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

中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)12-3838-17

论文引用格式:Song X G, Tan Y P, Guo F Q, Lu X F and Hei X H. 2025. Cross-modal feature fusion and detail-enhanced RGB-D salient object detection. Journal of Image and Graphics, 30(12):3838-3854(宋霄罡, 谭裕平, 郭富强, 鲁晓锋, 黑新宏. 2025. 跨模态特征融合与细节信息增强的RGB-D显著目标检测. 中国图象图形学报, 30(12):3838-3854)[DOI:10.11834/jig.240653]

跨模态特征融合与细节信息增强的RGB-D显著目标检测

宋霄罡[1,2*],谭裕平[1],郭富强[1],鲁晓锋[1,2],黑新宏[1,2]

1. 西安理工大学计算机科学与工程学院,西安　710048; 2. 人机共融智能机器人陕西省高校工程研究中心,西安　710048

摘　要:**目的**　RGB-D显著目标检测通过整合RGB图像和深度图像的互补信息,可以提高应对复杂和具有挑战性场景的显著目标检测(salient object detection,SOD)能力,取得了比RGB显著性检测模型更好的性能,受到高度关注。然而,现有RGB-D检测模型面临如何高效利用输入的多模态信息进行融合以及如何提高显著目标边缘检测精度等问题。为此,提出一种跨模态特征融合与边缘细节增强的RGB-D显著目标检测方法。**方法**　通过跨模态注意力融合增强模块(cross-modal attention fusion enhancement module,CAFEM)对不同模态特征进行注意力整合,使RGB图像和深度图像的互补信息充分融合,使模型充分利用多模态特征,从而提高模型的性能。但是两种模态的输入容易出现背景信息混淆、噪声增多、深度图质量低和目标轮廓提取困难的情况。为应对上述问题,提出一种卷积神经网络(convolutional neural network,CNN)低层特征引导的边缘特征提取模块(boundary feature extraction module,BFEM),通过通道注意力对低层特征携带的噪声进行过滤,然后使用低层细节特征引导跨模态融合特征进行聚焦解码以得到更加准确的显著图像。**结果**　在4个RGB-D显著目标检测数据集进行实验,与16种代表性方法进行定量和定性实验对比。在平均绝对误差(mean absolute error, MAE)指标上,本文方法相较于排名第2的方法,在4个数据集上分别提升6.9%、10.5%、9.7%和2.4%。结果表明,本文方法在各场景均有优异表现。**结论**　提出一种用于RGB-D显著目标检测的跨模态特征融合与细节信息增强网络(cross-modal feature fusion and detail-enhanced network,CFADNet),通过跨模态注意力融合增强模块(CAFEM),较好地实现了RGB特征与深度特征的融合。此外,构建了边缘特征提取模块(BFEM)提取低层细节特征,最终较为准确地定位显著物体并增强了边缘细节的清晰度。

关键词:显著性目标检测(SOD);注意力机制;跨模态;特征融合;边缘细节增强

Cross-modal feature fusion and detail-enhanced RGB-D salient object detection

Song Xiaogang[1,2*], Tan Yuping[1], Guo Fuqiang[1], Lu Xiaofeng[1,2], Hei Xinhong[1,2]

1. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an　710048, China;

2. Human Machine Integration Intelligent Robot Shaanxi Provincial University Engineering Research Center, Xi'an　710048, China

收稿日期:2024-11-07;**修回日期**:2025-04-21;**预印本日期**:2025-04-28

*通信作者:宋霄罡　songxg@xaut.edu.cn

基金项目:国家重点研发计划资助(2022YFB2602203);国家自然科学基金项目(52372418,62076201);陕西省重点研发计划资助(2023GXLH-043);西安理工大学硕士研究生创意创新种子基金项目

Supported by: National Key R&D Program of China(2022YFB2602203); National Natural Science Foundation of China (52372418, 62076201); Key R&D Program of Shaanxi Province, China(2023GXLH-043); Seed Fund for Creativity and Innovation of Postgraduates of Xi'an University of Technology

Abstract: **Objective** RGB-D salient object detection (SOD) combines complementary information from RGB and depth images, offering substantially enhanced performance in complex and challenging scenes compared to RGB-only models. This technique has gained considerable attention in the academic community due to its capability to effectively capture salient objects by leveraging visual and spatial information. However, existing RGB-D detection models face several key challenges. First, efficiently utilizing and fusing multi-modal information from RGB and depth inputs remains a difficult task due to the inherent differences between the two modalities. RGB images provide rich color and texture details but lack depth information, whereas depth maps offer spatial cues but are often noisy or of low quality. Second, achieving accurate boundary detection is particularly challenging in cluttered or noisy environments. Noisy depth maps and cluttered backgrounds can obscure object contours, making it difficult to predict sharp and precise boundaries. These challenges highlight the urgent need for a robust model that can effectively integrate RGB and depth information while simultaneously addressing noise and enhancing boundary precision. **Method** Aiming to address these challenges, a novel method, the cross-modal feature fusion and detail-enhanced RGB-D salient object detection network (CFADNet), is introduced. The proposed network incorporates two innovative modules: the cross-modal attention fusion enhancement module (CAFEM) and the boundary feature extraction module (BFEM). The CAFEM is designed to enhance the integration of RGB and depth features by leveraging attention mechanisms that emphasize the most informative aspects of each modality. Specifically, channel attention is applied to the RGB features to suppress noise and enhance critical color and texture details. Similarly, spatial attention is applied to the depth features to emphasize spatial regions that are relevant for salient object detection. This attention-based fusion mechanism ensures that the model effectively retains global semantic information from the depth map while preserving fine-grained details from the RGB image. The fusion process is structured in multiple layers, progressively integrating features at different scales to fully utilize the complementary strengths of RGB and depth modalities. In contrast, the BFEM is specifically designed to improve the accuracy of salient object boundaries. Accurate contour detection is crucial for generating high-quality saliency maps; thus, BFEM leverages low-level CNN features, which are rich in edge and texture information. These features are refined through channel attention, which filters out noise and irrelevant details, enhancing the clarity of boundary-related cues. The refined features are then used to guide cross-modal feature decoding, ensuring that the final saliency maps exhibit sharp and accurate boundaries. By combining the edge-extraction capabilities of low-level CNN features with the semantic richness of cross-modal features, BFEM notably improves boundary precision in RGB-D salient object detection. **Result** Aiming to evaluate the performance of CFADNet, extensive experiments are conducted on four widely used RGB-D salient object detection datasets: NJU2K, NLPR, STERE, and SIP. These datasets encompass a wide range of diverse and challenging scenes, making them ideal for evaluating the generalization capability of the proposed model. CFADNet is compared against 16 state-of-the-art RGB-D salient object detection methods, including DCF, CIRNet, and CAVER, using standard quantitative metrics such as mean absolute error (MAE), F-measure(F_β), and structural similarity (S_α). CFADNet demonstrated superior performance across all datasets, particularly excelling in the MAE metric. Specifically, this network outperformed the second-best method by 6.9%, 10.5%, 9.7%, and 2.4% on the NJU2K, NLPR, STERE, and SIP datasets, respectively. These substantial improvements highlight the effectiveness of the attention-based fusion strategy and edge refinement mechanisms. Furthermore, CFADNet consistently achieved higher F-measure and Sα scores, indicating that the model not only reduces pixel-level errors but also more accurately preserves the overall structure and shape of salient objects compared to competing methods. In addition to quantitative evaluations, qualitative comparisons are conducted to visually assess the performance of CFADNet in various challenging scenarios. Results show that the proposed method generates saliency maps with sharp and accurate boundaries, even in cases where salient objects exhibit complex edges or are embedded in cluttered and noisy backgrounds. This finding demonstrates the robustness of CFADNet in handling difficult scenes by effectively separating salient objects from their background while preserving fine boundary details. The visual results further confirm that CFADNet successfully captures global semantic information and local detail, ensuring accurate identification and clear isolation of salient objects from the background. **Conclusion** This paper presents CFADNet, a cross-modal feature fusion and detail-enhancement network for RGB-D SOD, designed to address the two major challenges: effective multimodal feature fusion and accurate boundary detection. CFADNet introduces two novel modules, the CAFEM and the BFEM. CFADNet effec-

tively integrates RGB and depth information while notably enhancing the precision of salient object boundaries. The attention mechanisms used in the CAF0EM enable the network to fully leverage the complementary information from RGB and depth modalities. Simultaneously, the BFEM module focuses on refining edge details, resulting in sharper and more accurate saliency predictions. Extensive experiments conducted on four benchmark datasets demonstrate that CFADNet consistently outperforms existing state-of-the-art methods, achieving superior performance across key evaluation metric, including MAE, F-measure, and structural similarity index. These findings highlight the robustness and strong generalization capability of CFADNet in diverse and challenging environments. By combining attention-based feature fusion with effective edge refinement, CFADNet emerges as a powerful and reliable solution for RGB-D salient object detection into complex scenarios. Future research could explore extending this approach to other multi-modal tasks, such as RGB-Thermal or multi-spectral image processing, where challenges related to multi-modal fusion and boundary detection are also prevalent. Additionally, optimizing the computational efficiency of CFADNet for real-time deployment represents a potential research direction, enabling its application in time-sensitive applications such as autonomous driving and robotics.

