中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-16

论文引用格式:Huang Rongmei, Yu Hong, Xie Caiyun, Dai Xiaofang, Chen Ying, Dai Jingjie, Hong Ruxia. RGB-D camouflaged object detection with state space model-guided multimodal fusion[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260176. (黄荣梅, 余宏, 张永选, 谢彩云, 陈颖, 戴靓婕, 洪如霞. 状态空间模型引导多模态融合的RGB-D伪装目标检测[J/OL]. 中国图象图形学报, XXXX:1-16. DOI: 10.11834/jig.260176.)[DOI:10.11834/jig.260176]

状态空间模型引导多模态融合的RGB-D伪装目标检测

黄荣梅,余宏,张永选,谢彩云,陈颖,戴靓婕,洪如霞*

豫章师范学院数学与计算机学院,南昌　330103

摘　要: **目的**　伪装目标检测(camouflaged object detection, COD)旨在从复杂场景中识别与背景高度融合的隐藏目标,在农业、医学等领域具有重要研究价值与应用潜力。针对现有方法受限于卷积神经网络(convolutional neural networks, CNN)有效感受野不足、Transformer计算复杂度高,以及仅依赖RGB图像、忽视深度几何先验等问题,开展本文研究。**方法**　提出一种状态空间模型引导多模态融合的RGB-D伪装目标检测方法。利用Depth Anything V2生成高质量伪深度图,输入参数共享编码器提取多模态金字塔特征;设计基于Mamba的多模态状态空间融合模块(multi-modality mamba fusion module, M3FM),实现RGB与深度特征双向互惠融合;构建基于多核非对称卷积的双向上下文混合卷积模块(dual-directional context mixture convolution, DCM-Conv)与多尺度解码器,在提取多感受野特征的同时控制参数量与计算开销。**结果**　在CAMO、COD10K、NC4K 3个伪装目标检测基准数据集进行实验,与11种代表性方法进行定量和定性对比。在平均绝对误差(mean absolute error, MAE)指标上,本文方法相较于排名第2的方法,在3个数据集上分别降低21.3%、17.4%和12.5%;同时在结构度量(structure measure, Sm)、增强对齐度量(enhanced alignment measure, Em)、加权F度量(weighted F-measure, wFm)上均取得最优值。模型参数量仅58.5M,计算复杂度(floating point operations, FLOPs)仅47.6G,精度与效率平衡优异。可视化结果表明,本文方法分割更准确、边界更清晰、细节保留更完整、背景误检更少。**结论**　提出状态空间模型引导多模态融合的RGB-D伪装目标检测方法MambaCOD。通过多模态状态空间融合模块M3FM有效实现RGB与深度特征双向互惠融合,利用Depth Anything V2提供高质量几何先验,并借助DCM-Conv模块增强多尺度上下文特征,可精准定位伪装目标并提升边界与细节清晰度。

关键词: 伪装目标检测;RGB-D;状态空间模型;多模态融合;深度特征

RGB-D camouflaged object detection with state space model-guided multimodal fusion

Huang Rongmei, Yu Hong, Xie Caiyun, Dai Xiaofang, Chen Ying, Dai Jingjie, Hong Ruxia*

School of Mathematics and Computer Science, Yuzhang Normal University, Nanchang 330103, China

Abstract: **Objective** Camouflaged object detection (COD) is a significant and challenging task in computer vision, which focuses on identifying and segmenting objects that are highly similar to complex backgrounds in terms of color, texture, and

收稿日期:2026-04-03;**修回日期**:2026-05-25

*通信作者:洪如霞 *lm199703295@163.com

基金项目:国家自然科学基金地区科学基金项目(62262042),江西省教育厅科学技术研究课题-课题青年项目(GJJ2502404)

Supported by: Regional Science Foundation Project of the National Natural Science Foundation of China (62262042); Science and Technology Research Project for Young Scholars, Department of Education of Jiangxi Province (GJJ2502404)

