中图法分类号:TP18; TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)02-0479-20

论文引用格式:Jia D, Zhao C, Zhang H X and Song H L. 2026. Lightweight RGB-D semantic segmentation network incorporating frequency-domain guidance. Journal of Image and Graphics, 31(2):0479-0498(贾迪, 赵辰, 张华修, 宋慧伦. 2026. 融合频域引导的RGB-D轻量级语义分割网络. 中国图象图形学报, 31(2):0479-0498)[DOI:10.11834/jig.250212]

融合频域引导的RGB-D轻量级语义分割网络

贾迪[1,2],赵辰[2*],张华修[2],宋慧伦[2]

1. 辽宁工程技术大学鄂尔多斯研究院,鄂尔多斯　017000; 2. 辽宁工程技术大学电子与信息工程学院,葫芦岛　125105

摘　要:**目的**　RGB-D语义分割通过融合RGB图像与深度图像的互补信息,能够有效提升复杂场景下的语义理解精度。然而,由于RGB与深度模态在特征表达上的本质差异,现有方法常依赖双编码器架构,导致特征对齐效率低、计算开销大,难以在模型性能与轻量化之间取得良好平衡。为此,提出一种融合频域引导的RGB-D轻量级语义分割网络,旨在提升跨模态特征融合效果并降低模型复杂度。**方法**　构建融合Transformer与卷积神经网络(convolutional neural network,CNN)结构的RGB-D Block,通过全局注意力分支与局部卷积分支并行设计,兼顾长距离依赖建模与细节提取;设计频域引导提示适配器,通过傅里叶变换提取频谱能量以动态生成提示向量,并与查询特征融合以加强跨层语义对齐能力;设计频谱引导动态卷积模块,将特征划分为低、中、高频3个频段,通过门控机制融合多频特征与空间域提取特征,提升多尺度建模能力;构建多尺度频域代理注意力模块,利用上阶段的频谱特征生成动态代理向量,在低成本下实现高效的跨层全局建模。**结果**　在RGB-D语义分割任务下的NYU Depth V2(New York University depth dataset V2)和SUN-RGBD(RGB-D scene understanding benchmark suite)数据集上,以不足主流方法一半的参数量分别实现了57.6%和52.8%的平均交并比;在RGB-L语义分割任务下的KITTI-360数据集上的实验进一步验证了对其他跨模态场景的鲁棒性。泛化测试中,模型在5个RGB-D显著目标检测数据集上于F-measure、E-measure、S-measure和平均绝对误差(mean absolute error,MAE)4项指标全面优于对比方法。此外,在NYU Depth V2上进行的消融实验也验证了各模块的独立价值与协同效益。**结论**　提出一种频域引导的轻量级RGB-D感知框架,在多个关键模块中引入频域引导机制,统一构建频谱感知的提示学习、动态卷积与代理注意力路径,显著提升了RGB-D语义分割任务中的语义对齐能力、信息融合效率与模型部署适应性,为轻量化多模态感知网络的设计提供了新思路。

关键词:RGB-D图像;RGB-D语义分割;频域建模;提示调优;动态卷积;代理注意力

Lightweight RGB-D semantic segmentation network incorporating frequency-domain guidance

Jia Di[1,2], Zhao Chen[2*], Zhang Huaxiu[2], Song Huilun[2]

1. Ordos Institute, Liaoning Technical University, Ordos　017000, China; 2. School of Electronic and Information Engineering, Liaoning Technical University, Huludao　125105, China

Abstract: **Objective** Semantic segmentation has emerged as a fundamental task in computer vision, aiming to assign a

收稿日期:2025-05-15;**修回日期**:2025-08-15;**预印本日期**:2025-08-23

*通信作者:赵辰　lntu_zhaochen@163.com

基金项目:国家自然科学基金项目(61601213);内蒙古自治区重点研发和成果转化计划项目(2025YFHH0047);辽宁省教育厅重点资助项目(LJ212410147003);辽宁工程技术大学鄂尔多斯研究院校地科技合作培育项目(YJY-XD-2023-003)

Supported by: National Natural Science Foundation of China(61601213); Science and Technology Plan Project of Inner Mongolia Autonomous Region(2025YFHH0047); Key Project of Liaoning Provincial Department of Education(LJ212410147003); University-Local Government Scientific and Technical Cooperation Cultivation Project of Ordos Institute-LNTU(YJY-XD-2023-003)

semantic label to each pixel for fine-grained scene understanding. With the increasing demand for intelligent systems such as autonomous driving, indoor robotics, and augmented reality, the ability to accurately parse complex visual scenes has become crucial. While RGB images provide abundant texture and color information, they often fail under challenging conditions such as poor illumination, occlusion, or cluttered environments. Depth maps capture geometric and structural cues that complement RGB images, making RGB-D semantic segmentation an attractive solution. However, integrating these two modalities is difficult because of their intrinsic representation differences. Mainstream approaches often adopt dual-encoder architectures, which separately extract modality-specific features before fusion. Although effective in certain cases, these models suffer from high computational overhead, redundant parameters, and suboptimal feature alignment. The challenge, therefore, lies in designing a framework that achieves a strong balance between segmentation accuracy and model efficiency. To address these limitations, this study introduces a frequency-guided lightweight RGB-D semantic segmentation network, which incorporates frequency-domain modeling into multiple key components. By doing so, the proposed method not only enhances cross-modal semantic alignment but also ensures adaptability for deployment in resource-constrained scenarios. **Method** The proposed framework is constructed around three novel modules, each designed to tackle a specific limitation of conventional RGB-D segmentation. First, we introduce a frequency-guided prompt adapter (FPA). Spectral energy distributions are extracted and transformed into dynamic prompt vectors by applying a Fourier transform to the input features. These prompts are fused with query features to encourage consistent semantic alignment across network layers. Unlike static fusion methods, FPA adaptively strengthens cross-layer information propagation and enhances semantic coherence between RGB and depth modalities. Second, we propose a spectrum-guided dynamic convolution (SDC) module, which explicitly leverages frequency-domain decomposition. Input features are divided into low-, mid-, and high-frequency bands, corresponding to global structures, boundary details, and fine-grained textures, respectively. These frequency-specific features are then integrated with spatial-domain features through a gating mechanism. This design simultaneously addresses local detail preservation and global context modeling, enabling robust multiscale representation learning. Third, we design a multiscale frequency-aware agent attention (MFAA) module. Traditional attention mechanisms, though powerful, incur high computational cost when applied to dense feature maps. MFAA alleviates this issue by dynamically generating compact agent vectors from spectral features of the previous stage. These vectors serve as efficient proxies to capture global dependencies while preserving contextual consistency across scales. In this way, MFAA reduces redundancy while strengthening long-range semantic modeling. The overall architecture adopts an encoder-decoder structure, in which Transformer and CNN branches are jointly employed in the encoder, and the frequency-guided modules bridge the gap between local precision and global reasoning. The decoder reconstructs pixel-level predictions while preserving cross-modal semantic alignment. **Result** Extensive experiments were conducted on multiple benchmark datasets to validate the effectiveness of the proposed framework. On the NYU Depth V2 and SUN-RGBD datasets, the network achieved 57.6% and 52.8% mIoU, respectively, using less than half the parameters of most state-of-the-art models. These results highlighted its superior efficiency-accuracy tradeoff. Beyond RGB-D tasks, we evaluated the network on the KITTI-360 dataset for RGB-L (RGB + LiDAR) semantic segmentation, where it demonstrated competitive performance with a mean IoU of 66.3%, surpassing recent baselines such as DFormer while maintaining a lightweight design. Generalization ability was tested on five RGB-D salient object detection datasets: NJU2K, NLPR, SIP, STERE, and DES. Across all four standard metrics (F-measure, E-measure, S-measure, and MAE), the proposed network consistently outperformed existing methods, achieving leading scores even under challenging conditions such as occlusion, reflective surfaces, or fine-grained object boundaries. Ablation studies on NYU Depth V2 confirmed the independent contributions and synergistic effects of each module: FPA improved cross-layer semantic propagation, SDC enhanced frequency-specific feature modeling, and MFAA strengthened multiscale global interactions. Importantly, the model achieved 46 frame/s inference speed on a single NVIDIA RTX 3090 GPU, making it suitable for real-time or resource-limited applications. **Conclusion** This paper presents a novel frequency-guided lightweight RGB-D perception framework that systematically integrates frequency-domain modeling into prompt learning, dynamic convolution, and agent attention. The proposed design addresses key challenges of RGB-D semantic segmentation, including modality misalignment, computational inefficiency, and insufficient global reasoning. By harmonizing spectral and spatial cues, the network achieves high segmentation accuracy with significantly reduced

parameters and computational costs. Experimental results across diverse datasets and tasks demonstrate not only its state-of-the-art performance but also its strong generalization to unseen modalities and scenarios. The contributions of this work are threefold: 1) an FPA that promotes consistent semantic alignment across layers, 2) an SDC module that enhances multi-scale modeling through explicit frequency decomposition, and 3) an MFAA mechanism that efficiently captures global context at low cost. Collectively, these innovations form a compact yet powerful framework for RGB-D perception. Beyond segmentation, the proposed methodology provides insights for designing lightweight multimodal neural networks that achieve an excellent trade-off between accuracy and efficiency, offering a reference for future research on deployable perception models.

Key words: RGB-D image; RGB-D semantic segmentation; frequency-domain modeling; prompt tuning dynamic convolution; agent attention

0 引　言

语义分割作为计算机视觉中的核心任务之一，旨在为图像中每一个像素赋予语义标签，从而实现对场景的精细解析(Sung等，2024)。随着深度学习的迅猛发展，语义分割技术已在自动驾驶、智能监控和医学影像分析等实际应用中展现出巨大潜力与广阔前景(Su等，2024)。

在复杂的室内场景中，由于存在低光照、遮挡频繁及物体形态多样等挑战，单纯依赖RGB图像进行语义分割往往难以捕捉深层结构与几何信息，导致分割精度显著下降(Minaee等，2022)。为解决上述问题，研究者逐步引入辅助模态信息融合策略，如融合RGB与深度图的RGB-D方法(赵经阳 等，2022)以及融合RGB与激光雷达的RGB-L(RGB+LiDAR)方法(贾迪 等，2022)。这些方法通过挖掘不同模态之间的互补优势，弥补了RGB单模态在表达能力上的不足。其中，RGB-D语义分割通过联合视觉与几何信息，显著增强了模型对细粒度特征的建模能力，已成为当前多模态语义分割研究的主要方向(Guo等，2024)。

早期RGB-D方法通常将深度图与RGB图像直接拼接为四通道输入送入浅层卷积网络。Couprie等人(2013)验证了多模态输入的可行性，但未考虑RGB与深度图在特征分布上的本质差异。为提升深度模态的表达能力，Gupta等人(2014)将单通道深度图转换为三通道的几何表征，通过HHA(horizontal disparity, height above ground, and angle with gravity)编码提升了分割性能，然而在训练数据有限的情况下，模型可能更倾向于学习RGB图像中的特征模式，而忽视深度信息的独特性。FuseNet(fuse network)(Hazirbas等，2017)提出双分支编码器—解码器结构，分别提取RGB与深度模态的层次特征，并在多层进行融合，从而更有效地整合模态间的互补信息。MM-RNNs(multimodal recurrent neural networks)(Fan等，2017b)则通过扩展传统RNN至多模态输入，利用共享“记忆”机制动态融合颜色与深度特征，增强了对物体类别和几何关系的区分能力。3DGNN(3D graph neural network)(Qi 等，2017)将RGB-D像素映射到三维空间构建K近邻图结构，通过图神经网络聚合局部几何与语义信息，显著提升了室内场景下的分割表现。然而，这些方法多依赖于局部感受野，难以建模全局上下文，且对非结构化数据的适应能力有限。