Key words: salient object detection(SOD); attention mechanism; cross-modal; feature fusion; edge detail-enhancement

0 引 言

视觉显著性检测利用算法模拟人类视觉，评估图像中不同部分的吸引力程度，称为显著性，最终生成显著性图像。显著性目标检测(salient object detection，SOD)专注于显著目标的分割，通过在图像处理算法中结合显著性目标检测技术，可以使算法得到优化，资源能够得到高效利用。在计算机视觉的各个领域，已经有很多算法模型将显著性检测技术作为它们的预处理操作，如图像检索(Gao等，2012)、照片裁剪(Wang等，2019)、场景分类(Ren等，2014)、语义分割(Wang等，2024)以及视频分割(Wang等，2021)等。除了在计算机视觉研究领域作为辅助研究和人工智能化系统中发挥作用之外，该技术已经成功应用于一些实际生活场景中，例如：医学图像分割(Jahanifar等，2019；Chen等，2021)、安防监控领域(赵兴科 等，2021)、卫星图像领域(Liang和Luo，2024)、智能驾驶领域(Ding等，2024)以及深地工程领域(张茹 等，2024a)。

自进入深度学习(Chen等，2018b；Hou等，2019；Li等，2021)时代以来，基于卷积神经网络(convolutional neural network，CNN)的RGB SOD框架(Mei等，2022；Wu等，2022b；Zhu等，2021)得到了大力发展，远远超过了基于手工特征的方法。然而，由于卷积操作的感受野受限，CNN理论上可以通过加深网络层数获得更大的感受野。但随着网络层数的加深，不断下采样导致了更多的特征丢失，并且无法预测出完整的显著目标以及锐利的边缘。此外，RGB输入虽然能够提供纹理细节、颜色特征以及对比度等信息，但是其缺少空间信息，并且受光照条件影响会使RGB图像变得模糊，提取到的纹理细节特征变少、噪声增加，导致预测精度下降，模型出现性能瓶颈。为了克服RGB图像质量变差的问题从而引入了深度信息，通过深度信息携带的空间特征建模全局关系，从而弥补RGB特征缺失的全局语义信息。

虽然在过去的几年里，已经提出各种基于RGB-D的显著目标检测模型，这些模型虽然引入了深度信息，但对于如何利用多模态信息以及如何增加边缘特征仍然存在着问题。近几年基于RGB-D的显著目标检测模型倾向于探索高效融合多模态特征的方法：Ji等人(2021)提出一个学习策略来校准原始深度图中的潜在偏差，以提高SOD性能，并提出一个简单的交叉融合模块，融合RGB和深度特征。Chen等人(2018a)提出一个互补感知的融合模块来集成跨模态和跨级特征表示。它可以通过显式地利用跨模态和水平的连接以及模态和水平的监督来有效地利用互补信息，以减少融合的模糊性。Fang等人(2024)提出一种模态净化模块和一种尺度统一模块来融合多模态特征。Chen等人(2023)提出三重编码器网络，通过多模态特征交互模块自适应评估模态重要性。Cheng等人(2023)引入了一个嵌入了ID信息的跨模态交互块(cross-modal interaction block，CMIB)。CMIB可以提取和融合不同模式的特征，交互作用有助于深度神经网络学习互补特征，减少语义差距。孙福明等人(2024)设计了CNN-Transformer网络架构，借助注意力机制学习深度图像和RGB图像之间的互补信息，并且将跨模态融合

特征输入到RGB分支中，以充分利用不同模态的特征信息。Cong等人（2023）通过跨模态点感知交互模块约束特征交互位置，并利用CNN细化单元缓解Transformer的块效应。

然而，上述方法将每一级的深度特征或经过融合的特征输入到RGB特征提取主干中，或将每一级融合特征单独输出，未与后续融合特征进行交互，直接使用解码器生成最终的显著性图像。这会影响特征提取主干对显著特征以及边缘细节信息的提取，降低对融合特征的利用率，导致SOD无法更好地泛化适应更多的场景，在一些背景复杂、边缘纹理较多的情况下的检测效果不理想。本文通过采用注意力融合增强的方式，将深度特征与RGB特征逐级融合，使每一级融合特征充分交互，不将融合特征输入到特征提取主干，而是直接输入到解码器进行解码，以降低融合后的特征对主干网络的负面影响，使主干网络专注于提取全局语义信息与局部细节信息，无须对融合特征进行参数的调整。

基于上述分析，本文提出一种跨模态特征融合与边缘细节增强的RGB-D显著目标检测方法，使用双分支Transformer进行特征提取，克服了CNN感受野受限导致的语义信息提取不足的问题。为了更好地融合利用多模态特征，并减少深度图质量差带来的负面影响，利用通道注意力机制对融合特征进行通道过滤，并使用深度图像提取空间注意力权重，对融合特征进行空间过滤。为了充分保留全局语义信息，本文将融合特征进行逐层拼接并利用自注意力机制建立逐层语义信息之间的依赖关系。由于Transformer提取到的局部细节信息较少，本文在网络的末端加入两层CNN特征提取网络，使用CNN提取低层细节信息，利用通道注意力机制将低层特征进行过滤后融入多模态融合特征中，从而丰富特征中的边缘细节，能够使预测出的显著目标完整且具有更锐利的边缘。

本文主要贡献如下：1）设计一个双分支Transformer网络架构，将Transformer提取的多模态全局语义特征逐步融合，最后通过CNN提取局部细节特征，丰富融合特征中的边缘细节以增强特征表示。2）设计跨模态注意力融合增强模块，兼顾RGB与深度特征。对于RGB通道间信息，通过通道注意力融合，深度信息的空间特征则通过空间注意力加权，以此实现全局语义、局部细节和通道间依赖信息的融合，并有效抑制噪声，确保模型在复杂场景中的精准预测。3）设计CNN引导的边缘特征提取模块，通过CNN提取低层细节特征，并利用通道注意力过滤噪声后将提取的细节特征逐步融合到显著特征中。使最终预测的特征中包含大量的语义，同时又有细节特征的补充。4）采用预训练的P2T（pyramid pooling transformer）作为骨干网络，在4个RGB-D显著目标检测测试数据集上的实验结果表明，本文提出的方法能够有效地利用RGB和深度信息，并通过提取低层细节信息提高显著目标检测的精度。

1 相关工作

1.1 RGB-D显著性目标检测

在过去的几年中，许多基于手工制作特征的传统RGB-D显著性目标检测模型已经被开发出来（Liang等，2018）。如早期工作（Ciptadi等，2013）中，专注于由RGB图像和深度图生成的布局和形状特征之间的交互建模，Peng等人（2014）开发了一个新的多阶段RGB-D模型，并构建了第一个大规模的RGB-D基准数据集：NLPR（National Laboratory of Pattern Recognition）。但是上述传统方法由于手工制作特征的表达能力有限，其显著目标检测性能不理想。为了解决这一问题，一些研究已经转向使用深度神经网络融合RGB-D特征（Zhang等，2020）。这些模型可以学习高级特征表示，探索RGB图像和深度线索之间的复杂相关性，以提高显著目标检测性能。张晴等人（2019）通过多尺度超像素分割、深度特征提取和多核增强学习，有效抑制复杂图像中的无关背景区域。Ji等人（2021）提出一个学习策略来校准原始深度图中的潜在偏差，以提高SOD性能，并提出一个简单而有效的交叉融合模块，融合RGB和深度特征。Chen等人（2018a）提出一个互补感知的融合模块来集成跨模态和跨级特征表示，它可以通过显式地利用跨模态和水平的连接以及模态和水平的监督来有效地利用互补信息，以减少融合的模糊性。Lee等人（2022）提出一种新的超像素原型采样网络体系结构，将输入的RGB图像和深度映射分割为组件超像素，生成组件原型，使该网络只对显著对象对应的原型进行采样，消除了非显著性对象的影响。跨模态视图混合变换器（Pang等，2023）引入了Transformer，从序列到序列的角度重新考虑

双模态SOD建模，从而获得了更好的可解释性，并构建了一个基于自顶而下的转换器的信息传播路径，由视图混合注意块增强来充分利用来自空间和通道视图的模态间和模态内信息。Zhang等人(2023)提出一种基于校准后融合的RGB-D两阶段SOD模型，同时考虑低质量图像和前景不一致图像对显著性检测的影响。在图像生成阶段，从原始输入的RGB-D图像对中选择高质量、前景一致的深度图像作为伪深度图像生成网络的监督信息；在推理阶段，校准不可靠的深度信息，然后从RGB-D图像中捕获更多的跨模态信息，用于最终的预测。

1.2 RGB-D的融合方法

早期融合的方法主要为输入融合和早期特征融合。RGB图像和深度图像直接拼接形成一个四通道输入，称之为输入融合(Ren等，2015)。RGB图像和深度图像输入到单独的网络进行特征提取，之后利用提取的低级特征拼接，然后输入后续网络进一步显著预测，称之为早期的特征融合。后期融合也分为后期特征融合和后期结果融合，采用两个并行网络流分别学习RGB和深度图像的高级特征，并将它们拼接起来，生成最终的显著预测，称之为后期特征融合(Han等，2018)。利用两个并行网络流分别获得RGB图像和深度图像的显著预测图，然后将两个显著预测图拼接起来，得到最终的预测图，称为后期结果融合(Ding等，2019)。为了有效地探索RGB图像与深度图像之间的相关性，学者们提出一种多尺度融合策略，例如，Cong等人(2023)引入了CNN辅助的Transformer架构，考虑到RGB模态和深度模态之间的先验相关性，设计了一个注意触发的跨模态点感知交互模块来探索具有位置约束的不同模态的特征交互，并设计了一个CNN诱导的细化单元，用于内容的细化和补充。这种方法将跨模态交互引入到多层中，可以提供额外的梯度来增强对深度流的学习，并使低级和高级表示之间的互补性得以探索。Zhang等人(2023)提出一种通过解耦动态卷积实现的交叉动态滤波网络，通过一个动态增强模块，利用全局上下文指导动态地增强了模态内特征，同时提出一个场景感知的动态融合模块来实现两种模式之间的动态特征选择。Chen等人(2023)提出三重编码器网络，通过多模态特征交互模块自适应评估模态重要性。叶欣悦等人(2024)通过设计新的互补信息交互模块和跨模态特征融合模块，实现了互补信息交互融合网络。然而，对于两种模态的输入导致的背景信息与前景混淆，噪声增多，以及目标轮廓提取困难的问题，目前缺乏统一的解决方案。

2 本文方法

通过整合RGB图像和深度图的互补信息，可以提高对复杂和具有挑战性的场景的显著目标检测能力。本文提出一种跨模态融合和细节信息增强的RGB-D显著性目标检测方法。首先，以P2T(Wu等，2023)为特征提取主干，分别对RGB图像与深度图像进行特征提取，并通过跨模态注意力融合增强模块(cross-modal attention fusion enhancement module，CAFEM)对两种模态特征进行融合，以挖掘两种模态中显著性特征的共性与互补特征。将融合特征输入到Transformer解码器进行解码，然后，通过构建的边缘特征提取模块(boundary feature extraction module，BFEM)生成边界引导特征，并对每层跨模态融合特征进行边缘细化，确保生成显著物边界的完整性。最后，本文设计了一个CNN低层特征引导模块，使用CNN提取的低级细节特征引导跨模态融合特征进行聚焦解码以得到更加准确的显著图像。

2.1 网络结构

所提跨模态特征融合与细节增强网络(cross-modal feature fusion and detail-enhanced network，CFADNet)架构如图1所示，该网络主要分为双分支Transformer特征提取、跨模态融合、边界特征提取、VGG16(Visual Geometry Group)(Simonyan和Zisserman，2015)低层细节特征提取以及聚焦预测部分。

首先，网络通过双分支的预训练P2T特征提取网络(*P2T*)对输入RGB图像深度图进行特征提取，表示为

$$\begin{cases} F_r^i = P2T\left(\boldsymbol{I}_{\text{rgb}}\right) \\ F_d^i = P2T\left(\boldsymbol{I}_{\text{depth}}\right) \end{cases} \tag{1}$$

式中，$\boldsymbol{I}_{\text{rgb}}$和$\boldsymbol{I}_{\text{depth}}$分别表示RGB和深度输入，$F_r^i$和$F_d^i$分别表示P2T提取的RGB特征和深度图像特征。RGB特征与深度特征存在模态差异，RGB特征包含大量的细节和纹理信息，深度特征包含更多的空间信息，可以很好地定位显著目标。因此，针对两种模态特征的特点设计了跨模态注意力融合增强模块

