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

中图法分类号:TP391;TP18　　**文献标识码**:A　　**文章编号**:1006-8961(2026)10-3618-18

论文引用格式:Chen X L, Zhong Z H and Shen Y J. 2026. A dual-modal salient object detection network for 360° omnidirectional images. Journal of Image and Graphics, 31(10):3618-3635(陈晓雷, 钟智华, 申玉杰. 2026. 双模态360度全景图像显著目标检测网络. 中国图象图形学报, 31(10):3618-3635)[DOI:10.11834/jig.250564]

双模态360度全景图像显著目标检测网络

陈晓雷*,钟智华,申玉杰

兰州理工大学微电子现代产业学院,兰州　730050

摘　要:**目的**　显著目标检测(salient object detection,SOD)旨在模拟人类视觉注意力机制,从图像或视频中识别并分割最显著的物体。尽管基于深度学习的2D SOD已取得显著进展,但面向360°全景图像的SOD因球面投影畸变和边界不连续性等问题,面临独特挑战。同时,虽然深度信息在2D RGB-D SOD中被证明能增强几何推理能力,但由于模态对齐困难、噪声敏感性以及缺乏针对ERP(equal rectangular projection)畸变的融合框架,其在360° SOD中的应用仍探索不足。对此,提出一种新颖的非对称双分支U-Net网络用于RGB-D 360° SOD。**方法**　该网络包含全景感知感受野模块(panoramic-aware receptive field module,PA-RFM)、注意力引导融合模块(attention-guided fusion module, AFM)和跨模态引导协同解码策略。PA-RFM通过经度—纬度—全局三重注意力机制缓解投影畸变;AFM实现自适应跨模态特征融合,有效利用深度信息;跨模态引导协同解码策略利用高分辨率RGB解码细节提升边界恢复精度。**结果**　在两个基准数据集(360-SOD和360-SSOD)上的大量主客观实验表明,本文方法性能优于现有10种代表性RGB-D 2D SOD先进方法和7种代表性RGB 360° SOD先进方法。在360-SOD数据集中,相比于性能第2的模型,MAE(mean absolute error)降低了13.7%,max-F提升4.64%,mean-F提升3.24%,Sm(structure-measure)提升1.96%,同时也在360-SSOD数据集中展示竞争优势。**结论**　本文提出的面向360°全景图像显著性检测的RGB-D网络,通过PA-RFM、AFM和跨模态引导协同解码三大模块协同优化,显著提升了检测精度与鲁棒性,同时验证了深度信息在360° SOD中的有效性与潜力。

关键词:显著目标检测(SOD);360°全景图像;RGB-D;跨模态融合;注意力机制

A dual-modal salient object detection network for 360° omnidirectional images

Chen Xiaolei*, Zhong Zhihua, Shen Yujie

School of Microelectronics Industry-Education Integration, Lanzhou University of Technology, Lanzhou　730050, China

Abstract: **Objective** Salient object detection (SOD) seeks to emulate human visual attention by locating and segmenting the most conspicuous objects in visual scenes. While deep-learning-based 2D RGB and RGB-D SOD methods have advanced substantially, 360° omnidirectional (panoramic) images introduce unique challenges, most notably spherical-to-plane projection distortions (e. g., equal rectangular projection (ERP) artifacts), severe polar-region deformation, and boundary discontinuities across the left-right wraparound. At the same time, depth cues are known to improve geometric reasoning in conventional RGB-D SOD, but their effective exploitation for 360° SOD is underexplored due to modality-alignment difficulties, noise in estimated depth, and the lack of fusion architectures tailored to panoramic projection effects. The objective of this work is to investigate and demonstrate how carefully designed cross-modal modeling and

收稿日期:2025-11-07;**修回日期**:2026-01-21;**预印本日期**:2026-01-28

*通信作者:陈晓雷　chenxl1703@lut.edu.cn

基金项目:国家自然科学基金项目(61967012)

Supported by: National Natural Science Foundation of China (61967012)

projection-aware feature processing can harness depth information to substantially improve salient object detection on ERP panoramic images, while remaining robust to depth estimation noise and projection artifacts. **Method** We propose an asymmetric dual-branch U-Net architecture for RGB-D 360° SOD that explicitly accounts for ERP projection properties and for cross-modal alignment. The model contains three novel components: 1) a panoramic-aware receptive field module (PA-RFM) that enhances direction-sensitive context modeling via a longitude-latitude-global triplet attention scheme and a ring-padding/cropping strategy to respect horizontal periodicity; 2) an attention-guided fusion module (AFM) that performs adaptive, dynamic weighting of RGB and depth channels using combined channel and spatial attention plus a lightweight modality weight generator to suppress noisy depth responses while preserving boundary cues; and 3) a cross-modal guided collaborative decoder in which a fusion-centric decoding path is explicitly guided by an independent high-resolution RGB decoding path via attentional gating, yielding superior edge and texture recovery. The RGB encoder is a pretrained SAM2 Hiera backbone with lightweight adapters to limit finetuning cost; the depth encoder is a ResNet-34 modified for single-channel input. PA-RFM employs a four-branch multiscale dilated design with a panoramic context aggregation module (PCAM) at each branch to merge longitude attention (via circular padding + horizontal convolutions), latitude attention (via latitude-weighted vertical convolutions that attenuate polar exaggeration), and global semantic attention (via latitude-weighted global average pooling (GAP) and channel recalibration). Training is performed in PyTorch on NVIDIA RTX 3090 hardware using AdamW ($lr = 1 \times 10^{-3}$, weight decay $= 5 \times 10^{-4}$), cosine annealing to 1×10^{-7}, batch size 4, up to 50 epochs, input resolution 512 × 1 024, and a structured loss composed of weighted BCE + weighted IoU with multilayer deep supervision. **Result** We evaluate the proposed model on two public panoramic SOD benchmarks augmented with depth maps produced by a state-of-the-art 360° depth estimator: 360-SOD (500 ERP images; 400 train/100 test) and 360-SSOD (1 105 ERP images; 850 train/255 test). Quantitatively, our method achieves MAE = 0.015 1, max-F = 0.838 8, mean-F = 0.815 0, max-E = 0.933 1, mean-E = 0.924 0, and Sm = 0.873 6 on 360-SOD (same scores on 360-SSOD), outperforming 10 representative RGB-D 2D SOD methods and seven representative RGB 360° SOD methods in objective metrics and visual quality. Relative to the strongest competing baseline, our mean absolute error(MAE) is reduced by 13.7%, max-F increases by 4.64%, mean-F by 3.24%, and Sm by 1.96%, demonstrating substantial gains particularly in boundary accuracy and structural consistency. Extensive qualitative comparisons show our model is markedly better at detecting objects in severely distorted polar regions, maintaining left-right boundary continuity, and recovering fine edges in cluttered or low-contrast scenes. Ablation studies confirm the contribution of each module: PA-RFM yields the largest single-module improvement (vs. RFB and a no-PA-RFM baseline), PCAM's three attention branches (longitude, latitude, global) act synergistically, and a ring-padding width of $k = 2$ provides the best tradeoff between boundary continuity and redundancy. AFM outperforms simple fusion strategies (add, concat, multiply) and several state-of-the-art (SOTA) fusion blocks (MobileSal, HIDANet, CPNet variants). The cross-modal guided decoder meaningfully improves edge/detail recovery over a variant that omits the RGB decoding path. We also test robustness to different depth estimators and find consistent improvements over RGB-only baselines; performance varies modestly with depth source, with Joint_360depth yielding the best overall results but not causing catastrophic degradation when replaced. **Conclusion** This study demonstrates that depth information can be effectively and robustly exploited for 360° omnidirectional SOD when combined with projection-aware receptive field design and modality-aware fusion and decoding strategies. The proposed PA-RFM addresses ERP-specific distortions (horizontal periodicity and polar exaggeration) via circular padding, latitude weighting and global recalibration; AFM adaptively reconciles noisy depth with RGB structure; and the cross-modal guided collaborative decoder preserves high-resolution RGB priors to restore fine boundaries and textures lost in fused representations. Together, these innovations lead to consistent and remarkable gains across standard panoramic SOD benchmarks, especially in polar regions and at projection boundaries. The approach is general and can be extended to other panoramic vision tasks and to setups with sensor-captured depth; future work will explore domain adaptation to real depth sensors, multiprojection fusion (ERP + CMP), and model compression for mobile/AR/VR deployment.

Key words: salient object detection(SOD); 360° omnidirectional image; RGB-D; cross-modal fusion; attention mechanism

0 引 言

随着全景视觉在虚拟现实(virtual reality, VR)/增强现实(augmented reality, AR)等应用中的日益普及,360°全景图像显著目标检测(360° salient object detection,简称360° SOD)逐渐成为视觉理解的重要方向。与常规2D图像不同,360°全景图像本质上是球形图像,需要通过等矩柱状投影(equal rectangular projection, ERP)、立方体投影(cube map projection, CMP)等技术转换为平面图像,才能进行后续处理,无论何种投影方式都不可避免地会引起图像畸变,这种投影畸变会严重影响SOD的效果。此外,360°全景图像还存在极区畸变、经纬向采样不均匀和边界不连续等复杂特性,导致显著区域的语义连贯性和空间一致性难以保持。

图像的深度信息可以反映图像的几何结构,具有内部一致性和光照不变性,能够为RGB图像提供辅助信息,从而帮助SOD模型有效区分前景和背景。在光照条件低、目标与背景对比度弱或纹理信息易混淆的情形下,深度信息仍能保持清晰的空间结构线索,有效辅助模型区分前景与背景并抑制由纹理导致的误检。然而,针对360°全景图像,将深度信息有效地与RGB信息融合并非易事:一方面,现有全景相机无法同时采集RGB信息和深度信息,已有的两个360° SOD公开数据集也只是RGB单模态数据集,只能采用深度估计方法为现有数据集增加深度信息;另一方面,模态间语义对齐与尺度一致性在360°全景图像上更难保证。尽管近年来已有大量RGB-D 2D SOD和若干RGB 360° SOD研究,但系统性地将深度信息引入360° SOD,并同时考虑ERP特性与跨模态鲁棒融合的研究未见文献报道。

针对以上问题,本文提出一种RGB-D 360° SOD网络。网络总体为非对称RGB-D双分支U-Net结构:为保证对全景图像畸变区域和大范围显著区域的表征能力,RGB分支采用了具备强全局建模能力与长程依赖捕获能力的SAM2-Hiera作为编码器。深度信息分支采用ResNet-34(He等,2016)专门提取单通道几何边界特征。为适配360°全景图像的几何特性,提出了全景感知感受野模块(panoramic-aware receptive field module, PA-RFM)以增强方向感知上下文建模能力并缓解极区畸变与边界不连续性;为实现跨模态融合,设计注意力引导融合模块(attention-guided fusion module, AFM)自适应地抑制深度噪声同时保留边界补偿能力;解码端采用以融合分支为主、独立RGB解码路径作为高分辨率细节引导的协同解码策略,通过多尺度注意力加权显著提升边缘与纹理恢复。