shape. This technique presents important research value and wide application potential in agricultural monitoring, medical imaging, ecological protection, military reconnaissance, and other fields. However, existing camouflaged object detection methods still face several critical limitations. On the one hand, convolutional neural networks (CNN) are restricted by insufficient effective receptive fields, making it difficult to capture global context information and long-range dependencies required for distinguishing camouflaged objects. On the other hand, vision Transformers rely on self-attention mechanisms with quadratic computational complexity, resulting in huge computational overhead and memory consumption, which makes it hard to balance detection accuracy and efficiency. In addition, most mainstream methods only use single-modal RGB images as input and ignore the rich geometric and spatial prior information contained in depth maps. The existing cross-modal fusion strategies are relatively simple and cannot fully exploit the complementary information between RGB and depth modalities, leading to poor detection performance in highly camouflaged scenes. Aiming at these problems, this paper conducts an in-depth study on RGB-D camouflaged object detection methods based on multi-modal state space models, so as to achieve accurate, efficient and robust camouflaged object detection by fusing appearance information and geometric priors. **Method** This paper proposes a novel RGB-D camouflaged object detection framework based on a multi-modal state space model, named MambaCOD. First, considering the lack of real depth maps in public COD datasets, the advanced visual foundation model Depth Anything V2 is employed to generate high-quality pseudo-depth maps from raw RGB images, providing reliable geometric structure priors and constructing stable RGB-D multi-modal input pairs. Second, a parameter-shared dual-branch encoder is designed to extract hierarchical multi-scale pyramid features from RGB images and depth maps respectively, ensuring the consistency of feature extraction and reducing redundant parameters. Third, a multi-modality Mamba fusion module (M3FM) based on the state space model is proposed to achieve bidirectional reciprocal feature fusion between RGB and depth modalities. This module integrates depth-wise separable convolution, 2D selective scanning (SS2D), and bidirectional 2D selective scanning (Bi-SS2D), which can model long-range global dependencies with linear complexity, break through the bottlenecks of traditional CNNs and Transformers, and fully mine complementary information between modalities. Fourth, a dual-directional context mixture convolution module (DCM-Conv) based on multi-kernel asymmetric convolution is constructed for the decoder stage. By channel splitting, cascaded vertical and horizontal asymmetric depth-wise separable convolutions, and channel mixing operations, this module extracts multi-receptive-field contextual features while effectively controlling the number of parameters and computational costs. On this basis, a progressive multi-scale decoder is built to fuse adjacent-scale features layer by layer and gradually output the final camouflaged object prediction mask. Finally, a hybrid loss function combining binary cross-entropy (BCE) loss, IoU loss, and structural similarity (SSIM) loss is adopted to jointly optimize the model from pixel accuracy, global structure, and boundary integrity. **Result** Comprehensive experiments are conducted on three challenging and widely used COD benchmark datasets: CAMO, COD10K, and NC4K. The proposed MambaCOD is compared with 11 state-of-the-art methods, including 6 RGB-based models and 5 RGB-D-based models. Quantitative results show that MambaCOD achieves the optimal performance on most key evaluation metrics, including structural measure (Sm), enhanced alignment measure (Em), weighted F-measure (wFm), and mean absolute error (MAE). Specifically, compared with the second-best method, the proposed method reduces the mean absolute error by 21.3%, 17.4%, and 12.5% on the three datasets respectively, and achieves the best values in all major metrics. Efficiency analysis indicates that the model has only 58.5M parameters and 47.6G FLOPs, which are significantly lower than most comparison methods; compared with FSPNet, the parameter quantity is reduced by 78.6%, and compared with the Samba model, FLOPs are decreased by 4.0%, achieving an excellent balance between accuracy and efficiency. Visual effect analysis demonstrates that MambaCOD generates segmentation masks highly consistent with ground truth, accurately restores the contours and fine details of camouflaged objects, effectively distinguishes targets from highly similar backgrounds, reduces background noise and false detections, and maintains complete segmentation for irregularly shaped and highly concealed targets. Ablation experiments verify that each core component, including M3FM and DCM-Conv, contributes effectively to performance improvement. Further experiments confirm that Depth Anything V2 provides higher-quality geometric priors than traditional depth generators such as DPT, and the proposed modules still maintain effectiveness under different backbone networks. **Conclusion** This paper proposes a lightweight and high-performance RGB-D camouflaged object detection framework MambaCOD based on a multi-

modal state space model. By introducing high-quality pseudo-depth maps generated by Depth Anything V2, the model enriches the geometric structure information of input features and strengthens the multi-modal fusion between RGB appearance and depth geometry. The designed M3FM module realizes efficient bidirectional cross-modal fusion based on Mamba, which effectively captures long-range dependencies with linear complexity and breaks through the limitations of traditional CNNs and Transformers. The DCM-Conv module constructs a high-efficiency multi-scale decoder through asymmetric multi-kernel convolution, further improving detection accuracy while controlling computational costs. Experimental results on multiple benchmark datasets show that the proposed method outperforms existing mainstream approaches in both detection performance and computational efficiency, achieving state-of-the-art COD results. The proposed method effectively solves the problems of insufficient feature representation, low cross-modal fusion efficiency, and unbalanced precision and efficiency in traditional methods, providing a new solution for high-precision and high-efficiency camouflaged object detection in complex scenes. In the future, we will focus on enhancing the model's robustness to noisy depth inputs, extending the framework to video camouflaged object detection, and exploring lightweight deployment on resource-constrained edge devices to expand the practical application scope of the model.

Key words: camouflage target detection; RGB-D; state space model; multimodal fusion; depth features

论文引用格式:[DOI:10.11834/jig.260176]

0 引　言

伪装目标检测(camouflaged object detection, COD)是计算机视觉领域一项极具挑战性的重要任务,其目标是从复杂背景中精确识别和分割那些颜色、纹理或形态与周围环境高度相似的目标。这类目标在自然环境中常通过进化机制完美融入背景,如变色龙、枯叶蝶、比目鱼等生物,以及各类军事伪装装备和设施,使得即使人类视觉系统也难以快速准确地定位它们。正因如此,自动化、高精度的伪装目标检测技术在多个国计民生领域展现出巨大的应用价值和广阔前景:在精准农业中,它可用于早期病虫害监测,从作物叶片背景中识别伪装性害虫(Zhang和Lv,2024);在医学影像分析中,它有助于从周围组织中分割早期病变区域(Xiao等,2024),如息肉、肿瘤等;在生态保护中,它能够实现野生动物种群监测与行为研究,特别是对具有保护色的珍稀物种(Li等,2023);在安防军事领域(Lv等,2023),它对于战场侦察、伪装目标识别具有不可替代的战略意义。