近年来，为克服传统方法在全局建模方面的局限，注意力机制被引入计算机视觉领域，并迅速成为研究热点。例如，ACNet(attention based network)(Hu 等，2019)采用双ResNet(residual network)分支分别提取RGB和深度图像特征，并通过引入注意力互补模块将其融合至第3个分支中，依据不同层级的特征信息动态调节通道权重，从而增强网络对关键区域的响应能力。CMX(cross-modal network for RGB-X)(Zhang等，2023a)基于Transformer架构，提出跨模态特征校正与融合模块，在通道和空间维度上对另一模态特征进行校准，并通过交叉注意力机制实现有效融合。PGDENet(progressive guided fusion and depth enhancement network)(Zhou 等，2023)提出逐步引导的融合策略，并设计了针对几何结构优化的深度增强模块。CMFormer(content-enhanced mask Transformer)(Bi 等，2024)引入内容增强的掩码注意力机制与多分辨率特征融合策略，从原始图像及其下采样版本中学习掩码查询，捕捉丰富的上下文信息，提升模型的语义理解能力。

TransD-Fusion(Wu等,2024)利用自注意力机制对单模态特征进行优化,并引入语义感知的位置编码,通过空间约束局部化注意力的计算范围,在保持结构连续性的同时减少冗余计算。CMPFFNet(cross-modal and progressive feature fusion network)(Zhou等,2024)设计基于加性注意的多模态对齐机制与点积注意力融合策略,有效提升了融合质量并抑制噪声干扰。尽管多数方法聚焦于分割精度的提升,网络的轻量化设计与推理效率也日益受到重视。例如,AsymFormer(Du等,2024)融合了Transformer与卷积神经网络(convolutional neural network,CNN)的非对称混合结构,在保证性能的同时有效降低了计算负担。Sigma(Wan等,2025)基于Mamba架构构建了双分支网络,能够捕捉长距离依赖关系并增强模态间的信息交互。然而,当前大多数双编码器结构仍面临显著的冗余计算与推理延迟,尽管这些方法在提升性能方面取得了显著进展,但普遍依赖于多编码器架构和大规模注意力计算,难以在模型性能与计算效率之间取得理想的平衡。

现有RGB-D语义分割方法在网络结构设计上通常可分为单分支、双分支与三分支架构。单分支模型(Zhang等,2023c)将深度信息作为RGB分支的辅助输入,融合于统一路径中;双分支模型(Chen等,2020)分别对RGB与深度模态进行特征提取,并通过交互模块实现信息整合;三分支模型(Zhou等,2021)则引入额外的融合通路,以增强模态间的协同能力,但同时也带来了更高的模型复杂度。DFormer(Yin等,2024)尝试通过联合RGB-D预训练策略,在保持性能的前提下提升推理效率,然而如何在不显著牺牲精度的情况下进一步压缩模型,仍是一个亟待解决的问题。当前主流方法普遍依赖于空间域卷积或Transformer模块分别提取RGB与深度特征,随后在浅层或解码阶段进行融合。这种处理方式在特征整合路径上缺乏全局建模能力,难以捕捉远距离的语义关联,导致模态间语义对齐不足,影响整体感知一致性。

为解决上述问题,本文提出一种面向RGB-D语义分割的频域引导轻量感知网络,从频域建模视角出发,构建端到端的多模态融合机制,以实现更高效、更鲁棒的语义表示能力。为全面验证所提方法的有效性,在两个核心任务上开展实验,包括RGB-D语义分割与RGB-L语义分割。此外,为评估模型的跨任务泛化能力,在RGB-D显著目标检测任务中进行扩展测试,进一步验证其可迁移性与适应性。本文主要贡献如下:1)提出一种频域引导提示适配器(frequency-guided prompt adapter,FPA),通过对输入特征进行傅里叶变换提取频谱能量,引导提示向量的动态生成过程,从而实现跨层语义感知与提示调优的深度融合。2)设计一种频谱引导动态卷积模块(spectrum-guided dynamic convolution,SDC),将特征划分为低频、中频与高频3个频段,分别建模全局结构、边缘信息与细节纹理,并通过门控机制融合多频段与空间域特征,有效提升多尺度建模能力。3)构建多尺度频域代理注意力模块(multi-scale frequency-aware agent attention,MFAA),基于前一阶段频谱特征动态生成代理向量,构建高效的跨层注意力通路,在提升语义一致性与上下文建模能力的同时降低计算成本。如图1所示,本文方法在多个基准数据集上取得了先进的性能,图中纵轴表示平均交并比(mean intersection over union,mIoU),横轴表示模型参数量。

1 相关工作

1.1 RGB-D显著目标检测

RGB-D显著目标检测(salient object detection,SOD)旨在联合利用RGB图像与深度图信息,精确识别图像中具有显著性的区域。近年来,提出了大量方法以提升该任务的检测性能(叶欣悦 等,2024)。例如,CIRNet(cross-modality interaction and refinement network)(Cong等,2022)引入多层次引导注意力机制,结合融合与细化模块,有效增强了多模态特征的协同表达能力。HiDANet(hierarchical depth awareness network)(Wu等,2023)利用分层的深度感知先验机制,促进了多粒度特征的建模与交互。最新的MambaSOD(Zhan等,2025)采用基于Mamba结构的双流网络设计,能够捕捉全局依赖关系,实现更为精细的显著区域定位。尽管上述方法在精度方面不断取得突破,多数RGB-D SOD框架仍存在参数规模庞大、推理速度缓慢等问题。为此,本文设计一种结构简洁、高效的频域感知模块组合,能够在保持检测精度的同时显著降低计算开销,较好地实现了精度与效率之间的均衡。

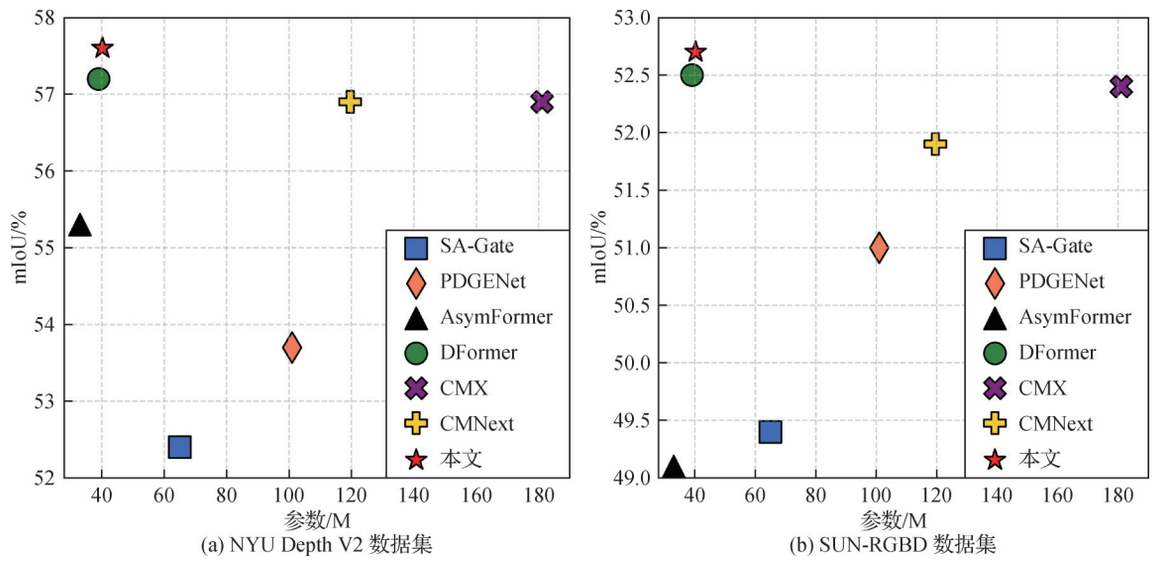


图 1　可视化对比实验

Fig. 1　Visual comparison experiment ((a) NYU Depth V2 dataset; (b) SUN-RGBD dataset)

1.2　注意力机制

注意力机制作为提升特征建模能力的重要手段，已在RGB-D语义分割与显著目标检测等多模态任务中得到广泛应用。其中，softmax注意力机制(Vaswani等，2017)通过计算特征之间的全局相关性，有效建模长距离依赖关系，显著提升了语义一致性，然而该注意力在高分辨率特征图上计算和内存开销极大。为降低计算成本，线性注意力机制(Katharopoulos等，2020)通过引入核函数对原始相关性进行近似，将计算复杂度由平方量级降至线性量级，从而提升了大规模场景下的推理效率，但其在建模细粒度语义关系方面的能力有所下降。为进一步优化注意力效率，代理注意力机制(Han等，2025)提出引入可学习的少量代理向量来参与注意力计算，有效减少了冗余权重的计算开销。现有方法普遍忽视了频域特征在注意力建模中的潜力，且在多尺度建模方面仍存在一定局限性。

2　方　法

2.1　框架概述

图2给出了本文网络的总体架构。整体结构遵循典型的编码器—解码器设计，其中编码器融合了Transformer与CNN模块，通过RGB-D Block对RGB图像和深度图进行联合特征提取与融合；多尺度频域代理注意力模块进行特征优化；解码器则用于将编码特征恢复至像素级语义预测。

RGB图像与深度图首先通过嵌入层进行初始特征提取与降维。该嵌入层由两个阶段的混合模块构成：第1阶段采用多分支结构以增强模态融合能力；第2阶段则通过残差增强提升特征表达的稳健性。对于输入张量$\boldsymbol{X}$，该混合块的实现为

$$\begin{aligned}\boldsymbol{Y} &= GELU\left(BN\left(Conv_{3\times3}(\boldsymbol{I})\right)+BN\left(Conv_{1\times1}(\boldsymbol{I})\right)\right)\\ \boldsymbol{Z} &= BN\left(Conv_{3\times3}(\boldsymbol{Y})\right)+\boldsymbol{Y}\end{aligned}\tag{1}$$

式中，$Conv_{k\times k}$为核大小为k的卷积操作，BN为批归一化层，$\boldsymbol{Z}$为编码后特征，进入分层编码器。

经过嵌入层的处理后，RGB特征和深度特征进入分层编码器，在原始图像的1/4，1/8，1/16，1/32的尺度下进行编码。分层编码器每个stage包含多个RGB-D块，RGB特征和深度特征分别执行跨模态全局建模与频谱引导卷积特征提取。同时，频域引导提示适配器基于傅里叶频谱能量动态生成提示向量，并与主干注意力融合，从而提升跨层语义一致性。此外，在编码器与解码器之间引入的多尺度频域代理注意力模块，能够基于前一阶段的频谱特征生成代理向量，引导解码器获取更具上下文一致性的融合特征。最后，整合后的特征进入解码器进行多尺度重构，生成最终的语义预测结果。