本文主要创新点和工作总结如下:1)提出了一个非对称双分支U-Net结构的RGB-D 360° SOD网络,该网络同时考虑了360°全景图像自身特性带来的挑战和深度信息对显著性目标检测性能的提高作用,进一步提高了360° SOD性能。2)设计了全景感知感受野模块PA-RFM,通过经度注意、纬度加权与全局语义注意的联合建模,有效缓解边界不连续与极区畸变问题。3)设计了注意力引导融合模块AFM,以注意力机制为核心驱动,结合通道与空间注意力、动态权重分配策略及多尺度特征建模,实现RGB特征与深度信息特征的高效对齐与融合。4)提出跨模态引导协同解码策略,通过独立的RGB解码路径为融合分支提供高分辨率结构先验,显著提升边缘与细节恢复效果。5)在两个公开的360° SOD数据集上评估了本文方法,并将其与现有的RGB-D 2D SOD和360° SOD方法进行比较。实验结果表明,本文方法的性能超过了现有代表性先进方法。

本文结构组织如下:第1节回顾相关工作(包括RGB-D SOD与360° SOD的最新进展);第2节详细介绍所提网络结构与各模块设计;第3节给出实验设置、本文方法与现有代表性先进方法的对比结果及消融分析;最后在第4节对本文工作进行总结并讨论未来研究方向。

1 相关工作

1.1 RGB-D 2D SOD方法

近年来,RGB 2D SOD虽然已取得较好性能,但在复杂背景与弱光场景下仍存在局限,因此研究者开始引入深度信息增强显著性检测。随着深度传感器的发展,早期RGB-D SOD工作(如Lang等,2012;Ciptadi等,2013;Ren等,2015;Cong等,2019)主要依赖手工特征,性能有限。深度学习出现后,大量基于卷积神经网络(convolutional neural network, CNN)的RGB-D 2D SOD方法取得明显进展,例如Wu等人(2022)利用轻量级CNN与隐式深度恢复实现高效

多尺度融合。但受限于卷积的局部感受野，CNN在建模长程依赖方面仍存在不足。

随着ViT（vision Transformer）（Dosovitskiy等，2021）的兴起，更强的跨区域建模能力被引入RGB-D SOD中。部分方法采用纯Transformer框架，如Liu等人（2021c）通过三元组Transformer嵌入与三分支解码器实现跨模态融合。也有方法结合CNN与Transformer，如Liu等人（2022）使用双流Swin Transformer进行多模态编码；Lee等人（2022）将Transformer与图卷积神经网络（graph convolutional network，GCN）结合以实现原型采样与鲁棒融合；Cong等人（2023b）提出CNN辅助的Transformer架构，通过注意力触发的跨模态点感知交互提升特征融合；Yin等人（2023）在编码器中加入RGB-D block以加强模态交互；Hu等人（2024）基于Swin Transformer设计跨模态融合与渐进解码结构，在无需额外增强模块的情况下获得良好性能。

1.2　RGB 360° SOD方法

近年来，已有部分研究开始探索360° SOD，但整体仍较有限。DDS（distortion-adaptive network with deep supervision）（Li等，2020）首次提出RGB 360° SOD网络，并引入失真自适应模块处理ERP畸变；LDNet（lightweight distortion-aware network）（Huang等，2023）基于失真感知与深度可分离卷积缓解投影失真；DATFormer（distortion-aware Transformer）（Zhao等，2023）利用Transformer建模ERP特性；ACoNet（adjacent coordination network）（陈晓雷等，2024）通过多分支结构实现跨尺度特征交互；DSANet（distortion semantic aggregation network）（陈晓雷 等，2025b）结合畸变自适应卷积与多尺度注意增强语义融合；DPNet（distortion-adaptive and position-aware network）（陈晓雷 等，2025a）采用ViT + CNN双编码器并加入可变形卷积以适应ERP几何畸变。在互补模态利用方面，FANet（features adaptation network）（Huang等，2020）同时利用ERP与CMP信息实现自适应特征融合，MPFR-Net（multi-projection fusion and refinement network）（Cong等，2023a）采用多CMP图像与动态加权策略保持目标结构完整。

2　本文模型

2.1　网络总体结构

本文提出的模型结构如图1所示，支持RGB图像与深度图并行输入。

RGB分支编码器采用SAM2（segment anything model 2）（Ravi等，2025）预训练的Hiera（Ryali等，2023）主干，输出多尺度特征$\boldsymbol{F}_i(i=1,2,3,4)$。Hiera的结构如图1左下所示，原始Hiera的大参数使得完全微调的计算代价很高，为了提高参数效率，本文采用了SAM2-UNet（segment anything 2-UNet）（Xiong等，2026）的适配器设计，冻结了Hiera的部分参数，在每个多尺度块之前插入了轻量级适配器。输出多尺度特征后通过全景感知感受野模块（PA-RFM）增强上下文建模以缓解极区畸变；由于深度图本身已具备良好的结构边界和空间位置信息，特征较为简洁直接，深度分支使用轻量化ResNet-34（34-layer residual network）提取深度特征，无需复杂增强模块以避免冗余。RGB与深度特征通过注意力引导融合模块（AFM）自适应整合模态互补信息并抑制噪声，生成融合特征用于多级解码。解码阶段，RGB与融合分支分别采用U-Net结构的上采样模块进行逐级上采样重建。

RGB分支的解码侧重于纹理细节与色彩信息还原，其结构如图1右下所示，最终输出3层辅助显著性预测图，为减少计算量与复杂度，仅对拥有最精细特征的输出预测$\boldsymbol{S}_4$进行监督。对于融合特征的解码，本文设计了一个跨模态引导协同解码策略，利用注意力机制将RGB解码特征引导融合至融合分支，突出显著区域与边缘信息，强化解码分辨能力。最终，融合分支的输出包含主显著图$\boldsymbol{S}_1$及两层多尺度侧输出$\boldsymbol{S}_2$、$\boldsymbol{S}_3$，实现主—辅协同监督策略，有效提升模型的显著性区域聚焦能力与整体检测鲁棒性。

2.2　全景感知感受野模块PA-RFM

本文为解决ERP格式下360°全景图像存在的空间几何畸变与上下文不均衡问题，设计了全景感知感受野模块PA-RFM，其结构如图2（a）所示，旨在同时建模方向敏感的局部上下文、多尺度语义依赖与结构一致性的全局感知能力。PA-RFM为四路并行结构，通过不同卷积核尺寸与空洞率的空洞卷积感知多尺度信息，每一路都利用本文设计的全景上下文聚合模块（panoramic context aggregation module，PCAM）提高全景感知能力，最终拼接融合并加残差以增强信息保留，从而兼顾多尺度显著目标与ERP图像几何特性。

其中全景上下文聚合模块PCAM作为PA-RFM

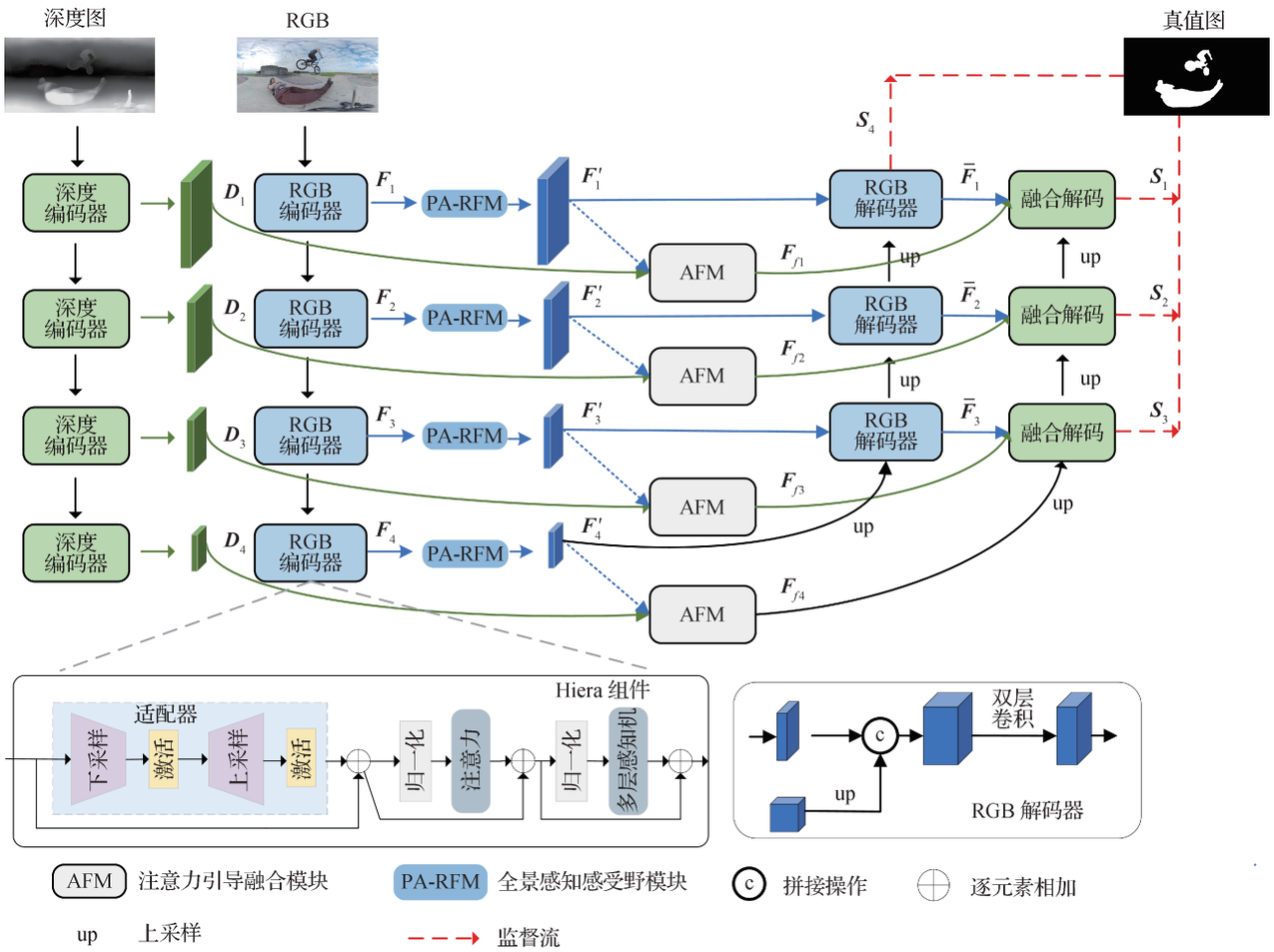


图1 本文模型网络结构

Fig. 1 The network architecture of the proposed model

的核心组件，专为ERP格式下的360°全景图像特性量身设计，结构如图2(b)所示，旨在从方向敏感性、空间均衡性与语义一致性3个维度对特征进行增强建模。该模块引入三重上下文建模机制：经度方向注意、纬度方向注意与全局语义注意，以充分刻画ERP投影下非均匀几何结构对显著性建模的影响，并融合为一个统一的上下文增强图以调节原始特征表示。