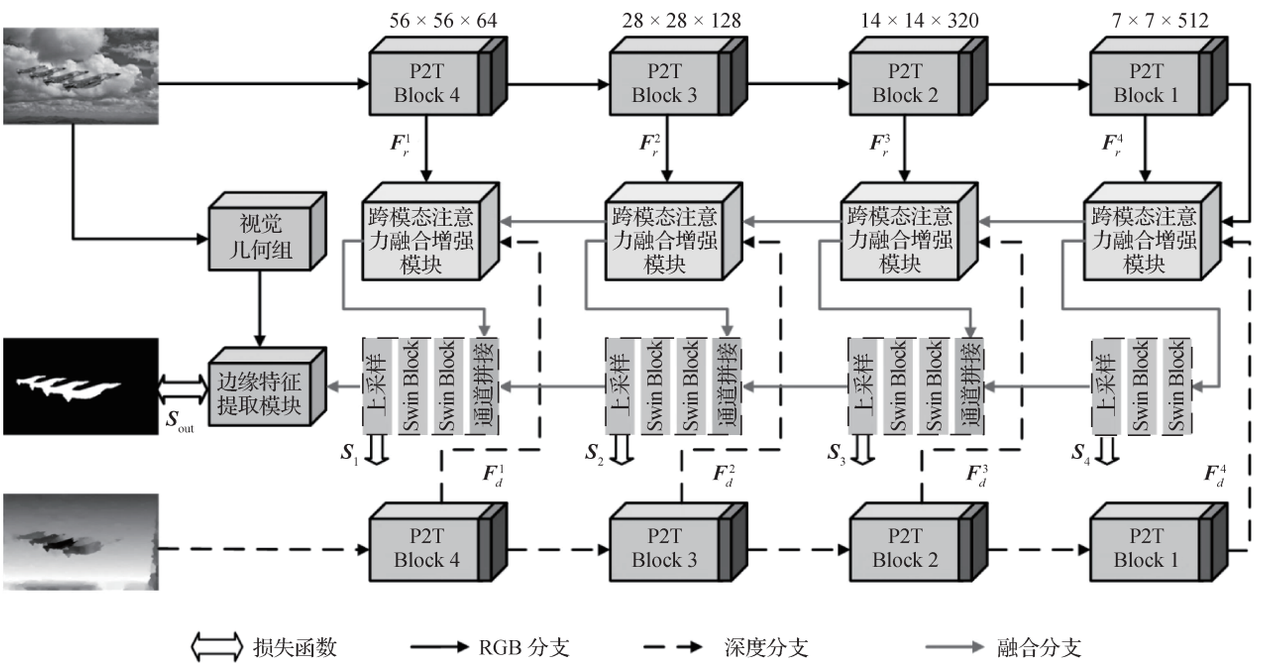

图 1　跨模态特征融合与细节信息增强网络架构

Fig. 1　Cross-modal feature fusion and detail-enhanced network (CFADNet) architecture

(CAFEM)对两部分特征进行注意力加权融合，并将融合后的多模态混合特征输出到Transformer解码器进行解码。

在网络末端将解码后的特征和VGG提取的细节特征进行聚焦定位细化，增加空间信息、语义信息和局部细节信息之间的融合交互，从而使模型能够充分利用多模态信息以及边缘预测的信息，输出完整且边缘轮廓清晰的预测图，以上过程可表示为

$$\begin{cases} \boldsymbol{Z}_i = CAFEM\left(\boldsymbol{F}_r^i, \boldsymbol{F}_d^i, \boldsymbol{Z}_{i-1}\right) \\ \boldsymbol{S}_i = Decoder\left(\boldsymbol{Z}_i, \boldsymbol{S}_{i+1}\right) \\ \boldsymbol{F}_{\text{vgg}}^{112}, \boldsymbol{F}_{\text{vgg}}^{224} = VGG\left(\boldsymbol{I}_{\text{rgb}}\right) \\ \boldsymbol{S}_{\text{out}} = BFEM\left(\boldsymbol{S}_i, \boldsymbol{F}_{\text{vgg}}^{112}, \boldsymbol{F}_{\text{vgg}}^{224}\right) \end{cases} \tag{2}$$

式中，$\boldsymbol{Z}_i$表示CAFEM模块融合特征的输出，$\boldsymbol{S}_i$表示Transformer解码器输出，$\boldsymbol{F}_{\text{vgg}}^{112}$和$\boldsymbol{F}_{\text{vgg}}^{224}$表示VGG前两层特征提取模块输出，形状分别为(112，112，128)和(224，224，64)，$\boldsymbol{S}_{\text{out}}$表示BFEM模块显著预测输出。最后，调整各个阶段的预测结果$\boldsymbol{S}_i$大小并计算Loss。

2.2　跨模态注意力融合增强模块

在提取了RGB模态和深度模态的多层次编码特征后，如何实现特征的充分融合交互是编码阶段需要关注的一个重要问题。深度特征包含丰富的空间信息，这对于显著目标的定位非常重要。RGB特征包含丰富的细节信息，这对于显著目标能否预测出清晰的边缘起决定作用。细节信息往往表现在特征通道之间的相关性，因此融合时首先将二者按通道拼接起来，然后利用拼接后的特征获得通道注意力权重，从而筛选出更重要的特征。深度特征自身就包含特征在空间上的相关性，因此利用深度特征提取空间注意力权重，对经过通道注意力筛选后的特征进行空间加权，加权后经过多层卷积，逐步将通道数压缩至输入通道数，接着将融合后的特征与前一层的融合特征拼接，这样在融合时能减少有效特征的丢失，拼接后进行自注意力加权，最后输出本层融合特征。所提出的跨模态融合模块结构如图2所示。

在跨模态注意力融合增强模块中，$\boldsymbol{F}_r^i$、$\boldsymbol{F}_d^i$、$\boldsymbol{Z}_{i-1}$和$\boldsymbol{Z}_i \in \mathbf{R}^{N \times C_i}$分别表示RGB输入特征、深度输入特征、前一层融合特征和当前层输出的融合特征，$\boldsymbol{F}_{rx}^i$和$\boldsymbol{F}_{dx}^i \in \mathbf{R}^{H \times W \times C_i}$分别表示变换形状后的RGB特征与深度特征，其中，$N = H \times W$，$C_i$表示输入和输出通道数，$i \in \{2,3,4\}$。首先将$\boldsymbol{F}_r^i$，$\boldsymbol{F}_d^i$的形状变换为($B$，$C$，$H$，$W$)，然后将变换后的特征按通道进行拼接，利用拼接后的特征获得通道注意力权重，并对拼接特征的通道进行加权，具体为

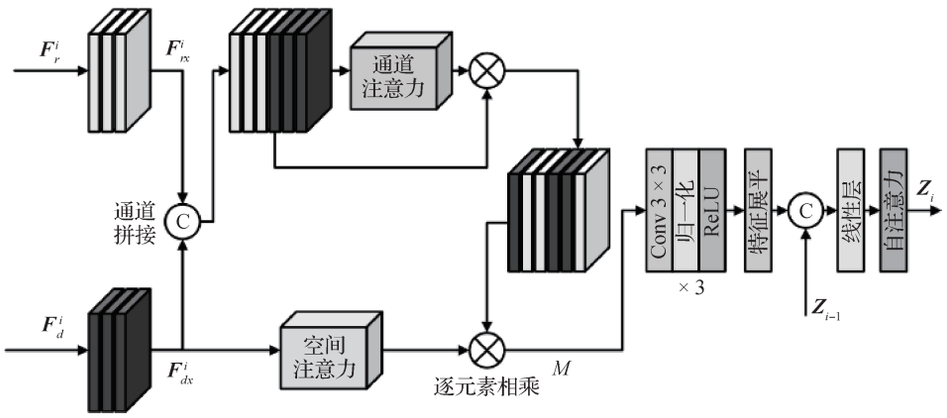

图2　跨模态注意力融合增强模块

Fig. 2　Cross-modal attention fusion enhancement module (CAFEM) structure

$$\begin{cases} \boldsymbol{F} = Concat\left(reshape\left(\boldsymbol{F}_r^i, \boldsymbol{F}_d^i\right)\right) \\ \boldsymbol{X}' = \sigma\left(C\left(Concat\left(Mean\left(\boldsymbol{F}\right), Max\left(\boldsymbol{F}\right)\right)\right)\right) \\ \boldsymbol{M} = \boldsymbol{X}' \otimes \boldsymbol{F} \end{cases} \tag{3}$$

式中，$\boldsymbol{F} \in \mathbf{R}^{H \times W \times Ci}$，*reshape* 表示将特征形状从（*B*, *N*, *C*）变换为（*B*, *C*, *H*, *W*），*Concat* 表示特征拼接操作，⊗表示像素级乘法，*Mean* 表示对每个通道进行平均值操作，*Max* 表示对每个通道进行最大值操作，$\boldsymbol{X}'$表示通道注意力权重。*C* 表示具有2个和1个通道的1 × 1卷积操作，σ表示sigmoid激活函数。然后，利用$\boldsymbol{F}_d^i$生成空间注意力权重，可以表示为

$$\boldsymbol{Y}' = \sigma\left(fc\left(avgpool\left(\boldsymbol{F}_d^i\right)\right) + fc\left(maxpool\left(\boldsymbol{F}_d^i\right)\right)\right) \tag{4}$$

式中，*avgpool* 表示全局平均池化操作，*maxpool* 表示最大池化操作，*fc* 表示1 × 1卷积、修正线性单元（rectified linear unit，ReLU）以及一个1 × 1卷积。在获得相应的权重之后，对经过通道注意力筛选的特征进行空间加权融合。具体过程为

$$\boldsymbol{M} = \boldsymbol{Y}' \otimes \boldsymbol{F} \tag{5}$$

式中，$\boldsymbol{M} \in \mathbf{R}^{H \times W \times C_i}$，空间加权融合后的特征输入到卷积层，增强特征的表征能力并将通道数减半，具体过程为

$$\boldsymbol{T}_x = C_{br}\left(\boldsymbol{M}\right) \tag{6}$$

式中，$\boldsymbol{T}_x \in \mathbf{R}^{H \times W \times C_i}$，$C_{br}$表示卷积核大小为3 × 3卷积操作，包括归一化和ReLU操作。最后，将T_x展平与前一层CAFEM模块的输出直接拼接，使得最终特征最大程度上保留原始融合特征，再通过两层线性层调整通道数，然后进行自注意力加权融合，得到最终融合特征。具体过程为

$$\boldsymbol{Z}_i = Att(Linear\left(Concat\left(Unfold\left(T_x\right), \boldsymbol{Z}_{i-1}\right)\right) \tag{7}$$