2.2　RGB-D信息处理块

针对RGB与深度模态在特征表达中的互补性与不一致性，在编码器中设计一种结构高效的RGB-D

Block，用于实现多模态特征的融合与校准，其结构如图3所示。该模块由两个并行子分支组成：全局注意力分支与CNN分支，分别负责跨模态全局建模与局部细节建模。

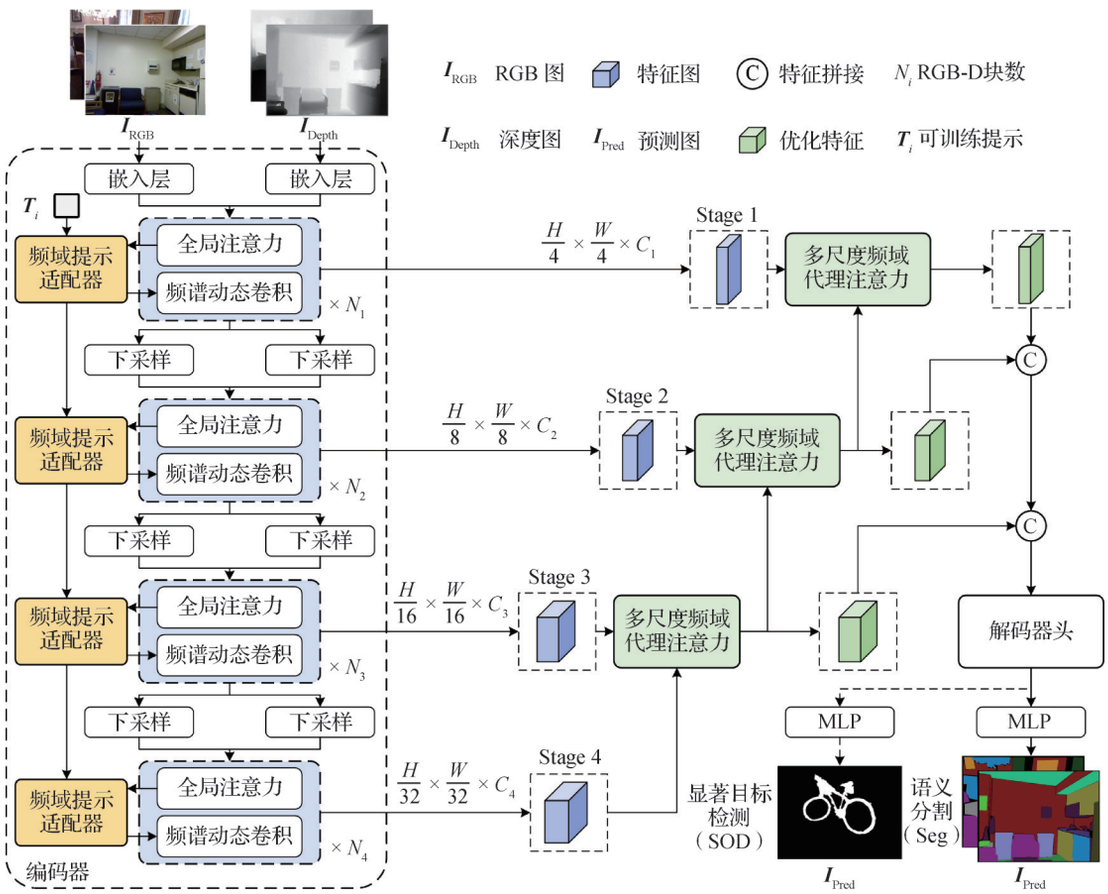


图2　总体网络架构

Fig. 2　Overall network architecture

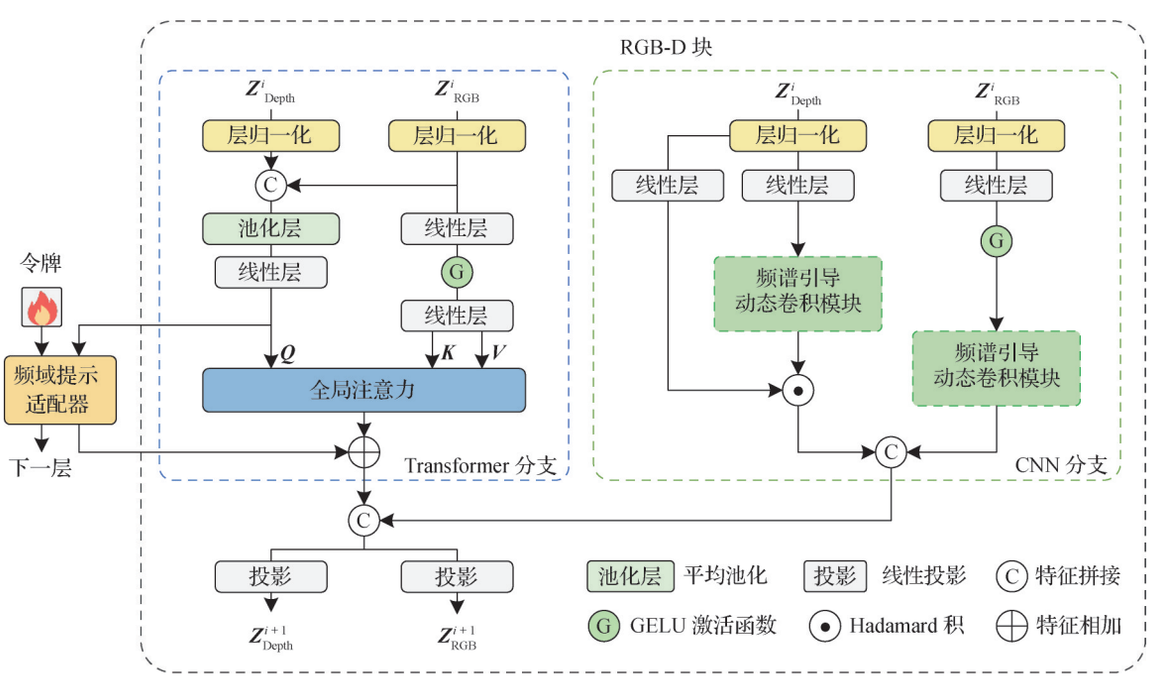


图3　RGB-D块架构

Fig. 3　RGB-D block architecture

全局注意力分支采用频域引导提示适配器，如图4所示，通过提示引导机制增强跨层信息传递能力，同时借助全局注意力模块建模长距离语义依赖。CNN分支则采用频谱引导动态卷积，分解输入特征为低频、中频和高频3个通道，分别提取不同频段的信息，再通过门控机制与空间特征融合，以实现多尺度精细建模。两个分支的输出在通道维度拼接，并通过线性变换生成更新后的RGB特征与深度特征，输入至下一阶段。

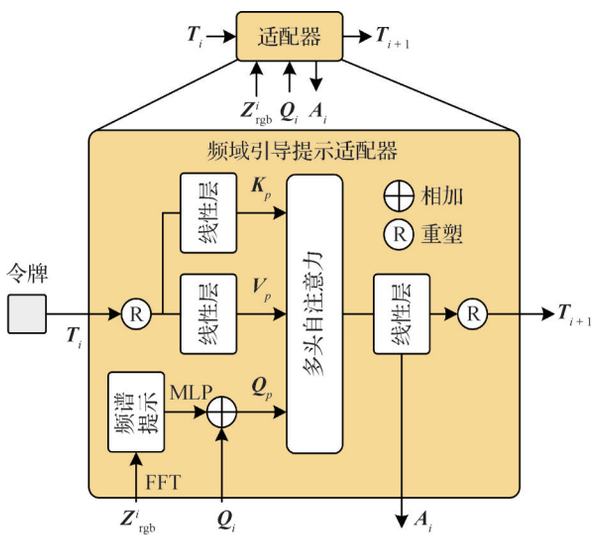


图4　频域引导提示适配器架构

Fig. 4　Frequency-guided prompt adapter

本模型的编码器由4个阶段组成，每个阶段包含 $N \in \{3, 3, 12, 3\}$ 个RGB-D块。给定一对RGB-D输入图像，经过上文中的嵌入层，获得特征 $\boldsymbol{Z}^0_{\mathrm{RGB}}, \boldsymbol{Z}^0_{\mathrm{Depth}}$。对于全局注意力分支，借鉴DFormer方法，将深度特征只与查询 $\boldsymbol{Q}$ 进行融合，而键值 $\boldsymbol{K}$、$\boldsymbol{V}$ 从RGB特征中进行提取。对于第 i 个RGB-D块，给定RGB特征 $\boldsymbol{Z}^i_{\mathrm{RGB}}$ 和深度特征 $\boldsymbol{Z}^i_{\mathrm{Depth}}$，生成查询 $\boldsymbol{Q}$ 的过程可以表述为

$$\boldsymbol{Q} = L(Pool(LN([\boldsymbol{Z}^i_{\mathrm{RGB}}, \boldsymbol{Z}^i_{\mathrm{Depth}}]))) \tag{2}$$

式中，$[\cdot,\cdot]$ 表示通道维拼接，$LN(\cdot)$ 表示层归一化，$Pool(\cdot)$ 表示平均池化，$L(\cdot)$ 表示线性变换。键 K 和值 V 则从RGB特征中生成，具体为

$$\boldsymbol{K}, \boldsymbol{V} = L(GELU(\boldsymbol{Z}^i_{\mathrm{RGB}})) \tag{3}$$

式中，$GELU$ 表示高斯误差线性单元激活函数。经过上面的处理，全局注意力计算为

$$\boldsymbol{A}_{\mathrm{main}} = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{C_d}}\right) \cdot \boldsymbol{V} \tag{4}$$

式中，C_d 为查询向量维度。

对于注意力分支，提示令牌、RGB和深度特征作为分支的输入，经过动态生成的键值对序列计算互补注意力，并通过维度重组和双线性上采样，将结果进行插值，以进行空间恢复与输出，最终结果的简要流程为

$$\begin{aligned} \boldsymbol{Z}^i_{\mathrm{cnn}} &= [\boldsymbol{Z}^i_{\mathrm{Depth}} \odot D(\boldsymbol{Z}^i_{\mathrm{Depth}}), D(\boldsymbol{Z}^i_{\mathrm{RGB}})] \\ \boldsymbol{Z}^i_{\mathrm{final}} &= [\boldsymbol{Z}^i_{\mathrm{cnn}}, Up(\boldsymbol{A}_{\mathrm{total}})] \end{aligned} \tag{5}$$

式中，$D(\cdot)$ 代表频谱引导动态卷积模块。

深度特征经过频谱引导动态卷积模块并跳跃连接进行Hadamard积，与经过频谱引导动态卷积模块的RGB特征进行拼接。全局注意力分支与CNN分支的结果按通道维度进行拼接成 $\boldsymbol{Z}^i_{\mathrm{final}}$，经线性投影产生新的RGB特征 $\boldsymbol{Z}^{i+1}_{\mathrm{RGB}}$ 和深度特征 $\boldsymbol{Z}^{i+1}_{\mathrm{Depth}}$，并送入下个RGB-D块处理。

2.3　频谱引导动态卷积

在多模态感知任务中，局部细节建模和全局结构感知同等重要。为此，设计一种频谱引导动态卷积模块，如图5所示，通过联合空间域与频域信息建模，充分捕获不同尺度、不同频率下的关键信息，同时控制计算复杂度，改善传统空间域卷积操作的局部感受野限制，从而提升下游语义分割任务的整体感知能力。