近年来,深度学习技术的飞速发展,尤其是卷积神经网络(convolutional neural networks, CNN)(He等,2016)和视觉Transformer(Liu等,2021)的兴起,显著推动了伪装目标检测领域的进步。基于CNN的方法通过平移不变性的卷积操作,具备局部感受野和层次化特征提取能力,能够有效捕捉目标的边缘、纹理等局部细节信息。代表性工作如SINet、PraNet等通过设计特定的感受野模块和边缘感知策略,在COD(Fan等,2020)任务上取得了突破性进展。然而,如图1(a)所示,CNN固有的局部性限制使其难以获取区分伪装目标与复杂背景所必需的全局上下文信息和长距离依赖关系——当目标与背景具有极其相似的纹理模式时,仅依靠局部特征往往导致误检和漏检。相比之下,视觉Transformer通过自注意力机制能够建模全局长距离依赖关系,捕捉全局语义信息,为COD任务提供了新的解决思路。代表性工作如高分辨率迭代反馈网络(high-resolution iterative feedback network, HitNet)(Hu等,2023)、不确定性引导Transformer推理(uncertainty-guided transformer reasoning, UGTR)(Yang等,2021)等通过引入自注意力机制,在COD任务上取得了优于CNN的性能,但其二次计算复杂度限制了实际应用效率。如图1(b)所示,Transformer的二次计算复杂度使其在处理高分辨率特征图时面临巨大的计算开销和显存负担,难以兼顾精度与效率。此外,现有主流方法大多仅利用单模态RGB图像数据,忽视了多模态信息融合可能带来的性能突破。

深度图像作为一种重要的视觉模态,蕴含丰富的三维几何结构和空间布局先验,能够为伪装目标检测提供关键的补充线索。RGB-D伪装目标检测(RGB-D COD)旨在同时利用RGB图像和深度图进行伪装目标分割,是COD任务的重要扩展方向。通过融合外观信息与几何先验,RGB-D COD有望在高度伪装场景中取得优于单模态方法的检测性能。研

究表明，即使伪装目标在RGB空间中与背景完美融合，其与背景之间往往仍存在细微的深度差异——这种差异可能源于物体与背景平面的微小高度差、空间位置的前后关系，或是物体自身厚度的变化。如图1(d)所示，这些深度线索在RGB图像中难以察觉，但在深度图像中可能变得相对明显。因此，如何有效融合RGB和深度信息，挖掘跨模态互补特征，构建高效的多模态伪装目标检测模型，是一个值得深入探索的研究领域。

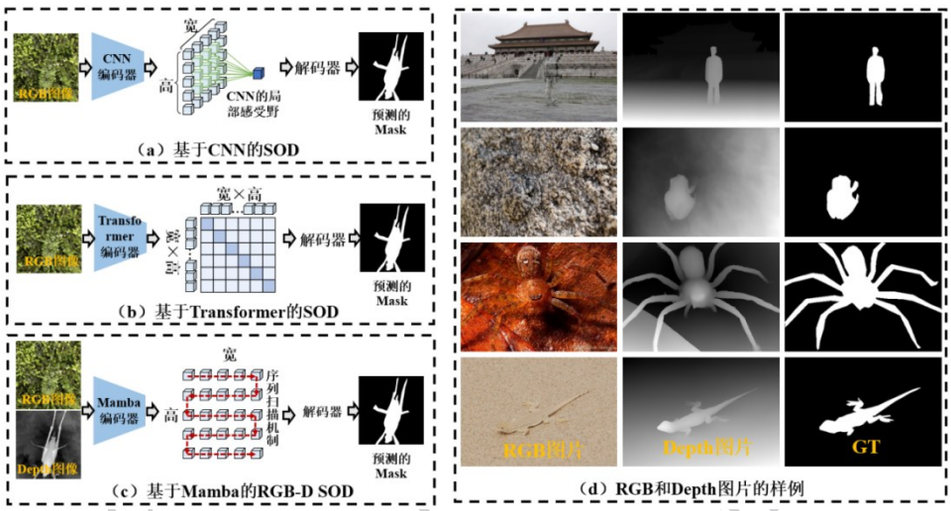

图1　本文研究动机示意图

Fig. 1　Schematic diagram of research motivation in this paper((a) CNN-based SOD;(b) Transformer-based SOD;(c) Mamba-based RGB-D SOD;(d) examples of RGB and depth images)