式中，$\boldsymbol{Z}_i \in \mathbf{R}^{N \times C_i}$，*Unfold* 表示特征在特定维度上被展平，*ConCat* 表示在最后一个维度上进行特征拼接，*Linear* 表示线性连接操作，*Att* 代表自注意力计算。

整体而言，本文方法根据RGB特征与深度特征的不同性质，通过对应的注意力进行加权融合，以学习具有不同层次的特征表系，并不断与前一层特征进行融合，充分学习并保留语义信息与细节信息。关于CAFEM的有效性以及对比实验的详细讨论见3.4实验部分。

2.3　边缘特征提取模块

为了补充Transformer提取特征的边缘细节，本文方法提出边缘特征提取模块（BFEM），在网络的末端将RGB图像输入到VGG特征提取主干的两边缘特征，并将此边缘特征与Transformer解码器的输出进行融合，如图3所示，本节将重点描述BFEM的详细结构及所提出的边缘特征获取以及融合细节。此外，将BFEM中的一些中间特征进行可视化的展示，具体操作为对于每一个特征，计算其在通道维度上的均值，将其转化为单通道图然后将其归一化，如图4所示。

在Transformer解码器的输出处，基本确定了显著性对象的主体，但由于Transformer的结构使特征提取感受野较大，所得到的显著性图像可能会出现边缘细节缺失等问题。为此，本文在解码器的末端提出一个CNN诱导的细化单元。CNN在特征提取时感受野逐步扩大，处理局部细节方面有着得天独厚的优势。并且该阶段的特征分辨率较高，在这个阶段使用卷积操作在参数数量和计算成本方面都更为合理。由于这一步的主要目的是对边缘细节的提取与细化，不需要引入完整的CNN编解码器网络，

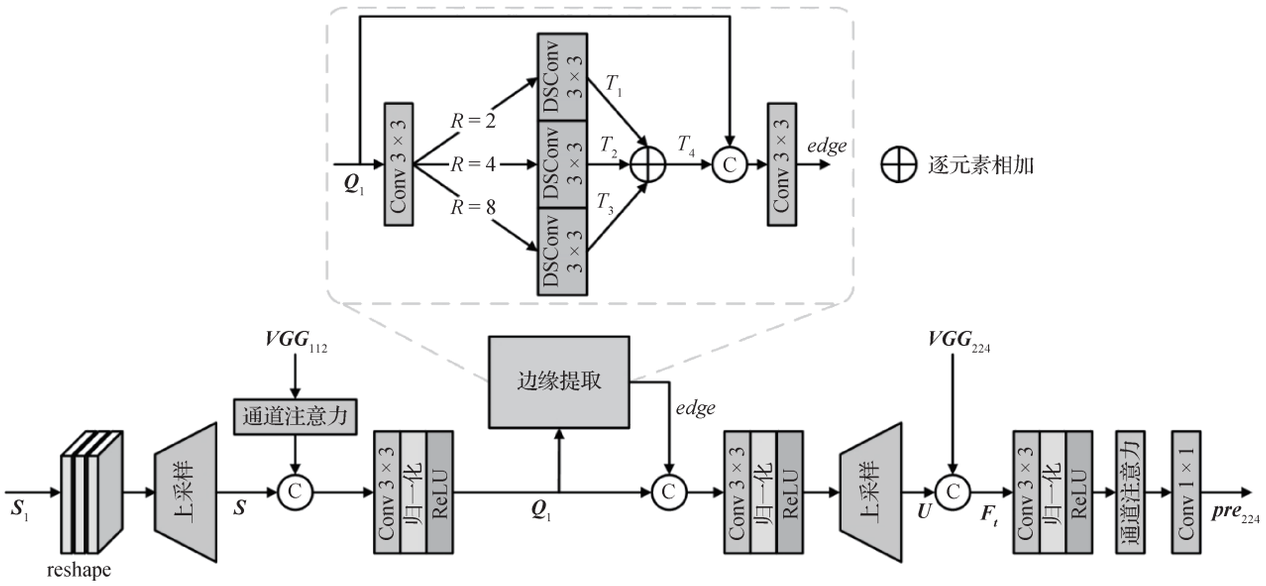

图3　边缘特征提取模块

Fig. 3　Boundary feature extraction module (BFEM) structure

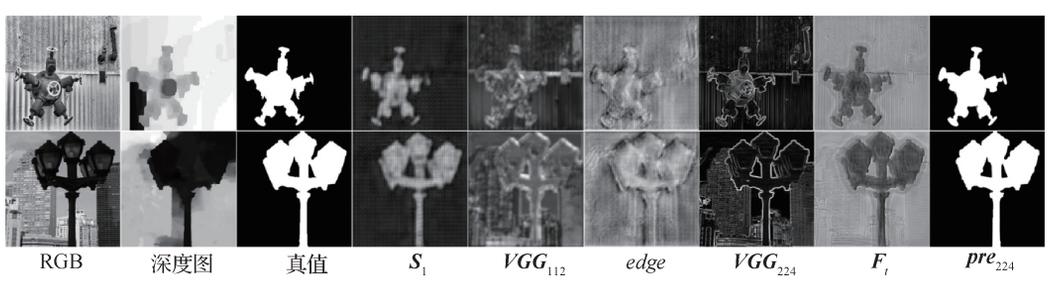

图4　边缘特征提取模块中间特征可视化

Fig. 4　Intermediate feature visualization of BFEM

因此使用VGG中纹理细节最为丰富的两层浅层特征进行边缘特征提取与特征融合，记为 $\boldsymbol{VGG}_{112}$ 和 $\boldsymbol{VGG}_{224}$。但由于低层特征同时包含大量的噪声，需要经过通道注意力进行筛选，因此将 $\boldsymbol{VGG}_{112}$ 通过通道注意力进行特征的筛选。首先，将来自最后一个Transformer解码器特征 $\boldsymbol{S}_1$ 的形状变换为 (B, C, H, W)，并上采样到与 $\boldsymbol{VGG}_{112}$ 相同的分辨率，然后将 $\boldsymbol{VGG}_{112}$ 特征通过通道注意力筛选并压缩通道数后与 $\boldsymbol{S}_1$ 拼接，并输入到卷积层进行细化融合，得到第1次融合的特征，具体为

$$\begin{cases} \boldsymbol{S} = Upsample\left(reshape\left(\boldsymbol{S}_1\right)\right) \\ \boldsymbol{VGG}_{112} = CA\left(\boldsymbol{VGG}_{112}\right) \\ \boldsymbol{Q}_1 = C_{br}\left(Concat\left(\boldsymbol{S}, \boldsymbol{VGG}_{112}\right)\right) \end{cases} \tag{8}$$

式中，$\boldsymbol{S}_1 \in \mathbf{R}^{N \times C_i}$，$\boldsymbol{Q}_1 \in \mathbf{R}^{H \times W \times C_i}$，*reshape* 表示从 (B, N, C) 恢复特征形状为 (B, C, H, W)，*Upsample* 表示双倍上采样操作，*CA* 表示通道注意力模块，*Concat* 表示特征拼接操作，C_{br} 代表3×3卷积、归一化和ReLU操作。然后，将融合后的特征 Q_1 输入到边缘提取层获得边缘特征后进行融合上采样，准备与VGG特征进行第2次融合，具体为

$$\begin{cases} \boldsymbol{edge} = EdgeExtraction\left(\boldsymbol{Q}_1\right) \\ \boldsymbol{U} = Upsample\left(C_{br}\left(Concat\left(\boldsymbol{Q}_1, edge\right)\right)\right) \end{cases} \tag{9}$$

式中，$\boldsymbol{edge} \in \mathbf{R}^{112 \times 112 \times 128}$，$\boldsymbol{U} \in \mathbf{R}^{224 \times 224 \times C_i}$，$C_{br}$ 代表3×3卷积、归一化和ReLU操作，*Concat* 表示特征拼接操作，*Upsample* 表示双倍上采样操作，*EdgeExtraction* 表示边缘提取操作，具体通过不同膨胀因子的空洞卷积提取多尺度特征，将多尺度特征相加后与输入特征拼接，充分地提取边缘细节特征，最后通过3×3卷积将通道数调整为输入通道数得到边缘特征，具体为

$$\begin{cases} \boldsymbol{T}_i = DSConv^{i}_{3\times3}\left(C_{3\times3}\left(\boldsymbol{Q}_j\right)\right) \\ \boldsymbol{T}_x = \sum_{i=1}^{3} i \\ \boldsymbol{edge} = C_{3\times3}\left(Concat\left(\boldsymbol{Q}_j, \boldsymbol{T}_x\right)\right) \end{cases} \tag{10}$$

式中，$\boldsymbol{T}_i, \boldsymbol{T}_x, \boldsymbol{Q}_j \in \mathbf{R}^{H\times W\times C_i}$表示边缘特征，*Concat*表示特征拼接操作，$DSConv^{i}_{3\times3}$表示空洞卷积操作，膨胀因子分别为2，4，8。将上采样后的特征$\boldsymbol{U}$与$\boldsymbol{VGG}_{224}$最终拼接后输入到融合卷积层进行融合，融合后输入到通道注意力层进行通道筛选并压缩通道，最后通过1×1卷积操作将通道数压缩为1得到最终的显著预测结果$\boldsymbol{pre}_{224}$。可以表示为

$$\begin{cases} \boldsymbol{F}_t = Concat\left(\boldsymbol{U}, \boldsymbol{VGG}_{224}\right) \\ \boldsymbol{pre}_{224} = C\left(CA\left(C_{br}\left(\boldsymbol{F}_t\right)\right)\right) \end{cases} \tag{11}$$

式中，$\boldsymbol{pre}_{224} \in \mathbf{R}^{224\times224\times1}$，$C_{br}$代表3×3卷积、归一化和ReLU操作，$C$表示将通道数压缩为1通道的1×1卷积操作，*CA*表示通道注意力模块。所提方法引入CNN提取边缘特征，且在仅使用了低层特征的情况下融合了大量的边缘细节，使最终的显著目标预测的边缘更清晰，细节更丰富。对该模块的有效性以及对比实验的详细讨论见4.3实验部分。

最后，在训练阶段，为了获得具有清晰边界的高质量显著性图像，本文采用混合损失函数训练网络，包括常用的二元交叉熵损失（binary cross-entropy loss，BCE）、测量结构相似度的SSIM（structure similarity index measure）以及交并比（intersection over union，IoU）损失（Cong等，2023）。将每个阶段显著目标的BCE损失、SSIM损失和IoU损失的总和作为显著目标的总损失，定义为

$$\begin{cases} L_{\text{sal}}\left(\boldsymbol{P}, \boldsymbol{G}\right) = L_{\text{bce}}\left(\boldsymbol{P}, \boldsymbol{G}\right) + L_{\text{ssim}}\left(\boldsymbol{P}, \boldsymbol{G}\right) + L_{\text{iou}}\left(\boldsymbol{P}, \boldsymbol{G}\right) \\ L_{\text{total}} = \sum_{i=1}^{4} \frac{1}{2^i} L_{\text{sal}}\left(\boldsymbol{S}_i, \boldsymbol{G}_i\right) + L_{\text{sal}}\left(\boldsymbol{S}_{\text{out}}, \boldsymbol{G}\right) \end{cases} \tag{12}$$

