误差(MAE)等等。主要指标说明如下：

Precision-Recall (P-R)曲线：将连续的显著图依据不同的阈值进行二值化处理，然后和真值掩膜进行比较计算，从而获取多个(Precision, Recall)点并绘制出曲线。所谓的准确性和精确性是指被预测为重要的像素中实际上属于前景的像素所占的百分比。所谓的召回率是指被预测为重要的前景像素所占的百分比。P-R 曲线从数值方面展现了在不同阈值时模型的性能权衡状况，如图 3 所示曲线越接近右上方就意味着性能越出色。

F-measure (F_β)：它指的是 Precision 与 Recall 的加权调和平均，其公式为：

$$F_\beta = \frac{\left(\beta^2 + 1\right)PR}{\beta^2 \cdot P + R} \tag{1}$$

我们将最大 F-measure (max F)用作评价标准，也就是对所有阈值进行遍历以获取最大值 F。F 值为 1.0 时，确保最大可能召回率的模型将同时提供最大程度的准确率和召回率。

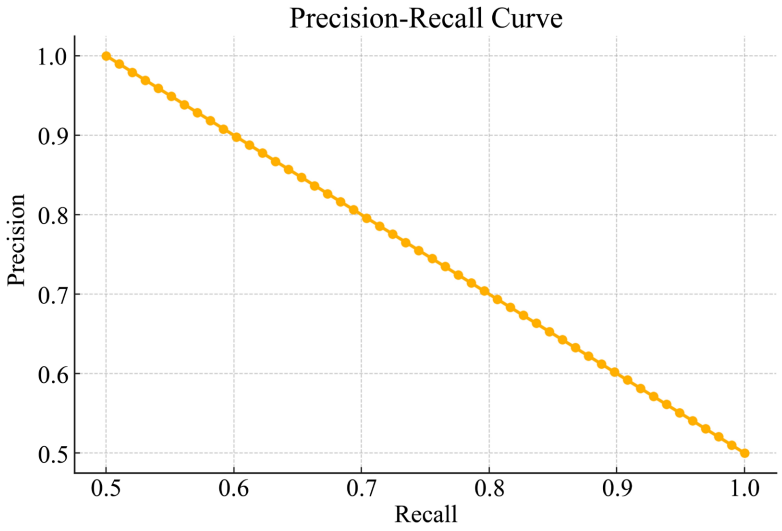

Figure 3. Precision recall curve
图 3. Precision-Recall 曲线图

S-measure (结构相似性度量)：此指标着重于预测显著图和真值在结构方面的相似状况，是由区域相似性 Sr 以及对象相似性 So 共同构成的。“形状匹配”标准是对结果保真度的综合度量，用于评估前景的整体布局以及目标细节的匹配情况。F-measure 仅仅着眼于像素的准确性，而 S-measure 则更为注重形状以及空间布局的一致性，如图 4 所示，其值越趋近于 1，意味着预测和真值在结构方面越相符。

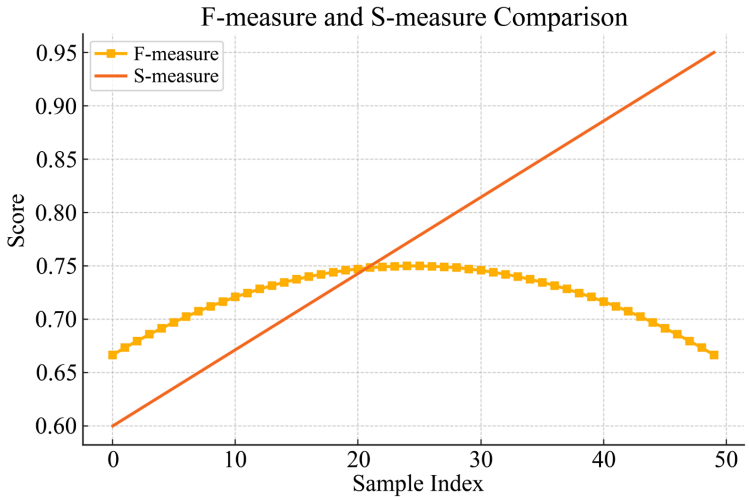

Figure 4. Comparison of F-measure and S-measure
图 4. F-measure 和 S-measure 对比图

Mean Absolute Error (MAE)：对预测显著图 S 与真值掩膜 G 之间逐像素差值的绝对值进行平均计算

$$MAE = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H}\left|S(x,y)-G(x,y)\right| \tag{2}$$