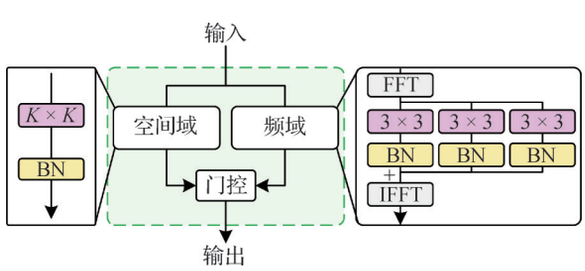


图5　频谱引导动态卷积

Fig. 5　Spectrum-guided dynamic convolution

具体而言，频谱引导动态卷积模块包括空间域建模分支与频域建模分支两部分。空间域分支专注于局部邻域信息的提取与局部上下文建模，而频域分支则面向不同频率成分的全局特征表达与多尺度融合。两条分支各自独立处理输入特征，在输出阶段进行逐元素融合，以兼顾局部细节与全局结构特征。

空间分支采用 $k \times k$ 大卷积核，在原始空间域直接建模局部细节与中尺度上下文关系。通过较大的卷积核扩大感受野，减少深层堆叠对特征传递的依

赖，提升局部纹理建模能力。输入特征$\boldsymbol{Z}_{\text{in}}$经空间卷积分支处理，得到空间特征响应。具体为

$$\boldsymbol{Y}_{\text{spatial}} = Conv_{k\times k}(\boldsymbol{Z}_{\text{in}}) \tag{6}$$

式中，k表示卷积核大小。频域分支通过对输入特征施加二维傅里叶变换，将其从空间域映射到频域，提取对应的频谱特征。具体为

$$\boldsymbol{F} = FFT(\boldsymbol{Z}_{\text{in}}) \tag{7}$$

在频域中，不同频率成分对应着图像中的不同结构信息。低频成分主要反映整体轮廓与全局背景信息，中频成分包含局部区域边界与粗略纹理信息，高频成分则与细节纹理、噪声等高频特性相关。为此，将频谱特征$\boldsymbol{F}$划分为低频、中频与高频3个频段，即

$$\boldsymbol{F}_{\text{low}}, \boldsymbol{F}_{\text{mid}}, \boldsymbol{F}_{\text{high}} = BandSplit(\boldsymbol{F}) \tag{8}$$

式中，$BandSplit(\cdot)$表示根据设定的频率阈值对原频谱进行区域划分。

在完成频段划分后，针对每一个频段，直接在频域上施加3×3的小卷积，以增强局部结构调制能力。对应地，各频段卷积过程分别表示为

$$\begin{cases} \boldsymbol{G}_{\text{low}} = BN(Conv_{3\times3}(\boldsymbol{F}_{\text{low}})) \\ \boldsymbol{G}_{\text{mid}} = BN(Conv_{3\times3}(\boldsymbol{F}_{\text{mid}})) \\ \boldsymbol{G}_{\text{high}} = BN(Conv_{3\times3}(\boldsymbol{F}_{\text{high}})) \end{cases} \tag{9}$$

经上述过程处理，网络可以捕获频率局部变化并增强特征细粒度表达能力。不同于在空间域进行卷积，在频域进行的小卷积可有效抑制无关频率成分的干扰，强化感兴趣频段的信息。将3类频段卷积后的输出特征逐元素加和，整合成统一的频谱特征表示。具体为

$$\boldsymbol{F}_{\text{fused}} = \boldsymbol{G}_{\text{low}} + \boldsymbol{G}_{\text{mid}} + \boldsymbol{G}_{\text{high}} \tag{10}$$

将融合后的频谱特征经过逆傅里叶变换，映射回空间域，得到频域建模分支的输出特征，具体为

$$\boldsymbol{Y}_f = IFFT(\boldsymbol{F}_{\text{fused}}) \tag{11}$$

逆变换确保了频域处理后的特征可以与空间域处理特征在同一域上进行有效对齐和融合。最后，空间域特征$\boldsymbol{Y}_{\text{spatial}}$与频域特征$\boldsymbol{Y}_f$在空间域中逐元素加和，得到频谱引导动态卷积模块的最终输出。具体为

$$\boldsymbol{Z}_{\text{out}} = \boldsymbol{Y}_{\text{spatial}} + \boldsymbol{Y}_f \tag{12}$$

通过联合空间域与频域特征，频谱引导动态卷积模块不仅保留了空间卷积对局部信息的强建模能力，同时引入频域多频段特征补充传统卷积对全局与结构性信息的感知盲点，在不显著增加计算成本的前提下，可以有效提升整体性能与鲁棒性。

2.4 多尺度频域代理注意力

经分层编码器的特征提取，将每层特征传向解码器，在深层次特征建模过程中，不同尺度特征在空间分辨率与语义粒度上存在显著差异。为此，提出多尺度频域代理注意力模块，如图6所示，通过频域引导机制动态生成代理向量，实现不同尺度特征之间的高效全局信息交互与上下文建模，图中$\boldsymbol{\Phi}$为特征映射函数。具体而言，多尺度频域代理注意力模块首先利用上一级尺度的输出特征进行频谱分析，通过二维傅里叶变换将空间特征映射至频率域，提取其全局能量分布。在频谱空间中，不同频率成分对应不同尺度的图像结构信息，为提取整体特征的概括性表示，进一步对频谱幅值进行空间均值池化得到频谱能量向量。具体为

$$\boldsymbol{E} = Pool(FFT(\boldsymbol{Z}_{i-1})) \tag{13}$$

式中，E能够有效编码上一尺度特征在频域下的重要性分布。采用多层感知机(multilayer perceptron，MLP)将频谱能量映射为一组多尺度动态代理向量，通过频域信息动态生成代理，提升跨尺度特征间的感知与适配能力。具体为

$$\boldsymbol{A} = MLP(\boldsymbol{E}) \tag{14}$$

分两阶段完成注意力建模，将代理向量$\boldsymbol{A}$作为查询，从当前尺度的键值对$\boldsymbol{K}$、$\boldsymbol{V}$中聚合信息，生成中间上下文特征$\boldsymbol{F}_A$。具体为

$$\boldsymbol{F}_A = Attn(\boldsymbol{A},\boldsymbol{K},\boldsymbol{V}) \tag{15}$$

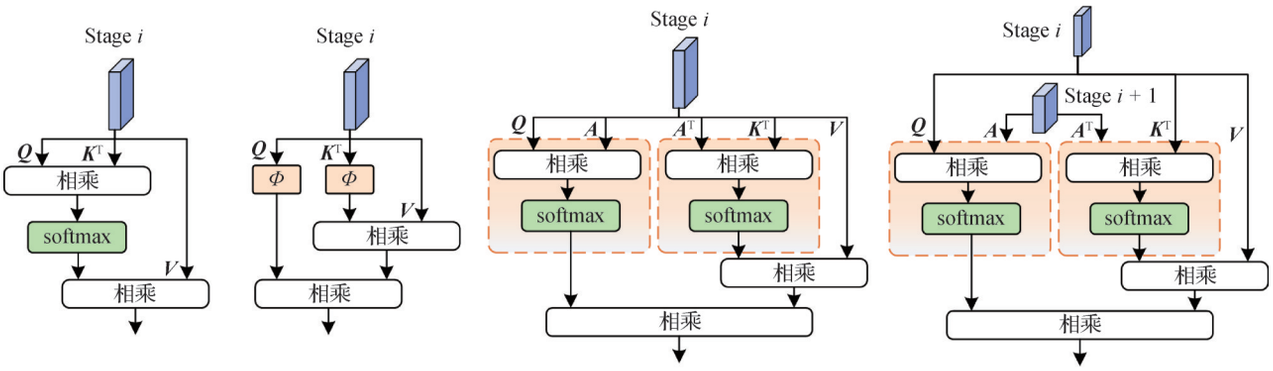
将当前尺度的查询向量$\boldsymbol{Q}$与来自上一次尺度产生的代理向量$\boldsymbol{A}$进行注意力计算，形成最终融合特征Y。具体为

$$\boldsymbol{Y} = Attn(\boldsymbol{Q},\boldsymbol{A},\boldsymbol{F}_A) \tag{16}$$

综上，通过引入频域信息驱动的代理机制，MFAA模块能够以极低的参数与计算开销，提高深层语义一致性建模与上下文特征补全能力。该模块作为编码器与解码器之间的特征桥接路径，不仅可提升全局建模能力，还具备良好的可插拔性，可广泛应用于多种分割网络中。

2.5 频域引导机制

近年来频域特征建模逐渐成为多模态任务中的一种有效补充手段。传统的空间卷积网络在处理局部结构与纹理方面具备较强能力，但在捕捉长距离



(a) sofmax 注意力　(b) 线性注意力　(c) 代理注意力　(d) 多尺度频域代理注意力

图 6　多尺度频域代理注意力

Fig. 6　Multi-scale frequency-aware agent attention ((a) softmax attention; (b) linear attention; (c) agent attention; (d) multi-scale frequency-aware agent attention)

依赖、全局轮廓结构以及跨尺度一致性方面存在先天劣势。频域变换可将图像由空间表示映射到频谱空间，在该域内，高频部分对应边缘与细节，低频部分对应全局轮廓与语义结构。利用频谱空间的能量分布特性，网络能够以更低的计算成本感知远距离语义变化与模态差异。

对于RGB-D任务而言，RGB模态往往在高频区域分布能量较多，表现出丰富的纹理与边缘信息；而深度图则集中在低频段，反映几何结构与空间连续性。这种固定的频域互补性为融合建模提供了明确的引导线索。本文提出的频域提示机制通过频谱变换构建语义一致性提示，频谱动态卷积模块则在不同频段实现分组感知与尺度控制，从而避免了空间卷积对模态耦合特征的混叠问题。相比空间域直接交互，频域引导机制可在不增加过多参数的情况下提升跨模态对齐效率，并具备更强的稳定性与鲁棒性。

3　实　验

3.1　评价指标

为全面评估本文在不同RGB-D任务中的性能，针对语义分割任务与显著目标检测任务分别采用多种主流评价指标评价不同网络的性能。

对于语义分割任务，主要采用平均交并比(mIoU)作为性能评估指标。mIoU是预测分割区域与真实标注区域的交集面积与并集面积之比，用于衡量模型在各类别上的分割准确性。对于显著目标检测任务，为多角度衡量预测显著图与真实标签之间的一致性，采用4个主流指标进行综合评估，分别为调和平均值F-measure(Margolin等，2014)、增强评价指标E-measure(Fan等，2018)、结构相似度指标S-measure(Fan等，2017a)和平均绝对误差(mean absolute error，MAE)(Perazzi等，2012)，以显示全面的结果。F-measure通过综合准确率与召回率，衡量检测结果与真实标签的一致性，强调模型在显著区域提取中的整体准确性。E-measure融合区域层次与像素级对齐特性，兼顾整体结构与局部细节，体现了显著图与真实掩膜的视觉一致性。S-measure从对象感知与区域感知两个方面评估预测图与真实图的结构相似性，可以反映出模型对显著目标完整性和布局一致性的建模能力。MAE直接计算预测显著图与真实掩膜之间的像素级平均误差，量化整体预测偏差，数值越小表明预测结果越接近真实标签。

3.2　数据集

在RGB-D语义分割任务中，采用NYU Depth V2 (New York University depth dataset V2)(Silberman等，2012)和SUN-RGBD(RGB-D scene understanding benchmark suite)(Song等，2015)两个具有代表性的RGB-D室内场景语义分割数据集。NYU Depth V2包含1 449对RGB-D样本图像，图像分辨率为640×480像素，其中包含795幅训练图像和654幅测试图像，有40个语义类别。SUN-RGBD共包含10 335幅RGB-D图像，其中5 285幅图像作为训练集，5 050幅图像作为测试集，含有37个语义类别，将输入大小调整为480×480像素。