式中，$\boldsymbol{G}$为真实值，$\boldsymbol{S}_i$为各阶段显著目标预测输出。

3 实验与分析

3.1 数据集与评价指标

采用NJU2K（Nanjing University 2K）（Ju等，2014）的1 485组图像和NLPR（Peng等，2014）的700组图像作为训练集。验证则采用NJU2K除用于训练外的其他图像作为验证数据集，测试采用4个使用较为广泛的公开基准数据集NJU2K、NLPR、SIP（salient person）（Fan等，2021）和STERE（stereo dataset）（Niu等，2012）。评估时采用常用的3项评价指标：1）平均绝对误差（mean absolute error，MAE）；2）F_β，评估模型的识别能力及查全能力，在本文中始终将β^2设置为0.3（Wang等，2022b），以强调准确率的重要性；3）S_α，评估显著图像区域感知和对象感知的空间结构相似性，在本文中始终将α设置为0.5（Zhao等，2019）。

3.2 实验细节

所提出的方法基于PyTorch（Paszke等，2019）库实现。在Nvidia RTX 3090 GPU（24 G显存）上进行训练与测试。本文采用的P2T版本为P2T-base，VGG版本为VGG-16，并且只采用了VGG16前两层的结构用于微调训练。在训练过程中，采用了Adam优化器（Kingma和Ba，2015），使用默认的超参数设置。为了标准化输入，将图像调整为224×224像素。此外，所有的深度图像都被规范化并复制成3个通道以适应输入的大小，所有图像采用随机旋转和水平翻转进行数据增强。本文使用批量大小为16训练网络，共进行了200个epoch。学习率设置为10^{-4}，每40个epoch衰减为原来的1/5。测试过程中，将各个阶段输出使用双线性插值将它们调整回原始尺寸。

3.3 与先进方法对比

3.3.1 定量评估

为了验证所提出的CFADNet模型的有效性，在NJU2K、NLPR、STERE和SIP数据集上进行定量评估，并与16种先进的RGB-D显著目标检测方法进行定量比较，结果如表1所示。对比方法包括DCF（depth calibration and fusion）（Ji等，2021）、CIRNet（cross-modality interaction and refinement network）（Cong等，2022）、CM-LCG（cross-modality long-range context information gathering）（Wang等，2022a）、AIL-Net（aggregate interactive learning network）（Wu等，2022a）、SPSN（superpixel prototype sampling network）（Lee等，2022）、TMFNet（three-input multi-level fusion network）（Zhou等，2022）、AFNet、JL-DCF（joint learning and densely cooperative fusion）（Fu等，2022）、EBFSP（employing bilinear fusion and saliency prior）（Huang等，2022）、CAVER（cross-modal view-mixed transformer）（Pang等，2023）、HINet（hierarchical interaction network）（Bi等，2023）、C2DFNet（criss-

cross dynamic filter network)(Zhang 等, 2023)、PICRNet(point-aware interaction and cnn-induced refinement network)(Cong 等, 2023)、DGFNet(depth-guided cross-modality fusion network)(Xiao 等, 2024)、FCFNet(feature calibrating and fusing network)(Zhang 等, 2024)以及RD3D(Chen 等, 2024)。

在所有模型中,CFADNet的3项指标在NJU2K、NLPR、STERE和SIP数据集上均取得优异成绩。特别是与排名第2的方法相比,CFADNet的MAE分别降低了6.9%、10.5%、9.7%和2.4%,并且S_α和F_β在4个数据集上都达到最优结果。

F_β曲线和PR曲线如图5和图6所示。可以看出,本文方法在4个数据集上的F_β曲线比其他模型的曲线更平坦,表明本文结果更接近二分类预测,且对阈值变化具有不变性。从PR曲线可以看出,本文方法在4个数据集上达到更高精度。

表2展示了不同方法在所有数据集的平均精度和平均召回率。表3展示了本文方法与开源对比方法的计算复杂度和参数量对比。可以看出,尽管复杂的网络结构导致了较高的参数量,但每秒浮点运算次数(floating point operations per second, FLOPs)维持在一个相对较低的值46.82 G。因此在计算资源受限的场景中,同样能实现计算效率与模型性能的良好平衡。此外,本文方法在所有数据集上保持了高精度,同时确保了高召回率。表明本文方法能够正确识别更多的正例,将更少的负例误判为正例,并尽可能找到所有显著对象。

综上所述,CFADNet在性能上展现出强大竞争力。与目前先进方法相比,在4个基准数据集上的所有指标上均表现出色,达到最优或次优结果,充分证明了CFADNet的优秀性能,同时表明其在处理不同场景下显著目标预测的准确性和稳定性。

3.3.2 定性评估

除定量评估对比外,为了直观展示CFADNet的

表1 定量评估

Table 1 Quantitative evaluation

方法	接收会议	NJU2K数据集			NLPR数据集			STERE数据集			SIP数据集		
		MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑
DCF	CVPR 21	0.035	0.902	0.912	0.021	0.891	0.924	0.039	0.885	0.902	0.051	0.875	0.876
CIRNet	TIP 22	0.035	0.927	0.925	0.023	0.924	0.933	0.038	0.914	0.917	0.052	0.896	0.888
CM-LCG	TIP 22	0.043	0.915	0.913	0.029	0.906	0.922	0.043	0.906	0.910	–	–	–
AILNet	ESWA 22	0.045	0.876	0.898	0.029	0.857	0.912	0.038	0.880	0.908	0.050	0.866	0.889
SPSN	ECCV 22	0.032	0.920	0.918	0.023	0.910	0.923	0.035	0.900	0.907	<u>0.042</u>	0.899	0.892
TMFNet	TETCI 22	0.041	0.882	0.910	0.027	0.867	0.921	–	–	–	0.057	0.853	0.874
AFNet	Nuecom22	0.032	0.928	0.926	0.020	0.925	<u>0.936</u>	0.034	0.918	0.918	0.043	<u>0.909</u>	<u>0.896</u>
JL-DCF	TPAMI 22	0.040	0.913	0.911	0.023	0.917	0.926	0.039	0.907	0.911	0.046	0.900	0.892
EBFSP	TMM 22	0.038	0.895	0.907	0.028	0.887	0.909	0.041	0.873	0.900	0.052	0.863	0.877
CAVER	TIP 23	0.031	0.925	0.921	0.020	0.921	0.929	0.033	0.912	0.913	<u>0.042</u>	0.902	0.893
HINet	PR 23	0.039	0.914	0.915	0.026	0.906	0.922	0.049	0.883	0.892	0.066	0.855	0.856
C2DFNet	TMM 23	0.039	0.909	0.908	0.022	0.917	0.928	0.038	0.897	0.902	0.053	0.877	0.872
PICRNet	ACMM 23	<u>0.029</u>	<u>0.931</u>	0.927	<u>0.019</u>	<u>0.928</u>	0.935	<u>0.031</u>	<u>0.920</u>	<u>0.921</u>	0.053	0.883	0.872
DGFNet	TMM 24	0.032	0.914	0.921	0.021	0.902	0.928	0.035	0.896	0.911	0.048	0.879	0.883
FCFNet	TCSVT 24	0.034	0.923	0.918	0.024	0.911	0.924	0.038	0.906	0.906	–	–	–
RD3D	TNNLS24	0.033	0.928	<u>0.928</u>	0.022	0.921	0.933	0.037	0.905	0.914	0.046	0.900	0.892
CFADNet(本文)	–	**0.027**	**0.933**	**0.930**	**0.017**	**0.934**	**0.939**	**0.028**	**0.923**	**0.925**	**0.041**	**0.910**	**0.897**

注:加粗、下划线字体分别表示各列最优、次优结果,“–”表示无相应数据,“↓”表示值越小越好,“↑”表示值越大越好。

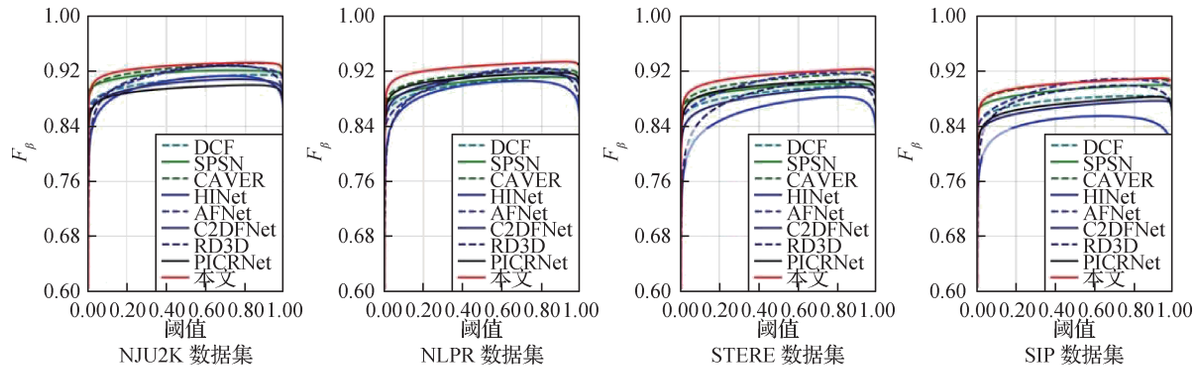

图5 F_β曲线

Fig. 5 F_β curves

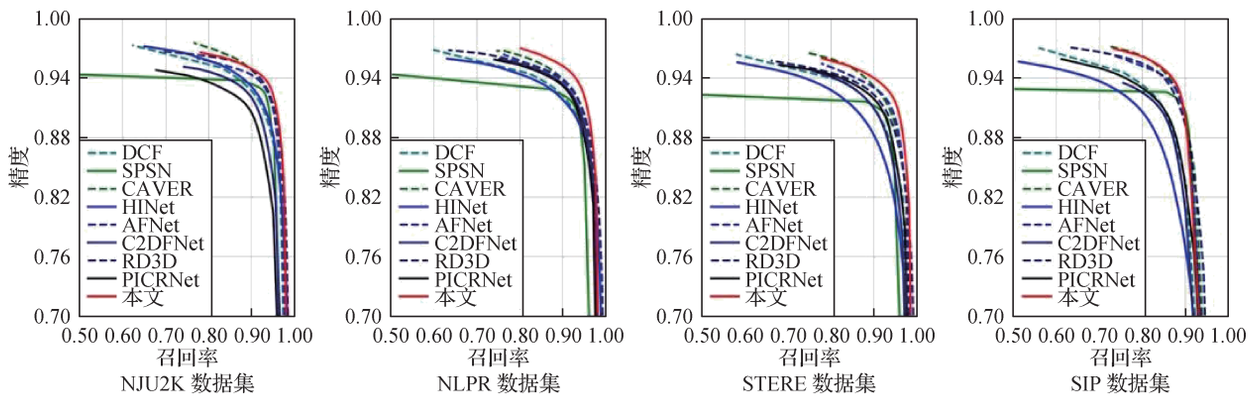

图6 精度—召回率曲线

Fig. 6 Precision-recall curves

表2 与其他方法的平均精度和平均召回率比较

Table 2 Quantitative comparison in terms of average precision and average recall with other methods