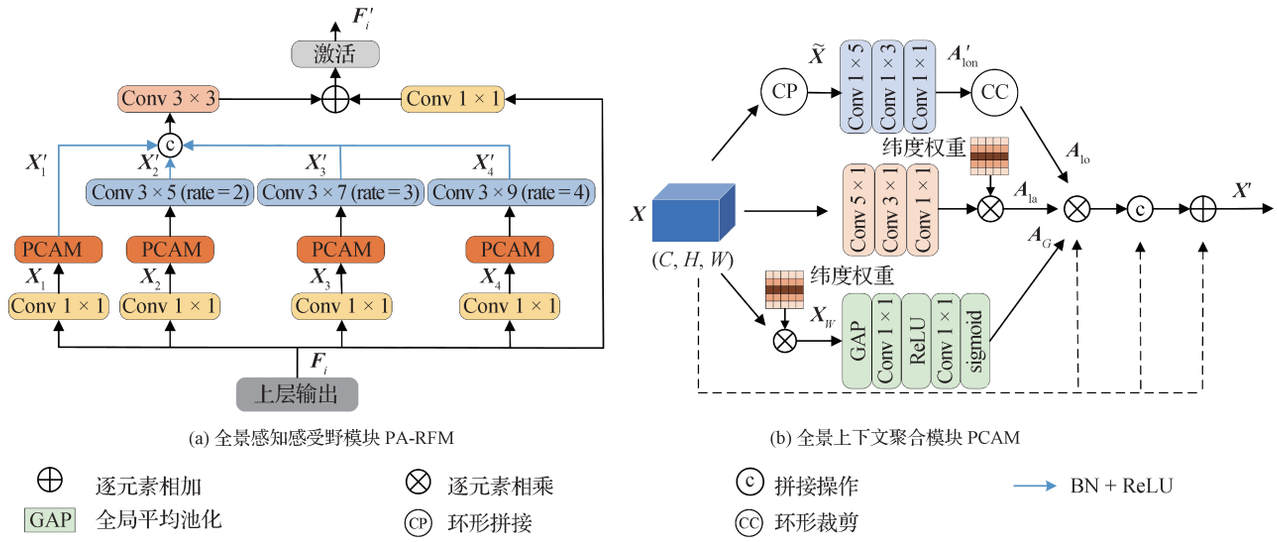
经度方向建模如图2(b)中的上部分支所示，首先针对360°全景图像在水平方向具备周期性边界的拓扑属性，引入环形填充与裁剪策略，以弥补常规卷积在边界处感受野断裂的问题，其效果如图3所示，设给定输入特征 $\boldsymbol{X} \in \mathbf{R}^{B \times C \times H \times W}$，$\mathbf{R}^{B \times C \times H \times W}$ 表示特征的维度。首先进行环形填充，基本原理如图3上部所示，从完整特征的最右侧和最左侧各提取 k 列特征(本文取 $k = 2$，既能保证卷积在边界处仍能感知完整的局部窗口并与全景图像的周期性拓扑对齐，同时避免不必要的填充带来的冗余计算，下文环形裁剪操作也取 $k = 2$)，再分别拼接到左侧和右侧，完成环形扩展，具体为

$$\tilde{\boldsymbol{X}} = concat(\boldsymbol{X}_{[\cdots,W-k:W]}, \boldsymbol{X}, \boldsymbol{X}_{[\cdots,0:k]}) \in \mathbf{R}^{B \times C \times H \times (W+2k)} \quad (1)$$

式中，$\tilde{\boldsymbol{X}}$ 表示进行环形裁剪后得到的特征图，$concat$ 表示拼接操作，$\boldsymbol{X}$ 表示输入的原始特征。该操作使模型在处理边缘区域时具备连续空间感知能力，有效提升水平结构一致性建模性能。对于填充后的特征 $\tilde{\boldsymbol{X}}$，设计了一个经度方向注意力机制，以增强其在水平方向的上下文感知能力。该模块依次采用核为1 × 5和1 × 3的卷积聚合横向信息，随后通过1 × 1卷积恢复通道数，并经sigmoid得到经度注意力权重图。

随后进行环形裁剪，基本原理如图3下部所示，对 $\boldsymbol{A}'_{\text{lo}}$ 在宽度维度裁剪去除左右各 k 列，恢复为与原始输入宽度相同的经度注意力图，具体为

$$\boldsymbol{A}_{\text{lo}} = \boldsymbol{A}'_{\text{lo}}[:, :, :, k:W+k] \in \mathbf{R}^{B \times C \times H \times W} \quad (2)$$



(a) 全景感知感受野模块 PA-RFM　　(b) 全景上下文聚合模块 PCAM

图 2　全景感知感受野模块 PA-RFM 结构图

Fig. 2　Structure of the panoramic-aware receptive field module (PA-RFM) ((a) PA-RFM; (b) PCAM)

式中，$\boldsymbol{A}_{\mathrm{lo}}$ 表示进行环形裁剪后得到的特征图，$\boldsymbol{A}'_{\mathrm{lo}}$ 表示进行卷积操作后的中间结果，k 表示裁剪宽度。该机制可有效建模跨水平方向的上下文依赖关系，从而提升模型对全景图像水平连续区域的理解能力。

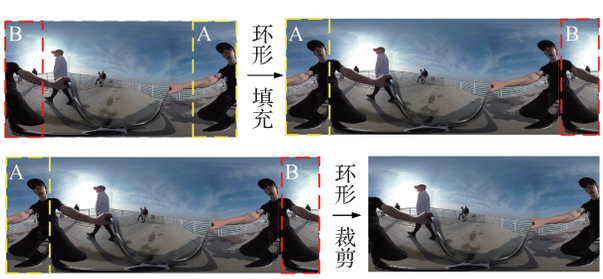


图 3　环形填充与裁剪示意图

Fig. 3　Illustration of ring padding and cropping

在以上经度方向建模中，主要解决了全景图像中边界不连续的问题，但是全景图像还有一个严重问题，就是在极区的严重畸变与失真。为解决这一问题，如图 2(b)中的中部分支所示，本文首先根据 ERP 图像的球面投影特性，对每一行像素计算对应的纬度角，具体为

$$\theta_i = -\frac{\pi}{2} + \frac{i\pi}{H-1}, i = 0, \cdots, H-1 \tag{3}$$

式中，θ_i 表示计算得到的纬度角，i 代表特征图高度方向上的像素行索引，H 代表 ERP 格式全景图像特征图的高度。以余弦函数生成初始权重向量，再乘以一个可学习标量 ∂。具体为

$$\omega_{\mathrm{la}}(i) = \partial \cos\theta_i \tag{4}$$

式中，ω_{la} 表示特征对应的纬度权重，∂ 表示可学习标量。这一权重能够衰减极区过度聚集的投影效应，强化赤道附近区域的特征响应。接着，对原始特征 $\boldsymbol{X} \in \mathbf{R}^{B\times C\times H\times W}$ 逐像素乘以 ω_{la}，先后施加 3 层竖向卷积以捕获多尺度南北上下文，具体方法如下：首先由 5 × 1 卷积及后续的批归一化和 ReLU (rectified linear unit) 激活提取宽度范围的垂直信息；紧接着用 3 × 1 卷积加 BN (batch normalization, BN) + ReLU 对该信息进行局部细化；最后以 1 × 1 卷积恢复原始通道数并通过 sigmoid 激活生成归一化的纬度注意力图。

将该注意力映射与预先计算的纬度权重相乘，可得最终的纬度注意力，具体为

$$\boldsymbol{A}_{\mathrm{la}} = \tilde{\boldsymbol{A}}_{\mathrm{la}} \otimes \omega_{\mathrm{la}} \in \mathbf{R}^{B\times C\times H\times W} \tag{5}$$

式中，$\boldsymbol{A}_{\mathrm{la}}$ 表示纬度注意力图，$\tilde{\boldsymbol{A}}_{\mathrm{la}}$ 表示进行 3 层竖向卷积后得到的特征。

在完成经度与纬度方向的局部上下文建模后，局部感受野虽然能有效捕捉跨边界与方向感知信息，但是依然难以覆盖整幅图像的语义全貌，尤其对于 360° 全景图像中跨区域的长距离依赖关系。为此，PCAM 模块进一步引入全局上下文建模机制，其结构如图 2(b)下部分支所示。具体而言，首先利用纬度权重对原始特征进行加权，具体为

$$\boldsymbol{X}_{\mathrm{w}} = \boldsymbol{X} \otimes \boldsymbol{w}_{\mathrm{la}} \tag{6}$$

式中，$\boldsymbol{X}_{\mathrm{w}}$ 表示经过纬度加权的特征，⊗表示逐元素相乘。随后，对 $\boldsymbol{X}_{\mathrm{w}}$ 执行全局平均池化，将空间维度压缩为 1 × 1，具体为

$$\boldsymbol{Z} = GAP\left(\boldsymbol{X}_{\mathrm{w}}\right) \in \mathbf{R}^{B\times C\times 1\times 1} \tag{7}$$

式中，$\boldsymbol{Z}$表示压缩后的特征，GAP表示全局平均池化。但在进入全局平均池化之前，先对特征图按纬度加权，这是因为在ERP投影下，两极像素代表的球面区域远小于赤道像素，若直接做全局平均池化，极区冗余信息会过度主导通道统计。接下来，通过两层1×1卷积与ReLU激活，构建通道级的语义重标定，再经sigmoid归一化生成全局语义注意力向量$\boldsymbol{A}_g \in [0,1]^{B\times C\times 1\times 1}$。最后将$\boldsymbol{A}_g$沿空间维度广播回$(H, W)$，得到与原始特征尺寸对齐的全局上下文图$\boldsymbol{A}_{\mathrm{G}} \in \mathbf{R}^{B\times C\times H\times W}$。以上处理过程可以表示为

$$\boldsymbol{A}_{\mathrm{G}} = expand\left(sigmoid\left(Conv_{1\times 1}\left(ReLU\left(Conv_{1\times 1}(\boldsymbol{Z})\right)\right)\right)\right) \tag{8}$$

式中，$\boldsymbol{A}_{\mathrm{G}}$表示计算得到的全局上下文图，$expand(\cdot)$表示将$(B, C, 1, 1)$的注意力向量广播到$(B, C, H, W)$，$ReLU$表示激活函数，$Conv_{1\times 1}$表示1×1的卷积，$sigmoid$表示sigmoid激活函数。

将3种上下文注意力图进行逐像素相乘融合，形成统一的上下文增强图$\boldsymbol{A}_{\mathrm{c}}$，具体为

$$\boldsymbol{A}_{\mathrm{c}} = \boldsymbol{A}_{\mathrm{lo}} \otimes \boldsymbol{A}_{\mathrm{la}} \otimes \boldsymbol{A}_{\mathrm{G}} \tag{9}$$

式中，$\boldsymbol{A}_{\mathrm{lo}}$、$\boldsymbol{A}_{\mathrm{la}}$、$\boldsymbol{A}_{\mathrm{G}}$分别表示上文中计算得到的经度注意力图、纬度注意力图以及全局上下文图。

随后将其与原始特征进行逐通道加权相乘并拼接，具体为

$$\boldsymbol{X}' = concat\left(\boldsymbol{X} \otimes \boldsymbol{A}_{\mathrm{c}}, \boldsymbol{X}\right) \tag{10}$$