MAE 直接展现了像素强度的平均偏差情况，如图 5 所示，数值越接近 0 越佳，可用以评估显著图在整体上的像素级精准程度。它是评估漏检和过检情况的一个重要指标。

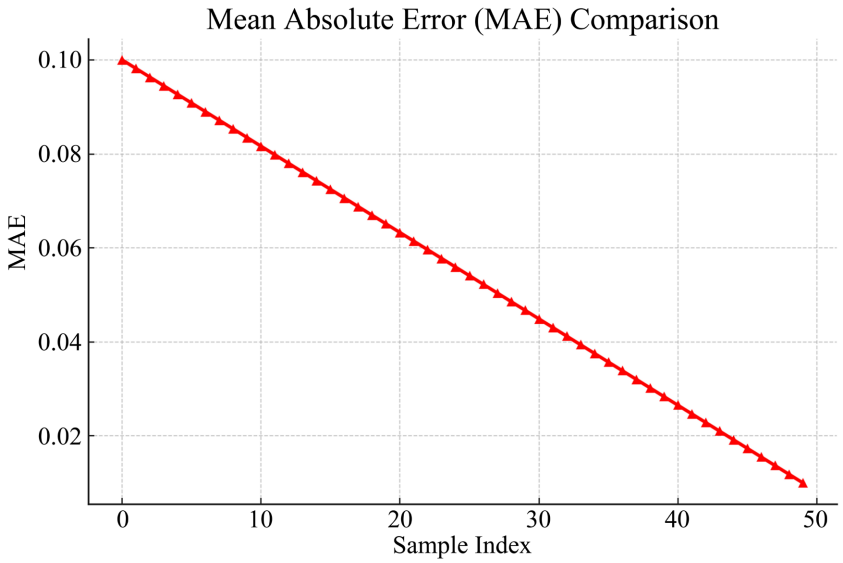

Figure 5. Mae comparison chart
图 5. MAE 对比图

本文重点阐述了上述几个较为常用的指标。最后，计算所有数据集每个指标的平均性能，这是模型性能最全面的指标。

4.3. 结果分析与可视化

为了更直观地凸显本文方法所具有的优势，我们在图 6 中呈现了不同场景下显著性检测结果的可视化对照。这是五个有代表性的案例。分别是常见案例(清晰的前景和简单的背景)、深度图中有噪声的案例(不准确的深度信息导致深度图不完整或不准确)、多个对象的案例(许多重要对象分散)、对比度低的案例(前景颜色接近背景颜色)以及小对象的案例(重要对象较小)。通过对比的手段选取了诸如 PopNet、BBSNet 等富有代表性的模型，并且将其与我们的成果展开对比。