在RGB-L语义分割任务中，使用KITTI-360 (Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)(Liao等，2023)数据集进

行评估。KITTI-360是一个郊区城市驾驶场景数据集，包含49 004幅训练图像和12 276幅验证图像，图像分辨率为1 408 × 376像素，覆盖19个语义类别。在RGB-D显著目标检测任务中，选用NJU2K（Nanjing University 2K RGB-D dataset）（Ju等，2014）、NLPR（National Laboratory of Pattern Recognition RGB-D salient object detection dataset）（Peng等，2014）、SIP（salient person dataset）（Fan等，2021）、STERE（stereoscopic image saliency dataset）（Niu等，2012）和DES（depth estimation based saliency dataset）（Ding等，2019）5个主流数据集进行评估。NJU2K包含1 985幅RGB-D图像；NLPR由Kinect采集的1 000对RGB-D图像构成，分辨率为640 × 480像素，覆盖室内外多种场景；SIP数据集包含929幅带有人物姿态变化的图像，分辨率为992 × 774像素；STERE包含1 000幅RGB-D图像；DES数据集相对较小，仅包含135幅室内场景图像。按照主流设定，采用NLPR-train中的700幅样本和NJU2K-train中的1 485幅样本作为训练集，并在5个数据集上统一进行测试：其中NJU2K-test为500幅，NLPR-test为300幅，SIP、STERE和DES分别包含929幅、1 000幅和135幅图像。

3.3 实验环境

实验均在单个NVIDIA 3090 GPU上进行测试，采用的深度学习框架为PyTorch 2.1.2。在训练过程中，采用随机水平翻转与随机缩放两种数据增强策略，以增强模型的泛化能力。对于RGB-D模块的全局注意力分支，选用在ImageNet（Russakovsky等，2015）上预训练的DFormer网络中的GAA（global-aware attention）模块进行初始化，并在此基础上进行微调。网络解码器部分，参考SegNeXt（Guo等，2022）的设计，采用轻量级的Hamburger模块（Geng等，2021）作为基础结构。优化器方面，选用AdamW（Loshchilov和Hutter，2019），权重衰减系数设置为0.01，学习率采用多项式衰减策略。损失函数采用标准的交叉熵损失，用于监督像素级语义预测。在RGB-D语义分割任务中，NYU Depth V2数据集的训练轮次设置为400轮，批量大小为8，初始学习率为6×10^{-5}。SUN-RGBD数据集的训练轮次设置为300轮，批量大小为16，初始学习率为8×10^{-5}。参考近期其他相关方法，在评估阶段采用多尺度翻转推理策略，缩放因子设置为[0.5, 0.75, 1, 1.25, 1.5]，以提升预测稳定性与精度。在RGB-D显著目标检测任务中，训练轮次设置为200轮，批量大小为16，初始学习率为2×10^{-4}，并在每60轮后按0.1的衰减率更新学习率，以保证训练过程的稳定收敛。

3.4 RGB-D语义分割实验对比与分析

在NYU Depth V2和SUN-RGBD两个主流RGB-D室内语义分割数据集上，将所提网络与16种先进方法进行了系统比较。结果如表1所示，本文模型在保持较低参数量的同时，取得最优的整体性能。

在NYU Depth V2数据集上，借助多尺度推理策略，本文提出的Large模型以40.8 M的训练参数实现了57.6%的mIoU，显著优于同等规模的轻量级网络。与代表性方法CMX-BS相比，本文方法在mIoU指标上提升了约0.7%，同时训练参数减少了约77%，充分验证了本文方法在压缩模型规模的同时保持表达能力的优势。此外，本文方法的Small版本模型仅使用20.3 M的训练参数实现了54.6%的mIoU，训练开销仅为多数主流方法的一半，进一步验证了本文提出的频谱引导动态卷积与频域代理注意力机制在轻量化部署中的适用性。在SUN-RGBD数据集上的实验结果亦表明所提方法具有较强的跨数据集泛化能力。尤其与近期SOTA（state-of-the-art）方法Sigma（Wan等，2025）相比，本文以不到其50%的参数量实现了更高的分割性能，达到52.8%的mIoU。该结果进一步验证了频域提示调优机制和跨层语义对齐策略在建模RGB与深度模态间互补信息方面的有效性。整体而言，本文网络能够在保持低计算开销与参数量的同时，展现出较高语义建模能力和稳定的跨模态融合性能，实现了性能与效率间的优良权衡。

为更直观地评估不同RGB-D语义分割方法的性能，图7给出了在NYU Depth V2数据集上的可视化分割结果对比。实验结果表明，本文方法在边界感知、小目标识别以及复杂材质理解等方面展现出显著优势。在第1行场景中，其他方法在白板与椅子之间的边界区域出现模糊，容易导致类别混淆，常将白板区域误判为周围的家具或墙体结构；本文方法能够准确辨识两者边界，明显增强了边缘结构的感知能力。在第2行样本中，场景中包含床、地板与枕头等目标。由于床单表面纹理复杂，部分方法在区分床与枕头时存在显著的区域漏检与标签混淆。本文所提出的频域与空间域特征融合策略，有效增

表1　RGB-D语义分割任务的定量对比实验

Table 1　Quantitative comparison experiments of RGB-D semantic segmentation task

方法	参数/M	NYU Depth V2		SUN-RGBD	
		输入尺寸/像素	mIoU/%	输入尺寸/像素	mIoU/%
ACNet(Hu等,2019)	116.6	480×640	48.3	530×730	48.1
SA-Gate(Chen等,2020)	110.9	480×640	52.4	530×730	49.4
TokenFusion(Wang等,2022b)	45.9	480×640	54.2	530×730	51.0†
MultiMAE(Bachmann等,2022)	95.2	640×640	56.0	640×640	51.1†
Omnivore(Girdhar等,2022)	95.7	480×640	54.0	–	–
NAS(Wang等,2023)	48.9	–	55.1	–	50.3
PGDENet(Zhou等,2023)	100.7	480×640	53.7	530×730	51.0
CMX-B2(Zhang等,2023a)	66.6	480×640	54.4	530×730	49.7
CMX-B5(Zhang等,2023a)	181.1	480×640	56.9	530×730	52.4
CMNext(Zhang等,2023b)	119.6	480×640	56.9	530×730	51.9†
SGACNet(Zhang等,2023c)	–	480×640	49.4	530×730	47.8
Dformer-L(Yin等,2024)	39.0	480×640	57.2	530×730	52.5
Dformer-S(Yin等,2024)	18.7	480×640	53.6	530×730	50.0
SMMCL(Dong和Yokoya,2024)	–	480×640	55.8	–	–
CFCI-Net(Zhou等,2025)	424.2	480×640	<u>57.3</u>	530×730	<u>52.7</u>
SGNet(Wang等,2024)	64.7	480×640	51.1	530×730	48.6
AsymFormer(Du等,2024)	33	480×640	55.3	530×730	49.1
Sigma(Wan等,2025)	69.8	480×640	57.0	480×640	52.4
本文(small)	20.3	480×640	54.6	530×730	50.5
本文(large)	40.8	480×640	**57.6**	530×730	**52.8**

注:加粗、下划线字体分别表示各列最优、次优结果,“–”表示未进行对应实验,“†”表示参考DFormer的结果。

强了模型对结构边界与细节纹理的建模能力,准确地保留了目标对象的完整轮廓,实现了更精细的像素级分割。在第7行厨房场景中,冰箱表面存在贴附纸张等细小目标。多数方法容易将其误识为背景或周边物体,导致细节缺失;相比之下,本文网络通过跨模态提示调优与多尺度特征增强机制,成功分离出纸张与冰箱的边界信息,体现出本文方法对低对比度小目标具有较高感知能力。此外,在具有镜面反射、地面材质复杂等因素的客厅场景中,其他方法往往面临语义不一致或纹理扰动的问题,而本文方法凭借频域引导的语义对齐机制与上下文建模能力,能够更精确地分辨反光桌面与周围物体之间的语义边界,保持各区域的连贯性与细节的完整性。

3.5　RGB-L语义分割实验对比与分析

为了进一步验证所提方法在不同模态组合下的适应性与泛化能力,本文在KITTI-360数据集上开展了RGB-L语义分割实验,并与当前主流方法进行了系统对比,结果如表2所示。

从表2可以看出,尽管本文方法最初是面向RGB-D场景设计,仍然能够在RGB-L场景下展现出优异的性能和出色的泛化能力。本文模型的mIoU指标为66.3%,在保持轻量参数规模的前提下,超越了DFormer(66.1%)等近期代表性方法。得益于频域引导机制的介入,有效缓解了RGB图与激光雷达深度图在特征表达上的分布偏移问题,提升了跨模态特征的融合鲁棒性,表现出良好的任务迁移适应性,进一步验证了频域感知机制在模态变化环境下

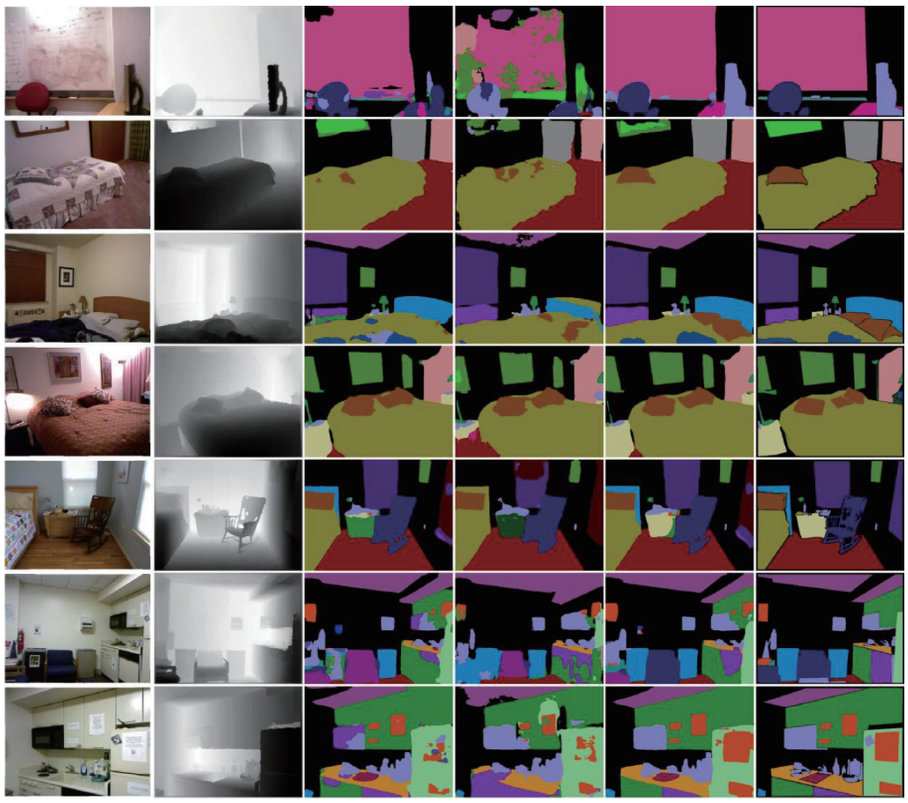


图7　NYU Depth V2数据集上的可视化对比图

Fig. 7　Visualization comparison on NYU Depth V2 dataset

表2　KITTI-360数据集上RGB-X分割的实验结果

Table 2　RGB-X segmentation experimental results on the KITTI-360 Dataset

方法	骨干网络	mIoU/%
PMF(Zhuang等,2021)	SalsaNext	54.5
TransFuser(Prakash等,2021)	RegNetY	56.6
TokenFusion(Wang等,2022b)	MiT-B2	54.6
HRFuser(Bröedermann等,2023)	HRFormer-T	48.7
CMX(Zhang等,2023a)	MiT-B2	64.3
CMNeXt(Zhang等,2023b)	MiT-B2	65.3
DFormer(Yin等,2024)	DFormer-L	<u>66.1</u>
本文	DFormer-L	**66.3**