式中，$\boldsymbol{X}'$表示经过全景上下文聚合模块最后得到的特征。

2.3 注意力引导融合模块AFM

本文为了实现RGB与深度模态之间的高效融合，设计了注意力引导融合模块AFM，结构如图4所示，该模块结合通道—空间注意力机制以及动态权重分配策略，以实现跨模态特征的有效增强与融合。具体而言，对于输入的RGB特征$\boldsymbol{F}'$和深度特征$\boldsymbol{D}$，首先通过通道注意力与空间注意力机制增强其表示能力。通道注意力机制建模了特征图不同通道在全局语义上的重要性，同时对质量较差或噪声较多的深度通道进行自适应抑制，通道注意力图$\boldsymbol{M}_{\mathrm{c}}$的计算过程为

$$\boldsymbol{M}_{\mathrm{c}} = sigmoid\left(Conv_{1\times 1}\left(Conv_{1\times 1}GAP(\boldsymbol{F}')\right)\right) \tag{11}$$

$$\dot{\boldsymbol{F}} = \boldsymbol{F}' \otimes \boldsymbol{M}_{\mathrm{c}} \tag{12}$$

式中，$\boldsymbol{M}_{\mathrm{c}}$表示得到的通道注意力图，$\boldsymbol{F}'$表示经过全景感知感受野模块处理后的特征，$\dot{\boldsymbol{F}}$表示将$\boldsymbol{F}'$与通道注意力图进行逐通道相乘得到的特征。接着，为进一步强调空间上的显著区域，引入空间注意力机制，计算空间注意力图$\boldsymbol{M}_{\mathrm{s}}$，具体为

$$\boldsymbol{M}_{\mathrm{s}} = \Phi\left(Conv_{7\times 7}\left(concat\left[Avg\left(\dot{\boldsymbol{F}}\right), Max\left(\dot{\boldsymbol{F}}\right)\right]\right)\right) \tag{13}$$

$$\breve{\boldsymbol{F}} = \boldsymbol{F}' \otimes \boldsymbol{M}_{\mathrm{s}} \tag{14}$$

式中，$\boldsymbol{M}_{\mathrm{s}}$表示得到的空间注意力图，$\Phi$表示ReLU + BN操作，$Conv_{7\times 7}$表示7×7的卷积，$Avg$表示逐通道平均池化操作，$Max$表示逐通道最大池化操作，$\breve{\boldsymbol{F}}$表示通过与空间注意力图进行逐元素相乘得到的特征，为实现模态间的信息选择性融合，本文设计了一个轻量级动态模态权重生成器，用于生成模态级别的权重向量，具体为

$$\breve{\boldsymbol{F}}_{\mathrm{c}} = concat\left(\breve{\boldsymbol{F}}_{\mathrm{R}}, \breve{\boldsymbol{F}}_{\mathrm{D}}\right) \tag{15}$$

$$\boldsymbol{W} = Softmax\left(Conv_{1\times 1}\left(GAP\left(\breve{\boldsymbol{F}}\right)\right)\right) \tag{16}$$

式中，$\breve{\boldsymbol{F}}_{\mathrm{c}}$表示将经过通道和空间注意力得到的RGB和深度信息特征进行融合后的特征，$\breve{\boldsymbol{F}}_{\mathrm{R}}$表示RGB分支的注意力图，$\breve{\boldsymbol{F}}_{\mathrm{D}}$表示深度分支得到的注意力图。

$$\boldsymbol{F}_{\mathrm{out}} = concat\left(\boldsymbol{W}_{\mathrm{R}} \otimes \breve{\boldsymbol{F}}_{\mathrm{R}}, \boldsymbol{W}_{\mathrm{D}} \otimes \breve{\boldsymbol{F}}_{\mathrm{D}}\right) \tag{17}$$

式中，$\boldsymbol{W}_{\mathrm{R}}$和$\boldsymbol{W}_{\mathrm{D}}$表示$\boldsymbol{W}$经通道拆分得到的模态特定权重，$\boldsymbol{F}_{\mathrm{out}}$表示两个模态的最终融合结果。

在融合特征生成后，为进一步提升特征表达能力，本文引入了残差连接与特征增强机制。将$\boldsymbol{F}_{\mathrm{out}}$送入一个1×1卷积层，得到增强特征。随后，与两个原始模态的注意力增强特征进行残差融合，并通过3×3卷积输出最终结果。

综上所述，AFM模块通过引入双重注意力机制与动态模态融合策略，在保持结构简洁的同时有效增强了多模态特征的表达与融合能力。

2.4 跨模态引导协同解码策略

在完成全景感知与跨模态融合设计后，核心挑战在于如何高效地将多尺度特征从编码端引导至解码端以精细还原显著区域。为此，本文提出一种跨模态引导协同解码策略，如图5所示，以融合分支为主干，动态汲取RGB分支的高分辨率细节信息。该策略通过逐层注意力加权实现信息重塑与多尺度一

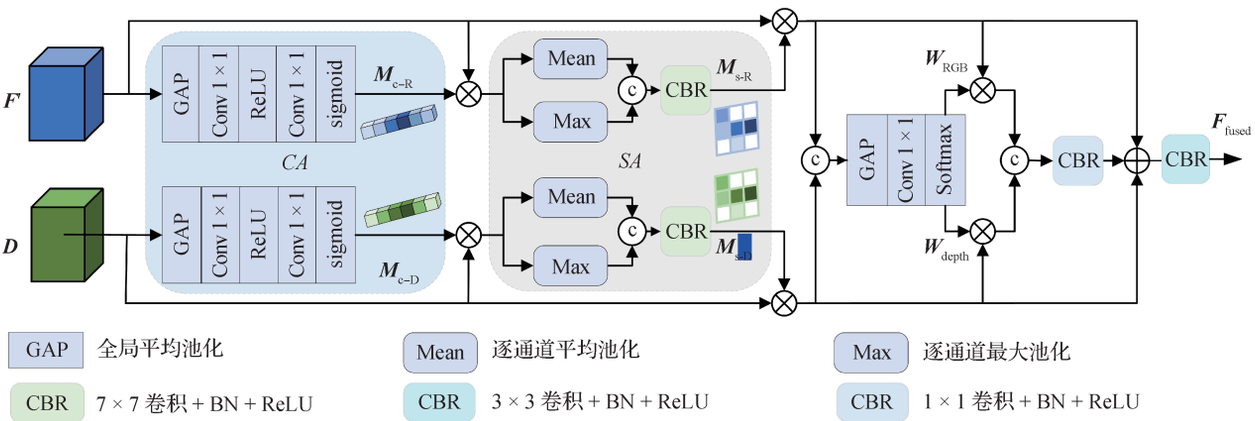


图4　注意力引导融合模块AFM结构图

Fig. 4　Structure of the attention-guided fusion module (AFM)

致性，同时引入独立的RGB解码路径缓解融合特征中RGB语义弱化问题。具体而言，每层融合特征先与上层上采样特征拼接并经过两次Conv-BN-ReLU生成初级解码特征 $\boldsymbol{X}_i$，再与RGB解码特征 $\bar{\boldsymbol{F}}_i$ 拼接，通过轻量注意力生成通道加权系数 $\boldsymbol{A}$，得到高级解码特征 $\boldsymbol{Y}_i$ 并进行特征融合，最后经3 × 3 Conv-BN-ReLU生成该层最终输出 $\boldsymbol{S}_i$。该设计有效增强了RGB特征与融合特征间的互补性，提升了显著区域的结构保留与边缘细节恢复能力，以上处理过程可以表示为

$$\boldsymbol{X}_i = \sigma\left(concat\left(\boldsymbol{F}_{fi}, upsampling\left(\boldsymbol{F}_{f(i+1)}\right)\right)\right),\ i = 1, 2, 3 \tag{18}$$

$$\boldsymbol{A} = sigmoid\left(Conv_{1\times1}ReLU\left(Conv_{1\times1}\left(concat\left(\boldsymbol{X}_i, \bar{\boldsymbol{F}}_i\right)\right)\right)\right) \tag{19}$$

$$\boldsymbol{Y}_i = \boldsymbol{A}_i \otimes \bar{\boldsymbol{F}}_i + (1 - \boldsymbol{A}) \otimes \boldsymbol{X}_i \tag{20}$$

$$\boldsymbol{S}_i = \Phi\left(Conv_{3\times3}\left(concat\left[\boldsymbol{Y}_i, \bar{\boldsymbol{F}}_i\right]\right)\right) \tag{21}$$

式中，$\boldsymbol{F}_{fi}$ 表示经过融合模块得到的特征，$\boldsymbol{F}_{f(i+1)}$ 表示上一层得到的融合特征，$\boldsymbol{A}$ 表示通道加权系数，$\boldsymbol{Y}_i$ 与 $\boldsymbol{X}_i$ 表示计算过程中的中间结果，$\bar{\boldsymbol{F}}_i$ 表示RGB解码特征，$i = 1, 2, 3$，σ 表示进行两次的 $\text{Conv}_{3\times3}$-Batch Normalization-ReLU操作，*upsampling* 表示双线性插值上采样。

最后，为RGB最底层特征 $\boldsymbol{S}_4$ 和每一层融合解码特征 $\boldsymbol{S}_1$、$\boldsymbol{S}_2$、$\boldsymbol{S}_3$ 附加侧输出监督，采用1 × 1卷积层进行显著图生成，并通过双线性插值上采样操作恢复至原图尺度以便于训练与评估。最终输出由最底层融合解码特征 $\boldsymbol{S}_1$ 生成，表示最终的预测结果。

2.5　损失函数

本文的损失函数为加权的交并比(intersection over union，IoU)损失 L_{wIoU} 和加权的二进制交叉熵(binary cross entropy，BCE)损失 L_{wBCE} 组合，定义为

$$L = L_{\text{wIoU}} + L_{\text{wBCE}} \tag{22}$$

为了增强训练有效性，对RGB分支解码的顶层输出和融合解码的所有输出应用深度监督，从而得到最终的总损失函数，具体为

$$L_{\text{total}} = L_{\text{RGB}}(\boldsymbol{G}, \boldsymbol{S}) + \sum_{i=1}^{3} L_f(\boldsymbol{G}, \boldsymbol{S}_i) \tag{23}$$

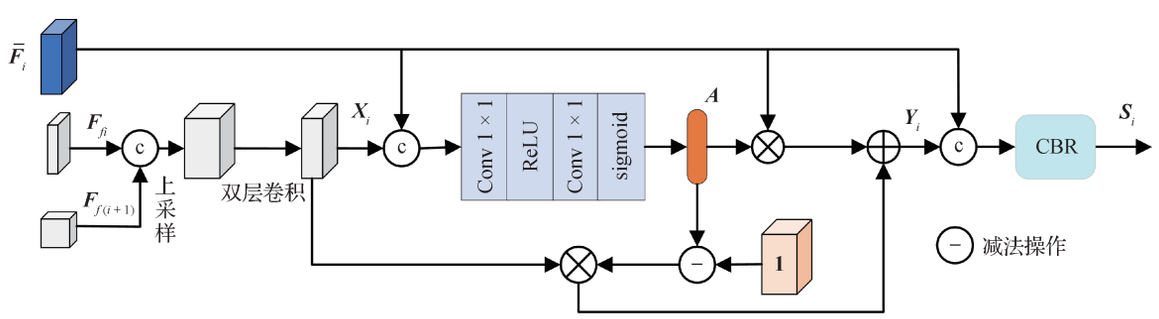


图5　跨模态引导协同解码策略结构图

Fig. 5　Structure of the cross-modal guided collaborative decoding strategy

式中，G表示真实标签，S表示RGB分支解码的顶层输出，S_i表示由融合解码生成的多级分割输出，L_{total}表示最终损失。

3 实验与分析

3.1 数据集及评价指标

本文使用360-SOD（Li等，2020）和360-SSOD（Ma等，2020）两个公开的360°全景图像数据集测试本文模型的性能，将Yun等人（2022）提出的360°全景图像深度估计方法，为上述两个360°全景图像数据集生成深度图子集。扩充后的360-SOD包含500幅高分辨率ERP全景图像及其对应的深度图，360-SSOD包含1 105幅高分辨率ERP全景图像及其对应的深度图。本文使用以下几个评价指标对所有模型的性能进行评估：平均绝对误差（mean absolute error，MAE）（Perazzi等，2012）、F-measure（F）（Achanta等，2009）、E-measure（Em）（Fan等，2018）和结构测度（structure-measure，Sm）（Fan等，2018）。