为应对上述挑战，本文提出一种状态空间模型引导多模态融合的RGB-D伪装目标检测方法，简称为MambaCOD。该方法的核心创新在于：充分利用深度几何先验增强模型对伪装目标的感知判别能力，同时通过高效的状态空间序列建模技术实现多模态特征的充分融合与交互。具体而言，本文主要贡献包括以下四个方面：首先，针对现有RGB-D数据集缺乏真实深度图的实际困境，本文采用先进的视觉基础模型Depth Anything V2(Yang等，2024)生成高质量的伪深度图，为RGB图像补充可靠的深度模态信息，构建高质量的RGB-D多模态输入对。该策略不仅解决了深度数据获取成本高的问题，还通过大规模预训练模型保证了生成深度图的质量和泛化能力。其次，本文创新性地引入基于状态空间模型的跨模态特征融合模块——双向跨模态扫描融合模块。该模块突破传统特征融合方式的局限，通过设计双向扫描路径，实现RGB→Depth和Depth→RGB两个方向的特征交互与互惠增强。得益于Mamba的线性复杂度特性，该模块能够在保持高效计算的同时，充分建模跨模态的长距离依赖关系，有效克服Transformer在特征融合时的计算瓶颈。最后，设计基于非对称多核卷积的多尺度特征解码器——多感受野上下文混合卷积模块。该模块采用多个不同尺寸的非对称卷积核(如1×3、3×1、1×5、5×1等组合)并行提取丰富的多感受野上下文特征，在显著减少参数量的同时保持强大的特征表达能力。在此基础上，构建层级递进的特征解码器，通过逐层融合邻接尺度特征，逐步将通道维度压缩至最终的伪装目标预测掩码。本文在四个公开的COD基准数据集(CAMO、COD10K、NC4K)上进行了全面的实验验证，证明了本文提出的MambaCOD方法在各项评价指标上均超越现有11种主流方法，达到先进的检测性能。消融实验进一步验证了各核心模块的有效性，可视化分析直观展示了方法在复杂场景下的优越性。

2　相关工作

2.1　伪装目标检测

伪装目标检测旨在从复杂背景中精准识别那些颜色、纹理或形态与周围环境高度融合的目标，是计算机视觉领域一项极具挑战性的任务。随着深度学

习技术的蓬勃发展，特别是卷积神经网络和视觉Transformer的兴起，COD领域取得了突破性进展。自2019年以来，研究者们提出了大量基于深度学习的COD模型。早期基于CNN的方法，如搜索识别网络（search identification network，SINet）（Fan等，2020）、并行反向注意力网络（parallel reverse attention network，PraNet）（Fan等，2020）等，通过设计特定的感受野模块（如空洞卷积、金字塔池化）来扩大特征提取的范围，并结合边缘感知策略增强目标边界的清晰度，取得了显著成效。然而，CNN固有的局部连接特性限制了其捕获全局上下文信息的能力，而全局语义理解对于区分与背景高度相似的伪装目标至关重要。为克服这一局限，视觉Transformer被引入COD任务。基于Transformer的方法，如HitNet（Hu等，2023）、UGTR（Yang等，2021）等，利用自注意力机制有效建模了图像中的长距离依赖关系，能够更好地理解场景上下文。但Transformer的二次计算复杂度使其在处理高分辨率特征图时面临巨大的计算开销和显存负担，难以在精度和效率之间取得平衡。近期研究也开始探索无监督/弱监督学习、以及利用大规模视觉语言模型等新型策略来解决数据稀缺和模型泛化能力的问题。

2.2 RGB-D伪装目标检测

RGB-D伪装目标检测（RGB-D COD）是近年来兴起的研究方向，其目标是在RGB图像的基础上，引入深度图作为辅助模态，利用几何结构信息增强对伪装目标的感知能力。与单模态COD相比，RGB-D COD更侧重于跨模态信息的互补与融合。现有代表性工作包括深度辅助伪装目标检测（depth-aided camouflaged object detection，DaCOD）（Wang等，2023；赖杰等，2024）、深度感知Segment Anything模型（segment anything model with depth perception，DSAM）（Yu等，2024）等。尽管基于RGB图像的COD方法已取得长足进步，但单模态信息固有的局限性依然存在——当伪装目标与背景在视觉外观上近乎完美融合时，仅依靠RGB线索往往难以可靠检测。深度图像作为RGB模态的重要补充，蕴含了丰富的三维几何结构、空间布局和表面朝向等先验信息。伪装目标与背景之间即便在颜色纹理上高度一致，也可能存在由物体厚度、前后遮挡关系或与背景平面的微小高度差所导致的深度突变。因此，将深度模态引入COD任务，通过挖掘RGB-D数据的互补性来增强模型的空间感知和判别能力，已成为当前的研究热点。并且深度图像在显著性目标检测领域也收到了广泛关注。（宋霄罡等，2025；叶欣悦等，2024）。

现有RGB-D COD方法的研究重点主要围绕两个核心问题：高质量的深度图获取与跨模态特征融合策略。在深度图获取方面，由于大规模RGB-D伪装数据集标注成本高昂，许多研究采用单目深度估计模型为现有RGB COD数据集生成伪深度图。例如，基于混合数据集的单目深度估计方法（mixed datasets for monocular depth estimation，MiDaS）（Ranftl等，2020）、密集预测Transformer（dense prediction transformer，DPT）（Ranftl等，2021）等模型被广泛用于此目的，而最新发布的Depth Anything V2（Yang等，2024）等视觉基础模型，凭借其强大的泛化能力和高质量的深度预测结果，为RGB-D COD（Wang等，2023）研究提供了更可靠的数据支持。在跨模态融合方面，研究者们提出了多样化的网络架构。Wang等人（2023）提出了多模态协同学习模块，旨在通过混合骨干网络从RGB和深度通道中协同学习深度特征。Liu等人（2024）引入了一种深度加权交叉的注意力融合模块，用于动态调整深度和RGB特征图上的融合权重。