注:加粗、下划线字体分别表示最优、次优结果。

的稳定建模能力。

3.6　泛化性实验

为全面验证本文网络的泛化能力,在NJU2K、NLPR、SIP、STERE和DES 5个主流RGB-D显著目标检测数据集上,与多种具有代表性的先进方法进行对比,其中包括DCF(depth calibration and fusion)(Ji等,2021)、CMINet(cross-stage multi-scale interaction network)(Zhang等,2021)、DCMF(discriminative cross-modality features)(Wang等,2022a)、SSLSOD(self-supervised pretraining for RGB-D salient object detection)(Zhao等,2022)、CIRNet(cross-modality interaction and refinement for RGB-D salient object detection)(Cong等,2022)、CAVER(cross-modal view-mixed Transformer for bi-modal salient object detection)(Pang等,2023)、MMRNet(multi-modal and multi-scale refined network for RGB-D salient object)(Fang等,2023)、PICR-Net(point-aware interaction and CNN-induced refinement network)(Cong等,2023)、HIDANet(hierarchical depth awareness network)(Wu等,2023)、AirSOD(Zeng等,2024)、GTransNet(Fang等,2024)以及MambaSOD(Zhan等,2025),相关定量对比结果如表3所示。

从整体上看,本文提出的频域引导轻量感知网络在所有数据集的4个评估指标(F-measure、E-measure、S-measure和MAE)上均取得了最优性能。具体而言,本文给出的Large模型在所有数据集上的

表 3　RGB-D SOD 任务的定量对比实验

Table 3　Quantitative comparison experiments of RGB-D SOD task

方法	参数量/M	NJU2K 数据集				NLPR 数据集				STERE 数据集			
		F	E	S	M	F	E	S	M	F	E	S	M
DCF	108.5	0.922	0.946	0.912	0.036	0.918	0.958	0.924	0.022	0.911	0.940	0.902	0.039
CMINet	–	0.940	0.954	0.929	0.028	0.931	0.959	0.932	0.020	0.925	0.946	0.918	0.032
DCMF	58.9	0.888	0.925	0.913	0.043	0.875	0.940	0.922	0.029	0.872	0.930	0.903	0.043
SSLSOD	74.2	0.902	0.939	0.903	0.044	0.899	0.949	0.915	0.028	0.882	0.929	0.887	0.047
CIRNet	103.2	0.927	0.925	0.925	0.035	0.924	0.955	0.934	0.027	0.913	0.928	0.916	0.037
CAVER	55.8	0.923	0.951	0.920	0.031	0.921	0.961	0.929	0.022	0.911	0.949	0.914	0.033
MMRNet	172	0.922	0.904	0.910	0.049	0.921	0.941	0.918	0.033	0.913	0.929	0.899	0.042
PICR-Net	112	0.923	0.954	0.922	0.032	0.930	0.971	0.936	0.019	0.910	0.951	0.913	0.033
HIDANet	525	0.922	0.953	0.920	0.031	0.924	0.964	0.929	0.021	0.894	0.940	0.897	0.042
AirSOD	2.4	0.918	0.944	0.908	0.039	0.923	0.963	0.924	0.023	0.900	0.939	0.895	0.043
GTransNet	431.6	0.921	0.926	0.922	0.028	0.908	0.961	0.928	0.019	0.895	0.928	0.908	0.032
MambaSOD	78.9	0.937	**0.963**	0.934	0.027	0.934	**0.973**	**0.941**	**0.017**	0.920	**0.955**	0.924	0.031
本文	26.2	**0.948**	0.962	**0.937**	**0.024**	**0.941**	0.967	**0.941**	**0.017**	**0.932**	0.953	**0.925**	**0.028**

方法	DES 数据集				SIP 数据集			
	F	E	S	M	F	E	S	M
DCF	0.910	0.941	0.905	0.024	0.899	0.916	0.876	0.052
CMINet	0.944	0.975	0.940	0.016	0.923	0.934	0.898	0.040
DCMF	0.901	0.967	0.880	0.023	–	–	–	–
SSLSOD	–	–	–	–	0.894	0.928	0.890	0.047
CIRNet	0.878	0.952	0.913	0.025	0.895	0.917	0.888	0.052
CAVER	–	–	–	–	0.906	0.933	0.893	0.042
MMRNet	–	–	–	–	0.902	0.921	0.882	0.049
PICR-Net	0.939	0.970	0.940	0.017	0.894	0.926	0.881	0.050
HIDANet	0.918	0.954	0.919	0.021	0.894	0.925	0.884	0.050
AirSOD	0.934	0.962	0.923	0.022	0.887	0.903	0.858	0.060
GTransNet	0.938	**0.982**	0.944	0.014	0.895	0.926	0.887	0.041
MambaSOD	0.935	0.969	0.937	0.019	0.914	0.940	0.904	0.040
本文	**0.955**	0.980	**0.946**	**0.013**	**0.934**	**0.943**	**0.910**	**0.034**

注：F、E、S、M 分别表示 F-measure、E-measure、S-measure、MAE。加粗、下划线字体分别表示各列最优、次优结果。“–”表示对应文献未进行对应实验。

表现均优于或相当于对比中的最优方法，可视化对比如图 8 所示。在 NJU2K 数据集中，本文网络的 MAE 为 0.024，领先其他方法，且在 F-measure 和 S-measure 指标上实现了进一步提升，表明本文网络对目标区域具有更强的识别能力。在 NLPR 数据集中，本文网络同样展现出领先的性能，指标 F-measure 达到 0.941、E-measure 达到 0.967、S-measure 达到 0.941、MAE 仅为 0.017，均超越了现有方法。尤其在背景抑制和边界保持方面，本文方法的优势更加明显，体现了频域建模对细粒度特征捕获的有效促进

图8　NJU2K数据集上的可视化对比图

Fig. 8　Visualization comparison on NJU2K dataset

作用。在STERE数据集实验中，本文网络通过深层特征优化与频域引导提示调优机制，实现了优异的显著性检测效果，F-measure指标为0.932，均优于对比方法，在S-measure和MAE上也有一定优势。

在数据量较小的DES数据集中，本文网络同样表现稳定，4项指标中有3项超过其他全部方法，充分说明本文网络在样本数量有限的情况下，依然能够保持较高的鲁棒性。在SIP数据集上，本文网络的指标F-measure为0.934、E-measure为0.943、S-measure为0.910、MAE降至0.034，相比Mamba-SOD等方法均有提升，再次验证了本文网络具有良好的跨场景泛化能力。综上，得益于频域引导提示调优、频谱引导动态卷积与多尺度频域代理注意力机制的协同优化，本文方法在RGB-D显著目标检测任务中实现了在精度与效率之间的平衡，展现出较高的性能优势与广泛的适应性。

图9展示了4组具有代表性的复杂场景样本，涵盖边界模糊、遮挡干扰以及复杂背景等挑战情形，直观反映了所提方法在RGB-D显著目标检测任务中的表现优势。例如，在第2组马的图像中，栅栏遮挡严重影响了前景目标的连贯性，HIDANet产生了较大区域的误判，而本文模型通过频域建模抑制了遮挡干扰，更好地保留了目标形状。在第3组战斗机样本中，由于尾翼与天空亮度接近，HIDANet分割区域发生粘连和轮廓偏移，本文方法则凭借频谱引导机制有效提取了结构边缘信息，实现了更精细的语义划分。总体来看，本文方法在多个复杂场景下均体现出更优的边界保持与细节表达能力。

3.7　消融研究

为验证本文网络的有效性，对RGB-D语义分割进行消融实验，所有实验在NYU Depth V2数据集上进行。本文网络由频域引导提示适配器（FPA）、频谱引导动态卷积模块（SDC）和多尺度频域代理注意力模块（MFAA）组成。为更好地理解每个组件的影响和贡献，分别从本文网络中取出这3个组件。

表4给出了在不同模块组合下Large版本模型的性能变化情况。其中，Exp. A作为基线，仅使用跨模态全局注意力。在此基础上，引入频域引导提示适配器（Exp. B），mIoU由55.3%提升至55.8%，验证了频域引导的提示调优能够有效促进跨层语义对

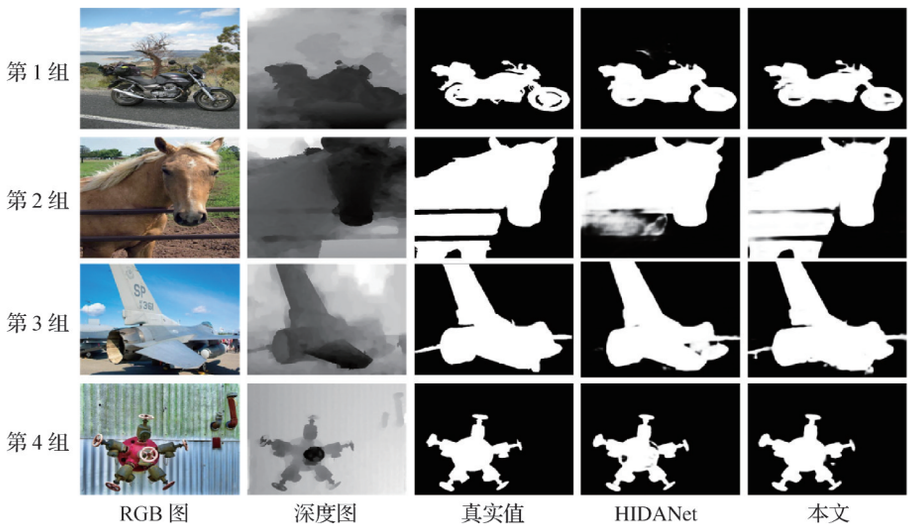


图9　复杂场景可视化对比

Fig. 9　Visualization comparison of complex scene

齐。仅引入频谱引导动态卷积模块(Exp. C)时，mIoU达到56.0%，说明频域动态卷积能够有效增强局部细节与全局结构的建模能力。单独引入多尺度频域代理注意力模块(Exp. D)时，mIoU达到56.3%，表明基于频谱特征生成的动态代理机制可以有效缓解不同尺度特征的不一致性。进一步地，联合使用其中两个模块(Exp. E、Exp. F或Exp. G)时，mIoU提升至56.9%～57.1%。当3个模块全部集成后，mIoU达到57.6%。这些结果验证了各模块间的互补性：FPA能够促进跨层语义传播，SDC提升了频谱特征表达能力，MFAA可以强化不同尺度特征间的交互。特别是在复杂场景和细节处理方面，本文网络不仅能够优化跨层特征交互，还提高全局特征建模和多尺度信息融合的能力，进一步验证了每个模块在结构设计上的互补性与协同增强能力。