3.2 实验设置

本文使用PyTorch训练所提出的模型，所有实验均在NVIDIA RTX 3090 GPU服务器上完成。编码器采用经过预训练的SAM2 Hiera-L模型和ResNet-34（34-layer residual network），为避免大型预训练骨干导致算法对比时的不公平性，本文冻结了Hiera的部分参数，在每个多尺度块之前插入了轻量级适配器（Xiong等，2026）。输入包括RGB图像与对应的深度图像，图像分辨率为512×1 024像素。本文使用AdamW优化器进行训练，最大学习率设置为1×10^{-3}，权重衰减设置为5×10^{-4}，为防止过拟合提供正则化约束。训练过程中，引入了余弦退火策略，将学习率逐步衰减至最小值1×10^{-7}，衰减周期设置为整个训练周期数（即每50个epoch衰减一次），batch size设置为4，最大训练周期数设为50。训练过程中采用第2.5节定义的结构化损失（加权BCE + 加权IoU）并结合多层深度监督，以提升模型收敛与特征学习效果。实验中数据集的划分如下：360-SOD数据集中400幅RGB图像及对应深度图用于训练，100幅图像及对应深度图用于测试；360-SSOD中850幅图像及对应深度图用于训练，255幅图像及对应深度图用于测试。训练过程中，使用随机垂直和水平翻转对数据集进行增强。

3.3 实验对比

本文将所提出的模型与10种代表性的RGB-D 2D SOD方法和7种代表性的RGB 360° SOD方法进行了对比。RGB-D 2D SOD方法包括VST（visual saliency Transformer）（Liu等，2021a）、TriTransNet（Liu等，2021c）、BTS-Net（bi-directional transfer-and-selection network）（Zhang等，2021）、SwinNet（Liu等，2022）、MobileSal（Wu等，2022）、SPSN（superpixel prototype sampling network）（Lee等，2022）、PICR-Net（point-aware interaction and CNN-induced refinement network）（Cong等，2023b）、HIDANet（hierarchical depth awareness network）（Wu等，2023）、DFormer-B（Yin等，2023）和CPNet（cross-modal fusion and progressive decoding network）（Hu等，2024），上述10种方法使用360-SOD和360-SSOD数据集中的RGB图像和本文生成的深度估计图像进行训练。RGB 360° SOD方法包括FANet（Huang等，2020）、MPFR-Net（Cong等，2023a）、LDNet（Huang等，2023）、DAT-Former（Zhao等，2023）、ACoNet（陈晓雷 等，2024）、DPNet（陈晓雷 等，2025a）和DSANet（陈晓雷 等，2025b），这些方法只使用360-SOD和360-SSOD数据集中的RGB图像进行训练。为了公平比较，所有模型均使用官方代码或者作者提供的显著图在同一软硬件环境中测试，并且都做了微调以获得最好结果。

所有方法在360-SOD数据集上的客观指标对比如表1所示。可以看到，本文方法相较于现有的RGB-D 2D SOD方法和RGB 360° SOD方法取得了最好的性能，尤其相对于RGB 360° SOD方法，本文方法展示出了较大程度的提高。相对于所有对比方法中的次优值，本文方法的MAE指标降低了13.7%，max-F提高了4.64%，mean-F提高了3.24%，max-Em提高了0.42%，mean-Em提高了0.39%，Sm提高了1.96%。

在360-SOD数据集上的主观结果比较如图6所示，本文选取了若干具有挑战性和代表性的场景进行比较。图6是本文方法与10种代表性RGB-D 2D SOD方法和5种代表性RGB 360° SOD方法的对比，从图中可以看出，本文方法生成的显著性目标图比其他方法更接近真值图。例如：第1行和第7行的场景中，对于存在严重畸变的极区，只有本文方法正确检测到了极区物体，反映出本文方法对于严重畸变区域的检测效果是优秀的；在具有复杂背景的场景

中，例如第3、4、5、8行，本文方法能够准确识别目标主体和边界并完成分割，优于其他方法；在投影边界不连续场景中，例如第2行，ERP投影导致图像边界处的显著目标出现结构断裂，本文方法在该场景下的检测结果优于现有方法；另外，当图像中存在多个检测目标时，例如第6行场景，本文方法能够精准定位所有显著目标并实现精细分割，而其他方法则存在部分显著目标遗漏的情况。

所有方法在360-SSOD数据集的客观指标对比如表2所示。整体来看，本文方法综合性能最好，其各项指标上的表现同样具有竞争力。

360-SSOD数据集上的主观结果比较如图7所示，从图7可以看出，与目前代表性RGB-D 2D SOD先进方法和RGB 360° SOD先进方法相比，本文方法在畸变极区（例如图7第2行图像）的检测效果优于其他方法，另外对低对比度场景（例如图7第1、5行中的图像）和具有复杂背景的场景（例如图7第3、6、7行中的图像），本文方法能够有效抑制背景或突出前景，实现了更精细的显著目标检测。

图8和图9展示了不同模型在360-SOD和360-SSOD数据集上的P-R（precision-recall）曲线和F-measure曲线，从图中可以看到，本文模型（红实线）在两个数据集上的表现均为最优，进一步证明了本文模型的优越性能。

3.4　消融实验

本节将通过删除或者替换本文模型的不同组件来研究各组件的有效性，主要验证方法与组件为：不同深度估计方法的影响分析、全景感知感受野模块PA-RFM的有效性、注意力引导融合模块AFM的有效性以及跨模态引导协同解码策略的有效性。本文在360-SOD数据集上分别对以上方法和组件进行消融实验。

3.4.1　不同深度估计方法的影响分析

在相同的网络结构（均使用SAM2-Hiera作为编

表1　不同方法在360-SOD数据集上的客观指标对比

Table 1　Comparison of objective metrics of different methods on the 360-SOD dataset

	方法	MAE ↓	max-F ↑	mean-F ↑	max-Em ↑	mean-Em ↑	Sm ↑
RGB-D 2D SOD	VST（2021）	0.022 9	0.735 3	0.691 9	0.884 6	0.853 7	0.814 7
	TriTransNet（2021）	0.021 7	0.774 0	0.732 1	0.902 6	0.897 0	0.820 4
	BTS-Net（2021）	0.024 0	0.723 0	0.685 3	0.878 0	0.803 2	0.797 2
	SwinNet（2022）	0.028 3	0.706 3	0.666 1	0.864 0	0.793 2	0.778 3
	MobileSal（2022）	0.026 8	0.613 6	0.597 3	0.822 2	0.782 7	0.742 0
	SPSN（2022）	0.019 3	0.774 8	0.758 4	0.893 8	0.877 9	0.831 4
	PICR-Net（2023）	0.020 4	0.773 4	0.758 3	0.900 5	0.894 3	0.830 6
	HIDANet（2023）	0.019 1	0.786 2	0.773 0	0.908 6	0.903 9	0.847 3
	DFormer-B（2024）	0.017 5	0.799 9	0.775 8	0.929 2	0.920 4	0.856 8
	CPNet（2024）	0.018 7	0.794 2	0.789 4	0.903 9	0.894 6	0.840 5
RGB 360° SOD	FANet（2020）	0.024 9	0.684 6	0.662 8	0.863 0	0.822 5	0.769 7
	MPFRNet（2023）	0.019 1	0.765 1	0.754 9	0.884 8	0.874 4	0.841 8
	LDNet（2023）	0.028 9	0.656 2	0.639 1	0.865 5	0.841 4	0.767 9
	DATFormer（2023）	0.018 6	0.764 7	0.749 0	0.887 8	0.877 3	0.840 8
	ACoNet（2024）	0.018 1	0.789 3	0.781 5	0.914 1	0.904 3	0.849 3
	DPNet（2025）	0.019 0	0.801 6	0.786 4	0.920 7	0.909 6	0.848 6
	DSANet（2025）	0.019 8	0.784 2	0.770 3	0.908 1	0.898 8	0.846 3
RGB-D 360° SOD	本文	**0.015 1**	**0.838 8**	**0.815 0**	**0.933 1**	**0.924 0**	**0.873 6**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。