2.3 状态空间模型

状态空间模型作为一种描述动态系统的基本数学框架，其历史可追溯至20世纪60年代的卡尔曼滤波器。近年来，随着深度学习技术的发展，状态空间模型（state space model，SSM）在序列建模领域重新焕发生机。其优势在于能够在保持线性计算复杂度的同时有效捕获长距离依赖关系。最具有代表性的工作是Mamba（Gu和Dao，2024）。Mamba在视觉任务中的扩展应用已成为近期研究的热点。研究者们针对图像数据的二维结构特点，提出了多种改进方案。例如，视觉Mamba模型（Vision Mamba，ViM）（Zhu等，2024）和VMamba（Liu等，2024）通过设计不同的扫描策略，将一维序列建模能力拓展至二维空间，使模型能够全面感知图像的局部和全局上下文信息。在医学图像处理领域，U形Mamba网络（U-shaped Mamba，U-Mamba）（Ma等，2024）、Mamba-UNet（Wang等，2024）、分割Mamba模型（Segmentation Mamba，SegMamba）（Xing等，2025）等工作将Mamba与U-Net架构相结合，分别在2D/3D医学图像

分割、图像配准、图像融合等任务中取得了优越性能。

3 本文方法

3.1 整体网络结构

为有效解决CNN和视觉Transformer在SOD领域的瓶颈,并提升模型对伪装目标的空间感能力,本文提出了状态空间模型引导多模态融合的RGB-D伪装目标检测网络,如图2所示。整体遵循编码器-跨模态融合-多尺度特征解码的端到端架构,核心由双分支编码器、跨模态状态空间融合模块与多尺度特征解码器三部分构成,整体流程如下:

首先,将RGB图像$I_r \in \mathbb{R}^{3\times H\times W}$与深度图像$I_d \in \mathbb{R}^{1\times H\times W}$作为模型双输入模态,其中H和W分为表示输入图片的宽和高。通过参数共享的双分支编器分别提取RGB模态特征$\{F_i^r\}_{i=1}^4 \in \mathbb{R}^{C_i\times H/2^{i+1}\times w/2^{i+1}}$与深度模态特征,$\{F_i^d\}_{i=1}^4 \in \mathbb{R}^{C_i\times H/2^{i+1}\times W/2^{i+1}}$实现单模态深层语义特征与空间细节特征的初步挖掘;其次,引入多模态状态空间融合模块(multi-modality mamba fusion module, M3FM),通过跨模态的状态空间融合与双向的状态空间融合等组件,完成RGB与深度模态特征的跨模态交互、互补与融合,生成多模态融合特征$\{f_i\}_{i=1}^4 \in \mathbb{R}^{C_i\times H/2^{i+1}\times W/2^{i+1}}$,解决单一模态特征在伪装场景下的表征局限性;最后,设计多尺度特征解码器,通过逐层堆叠双向混合上下文卷积(dual-directional context mixture convolution, DCM-Conv)模块实现融合特征的多尺度解码与上采样,逐步恢复目标空间细节,最终输出与输入尺寸一致的伪装目标预测掩码$\boldsymbol{M}_{pre}$。

整个网络通过参数共享的编码器保证双模态特征提取的一致性,依托跨模态状态空间融合模块挖掘模态间互补信息,再经多尺度解码器实现精细的伪装目标分割,为RGB-D伪装目标检测任务提供鲁棒的多模态特征表征与精准的分割结果。

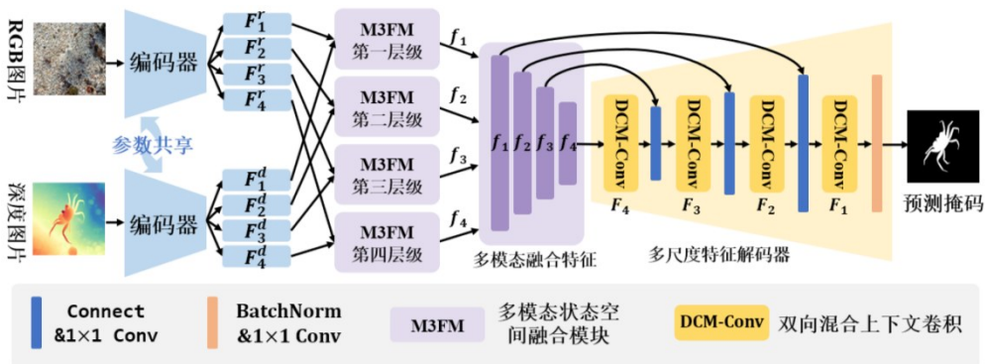

图2 状态空间模型引导多模态融合的RGB-D伪装目标检测网络结构图

Fig. 2 Structural diagram of RGB-D camouflage target detection network based on multimodal state space model

3.2 多模态状态空间融合模块

为了解决单一模态在伪装场景下表征能力受限的问题,本文提出了一种基于状态空间模型(SSM)的多模态状态空间融合模块(M3FM)。该模块旨在通过跨模态交互和双向特征扫描,充分挖掘RGB和深度模态之间的互补信息,从而生成判别力更强的多模态融合特征。如图3(a)所示,M3FM模块主要由深度可分离卷积(depthwise separable convolution, DWConv)、二维选择性扫描(2D selective scan, SS2D)、多层感知机(multi-layer perceptron, MLP)以及双向二维选择性扫描(bidirectional 2D selective scan, Bi-SS2D)等核心组件构成。