为研究频谱增强提示融合模块(FPA)在不同编码器阶段(Stage)的效果，表5列出了在不同位置插入FPA的实验结果。与基线(57.1%)相比，逐层加入FPA可以提高性能，在第1→3阶段达到57.3%，当FPA跨越所有4个阶段时达到57.6%，表明FPA能够有效促进不同层间的信息交互并增强特征的表达能力。如果不使用频域引导，仅依赖空间域提示，性能下降0.2%。此外，在仅使用Token而不跨层传播的情况下，性能有所下降，这进一步说明FPA跨层机制在特征融合与信息传递中的重要性。综合来看，通过在所有Stage中引入FPA，可以在不过多增加模型计算开销的同时提高分割性能。

表4　网络中不同模块作用的消融研究

Table 4　Ablation study of different modules for the network

实验	FPA	SDC	MFAA	参数量/M	mIoU/%
Exp.A	–	–	–	38.0	55.3
Exp.B	√	–	–	38.2	55.8
Exp.C	–	√	–	39.8	56.0
Exp.D	–	–	√	38.8	56.3
Exp.E	√	√	–	40.0	56.9
Exp.F	√	–	√	39.0	56.7
Exp.G	–	√	√	40.6	57.1
本文	√	√	√	40.8	57.6

注："√"表示使用该模块，"–"表示未采用。

表5　FPA模块位置的消融研究

Table 5　Ablation study of FPA module location

FPA位置	参数量/M	mIoU/%
w/o FPA	40.61	57.1
Stage 1	40.66	57.2
Stage 1→2	40.71	57.2
Stage 1→3	40.76	57.3
Stage 1→4(w/o无跨层)	41.23	57.5
Stage 1→4(空间域引导)	40.81	57.4
本文(频域引导)	40.81	57.6

如表6所示，使用不同的架构设置来评估频谱引导动态卷积模块(SDC)，给出单分支训练、结合频

域训练以及SDC区分频段的训练性能对比。单分支配置下，仅空间域训练加入3×3卷积，mIoU为56.4%，引入7×7大卷积核后，性能提升0.5%。使用空间域双分支训练后，指标提升至57.2%，验证了多尺度建模的必要性。将空间域次分支改为频域上处理，可以将性能提高至57.4%。而完整的SDC模块可以达到57.6%，模型在复杂纹理区域的效果更好，分割精度提升显著，更有效地捕获全局信息。

表6　SDC模块分支的消融研究

Table 6　Ablation study of SDC module branches

方法	参数量/M	mIoU/%
仅空间域(K=3)	0.9	56.4
仅空间域(K=7)	0.5	56.9
空间域(K=7)+空间域(K=3)	1.2	57.2
空间域(K=7)+频域(K=3)	1.2	57.4
空间域(K=7)+频域三频段(K=3)	1.8	57.6

为进一步分析SDC的设计细节，通过消融实验研究频域分支参数以及融合方式对性能的影响，结果如表7所示。首先，将频域分支的卷积核大小设为1×1，实验结果mIoU达到57.2%，说明小内核能够在一定程度上捕获频域局部特征。当将频域卷积核扩大为3×3时，保持快速傅里叶变换(fast Fourier transform，FFT)和门控相加策略，mIoU进一步提升至57.6%。若将频域变换由FFT替换为离散余弦变换(discrete cosine transform，DCT)，mIoU下降至57.0%，说明DCT在频率表示过程中由于缺失相位信息，导致特征表达能力下降。此外，当将门控融合策略改为直接加和，mIoU下降至57.2%，表明门控机制在分支特征融合过程中起到了有效的动态权重调节作用。综上，采用FFT、小内核卷积和门控融合的设计方案，能够最大程度发挥频谱建模优势，进一步提升分割精度和模型表现。

表7　SDC模块频域分支的消融研究

Table 7　Ablation study of SDC frequency domain branchs

频域核大小	频域变换	融合方式	mIoU/%
1×1	FFT	门控相加	57.2
3×3	DCT	门控相加	57.0
3×3	FFT	直接相加	57.2
3×3	FFT	门控相加	57.6

表8给出了多尺度频域代理注意力模块(MFAA)与其他注意力机制在特征建模能力上的差异。传统softmax注意力与线性注意力分别取得了56.8%和56.6%的mIoU，表明直接通过键值关系建模存在一定的上下文捕获能力，但受限于缺乏代理增强机制，表现相对有限。引入基于同尺度特征生成的代理注意力，采用不同代理生成方式(池化、卷积)后，mIoU提升至57.3%，验证了代理机制对语义建模的促进作用，其中局部卷积感知生成的代理效果最佳。在此基础上，本文提出的MFAA通过频域幅值池化生成代理向量，并引入跨尺度协作机制($\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$来自当前尺度，$\boldsymbol{A}$来自下一尺度)，最终使得mIoU提升至57.6%，明显优于其他注意力机制。

表8　MFAA模块的消融研究

Table 8　Ablation study of MFAA module

注意力类型	代理生成方式	代理来源	mIoU/%
softmax注意力	-	同尺度	56.8
线性注意力	-	同尺度	56.6
代理注意力	池化	同尺度	57.3
代理注意力	卷积	同尺度	57.3
多尺度频域代理注意力	频域幅值池化	跨尺度	57.6

注："-"表示无。

3.8　模型参数与FPS分析

如表9所示，为全面评估本文网络在轻量化设计与推理效率方面的性能优势，选取多种帧速率(frame per second，FPS)超过30的轻量级RGB-D语义分割模型进行对比，涵盖可训练参数量、推理参数量以及在实际设备上的推理速度(FPS)。PGDENet设计逐步引导策略与几何增强模块，强调融合过程中对深度结构信息的利用，但模块堆叠复杂，解码器计算开销大。AsymFormer通过非对称主干结构实现轻量建模，通过CNN和Transformer分配模态处理任务，融合能力有限。相比之下，本文方法采用统一融合主干，结合频域提示、频谱卷积与代理注意力机制，有效提升了特征对齐与多尺度语义建模能力。所有实验均在单个NVIDIA 3090 GPU上完成，从结果来看，本文方法能够在保持较低推理参数量的同时，实现46帧/s的推理速度，展现出其在资源受限应用场景中良好的适用性。

此外，表10展示了本文模块设计的计算效率。

表 9　模型参数与 FPS 分析

Table 9　Model parameters and FPS analysis

方法	参数量/M	FPS/(帧/s)	mIoU/%
SA-Gate(Chen 等,2020)	65	35	50.2
ESANet(Seichter 等,2021)	34	32	51.6
FRNet(Zhou 等,2022)	86	39	53.6
PGDENet(Zhou 等,2023)	101	32	53.7
DFormer-S(Yin 等,2024)	18.7	50	53.6
AsymFormer(Du 等,2024)	33	65	54.1
本文(small)	19.3	46	54.6

在基线网络的基础上,引入 FPA 仅增加约 0.2 M 参数量与每秒浮点运算次数(floating point operations per second,FLOPs)0.7 G,显存开销增长 0.11 GB,但 mIoU 提升了 0.5;进一步加入 SDC 后,参数量与 FLOPs 分别增长约 1.8 M 与 1.8 G,显存增加 0.23 GB,带来 1.1 个百分点的精度提升;在此基础上引入 MFAA,参数量仅增加 0.8 M,FLOPs 增加 1.7 G,显存增长 0.15 GB,却带来了额外 0.7 个百分点的性能提升。整体来看,3 个模块的叠加仅使参数量较基线增加 7.4%、FLOPs 增加 6.4%、显存占用增加 9.6%,却使得 mIoU 提升了 2.3,体现出较高的计算效率。

表 10　模块复杂度分析

Table 10　Module complexity analysis

模块	参数量/M	FLOPs/G	mIoU/%
基线	38.0	65.7	55.3
+FPA	38.2	66.4	55.8
+SDC	40.0	68.2	56.9
+MFAA 本文(large)	40.8	69.9	57.6

4　结　论

本文提出一种频域引导的轻量级 RGB-D 感知网络,旨在提升多模态特征对齐与语义建模效率,解决现有方法参数冗余与融合能力不足的问题。首先,设计频域引导提示适配器,提升跨层语义特征的一致性与传播效率;其次,提出一种频谱引导动态卷积模块,在融合空间域与频域特征的同时实现了高效的多尺度特征建模;最后,构建一种多尺度频域代理注意力模块,以低开销方式增强不同尺度特征之间的语义交互与全局建模能力。本文方法在多个 RGB-D、RGB-L 语义分割数据集上,能够以较低参数量实现了优异的分割性能,并在 RGB-D 显著目标检测任务中的 5 个数据集展现了良好的泛化能力。研究结果表明,与相关方法相比,本文方法在结构复杂的场景中分割结果更为准确,为多模态感知网络的融合与轻量化设计提供了新思路与实践依据。

参考文献(References)

Bachmann R, Mizrahi D, Atanov A and Zamir A. 2022. MultiMAE: multi-modal multi-task masked autoencoders//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 337-354 [DOI: 10.1007/978-3-031-19836-6_20]

Bi Q, You S D and Gevers T. 2024. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 819-827 [DOI: 10.1609/aaai.v38i2.27840]

Bröedermann T, Sakaridis C, Dai D X and Van Gool L. 2023. HRFuser: a multi-resolution sensor fusion architecture for 2D object detection//Proceedings of the 26th IEEE International Conference on Intelligent Transportation Systems (ITSC). Bilbao, Spain: IEEE [DOI: 10.1109/ITSC57777.2023.10422432]

Chen X K, Lin K Y, Wang J B, Wu W, Qian C, Li H S, et al. 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 561-578 [DOI: 10.1007/978-3-030-58621-8_33]

Cong R M, Lin Q W, Zhang C, Li C Y, Cao X C, Huang Q M, et al. 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, et al. 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: Association for Computing Machinery: 406-416 [DOI: 10.1145/3581783.3611982]

Couprie C, Farabet C, Najman L and LeCun Y. 2013. Indoor semantic segmentation using depth information [EB/OL]. [2025-05-01]. https://arxiv.org/pdf/1301.3572.pdf

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]

Dong X Y and Yokoya N. 2024. Understanding dark scenes by contrasting multi-modal observations//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 829-839 [DOI: 10.1109/WACV57701.2024.00089]

Du S Q, Wang W X, Guo R Z, Wang R S and Tang S J. 2024. AsymFormer: asymmetrical cross-modal representation learning for mobile platform real-time RGB-D semantic segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 7608-7615 [DOI: 10.1109/CVPRW63382.2024.00756]

Fan D P, Cheng M M, Liu Y, Li T and Borji A. 2017a. Structure-measure: a new way to evaluate foreground maps//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4558-4567 [DOI: 10.1109/ICCV.2017.487]

Fan D P, Gong C, Cao Y, Ren B, Cheng M M and Borji A. 2018. Enhanced-alignment measure for binary foreground map evaluation//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 698-704 [DOI: 10.5555/3304415.3304515]

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]

Fan H, Mei X, Prokhorov D and Ling H B. 2017b. RGB-D scene labeling with multimodal recurrent neural networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA: IEEE: 203-211 [DOI: 10.1109/CVPRW.2017.31]

Fang X, Jiang M F, Zhu J C, Shao X L and Wang H P. 2023. M^2RNet: multi-modal and multi-scale refined network for RGB-D salient object detection. Pattern Recognition, 135: #109139 [DOI: 10.1016/j.patcog.2022.109139]