方法	NJU2K数据集		NLPR数据集		STERE数据集		SIP数据集	
	$Prec_{avg}$	$Recall_{avg}$	$Prec_{avg}$	$Recall_{avg}$	$Prec_{avg}$	$Recall_{avg}$	$Prec_{avg}$	$Recall_{avg}$
DCF	0.908	0.917	0.898	0.922	0.888	0.918	0.902	0.847
SPSN	0.918	0.924	0.901	0.924	0.893	0.924	0.901	**0.894**
AFNet	0.916	0.925	0.907	0.924	0.900	<u>0.931</u>	<u>0.913</u>	0.874
CAVER	<u>0.924</u>	<u>0.929</u>	<u>0.917</u>	<u>0.924</u>	<u>0.904</u>	0.930	**0.922**	0.876
HINet	0.909	0.908	0.896	0.909	0.868	0.892	0.885	0.808
PICRNet	0.900	0.894	0.911	0.921	0.869	0.873	0.887	0.864
C2DFNet	0.910	0.898	0.911	0.919	0.887	0.913	0.887	0.858
RD3D	0.917	0.925	0.904	0.923	0.884	0.926	<u>0.907</u>	0.872
CFADNet(本文)	**0.927**	**0.937**	**0.926**	**0.938**	**0.909**	**0.946**	**0.922**	<u>0.882</u>

注:加粗、下划线字体分别表示各列最优、次优结果。$Prec_{avg}$表示平均精度,$Recall_{avg}$表示平均召回率。

卓越性能,在4个测试数据集上抽取具有挑战性的几个场景进行显著目标检测测试的可视化细节对比,其中所对比其他方法的预测图像是基于作者提供的开源代码生成或来自作者开源仓库中提供的测试图像,对比结果如图7所示。结果表明,在各种复杂场景下,CFADNet都能够保证预测的精度。第1、3、4、9行展示了图像中包含的丰富细节,并且显著目标边缘复杂不规则。与其他方法相比,CFADNet

表3　与其他方法计算复杂度和参数量对比

Table 3　Comparison of computational complexity and parameter size with other methods

方法	FLOPs/G	参数量/M
DCF	55.48	107.29
CIRNet	156.34	82.08
TMFNet	–	266.7
AFNet	130.02	258.13
HINet	389.7	98.9
C2DFNet	22.047	47.52
DGFNet	74.89	42.14
RD3D	57.8	47.14
CFADNet(本文)	46.82	203.88

注:“–”表示无相应数据。

能够正确分割显著区域,并且预测出清晰锐利的边缘轮廓,证明了CFADNet可以有效捕捉局部细节特征。在细节信息的帮助下,网络能够准确识别显著物体的边缘轮廓。第2、7、8行展示了目标主体区域不连续或部分区域占比较小的图像。结果显示,本文方法能精确分割出完整的显著物体,证明了CFADNet能够有效捕获全局语义信息,并对全局语义信息进行整合筛选,最终正确地预测出完整的显著目标。第5、6行展示了显著目标主体与背景相似或嵌入在背景中,本文方法能够正确地区分前景与背景,从而将显著目标从背景中分割出来。综上所述,在这些具有挑战性的场景中,CFADNet能够产生高精度的显著性检测结果。

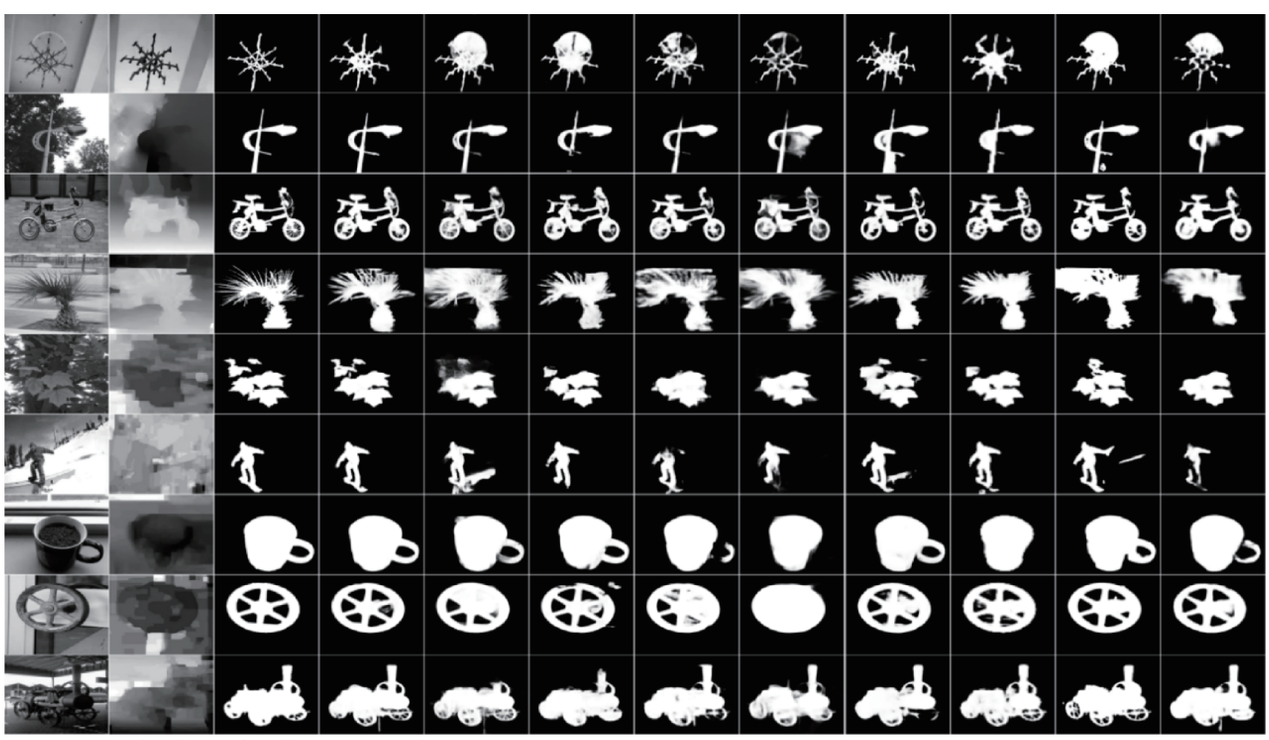

图7　本文方法与现有先进方法的直观对比

Fig. 7　Visual comparison of our method with the state-of-the-art methods

3.4　消融实验

使用3个常用的具有挑战性的测试数据集NJU2K、NLPR和STERE进行评估。从3项评价指标和可视化对比讨论CAFEM和BFEM在跨模态特征融合、边缘提取以及噪声抑制的有效性。同时,研究不同的边缘提取方法得到边缘特征对显著物体预测的影响。

3.4.1　跨模态注意力融合增强模块的有效性

为探究最佳的跨模态特征融合模块的设计,验证本文提出的跨模态注意力融合增强模块的有效性,进行了两组对照实验。第1组对照实验通过移除CAFEM中的通道注意力以及空间注意力机制,直接拼接两种模态的特征进行卷积融合,通过自注意力特征中语义和细节信息之间的依赖关系后输出,

训练剩余的网络进行对照实验。第2组实验将CAFEM模块替换为PICRNet(Cong等,2023)中提出的CmPI(cross-modality pointaware interaction)融合方法,对两种模态特征进行融合。

为确保实验结果的准确性和公平性,所有实验都在相同的环境下进行,具体而言,每个阶段提取的RGB特征和深度特征不再通过CAFEM模块进行融合,而是直接堆叠并输入到卷积模块进行融合或直接输入到CmPI融合模块进行融合。

实验结果如表4所示。从表中3个测试数据集上的所有评价指标可看出,引入CAFEM后,预测精度明显提高,RGB特征与深度特征得到了充分的融合利用,帮助了网络提取更丰富的多尺度细节与语义信息,此外融合语义信息与细节信息的同时还抑制了噪声的引入,使最终的预测精度得到了显著的提高,这证明了所提出的跨模态注意力融合增强模块是有效的。

3.4.2 边缘特征提取模块的有效性

为了证明本文方法提出的边缘特征提取模块的有效性,通过移除BFEM的边缘提取与融合部分并训练剩余网络来研究边缘细化的影响。即不使用空洞卷积提取多尺度边缘特征来融合补充到显著特征中,而是直接将VGG提取的低层特征与显著特征进行拼接卷积融合后直接预测,不对VGG的低层特征进行噪声的过滤。为保证公平性和准确性,每组实验的训练环境和超参设置均保持一致,评估结果如表5所示,引入边缘提取与融合后,性能有了明显的提升。此外,本文对有无边缘特征提取与融合的两种预测结果进行可视化比较,如图8所示。其中,$\boldsymbol{S}_1$表示未提取边缘及未过滤噪声的结果,$\boldsymbol{S}_2$表示使用BFEM提取边缘并过滤噪声的结果。$\boldsymbol{S}_{1_F}$和$\boldsymbol{S}_{2_F}$为两种方式在BFEM模块中最终输出前一层特征$\boldsymbol{F}_t$的可视化结果。可以看到,通过通道注意力对低层特征进行噪声过滤以及使用空洞卷积进行边缘特征提取后,融合到显著特征中进行预测,生成的预测图拥有了清晰的边界并且预测错误区域明显降低。表明BFEM可以有效过滤VGG提取低层特征的噪声并将局部细节特征充分融合到显著特征中得到精准的预测结果。综上所述,通过引入边缘细化模块,CFAENet能够更好地融合多种特征,准确提取目标多尺度下的细节信息,从而获得更好的预测结果。

3.4.3 混合损失函数的有效性

为了验证本文方法中混合损失函数的有效性,设计了4组对照实验,结果如表6所示。实验结果表明,单独使用BCE损失时,模型在3个数据集上仅取得基础性能;当进一步引入SSIM损失后,精度有所提升,表明结构相似度损失在保持前景整体结构和抑制背景噪声方面发挥了积极作用;加入IoU损失之后,模型性能进一步提升,说明IoU损失能够有效增强目标区域的覆盖性,并改善边界一致性。最后,