M3FM模块接收来自编码器的RGB特征$\boldsymbol{F}_i^r$和深度特征$\boldsymbol{F}_i^d$作为输入。首先,两个模态的特征分别通过一个深度可分离卷积层(DWConv),在减少参数量的同时初步提取局部空间特征。随后,为了捕捉长距离依赖关系,卷积后的特征被送入二维选择性扫描(SS2D)单元。在SS2D单元中,特征首先经过一个MLP层进行线性变换,然后通过一个卷积层进一步提取特征。接着,利用SS2D机制对特征进行处理。SS2D机制如图3(b)所示,它通过对特征图进行不同方向的扫描(如水平、垂直、对角线等),将二维

图像特征转换为一维序列，利用SSM的序列建模能力捕捉全局上下文信息。扫描后的序列通过Mamba S6(selective scan state-space module)进行处理，最后通过交叉合并(cross-merge)操作重新组合成二维特征图。

经过SS2D处理后的RGB和深度特征再次通过一个MLP层，并进行逐元素相加融合。融合后的特征被视为一个整体，送入双向二维选择性扫描(Bi-SS2D)模块进行深度的跨模态的双向交互。最后，Bi-SS2D模块的输出与经过MLP处理后的原始RGB和深度特征分别进行融合，生成最终的多模态融合特征f_i。

SS2D是M3FM模块捕捉全局上下文信息的关键组件。其核心思想是将二维图像特征图转换为一维序列，利用SSM的选择性扫描机制进行高效的长范围序列建模。如图3(b)所示，SS2D的首先将输入到的多模态特征($\boldsymbol{F}_i^r$和$\boldsymbol{F}_i^d$)进行交叉交互，实现特征级的跨模态的信息传递：

$$\widetilde{\boldsymbol{F}_i^r} = \boldsymbol{F}_i^r + \boldsymbol{F}_i^r \times \boldsymbol{F}_i^d \tag{1}$$

$$\widetilde{\boldsymbol{F}_i^d} = \boldsymbol{F}_i^d + \boldsymbol{F}_i^r \times \boldsymbol{F}_i^d \tag{2}$$

式中，$\widetilde{\boldsymbol{F}_i^r}$和$\widetilde{\boldsymbol{F}_i^d}$表示经过交叉交互的RGB和深度特征。以RGB特征$\widetilde{\boldsymbol{F}_i^r}$为例，然后经过多方向扫描机制将特征$\widetilde{\boldsymbol{F}_i^r}$沿着四个不同的方向展开为一维序列：从左到右，从右到左，从上到下，从下到上，具体如图3(b)所示。

通过这种多向扫描，模型可以从不同的角度捕捉图像的空间关系。展平后的四个序列分别送入四个独立的Mamba S6块进行处理。Mamba S6模块基于选择性状态空间模型，能够有效地捕捉序列中的长距离依赖关系。Mamba S6模块内部包含线性变换层和核心的SSM单元。SSM单元通过一组可学习的参数(如A, B, C, Δ)来建模序列的动态变化。具体而言，离散化的状态空间方程如下：

$$\bar{A} = \exp(\Delta A) \tag{3}$$

$$\bar{B} = \Delta A^{-1}\left(\exp(\Delta A) - I\right)\cdot \Delta B \tag{4}$$

$$\boldsymbol{h}_t = \bar{A}\boldsymbol{h}_{t-1} + \bar{B}\boldsymbol{X}_t \tag{5}$$

$$\boldsymbol{y}_t = C\boldsymbol{h}_t \tag{6}$$

式中，$\boldsymbol{X}_t$是输入序列，在本文中指代$\widetilde{\boldsymbol{F}_i^r}$和$\widetilde{\boldsymbol{F}_i^d}$经过四个不同方向扫描后的序列化特征，$\boldsymbol{h}_t$是隐藏状态，$y_t$是输出序列。通过这种方式，Mamba S6块能够有效地捕捉长距离依赖关系，。

四个Mamba S6块的输出序列y^1, y^2, y^3和y^4被重新组合并还原为二维特征图。具体而言，首先将每个序列还原为二维形状$\boldsymbol{Y}^{(k)}\ \mathbb{R}^{C_i \times H/2^{i+1} \times W/2^{i+1}}$，然后将它们进行逐元素相加：

$$\boldsymbol{Y}_E = \sum_{k=1}^{4} \boldsymbol{Y}^{(k)} \tag{7}$$

式中，$\boldsymbol{Y}_E$作为SS2D单元的输出。在本文中，$\boldsymbol{Y}_E$可以指代图3(a)中RGB和Depth SS2D的最终输出$\boldsymbol{Y}_E^R$和$\boldsymbol{Y}_E^D$。通过SS2D机制，M3FM模块能够有效地利用SSM处理二维图像数据，捕捉全局上下文信息，从而增强模型对伪装目标的表征能力。

接下来，本文通过Bi-SS2D实现更生成的跨模态交互和双向信息通信。在上述的RGB和Depth SS2D中，M3FM实现了多模态的特征级的交互，而Bi-SS2D通过级联RGB和Depth特征去实现token-level的模态信息交互。如图1(c)所示，Bi-SS2D模块接增强后的RGB和深度特征($\boldsymbol{Y}_E^R$和$\boldsymbol{Y}_E^D$)，然后经过MLP操作和特征级拼接得到Bi-SS2D的输入特征：

$$\boldsymbol{F}_i^M = CT\left(MP\left(\boldsymbol{Y}_E^R\right), ML\left(\boldsymbol{Y}_E^D\right)\right) \tag{8}$$