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

Geng Z Y, Guo M H, Chen H X, Li X, Wei K and Lin Z C. 2021. Is attention better than matrix decomposition? [EB/OL]. [2025-05-01]. https://arxiv.org/pdf/2109.04553.pdf

Girdhar R, Singh M, Ravi N, van der Maaten L, Joulin A and Misra I. 2022. Omnivore: a single model for many visual modalities//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16081-16091 [DOI: 10.1109/CVPR52688.2022.01563]

Guo M H, Lu C Z, Hou Q B, Liu Z N, Cheng M M and Hu S M. 2022. SegNeXt: rethinking convolutional attention design for semantic segmentation//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #84 [DOI: 10.5555/3600270.3600354]

Guo X Y, Ma W, Liang F F and Mi Q. 2024. Dual-modal non-local context guided multi-stage fusion for indoor RGB-D semantic segmentation. Expert Systems with Applications, 255: #124598 [DOI: 10.1016/j.eswa.2024.124598]

Gupta S, Girshick R, Arbeláez P and Malik J. 2014. Learning rich features from RGB-D images for object detection and segmentation//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 345-360 [DOI: 10.1007/978-3-319-10584-0_23]

Han D C, Ye T Z, Han Y Z, Xia Z F, Pan S Y, Wan P F, et al. 2025. Agent attention: on the integration of softmax and linear attention//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 124-140 [DOI: 10.1007/978-3-031-72973-7_8]

Hazirbas C, Ma L N, Domokos C and Cremers D. 2017. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture//Proceedings of the 13th Asian Conference on Computer Vision. Taipei, China: Springer: 213-228 [DOI: 10.1007/978-3-319-54181-5_14]

Hu X X, Yang K L, Fei L and Wang K W. 2019. ACNet: attention based network to exploit complementary features for RGB-D semantic segmentation//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 1440-1444 [DOI: 10.1109/ICIP.2019.8803085]

Ji W, Li J J, Yu S, Zhang M, Piao Y R, Yao S Y, et al. 2021. Calibrated RGB-D salient object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9466-9476 [DOI: 10.1109/CVPR46437.2021.00935]

Jia D, Wang Z T, Li Y Y, Jin Z Y, Liu Z Y and Wu S. 2022. Multi-stage guidance network for constructing dense depth map based on LiDAR and RGB data. Journal of Image and Graphics, 27(2): 435-446 (贾迪, 王子滔, 李宇扬, 金志楊, 刘泽洋, 吴思. 2022. 结合LiDAR与RGB数据构建稠密深度图的多阶段指导网络. 中国图象图形学报, 27(2): 435-446) [DOI: 10.11834/jig.210465]

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. Paris, France: IEEE: 1115-1119 [DOI: 10.1109/ICIP.2014.7025222]

Katharopoulos A, Vyas A, Pappas N and Fleuret F. 2020. Transformers are RNNs: fast autoregressive transformers with linear attention//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: JMLR.org, #478 [DOI: 10.5555/3524938.3525416]

Liao Y Y, Xie J and Geiger A. 2023. KITTI-360: a novel dataset and benchmarks for urban scene understanding in 2D and 3D. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3292-3310 [DOI: 10.1109/TPAMI.2022.3179507]

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization

[EB/OL]. [2017-11-14]. https://arxiv.org/pdf/1711.05101.pdf

Margolin R, Zelnik-Manor L and Tal A. 2014. How to evaluate foreground maps//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2014.39]

Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N and Terzopoulos D. 2022. Image segmentation using deep learning: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3523-3542 [DOI: 10.1109/TPAMI.2021.3059968]

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. Providence, USA: IEEE: 454-461 [DOI: 10.1109/CVPR.2012.6247708]

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. Zurich, Switzerland: Springer: 92-109 [DOI: 10.1007/978-3-319-10578-9_7]

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]

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]

Prakash A, Chitta K and Geiger A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7073-7083 [DOI: 10.1109/CVPR46437.2021.00700]

Qi X J, Liao R J, Jia J Y, Fidler S and Urtasun R. 2017. 3D graph neural networks for RGBD semantic segmentation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5209-5218 [DOI: 10.1109/ICCV.2017.556]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, et al. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Seichter D, Köhler M, Lewandowski B, Wengefeld T and Gross H M. 2021. Efficient RGB-D semantic segmentation for indoor scene analysis//Proceedings of 2021 IEEE International Conference on Robotics and Automation. Xi'an, China: IEEE: 13525-13531 [DOI: 10.1109/ICRA48506.2021.9561675]

Silberman N, Hoiem D, Kohli P and Fergus R. 2012. Indoor segmentation and support inference from RGBD images//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 746-760 [DOI: 10.1007/978-3-642-33715-4_54]

Song S R, Lichtenberg S P and Xiao J X. 2015. SUN RGB-D: a RGB-D scene understanding benchmark suite//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 567-576 [DOI: 10.1109/CVPR.2015.7298655]

Su J W, Luo Z M, Lian S, Lin D Z and Li S Z. 2024. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. Medical Image Analysis, 94: #103111 [DOI: 10.1016/j.media.2024.103111]

Sung C, Kim W, An J, Lee W, Lim H and Myung H. 2024. Contextrast: contextual contrastive learning for semantic segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3732-3742 [DOI: 10.1109/CVPR52733.2024.00358]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

Wan Z F, Zhang P P, Wang Y H, Yong S L, Stepputtis S, Sycara K, et al. 2025. Sigma: Siamese Mamba network for multi-modal semantic segmentation//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 1734-1744 [DOI: 10.1109/WACV61041.2025.00176]

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 N, Zhuo T, Zhang X W, Sun M J, Yin H L, Xing Y H, et al. 2023. Automatic network architecture search for RGB-D semantic segmentation//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: Association for Computing Machinery: 3777-3786 [DOI: 10.1145/3581783.3612288]

Wang Y K, Chen X H, Cao L L, Huang W B, Sun F C and Wang Y H. 2022b. Multimodal token fusion for vision transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12176-12185 [DOI: 10.1109/CVPR52688.2022.01187]

Wang Z X, Yan Z Q and Yang J. 2024. SGNet: structure guided network via gradient-frequency awareness for depth map super-resolution//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: #647 [DOI: 10.1609/aaai.v38i6.28395]

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]

Wu Z W, Zhou Z Y, Allibert G, Stolz C, Demonceaux C and Ma C. 2024. Transformer fusion for indoor RGB-D semantic segmentation. Computer Vision and Image Understanding, 249: #104174 [DOI: 10.1016/j.cviu.2024.104174]

Ye X Y, Zhu L, Wang W W and Fu Y. 2024. RGB_D salient object detection algorithm based on complementary information interac-

tion. 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]

Yin B W, Zhang X Y, Li Z Y, Liu L, Cheng M M and Hou Q B. 2024. DFormer: rethinking RGBD representation learning for semantic segmentation [EB/OL]. [2025-05-01]. https://arxiv.org/pdf/2309.09668.pdf

Zeng Z H, Liu H J, Chen F L and Tan X H. 2024. AirSOD: a lightweight network for RGB-D salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 34(3): 1656-1669 [DOI: 10.1109/TCSVT.2023.3295588]

Zhan Y, Zeng Z H, Liu H J, Tan X H and Tian Y L. 2025. Mamba-SOD: dual Mamba-driven cross-modal fusion network for RGB-D salient object detection. Neurocomputing, 631: #129718 [DOI: 10.1016/j.neucom.2025.129718]

Zhang J, Fan D P, Dai Y C, Yu X, Zhong Y R, Barnes N, et al. 2021. RGB-D saliency detection via cascaded mutual information minimization//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4318-4327 [DOI: 10.1109/ICCV48922.2021.00430]

Zhang J M, Liu H Y, Yang K L, Hu X X, Liu R P and Stiefelhagen R. 2023a. CMX: cross-modal fusion for RGB-X semantic segmentation with transformers. IEEE Transactions on Intelligent Transportation Systems, 24(12): 14679-14694 [DOI: 10.1109/TITS.2023.3300537]

Zhang J M, Liu R P, Shi H, Yang K L, Reiß S, Peng K Y, et al. 2023b. Delivering arbitrary-modal semantic segmentation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1136-1147 [DOI: 10.1109/CVPR52729.2023.00116]

Zhang Y, Xiong C Y, Liu J J, Ye X H and Sun G D. 2023c. Spatial information-guided adaptive context-aware network for efficient RGB-D semantic segmentation. IEEE Sensors Journal, 23(19): 23512-23521 [DOI: 10.1109/JSEN.2023.3304637]

Zhao J Y, Yu C Q and Sang N. 2022. RGB-D semantic segmentation: depth information selection. Journal of Image and Graphics, 27(8): 2473-2486 (赵经阳, 余昌黔, 桑农. 2022. RGB-D语义分割: 深度信息的选择使用. 中国图象图形学报, 27(8): 2473-2486) [DOI: 10.11834/jig.210061]

Zhao X Q, Pang Y W, Zhang L H, Lu H C and Ruan X. 2022. Self-supervised pretraining for RGB-D salient object detection//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI: 3463-3471 [DOI: 10.1609/aaai.v36i3.20257]

Zhou H, Qi L, Wan Z L, Huang H and Yang X. 2021. RGB-D co-attention network for semantic segmentation//Proceedings of the 15th Asian Conference on Computer Vision. Kyoto, Japan: Springer: 514-530 [DOI: 10.1007/978-3-030-69525-5_31]

Zhou H, Yang X, Qi L, Chen H J, Huang H and Qin H D. 2025. CFCI-Net: cross-modality feature calibration and integration network for RGB-D semantic segmentation. IEEE Transactions on Intelligent Vehicles, 10(3): 2049-2063 [DOI: 10.1109/TIV.2024.3442915]

Zhou W J, Xiao Y X, Yan W Q and Yu L. 2024. CMPFFNet: cross-modal and progressive feature fusion network for RGB-D indoor scene semantic segmentation. IEEE Transactions on Automation Science and Engineering, 21(4): 5523-5533 [DOI: 10.1109/TASE.2023.3313122]

Zhou W J, Yang E Q, Lei J S, Wan J and Yu L. 2023. PGDENet: progressive guided fusion and depth enhancement network for RGB-D indoor scene parsing. IEEE Transactions on Multimedia, 25: 3483-3494 [DOI: 10.1109/TMM.2022.3161852]

Zhou W J, Yang E Q, Lei J S and Yu L. 2022. FRNet: feature reconstruction network for RGB-D indoor scene parsing. IEEE Journal of Selected Topics in Signal Processing, 16(4): 677-687 [DOI: 10.1109/JSTSP.2022.3174338]

Zhuang Z W, Li R, Jia K, Wang Q C, Li Y Q and Tan M K. 2021. Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16260-16270 [DOI: 10.1109/ICCV48922.2021.01597]

作者简介

贾迪,男,教授,主要研究方向为摄影测量、三维视觉与空间感知、多传感器融合、机器人控制。

E-mail: lntu_jiadi@163.com

赵辰,通信作者,男,硕士研究生,主要研究方向为RGB-D语义分割、计算机视觉和单目深度估计。

E-mail: lntu_zhaochen@163.com

张华修,男,硕士研究生,主要研究方向为深度估计、SLAM、计算机视觉和RGB-D语义分割。E-mail: lntu_zhx@163.com

宋慧伦,女,硕士研究生,主要研究方向为单目深度估计、计算机视觉和三维点云配准。

E-mail: lntu_songhuilun@163.com