表4 CAFEM与不同融合方法的对比实验

Table 4 Comparison experiment of CAFEM with other fusion methods

方法	NJU2K数据集			NLPR数据集			STERE数据集		
	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑
w/o CAFEM	0.029	0.928	0.926	0.018	0.933	0.937	0.030	0.920	0.923
CmPI	0.030	0.929	0.925	0.018	0.931	0.938	0.032	0.915	0.919
CAFEM	**0.027**	**0.933**	**0.930**	**0.017**	**0.934**	**0.939**	**0.028**	**0.923**	**0.925**

注:加粗字体表示各列最优结果。w/o表示不使用。CmPI为Cong等人(2023)在PICRNet中提出的融合方法。"↑"和"↓"同表1。

表5 BFEM的消融实验

Table 5 Ablation studies on the BFEM

方法	NJU2K数据集			NLPR数据集			STERE数据集		
	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑
w/o EE	0.029	0.927	0.926	0.017	0.933	0.938	0.03	0.921	0.922
BFEM	**0.027**	**0.933**	**0.93**	**0.017**	**0.934**	**0.939**	**0.028**	**0.923**	**0.925**

注:加粗字体表示各列最优结果。w/o EE表示不使用BFEM中的edge extraction模块。"↓"和"↑"同表1。

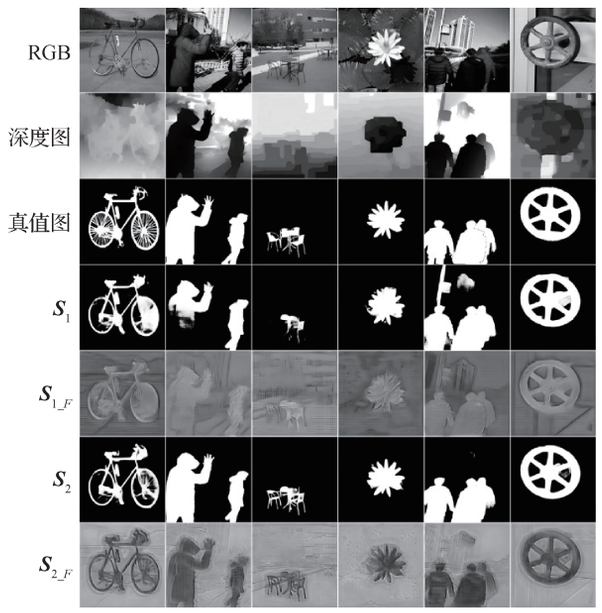

图8　采用不同边缘特征提取方法得到的预测图细节对比

Fig. 8　Comparison of details of prediction maps obtained by different edge feature extraction method

将多阶段监督策略替换为仅在最终预测结果上进行监督的单阶段策略。结果表明，尽管仍采用混合损失，单阶段策略下模型精度较多阶段监督策略显著下降。这一对比充分证明了本文训练策略的合理性和有效性。

3.5　局限性讨论

虽然本文模型已取得令人满意的结果，但在一些具有挑战性的场景中仍存在一些预测失败的情况。本节在4个测试数据集上分析了一些失败案例。图9展示了本文提出模型的6个失败案例。在第1、3个案例中，本文模型将石像后面的人错误识别为显著物体，并且只识别了半个，同时将与显著目标相连的下部分错误判断为显著目标。这是因为模型可能受深度图像提取特征的影响，由于深度图像层次分明，导致模型误判。在第2、4个案例中，显著目标的部分细长不明显或偏离视觉中心，使得模型很容易将这部分目标判断为非显著。在第5、6个案例中，显著目标与非显著目标过度平缓且连接在一起，本文模型错误地将其识别为显著物体。

表6　混合损失函数的消融实验

Table 6　Ablation studies on the loss function

方法	NJU2K数据集			NLPR数据集			STERE数据集		
	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑	MAE ↓	F_β ↑	S_α ↑
F_{BCE}	0.034	0.910	0.921	0.020	0.905	0.911	0.036	0.903	0.910
$F_{\mathrm{BCE}}+F_{\mathrm{SSIM}}$	0.030	0.925	0.927	0.019	0.925	0.935	0.031	0.921	0.921
$F_{\text{混合}}$	0.032	0.917	0.924	0.021	0.918	0.928	0.033	0.912	0.917
$F_{\mathrm{BCE}}+F_{\mathrm{SSIM}}+F_{\mathrm{IoU}}$	**0.027**	**0.933**	**0.930**	**0.017**	**0.934**	**0.939**	**0.028**	**0.923**	**0.925**

注：加粗字体表示各列最优结果。$F_{\text{混合}}$表示采用混合损失设计，但是仅对最终输出使用。“↑”和“↓”同表1。

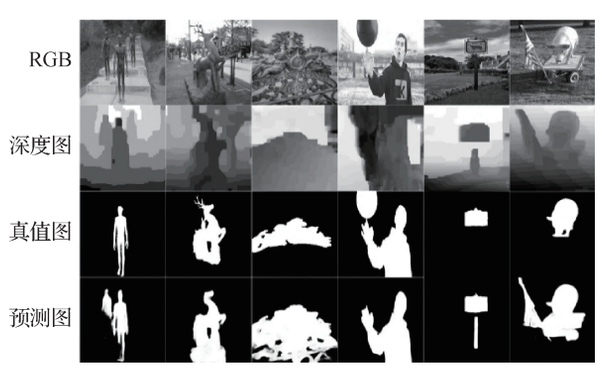

图9　失败案例

Fig. 9　Some failure cases of our method

上述失败案例表明，本文方法在图像背景出现多个相似显著物体且深度图像层次鲜明时表现不佳，仍有很大改进空间。本文认为可以通过判断深度图像的质量，训练出一组权重，当深度图像质量高、图像层次鲜明则可以用来确定显著目标具体位置，若深度图像质量较低，分辨不出显著目标时要尽可能将权重偏向RGB图像分支，多使用RGB图像提取的特征进行预测。针对显著目标与背景相似且黏连的问题，后续可以通过扩大类似场景的训练数据集来缓解。在未来工作中，将对这些问题进行研究。

4　结　论

本文提出一种用于RGB-D的显著目标检测网

络CFADNet。为了高效利用多模态特征，设计了一种跨模态注意力融合增强模块（CAFEM）。该模块考虑了RGB特征与深度特征的特点，借助RGB携带的通道间信息，通过通道注意力进行筛选融合。同时，利用深度信息携带的空间信息，采用空间注意力对融合特征进行空间加权，使融合特征能够包含全局语义信息、局部细节信息和通道间依赖信息，并在一定程度上抑制引入的噪声，有助于模型在复杂场景下保持预测精度。此外，构建了一个边缘特征提取模块（BFEM），通过CNN提取低层细节特征，并利用通道注意力过滤噪声后将这些细节特征逐步融合到显著特征中，使最终预测的特征既包含语义信息，又有细节特征的补充。在RGB-D显著目标检测基准数据集上进行的大量综合评估实验表明，CFADNet能够较好地融合并利用多模态特征，提取丰富的语义和细节特征，对显著物体进行准确的定位，并预测出较为清晰的边缘。

参考文献（References）

Bi H B, Wu R W, Liu Z Q, Zhu H H, Zhang C and Xiang T Z. 2023. Cross-modal hierarchical interaction network for RGB-D salient object detection. Pattern Recognition, 136: #109194 [DOI: 10.1016/j.patcog.2022.109194]

Chen H and Li Y F. 2018a. Progressively complementarity-aware fusion network for RGB-D salient object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3051-3060 [DOI: 10.1109/CVPR.2018.00322]

Chen J N, Lu Y Y, Yu Q H, Luo X D, Adeli E, Wang Y, Lu L, Yuille A L and Zhou Y Y. 2021. TransUNet: transformers make strong encoders for medical image segmentation [EB/OL]. [2024-11-07]. https://arxiv.org/pdf/2102.04306.pdf

Chen Q, Zhang Z X, Lu Y Y, Fu K R and Zhao Q J. 2024. 3-D convolutional neural networks for RGB-D salient object detection and beyond. IEEE Transactions on Neural Networks and Learning Systems, 35(3): 4309-4323 [DOI: 10.1109/TNNLS.2022.3202241]

Chen S H, Tan X L, Wang B and Hu X L. 2018b. Reverse attention for salient object detection//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 236-252 [DOI: 10.1007/978-3-030-01240-3_15]

Chen T Y, Xiao J, Hu X G, Zhang G F and Wang S J. 2023. Adaptive fusion network for RGB-D salient object detection. Neurocomputing, 522: 152-164 [DOI: 10.1016/j.neucom.2022.12.004]

Cheng X L, Zheng X, Pei J L, Tang H, Lyu Z and Chen C B. 2023. Depth-induced gap-reducing network for RGB-D salient object detection: an interaction, guidance and refinement approach. IEEE Transactions on Multimedia, 25: 4253-4266 [DOI: 10.1109/TMM.2022.3172852]

Ciptadi A, Hermans T and Rehg J M. 2013. An in depth view of saliency//Proceedings of 2013 British Machine Vision Conference (BMVC). Bristol, UK: BMVC: #112 [DOI: 10.5244/C.27.112]

Cong R M, Lin Q W, Zhang C, Li C Y, Cao X C, Huang Q M and Zhao Y. 2022. CIR-Net: cross-modality interaction and refinement for RGB-D salient object detection. IEEE Transactions on Image Processing, 31: 6800-6815 [DOI: 10.1109/TIP.2022.3216198]

Cong R M, Liu H Y, Zhang C, Zhang W, Zheng F, Song R and Kwong S. 2023. Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 406-416 [DOI: 10.1145/3581783.3611982]

Ding N, Zhang C and Eskandarian A. 2024. SalienDet: a saliency-based feature enhancement algorithm for object detection for autonomous driving. IEEE Transactions on Intelligent Vehicles, 9(1): 2624-2635 [DOI: 10.1109/TIV.2023.3287359]

Ding Y, Liu Z, Huang M K, Shi R and Wang X Y. 2019. Depth-aware saliency detection using convolutional neural networks. Journal of Visual Communication and Image Representation, 61: 1-9 [DOI: 10.1016/j.jvcir.2019.03.019]

Fan D P, Lin Z, Zhang Z, Zhu M L and Cheng M M. 2021. Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems, 32 (5): 2075-2089 [DOI: 10.1109/TNNLS. 2020.2996406]

Fang X, Jiang M F, Zhu J C, Shao X L and Wang H P. 2024. Group-TransNet: group transformer network for RGB-D salient object detection. Neurocomputing, 594: #127865 [DOI: 10.1016/j.neucom.2024.127865]