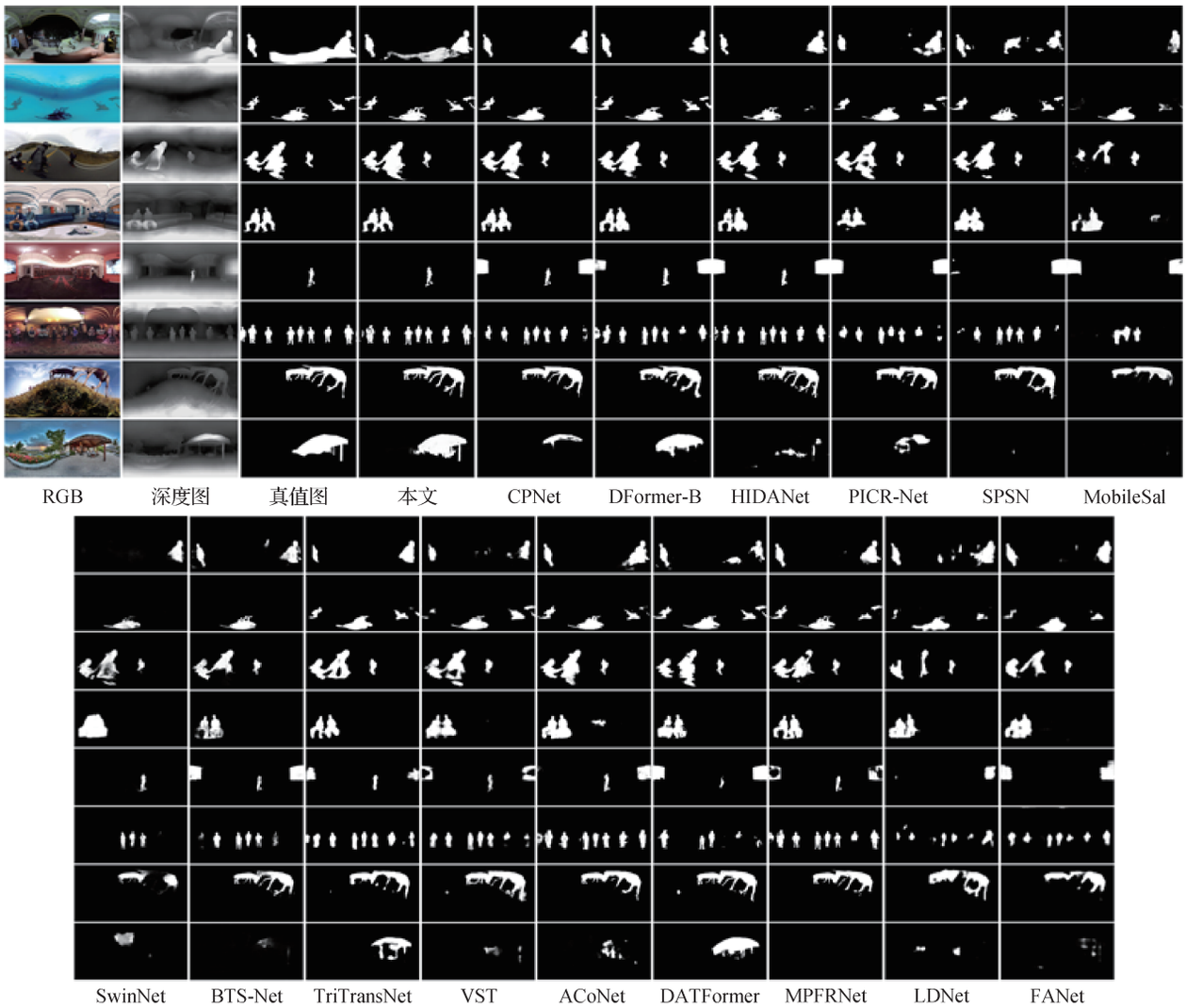


图6　不同方法在360-SOD数据集上的主观结果比较

Fig. 6　Qualitative comparison of different methods on the 360-SOD dataset

码器)、训练配置与数据预处理条件下,对4种主流的360°深度估计方法(Joint_360depth(Yun等,2022)(本文采用的方法)、DA(depth anywhere)(Wang和Liu,2025)、EGFormer(Yun等,2023)和CRF360D(Cao和Wang,2025))分别生成深度图并重新训练模型;同时,以仅使用RGB的w/o depth方案作为对照。表3的结果显示,无论采用哪一种深度估计方法,将深度模态引入模型后,各项指标均显著优于单模态(RGB-only),并且全部优于表1中未使用深度信息的RGB 360° SOD方法。这说明深度特征能够稳定辅助全景显著目标检测,从整体上提升检测精度和边缘完整度。同时值得注意的是,在RGB-only情况下,本文模型的性能仅处于现有RGB 360° SOD方法的中等偏上水平,说明SAM2-Hiera的表征能力并不足以在缺乏深度模态时解决全景投影畸变和结构歧义等核心难题,这证明了本文模型的优越性主要来自于网络结构和各子模块设计,而非由大型预训练骨干网络带来。

3.4.2　全景感知感受野模块PA-RFM的有效性

为验证本文提出的全景感知感受野模块PA-RFM的有效性,设计了3种对比方案:1)with RFB:采用经典感受野模块RFB(Liu等,2018)替换PA-RFM;2)w/o PA-RFM:以1×1卷积替代PA-RFM,即完全去除该模块;3)with PA-RFM:使用本文提出的PA-RFM模块。

客观指标对比如表4所示,加入PA-RFM后各项指标均达到最优,较其他替代方案表现更佳。主观可视化结果如图10所示,进一步验证了该模块的有效性:第2行显示其在极区畸变区域具有更准确的显著性定位;第3行则表明其在处理跨边界连续

表 2　360-SSOD数据集上各模型客观指标对比

Table 2　Comparison of objective metrics of different models on the 360-SSOD dataset

	方法	MAE ↓	max-F ↑	mean-F ↑	max-Em ↑	mean-Em ↑	Sm ↑
RGB-D 2D SOD	VST(2021)	0.028 0	0.618 9	0.590 0	0.851 1	0.797 9	0.758 3
	TriTransNet(2021)	0.028 3	0.626 3	0.622 5	0.857 8	0.837 2	0.745 1
	BTS-Net(2021)	0.034 0	0.619 0	0.564 3	0.828 9	0.755 6	0.744 7
	SwinNet(2022)	0.029 2	0.650 4	0.621 7	0.862 2	0.816 0	0.768 9
	MobileSal(2022)	0.032 4	0.507 8	0.489 1	0.765 2	0.701 6	0.687 3
	DFM-Net(2022)	0.040 7	0.574 7	0.508 4	0.855 5	0.743 3	0.719 0
	SPSN(2022)	0.028 3	0.637 9	0.628 8	0.844 6	0.810 9	0.738 3
	PICR-Net(2023)	0.029 9	0.609 9	0.601 3	0.850 9	0.824 7	0.732 0
	HIDANet(2023)	0.028 1	0.661 3	0.640 9	0.862 3	0.854 7	0.778 2
	DFormer-B(2024)	0.027 3	0.676 2	0.660 3	0.875 8	0.848 2	0.784 6
	CPNet(2024)	0.027 8	0.675 2	0.667 6	**0.877 8**	**0.864 9**	0.774 8
RGB 360° SOD	FANet(2020)	0.041 5	0.561 9	0.533 5	0.802 7	0.770 3	0.711 9
	MPFRNet(2023)	–	–	–	–	–	–
	LDNet(2023)	0.034 2	0.586 2	0.567 2	0.839 0	0.818 7	0.724 5
	DATFormer(2023)	**0.025 2**	0.619 6	0.605 0	0.832 5	0.792 0	0.752 7
	ACoNet(2024)	0.028 8	0.664 1	0.656 4	0.869 5	0.863 2	0.779 6
RGB-D 360° SOD	本文	0.027 3	**0.693 8**	**0.670 6**	0.875 9	0.856 4	**0.787 1**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。“–”表示没有源代码无法测试。

物体时能更好地保持目标完整性。同时对全景上下文聚合模块(PCAM)及其3类注意力机制(经度注意Lon、纬度注意Lat、全局注意Glob)进行了消融实验。以去除PCAM的完整模型(No. 1)为基线，逐步加入不同注意力分支，其结果如表5所示。单独加入任一注意力机制(No. 2—No. 4)均带来显著提升；同时使用两种机制(No. 5—No. 7)可获得更优表现，说明它们具备互补性；3种机制同时使用时(No. 8)，在max-E以外的其他指标达到最佳，证明3类注意力存在协同作用，可有效增强全景特征建模与显著性检测性能。

3.4.3　注意力引导融合模块AFM的有效性

为验证所提出的注意力引导融合模块AFM的有效性，本文在统一网络架构下设计了5组对比实验，结果如表6所示，表中No. 1代表单独使用RGB模态(RGB)；No. 2代表RGB模态与深度模态逐元素相加(add)融合；No. 3代表RGB模态与深度模态通道拼接(concat)融合；No. 4代表RGB模态与深度模态逐元素乘法(multi)融合；No. 5代表使用本文提出的AFM模块。由表6客观指标和图11主观结果对比可以看出，相较于RGB单模态，加入深度模态后，add、concat、multi、AFM这4种双模态模型性能均有显著提升，这充分证明了深度信息对360 RGB-D SOD检测具有重要价值，即便在采用最简单的融合策略时亦能显著提升性能。进一步地，采用本文提出的AFM模块融合RGB与深度特征时，性能在所有指标上均优于其他融合方式，说明AFM能够更充分地挖掘跨模态的互补性，从而获得更高的检测精度与鲁棒性。

为了进一步证明AFM的优越性，将AFM替换为其他SOTA(state-of-the-art)方法的特征融合模块，包括MobileSal(Wu等，2022)、HIDANet(Wu等，2023)和CPNet(Hu等，2024)。如表7所示，AFM模块比其他方法中使用的模块性能更好。