式中，$\boldsymbol{F}_i^M$表示Bi-SS2D的输入特征。CT和MP分别表示特征拼接操作和MLP操作。

随后，输出的$\boldsymbol{F}_i^M$分别与RGB和深度特征进行相乘从而得到增强后的RGB特征与深度特征，并经过MLP操作后进行相加，从而得到M3FM模块的输出特征f_i。

Bi-SS2D模块包含两个并行的扫描分支：前向扫描(forward scanning)和后向扫描(backward scanning)。在每个分支中，RGB和深度特征首先被展平成序列，然后沿着相反的方向进行扫描。这种双向扫描机制能够同时捕捉序列的前向和后向上下文信息，从而更全面地理解图像的空间结构。前向和后向扫描分支的输出通过MLP层进行融合，最后通过cross-merge操作重新组合成二维特征图。Bi-SS2D模块通过双向扫描和SSM建模，增强了模型对伪装目标的空间感能力，有效地克服了传统CNN和视觉Transformer在处理复杂空间关系时的局限性。

3.3 双向混合上下文卷积

近年来，为了追赶Transformer在捕获长距离依赖关系方面的优势，许多先进的CNN模型开始尝试

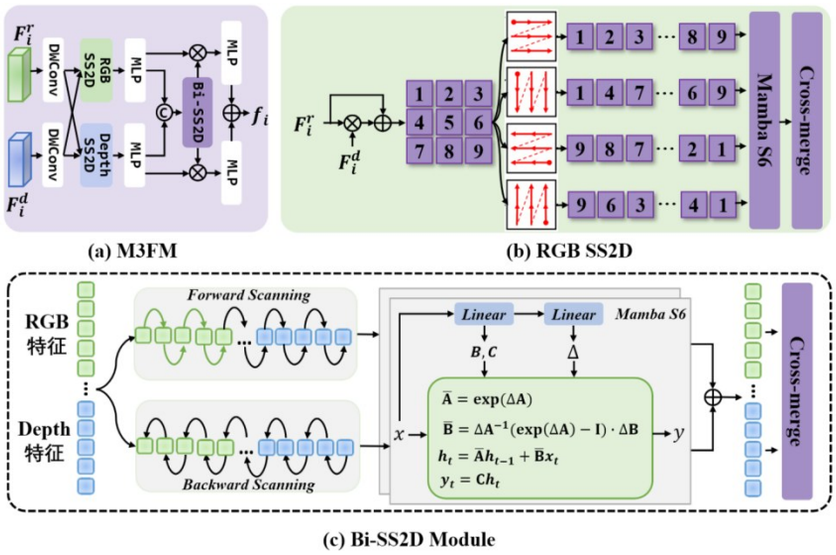

图3　多模态状态空间融合模块

Fig. 3　Multimodal state space fusion module((a)M3FM;(b)RGB SS2D;(c)Bi-SS2D Module)

使用大尺寸卷积核来扩大感受野,比如RepLKNet和PeLK。然而,直接使用大卷积核(如7×7, 9×9)会带来参数量和计算量的增长,导致模型难以训练且推理速度缓慢。为了在扩大感受野和控制参数量之间取得平衡,本文提出了一种双向混合上下文卷积模块(DCM-Conv)。如图4所示,DCM-Conv模块采用"分而治之"的策略,通过特征拆分、双向非对称深度可分离卷积、通道混合以及特征拼接等步骤,实现了高效的多尺度特征提取和多向上下文信息融合,在保持较低计算复杂度的同时,获得了具有大感受野的丰富上下文特征。

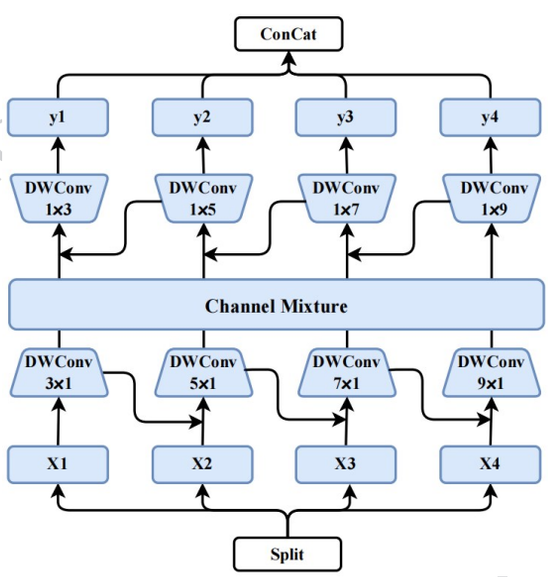

图4　双向混合上下文卷积

Fig. 4　Dual-directional Context Mixture Convolution

DCM-Conv作为多尺度特征解码器的核心部件,被使用去增强解码器特征的大感受野的上下文特征增强,其输入特征为融合后的多模态特征f_i。本文以f_4作为例子去描述整个DCM-Conv的特征流。首先,输入特征f_4沿着通道维度被均匀拆分为四个子特征图$\boldsymbol{X1}$, $\boldsymbol{X2}$, $\boldsymbol{X3}$,和$\boldsymbol{X4}$,每个子特征图的维度为$\mathbb{R}^{C_4 \times H/32 \times W/32}$。这一步旨在降低后续卷积操作的计算复杂度,并为多尺度特征提取奠定基础:

$$\boldsymbol{X1}, \boldsymbol{X2}, \boldsymbol{X3}, \boldsymbol{X4} = Split(f_4) \tag{9}$$