Fu K P, Fan D P, Ji G P, Zhao Q J, Shen J B and Zhu C. 2022. Siamese network for RGB-D salient object detection and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9): 5541-5559 [DOI: 10.1109/TPAMI.2021.3073689]

Gao Y, Wang M, Tao D C, Ji R R and Dai Q H. 2012. 3-D object retrieval and recognition with hypergraph analysis. IEEE Transactions on Image Processing, 21(9): 4290-4303 [DOI: 10.1109/TIP.2012.2199502]

Han J W, Chen H, Liu N, Yan C G and Li X L. 2018. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics, 48 (11): 3171-3183 [DOI: 10.1109/TCYB.2017.2761775]

Hou Q B, Cheng M M, Hu X W, Borji A, Tu Z W and Torr P H S. 2019. Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(4): 815-828 [DOI: 10.1109/TPAMI.2018.2815688]

Huang N C, Yang Y, Zhang D W, Zhang Q and Han J G. 2022. Employing bilinear fusion and saliency prior information for RGB-D salient object detection. IEEE Transactions on Multimedia, 24: 1651-1664 [DOI: 10.1109/TMM.2021.3069297]

Jahanifar M, Tajeddin N Z, Asl B M and Gooya A. 2019. Supervised saliency map driven segmentation of lesions in dermoscopic images. IEEE Journal of Biomedical and Health Informatics, 23(2): 509-518 [DOI: 10.1109/JBHI.2018.2839647]

Ji W, Li J J, Yu S, Zhang M, Piao Y, Yao S Y, Bi Q, Ma K, Zheng Y F, Lu H C and Cheng L. 2021. Calibrated RGB-D salient object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 9466-9476 [DOI: 10.1109/CVPR46437.2021.00935]

Ju R, Ge L, Geng W J, Ren T W and Wu G S. 2014. Depth saliency based on anisotropic center-surround difference//Proceedings of 2014 IEEE International Conference on Image Processing (ICIP). Paris, France: IEEE: 1115-1119 [DOI: 10.1109/ICIP. 2014. 7025222]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-15

Lee M, Park C, Cho S and Lee S. 2022. SPSN: superpixel prototype sampling network for RGB-D salient object detection//Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer: 630-647 [DOI: 10.1007/978-3-031-19818-2_36]

Li J, Su J M, Xia C Q, Ma M C and Tian Y H. 2021. Salient object detection with purificatory mechanism and structural similarity loss. IEEE Transactions on Image Processing, 30: 6855-6868 [DOI: 10.1109/TIP.2021.3099405]

Liang B C and Luo H L. 2024. MEANet: an effective and lightweight solution for salient object detection in optical remote sensing images. Expert Systems with Applications, 238: #121778 [DOI: 10.1016/j.eswa.2023.121778]

Liang F F, Duan L J, Ma W, Qiao Y H, Cai Z and Qing L. 2018. Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing, 275: 2227-2238 [DOI: 10.1016/j.neucom.2017.10.052]

Mei H Y, Liu Y Y, Wei Z Q, Zhou D S, Wei X P, Zhang Q and Yang X. 2022. Exploring dense context for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 32(3): 1378-1389 [DOI: 10.1109/TCSVT.2021.3069848]

Niu Y Z, Geng Y J, Li X Q and Liu F. 2012. Leveraging stereopsis for saliency analysis//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE: 454-461 [DOI: 10.1109/CVPR.2012.6247708]

Pang Y W, Zhao X Q, Zhang L H and Lu H C. 2023. CAVER: cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Transactions on Image Processing, 32: 892-904 [DOI: 10.1109/TIP.2023.3234702]

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z M, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J J and Chintala S. 2019. PyTorch: an imperative style, high-performance deep learning library//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #721

Peng H W, Li B, Xiong W H, Hu W M and Ji R R. 2014. RGBD salient object detection: a benchmark and algorithms//Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer: 92-109 [DOI: 10.1007/978-3-319-10578-9_7]

Ren J Q, Xiaojin Gong N, Yu L, Wenhui Zhou N and Yang M Y. 2015. Exploiting global priors for RGB-D saliency detection//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Boston, USA: IEEE: 25-32 [DOI: 10.1109/CVPRW.2015.7301391]

Ren Z X, Gao S H, Chia L T and Tsang I W H. 2014. Region-based saliency detection and its application in object recognition. IEEE Transactions on Circuits and Systems for Video Technology, 24(5): 769-779 [DOI: 10.1109/TCSVT.2013.2280096]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR

Sun F M, Hu X H, Wu J Y, Sun J and Wang F S. 2024. RGB-D salient object detection based on cross-modal interactive fusion and global awareness. Journal of Software, 35(4): 1899-1913 (孙福明, 胡锡航, 武景宇, 孙静, 王法胜. 2024. 跨模态交互融合与全局感知的RGB-D显著性目标检测. 软件学报, 35(4): 1899-1913) [DOI: 10.13328/j.cnki.jos.006833]

Wang F Y, Pan J S, Xu S K and Tang J H. 2022a. Learning discriminative cross-modality features for RGB-D saliency detection. IEEE Transactions on Image Processing, 31: 1285-1297 [DOI: 10.1109/TIP.2022.3140606]

Wang W G, Lai Q X, Fu H Z, Shen J B, Ling H B and Yang R G. 2022b. Salient object detection in the deep learning era: an in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 3239-3259 [DOI: 10.1109/TPAMI. 2021. 3051099]

Wang W G, Shen J B and Ling H B. 2019. A deep network solution for attention and aesthetics aware photo cropping. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7): 1531-1544 [DOI: 10.1109/TPAMI.2018.2840724]

Wang W G, Shen J B, Lu X K, Hoi S C H and Ling H B. 2021. Paying attention to video object pattern understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(7): 2413-2428 [DOI: 10.1109/TPAMI.2020.2966453]

Wang W G, Sun G L and Van Gool L. 2024. Looking beyond single images for weakly supervised semantic segmentation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(3): 1635-1649 [DOI: 10.1109/TPAMI.2022.3168530]

Wu J Y, Sun F M, Xu R, Meng J and Wang F S. 2022a. Aggregate interactive learning for RGB-D salient object detection. Expert Systems with Applications, 195: #116614 [DOI: 10.1016/j.eswa.2022.116614]

Wu Y H, Liu Y, Zhan X and Cheng M M. 2023. P2T: pyramid pooling transformer for scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11): 12760-12771 [DOI: 10.1109/TPAMI.2022.3202765]

Wu Y H, Liu Y, Zhang L, Cheng M M and Ren B. 2022b. EDN: salient object detection via extremely-downsampled network. IEEE Transactions on Image Processing, 31: 3125-3136 [DOI: 10.1109/TIP.2022.3164550]

Xiao F, Pu Z D, Chen J Q and Gao X P. 2024. DGFNet: depth-guided cross-modality fusion network for RGB-D salient object detection. IEEE Transactions on Multimedia, 26: 2648-2658 [DOI: 10.1109/TMM.2023.3301280]

Ye X Y, Zhu L, Wang W W and Fu Y. 2024. RGB_D salient object detection algorithm based on complementary information interaction. Journal of Image and Graphics, 29(5): 1252-1264 (叶欣悦, 朱磊, 王文武, 付云. 2024. 互补特征交互融合的RGB_D实时显著目标检测. 中国图象图形学报, 29(5): 1252-1264) [DOI: 10.11834/jig.230583]

Zhang R, Lyu Y, Zhang Z T, Ren L, Xie J, Zhang A L, Yan Z W and Mi O. 2024. Development and prospect of multidimensional information perception and intelligent construction in deep earth engineering. Journal of China Coal Society, 49(3): 1259-1290 (张茹, 吕游, 张泽天, 任利, 谢晶, 张安林, 严志伟, 米欧. 2024. 深地工程多维信息感知与智能建造的发展与展望. 煤炭学报, 49(3): 1259-1290) [DOI: 10.13225/j.cnki.jccs.2023.1439]

Zhang M, Ren W S, Piao Y, Rong Z K and Lu H C. 2020. Select, supplement and focus for RGB-D saliency detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 3469-3478 [DOI: 10.1109/CVPR42600.2020.00353]

Zhang M, Yao S Y, Hu B Q, Piao Y and Ji W. 2023. C^{2}DFNet: criss-cross dynamic filter network for RGB-D salient object detection. IEEE Transactions on Multimedia, 25: 5142-5154 [DOI: 10.1109/TMM.2022.3187856]

Zhang Q, Li Y, Li W J, Lin J J, Xiao M and Chen F Y. 2019. Salient object detection via deep features and multiple kernel boosting learning. Journal of Image and Graphics, 24(7): 1096-1105 (张晴, 李云, 李文举, 林家骏, 肖莽, 陈飞云. 2019. 融合深度特征和多核增强学习的显著目标检测. 中国图象图形学报, 24(7): 1096-1105) [DOI: 10.11834/jig.180224]

Zhang Q, Qin Q, Yang Y, Jiao Q and Han J G. 2024. Feature calibrating and fusing network for RGB-D salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 34(3): 1493-1507 [DOI: 10.1109/TCSVT.2023.3296581]

Zhao J X, Liu J J, Fan D P, Cao Y, Yang J F and Cheng M M. 2019. EGNet: edge guidance network for salient object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 8778-8787 [DOI: 10.1109/ICCV.2019.00887]

Zhao X K, Li M L, Zhang G, Li N and Li J S. 2021. Object detection method based on saliency map fusion for UAV-borne thermal images. Acta Automatica Sinica, 47(9): 2120-2131 (赵兴科, 李明磊, 张弓, 黎宁, 李家松. 2021. 基于显著图融合的无人机载热红外图像目标检测方法. 自动化学报, 47(9): 2120-2131) [DOI: 10.16383/j.aas.c200021]

Zhou W J, Pan S J, Lei J S and Yu L. 2022. TMFNet: three-input multilevel fusion network for detecting salient objects in RGB-D images. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(3): 593-601 [DOI: 10.1109/TETCI.2021.3097393]

Zhu X Z, Su W J, Lu L W, Li B, Wang X G and Dai J F. 2021. Deformable DETR: deformable transformers for end-to-end object detection//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR: 894-910

作者简介

宋霄罡，男，副教授，主要研究方向为计算机视觉和无人自主导航系统。E-mail: songxg@xaut.edu.cn

谭裕平，男，博士研究生，主要研究方向为显著目标检测和多模态融合。E-mail: 1179961190@qq.com

郭富强，男，硕士研究生，主要研究方向为人工智能和目标检测。E-mail: 3200441274@stu.xaut.edu.cn

鲁晓锋，男，教授，主要研究方向为模式识别和图像处理。E-mail: luxiaofeng@xaut.edu.cn

黑新宏，男，教授，主要研究方向为计算机视觉和人工智能。E-mail: heixinhong@xaut.edu.cn