3.4.4　跨模态引导协同解码策略的有效性

为验证所提出的跨模态引导协同解码策略的有效性，将完整模型(本文)与移除RGB解码路径的变体(w/o RGB)进行对比，结果如表8所示。可以看

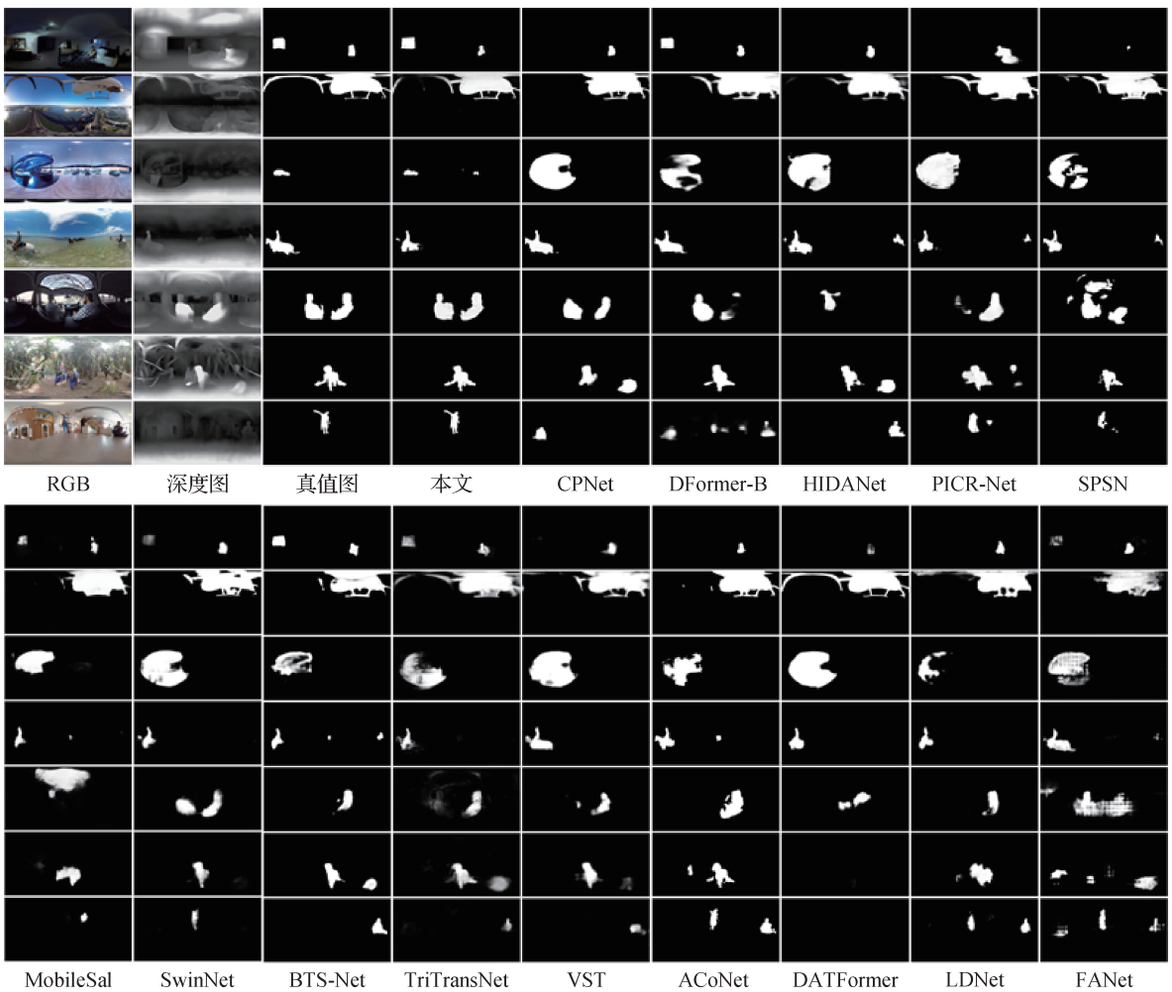


图7　不同方法在360-SSOD数据集上的主观结果比较

Fig. 7　Qualitative comparison of different methods on the 360-SSOD dataset

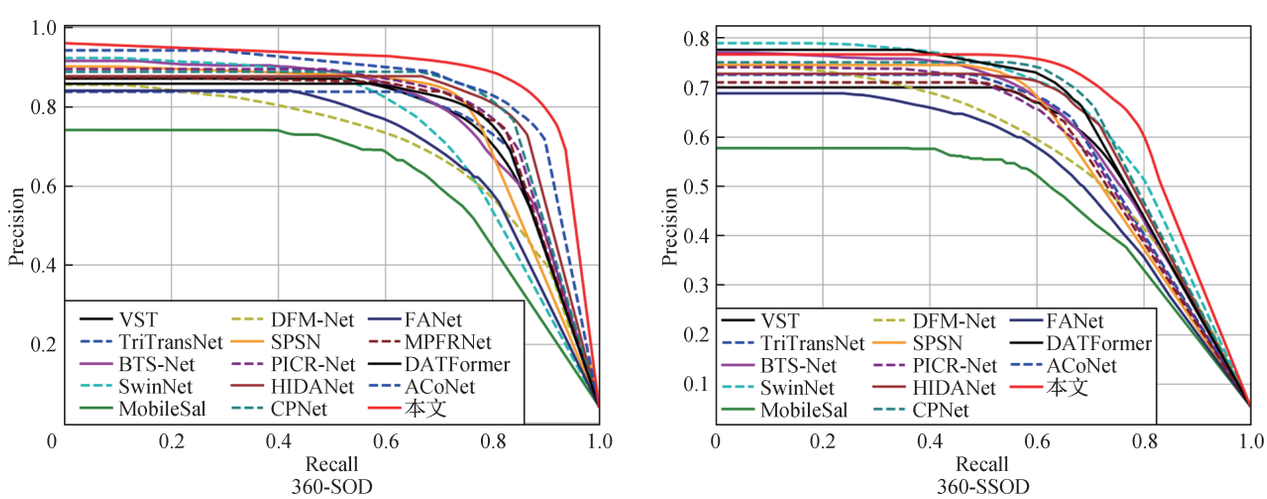


图8　不同模型在360-SOD和360-SSOD数据集上的P-R曲线比较

Fig. 8　Comparison of P-R curves of different models on the 360-SOD and 360-SSOD datasets

到，去掉RGB解码路径后，实验结果显示模型在多项评测指标上均有轻微下降，说明在缺少RGB引导时模型的边缘/细节恢复以及整体结构保持能力均受到影响。且由图12所展示的主观对比可以看出，RGB解码路径能够在融合分支解码过程中提供高分辨率的结构先验和细节补偿，有效缓解融合特征

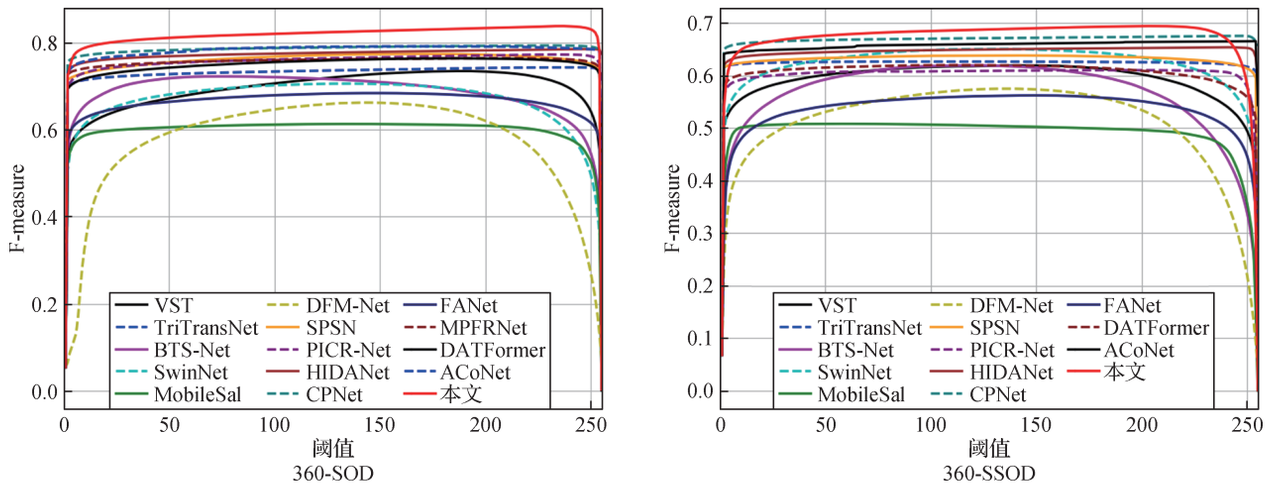


图 9　不同模型在 360-SOD 和 360-SSOD 数据集上的 F-measure 曲线比较

Fig. 9　Comparison of F-measure curves of different models on the 360-SOD and 360-SSOD datasets

表 3　使用不同深度估计方法的客观指标对比

Table 3　Comparison of objective metrics using different depth estimation methods

模型方法	深度估计方法	MAE ↓	max-F ↑	mean-F ↑	max-Em ↑	mean-Em ↑	Sm ↑
本文模型	w/o depth (RGB only)	0.022 0	0.730 4	0.702 1	0.852 6	0.841 4	0.809 6
	DA(2025)	0.016 0	0.830 0	0.806 4	0.921 1	0.915 4	0.873 1
	EGFormer(2023)	0.016 3	0.824 3	0.807 9	0.920 5	0.909 5	0.864 3
	CRF360D(2025)	0.017 1	0.817 9	0.798 0	0.920 4	0.910 1	0.872 5
	Joint_360depth(2022)(本文使用方法)	**0.015 1**	**0.838 8**	**0.815 0**	**0.933 1**	**0.924 0**	**0.873 6**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。w/o 表示不使用。

表 4　全景感知感受野模块 PA-RFM 有效性的客观指标对比

Table 4　Comparison of objective metrics for the effectiveness of the panoramic-aware receptive field module (PA-RFM)

方法	MAE ↓	max-F ↑	mean-F ↑	max-E ↑	mean-E ↑	Sm ↑
with RFB	0.016 2	0.817 3	0.800 9	0.917 7	0.907 3	0.870 3
w/o PA-RFM	0.016 6	0.814 0	0.801 1	0.920 1	0.900 8	0.861 0
with PA-RFM	**0.015 1**	**0.838 8**	**0.815 0**	**0.933 1**	**0.924 0**	**0.873 6**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。with 表示使用，w/o 表示不使用。

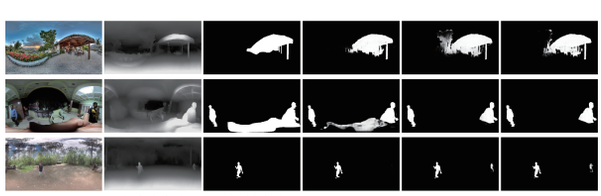


图 10　全景感知感受野模块 PA-RFM 有效性的主观结果对比

Fig. 10　Qualitative comparison demonstrating the effectiveness of the panoramic-aware receptive field module

中 RGB 语义弱化问题，显著提升模型的边缘区域与细节纹理检测能力。

4　结　论

针对 360° SOD 面临的投影畸变、边界不连续及跨模态融合等问题，本文提出了一种 RGB-D 360° SOD 方法。该方法包含 3 个协同工作的核心模块：PA-RFM 模块通过显式编码经纬度信息并结合全局注意力，缓解极区畸变与边缘断裂，增强几何感知；AFM 模块利用动态权重与双重注意力机制，抑制深度噪声，强化 RGB-D 互补表征；跨模态引导协同解

表5　PCAM模块的有效性及3种注意力不同组合的客观指标对比

Table 5　Comparison of objective metrics for the effectiveness of the PCAM module and different combinations of three attention mechanisms

NO.	Lon	Lat	Glob	MAE ↓	max-F ↑	mean-F ↑	max-E ↑	mean-E ↑	Sm ↑
1	–	–	–	0.021 5	0.747 1	0.719 7	0.872 0	0.859 7	0.820 1
2	√	–	–	0.019 3	0.796 6	0.779 2	0.888 3	0.876 3	0.843 1
3	–	√	–	0.018 1	0.801 0	0.783 9	0.914 7	0.908 5	0.856 8
4	–	–	√	0.018 6	0.809 6	0.799 6	0.927 4	0.914 2	0.839 3
5	√	√	–	0.016 6	0.816 4	0.780 0	0.924 8	0.909 3	0.868 3
6	√	–	√	0.017 1	0.829 9	0.801 3	**0.934 0**	0.919 9	0.854 7
7	–	√	√	0.017 4	0.832 7	0.811 4	0.929 3	0.918 4	0.855 3
8	√	√	√	**0.015 1**	**0.838 8**	**0.815 0**	0.933 1	**0.924 0**	**0.873 6**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。“√”和“–”分别表示使用和未使用对应注意力机制。

表6　注意力引导融合模块AFM有效性的客观指标对比

Table 6　Comparison of objective metrics for the effectiveness of the attention-guided fusion module（AFM）

NO.	RGB	depth	融合方法	MAE ↓	max-F ↑	mean-F ↑	max-E ↑	mean-E ↑	Sm ↑
1	√	–	–	0.022 0	0.730 4	0.702 1	0.852 6	0.841 4	0.809 6
2	√	√	add	0.020 8	0.744 5	0.712 8	0.879 9	0.862 4	0.818 5
3	√	√	concat	0.015 7	0.821 7	0.806 6	0.918 5	0.911 4	0.831 4
4	√	√	multi	0.016 6	0.814 5	0.809 5	0.901 9	0.881 4	0.845 0
5	√	√	AFM	**0.015 1**	**0.838 8**	**0.815 0**	**0.933 1**	**0.924 0**	**0.873 6**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。“√”和“–”分别表示使用和未使用。

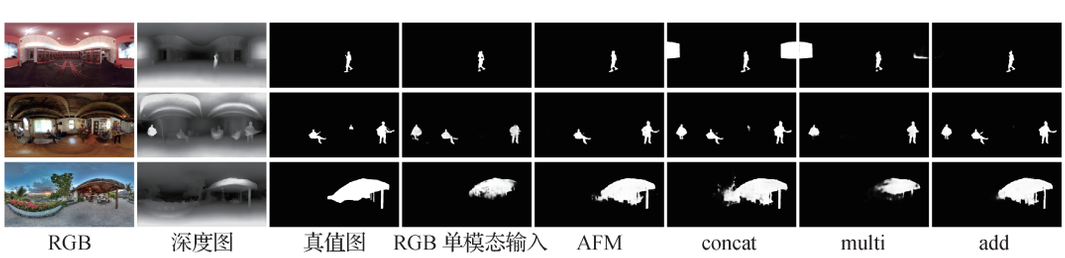


图11　注意力引导融合模块AFM有效性的主观结果对比

Fig. 11　Qualitative comparison demonstrating the effectiveness of the attention-guided fusion module（AFM）

表7　AFM模块与其他SOTA方法的特征融合模块客观指标对比

Table 7　Comparison of objective metrics between the AFM module and feature fusion modules of other SOTA methods