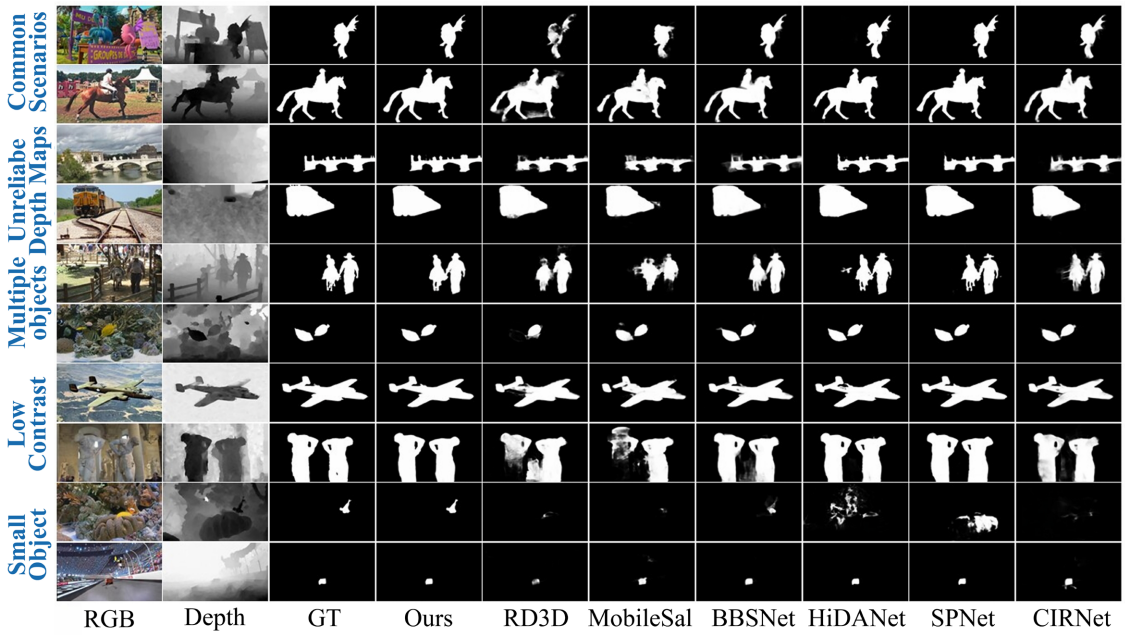

Figure 6. Comparison of significance detection results of different methods in various typical scenarios
图 6. 不同方法在各种典型场景下的显著性检测结果对比

由图 6 能够清晰地了解到，我们的方法在各类具有挑战性的场景中都获得了更为理想的显著图输出结果。然后计算平均线，该计算结果作为目标径向分布的平均值给出。比如在第一行的骑马图像里，我们的显著图能出色地把马及骑手从背景中分离出来，然而部分对比方法却忽略了马腿等一些细节区域。在这第二个场景中，深度图的很大一部分是错误的，因此其他方法可能会将一些背景错误地分类为显著区域(PopNet 结果中桥下的错误突出显示)，而我们的方法借助多尺度语义和显著性引导机制，能够成功避免深度图的影响，并主要突出桥梁的主要结构。这个结果更接近真实情况。在多目标的场景(第三行)里，画面中存在三个人物，我们的方法可以同步检测出所有人的轮廓，并且把他们相互区分开，然而有些方法不是漏检了其中的某个人，就是把几个人的前景混为一体。我们认为，我们引入的跨像素自注意力机制使模型能够独立对待每个对象，这就是它在包含许多对象的群组场景中表现更好的原因。

定量结果与定性分析均说明，本文所提的多特征融合模型在突出前景完整性以及抑制背景干扰方面有着优势，能看出，在较大图引导下，尺度间的融合给出了更连贯且更少碎片化的区域，注意力机制提高了前景和背景的对比度，让较大对象更突出，多模态特征融合使模型在 RGB 信息匮乏时，仍能有效借助深度等辅助信息，提升对复杂场景的适应能力。实验结果显示，虽模型参数较大，但融合策略在相对

简单网络中运行良好，融合策略与模型无关且有可移植性，确实我们也发现模型存在一定缺陷：在极度繁杂背景中，可能会出现少许误检情况，对于深度图严重扭曲甚至失效的状况，模型发挥的作用有限，不过在本文的较大性检测任务中，我们的方法呈现出了出色性能和鲁棒性。

诚然，我们同样发觉模型存在一定的缺陷：处于极为复杂的环境中，或许会有少量的误检情况；倘若深度图严重扭曲(乃至彻底失效)，模型所起到的改进作用比较有限。总之，本文提出的显著度检测方法在广泛的标准显著度检测任务中显示出高度的一致性和鲁棒性。