方法	MAE ↓	max-F ↑	mean-F ↑	max-E ↑	mean-E ↑	Sm ↑
MobileSal	0.015 9	0.833 0	0.813 4	0.926 0	0.916 4	0.863 4
HIDANet	0.016 3	0.837 3	0.811 2	0.932 3	0.923 1	0.873 4
CPNet	0.016 5	0.835 5	0.810 0	0.928 2	0.916 3	0.866 3
本文	**0.015 1**	**0.838 8**	**0.815 0**	**0.933 1**	**0.924 0**	**0.873 6**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。

表 8　跨模态引导协同解码策略有效性的客观指标对比

Table 8　Comparison of objective metrics for the effectiveness of the cross-modal guided collaborative decoding strategy

方法	MAE ↓	max-F ↑	mean-F ↑	max-E ↑	mean-E ↑	Sm ↑
w/o RGB	0.016 9	0.832 8	0.805 1	0.925 5	0.915 7	0.872 4
本文	**0.015 1**	**0.838 8**	**0.815 0**	**0.933 1**	**0.924 0**	**0.873 6**

注：加粗字体表示各列最优结果。“↑”表示值越大越好，“↓”表示值越小越好。w/o RGB表示移除RGB解码。

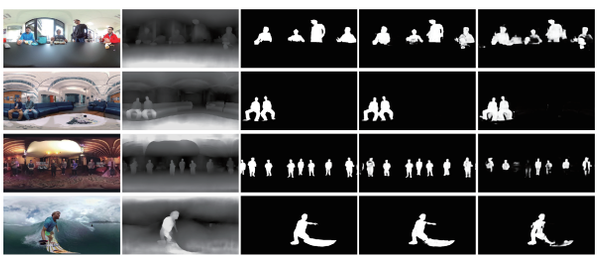


图 12　跨模态引导协同解码策略的有效性主观结果对比

Fig. 12　Qualitative comparison demonstrating the effectiveness of the cross-modal guided collaborative decoding strategy

码策略引入独立RGB路径作为高分辨率先验，在融合解码中有效恢复细节，克服传统方法的细节弱化问题。在两个数据集上的系统性实验与消融结果表明，相比现有代表性先进方法，本文方法在检测精度和鲁棒性方面均取得了显著提升，证明了深度模态在360° SOD中的应用潜力。需要说明的是，使用深度估计方法生成的深度图与使用深度相机拍摄的深度图还有一定的差距，且不同深度估计方法之间的性能存在差异，故本文模型的效果可能会因采用的深度图不同而有所变化。但是从大量实验来看，这种变化在可控范围内，不会造成较大程度的检测性能变化。未来研究将从以下几个方向展开：1）结合真实深度传感器数据开展跨域自适应学习；2）探索多投影视角联合建模，以进一步缓解极区畸变问题；3）研究模型蒸馏与网络压缩策略，实现模型轻量化与移动端友好部署。

参考文献（References）

Achanta R, Hemami S, Estrada F and Susstrunk S. 2009. Frequency-tuned salient region detection//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 1597-1604 [DOI: 10.1109/CVPR.2009.5206596]

Cao Z D and Wang L. 2025. Monocular 360 depth estimation via spherical fully-connected CRFs. IEEE Robotics and Automation Letters, 10(2): 1409-1416 [DOI: 10.1109/LRA.2024.3518109]

Chen X L, Du Z L, Zhang X G and Wang X. 2025a. Distortion-adaptive and position-aware network for salient object detection in 360° omnidirectional image. Journal of Image and Graphics, 30(8): 2758-2774 (陈晓雷, 杜泽龙, 张学功, 王兴. 2025a. 畸变自适应与位置感知的360°全景图像显著目标检测网络. 中国图象图形学报, 30(8): 2758-2774) [DOI: 10.11834/jig.240592]

Chen X L, Wang X, Zhang X G and Du Z L. 2024. Adjacent coordination network for salient object detection in 360 degree omnidirectional images. Journal of Electronics and Information Technology, 46(12): 4529-4541 (陈晓雷, 王兴, 张学功, 杜泽龙. 2024. 面向360度全景图像显著目标检测的相邻协调网络. 电子与信息学报, 46(12): 4529-4541) [DOI: 10.11999/JEIT240502]

Chen X L, Zhang X G, Du Z L and Wang X. 2025b. Distortion semantic aggregation network for salient object detection in 360° omnidirectional images. Journal of Image and Graphics, 30(7): 2451-2467 (陈晓雷, 张学功, 杜泽龙, 王兴. 2025b. 面向360°全景图像显著目标检测的畸变语义聚合网络. 中国图象图形学报, 30(7): 2451-2467) [DOI: 10.11834/jig.240371]

Ciptadi A, Hermans T and Rehg J M. 2013. An in depth view of saliency//Proceedings of 2013 British Machine Vision Conference. Bristol, UK: BMVA Press: 112.1-112.11 [DOI: 10.5244/C.27.76]

Cong R M, Huang K, Lei J J, Zhao Y, Huang Q M and Kwong S. 2023a. Multi-projection fusion and refinement network for salient object detection in 360° omnidirectional image. IEEE Transactions on Neural Networks and Learning Systems, 35(7): 9495-9507 [DOI: 10.1109/TNNLS.2022.3233883]

Cong R M, Lei J J, Fu H Z, Hou J H, Huang Q M and Kwong S. 2019. Going from RGB to RGBD saliency: a depth-guided transformation model. IEEE Transactions on Cybernetics, 50(8): 3627-3639

Cong R M, Liu H Y, Zhang C, Zhang W, Zheng F, Song R, et al. 2023b. 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]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Fan D P, Gong C, Cao Y, Ren B, Cheng M M and Borji A. 2018. Enhanced-alignment measure for binary foreground map evalua-

tion//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: ijcai.org: 698-704 [DOI: 10.24963/ijcai.2018/97]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu X H, Sun F M, Sun J, Wang F S and Li H J. 2024. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. International Journal of Computer Vision, 132(8): 3067-3085 [DOI: 10.1007/s11263-024-02020-y]

Huang M K, Li G Y, Liu Z and Zhu L C. 2023. Lightweight distortion-aware network for salient object detection in omnidirectional images. IEEE Transactions on Circuits and Systems for Video Technology, 33(10): 6191-6197 [DOI: 10.1109/TCSVT.2023.3253685]

Huang M K, Liu Z, Li G Y, Zhou X F and Le Meur O. 2020. FANet: features adaptation network for 360° omnidirectional salient object detection. IEEE Signal Processing Letters, 27: 1819-1823 [DOI: 10.1109/LSP.2020.3028192]

Lang C Y, Nguyen T V, Katti H, Yadati K, Kankanhalli M and Yan S C. 2012. Depth matters: influence of depth cues on visual saliency//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 101-115 [DOI: 10.1007/978-3-642-33709-3_8]

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. Tel Aviv, Israel: Springer: 630-647 [DOI: 10.1007/978-3-031-19818-2_36]

Li J, Su J M, Xia C Q and Tian Y H. 2020. Distortion-adaptive salient object detection in 360° omnidirectional images. IEEE Journal of Selected Topics in Signal Processing, 14(1): 38-48 [DOI: 10.1109/JSTSP.2019.2957982]

Liu N, Zhang N, Wan K Y, Shao L and Han J W. 2021a. Visual saliency transformer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4702-4712 [DOI: 10.1109/ICCV48922.2021.00468]

Liu S T, Huang D and Wang Y H. 2018. Receptive field block net for accurate and fast object detection//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 404-419 [DOI: 10.1007/978-3-030-01252-6_24]

Liu Z Y, Tan Y C, He Q and Xiao Y. 2022. SwinNet: swin transformer drives edge-aware RGB-D and RGB-T salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 32(7): 4486-4497 [DOI: 10.1109/TCSVT.2021.3127149]

Liu Z Y, Wang Y, Tu Z Z, Xiao Y and Tang B. 2021c. TriTransNet: RGB-D salient object detection with a triplet transformer embedding network//Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM: 4481-4490 [DOI: 10.1145/3474085.3475601]

Perazzi F, Krähenbühl P, Pritch Y and Hornung A. 2012. Saliency filters: contrast based filtering for salient region detection//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 733-740 [DOI: 10.1109/CVPR.2012.6247743]

Ravi N, Gabeur V, Hu Y T, Hu R H, Ryali C, Ma T Y, et al. 2025. SAM 2: segment anything in images and videos//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Ren J Q, Gong X J, Yu L, Zhou W H 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. Boston, USA: IEEE: 25-32 [DOI: 10.1109/CVPRW.2015.7301391]

Ryali C, Hu Y T, Bolya D, Wei C, Fan H Q, Huang P Y, et al. 2023. Hiera: a hierarchical vision transformer without the bells-and-whistles//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: #1224

Wang N H and Liu Y L. 2025. Depth anywhere: enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #4056 [DOI: 10.5555/3737916.3741972]

Wu Y H, Liu Y, Xu J, Bian J W, Gu Y C and Cheng M M. 2022. MobileSal: extremely efficient RGB-D salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 10261-10269 [DOI: 10.1109/TPAMI.2021.3134684]

Wu Z W, Allibert G, Meriaudeau F, Ma C and Demonceaux C. 2023. HiDAnet: RGB-D salient object detection via hierarchical depth awareness. IEEE Transactions on Image Processing, 32: 2160-2173 [DOI: 10.1109/TIP.2023.3263111]

Xiong X Y, Wu Z H, Tan S Y, Li W X, Tang F L, Chen Y, et al. 2026. SAM2-UNet: segment anything 2 makes strong encoder for natural and medical image segmentation. Visual Intelligence, 4(1): #2 [DOI: 10.1007/s44267-025-00106-w]

Yin B W, Zhang X Y, Li Z Y, Liu L, Cheng M M and Hou Q B. 2023. Dformer: rethinking RGBD representation learning for semantic segmentation//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Yun I, Lee H J and Rhee C E. 2022. Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press: 3224-3233 [DOI: 10.1609/aaai.v36i3.20231]

Yun I, Shin C Y, Lee H, Lee H J and Rhee C E. 2023. EGFormer: equirectangular geometry-biased transformer for 360 depth estimation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 6078-6089 [DOI: 10.

1109/ICCV51070.2023.00561］

Zhang W B，Jiang Y，Fu K R and Zhao Q J. 2021. BTS-Net：bi-directional transfer-and-selection network for RGB-D salient object detection//Proceedings of 2021 IEEE International Conference on Multimedia and Expo. Shenzhen，China：IEEE：1-6［DOI：10.1109/ICME51207.2021.9428263］

Zhao Y J，Zhao L C，Yu Q，Sheng L，Zhang J and Xu D. 2023. Distortion-aware transformer in 360° salient object detection//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa，Canada：ACM：499-508［DOI：10.1145/3581783.3612025］

Ma G X，Li S，Chen C，Hao A M and Qin H. 2020. Stage-wise salient object detection in 360 omnidirectional image via object-level semantical saliency ranking. IEEE Transactions on Visualization and Computer Graphics，26（12）：3535-3545［DOI：10.1109/TVCG.2020.3023636］

作者简介

陈晓雷，男，教授，硕士生导师，主要研究方向为人工智能与计算机视觉。E-mail：chenxl703@lut.edu.cn

钟智华，男，硕士研究生，主要研究方向为全景图像显著性目标检测。E-mail：1603524098@qq.com

申玉杰，女，硕士研究生，主要研究方向为全景图像显著性目标检测。E-mail：1455922810@qq.com