5. 总结与展望

在复杂场景中，由于显著性检测往往受到单一特征信息的制约，本文给出了一种基于多特征融合的图像显著性检测手段。该模型将显著图与注意力机制相结合，并通过多尺度和多模态网络框架，能够成功融合颜色和深度等各种信息，突出整体画面并保持细节。实验结果显示，此方法在多个公共的 RGB-D 数据集中都优于当下的主流方法，于 Precision-Recall 曲线、F-measure、S-measure 以及 MAE 等指标方面获得了更好的表现。尤其是在前景背景对比度不高、存在众多目标、深度信息含有噪声等充满挑战的状况下，我们的方法仍旧展现出强大的鲁棒性。显然，这是一个重要的结果，证实了多特征融合策略在确定显著性方面的有效性和优越性。

不过，本文的工作仍有部分缺陷以及需要继续完善的地方。该模型相当复杂，其高复杂性主要源于多尺度子程序和融合分支。这在某种程度上对模型在实时性需求更为严苛或者硬件资源有限的场景中的部署造成了制约。作为未来的工作，可以研究最轻量的网络结构，例如共享一些模型参数、修剪模型的某些部分，或者用更高效的转换节点替换，以在保持性能的同时减小模型的大小。其次，倘若深度信息缺失或者有误，那么模型进行显著性判断时仍有可能受到干扰。该研究团队的下一个任务将是找到一种对深度噪声不太敏感的融合机制。也有可能结合其他类型的信息(例如红外和偏振数据)来弥补深度的不足。

本文研究工作说明，把多尺度与多模态特征相融合，同时引入注意力机制，可有效提升较大性检测性能，随着图像中可提取特征数量增多，从图像里挑选能提取的特征变得日益显著，未来会在该方向深入研究，提出如模型分解、弱监督学习以及从视频序列中开展大规模特征提取等新想法，持续改进较大检测技术，为计算机视觉领域相关任务给予有力支撑。

基金项目

江苏省大学生创新训练计划项目(xcx2024189)。

致　谢

本文研究与撰写时，获诸多人士大力协助与支持，于此向诸位致以诚挚谢意，极感谢队友秦培鑫，因他认真投入且积极协作，论文才得顺利完成，周文婧、时晶晶、李辉三位队员于研究背后默默给予无私支持与帮助，他们在资料整理、实验调试以及数据分析方面的贡献，为本文顺利完成奠定坚实根基。本文完成亦离不开导师李子龙老师悉心指导与细致帮助，他在论文选题、理论梳理、实验设计及成果分析等环节，均给出宝贵建议与无私指导，再次向所有给予本文帮助与支持之人表达诚挚谢意！

参考文献

[1] Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S. (2009) Frequency-Tuned Salient Region Detection. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 1597-1604. https://doi.org/10.1109/cvpr.2009.5206596

[2] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[3] Itti, L., Koch, C. and Niebur, E. (1998) A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1254-1259. https://doi.org/10.1109/34.730558

[4] Harel, J., Koch, C. and Perona, P. (2007) Graph-Based Visual Saliency. In: Schölkopf, B., Platt, J. and Hofmann, T., Eds., *Advances in Neural Information Processing Systems* 19, The MIT Press, 545-552. https://doi.org/10.7551/mitpress/7503.003.0073

[5] Klein, D.A. and Frintrop, S. (2011) Center-Surround Divergence of Feature Statistics for Salient Object Detection. 2011 *International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 2214-2219. https://doi.org/10.1109/iccv.2011.6126499

[6] Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z. and Torr, P. (2017) Deeply Supervised Salient Object Detection with Short Connections. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5300-5309. https://doi.org/10.1109/cvpr.2017.563

[7] Liu, N., Zhang, N., Wan, K., Shao, L. and Han, J. (2021) Visual Saliency Transformer. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 4722-4732. https://doi.org/10.1109/iccv48922.2021.00468