拆分后的子特征图分别进入两个并行的分支进行处理:垂直分支($n \times 1$的非对称卷积)和水平分支($1 \times n$的非对称卷积)。这两个分支利用不同尺寸的非对称深度可分离卷积(DWConv)来捕获多尺度的空间上下文信息。具体而言,子特征图$\boldsymbol{X1}$, $\boldsymbol{X2}$, $\boldsymbol{X3}$,和$\boldsymbol{X4}$分别通过尺寸3×1, 5×1, 7×1, 9×1的垂直非对称深度可分离卷积层。并且,垂直分支采用了级联结构,即前一个卷积层的输出与当前分支的输入特征图进行逐元素相加后,再送入下一个卷积层,这种级联结构有助于增强特征的流动和复用:

$$\boldsymbol{X1}' = DC_{3\times1}(\boldsymbol{X1}) \tag{10}$$

$$\boldsymbol{X2}' = DC_{5\times1}(\boldsymbol{X2} + \boldsymbol{X1}') \tag{11}$$

$$\boldsymbol{X3}' = DC_{7\times1}(\boldsymbol{X3} + \boldsymbol{X2}') \tag{12}$$

$$\boldsymbol{X4}' = DC_{9\times1}(\boldsymbol{X4} + \boldsymbol{X3}') \tag{13}$$

式中,$\boldsymbol{X1}'$,$\boldsymbol{X2}'$,$\boldsymbol{X3}'$和$\boldsymbol{X4}'$表示经过垂直分支的特征提取之后的特征。在垂直分支和水平分支之间,引

入了一个通道混合层。$DC_{k\times1}$表示一个卷积核为$k\times1$的DWConv。该层旨在促进不同子特征图之间的信息交流，打破通道间的壁垒，增强特征的判别力：

$$Z1,Z2,Z3,Z4 = CM(X1',X2',X3',X4') \tag{14}$$

式中，CM表示通道混合操作，由一个1×1的普通卷积核通道注意力构成，$Z1$, $Z2$, $Z3$, 和$Z4$表示通道混合后的四个子特征。

水平分支与垂直分支一样采用了级联操作，不过不同的是水平分支的级联操作是从大核的非对称卷积到小核的非对称卷积，目的是为了增强不同子特征的上下文感受野信息量：

$$y4 = DC_{1\times9}(Z4) \tag{15}$$

$$y3 = DC_{1\times7}(Z3 + y4) \tag{16}$$

$$y2 = DC_{1\times5}(Z2 + y3) \tag{17}$$

$$y1 = DC_{1\times3}(Z1 + y2) \tag{18}$$

式中，y1，y2，y3和y4表示经过水平分支交互后的四个子特征，然后经过特征拼接操作得到DCM-Conv的输出特征。$DC_{1\times k}$表示一个卷积核为$1\times k$的DWConv。

两种不同的级联顺序在垂直和水平方向上形成了互补。垂直分支倾向于“由局部到全局”地构建特征，而水平分支倾向于“由全局到局部”地精炼特征。这种双向、多尺度的特征构建和精炼过程，使得DCM-Conv模块能够生成更全面、更判别力的上下文特征，从而有效提升伪装目标检测的性能。

3.4 损失函数

本文采用二元交叉熵损失(binary cross entropy loss, BCE Loss)、交并比损失(intersection over union loss, IoU Loss)和结构相似性损失(structural similarity loss, SSIM Loss)来联合训练伪装目标检测的模型。BCE损失广泛应用于二分类和图像分割任务中。IoU损失侧重于关注较大的前景区域，以提升对大尺寸目标的分割性能。SSIM损失则用于衡量预测结果与真实标签之间的结构相似性。本文结合这三种不同损失函数的优势，构建了一个混合损失函数L_H：

$$L_H = L_B(M_{pre}, G) + L_I(M_{pre}, G) + L_S(M_{pre}, G) \tag{19}$$

式中，M_{pre}和G分别表示预测的伪装目标的掩码和对应的标签(ground truth)。L_B，L_I和L_S分别表示BCE损失、IoU损失和SSIM损失。

4 实 验

4.1 数据集和评估指标

遵循先前的研究工作(Zhou等，2014；Cong等，2023)，本文使用三个基准数据集来评估提出的MambaCOD模型，包括伪装目标数据集(camouflaged object dataset, CAMO)(Le等，2019)、野外伪装目标检测数据集(camouflaged object detection in the wild dataset, COD10K)(Fan等，2020)，和自然伪装4K数据集(natural camouflaged 4K dataset, NC4K)(Lv等，2023)。具体地，本文从CAMO中选取1000个样本，从COD10K中选取3040个样本作为训练集来训练模型；CAMO、COD10K和NC4K的剩余样本则作为测试集用于评估模型性能。为了保证公平比较，本文采用通用的评估指标来衡量不同方法的性能，包括结构度量(Sm)、增强对齐度量(Em)、加权F度量(wFm)和平均绝对误差(MAE)。

4.2 实验细节

本文遵循大多数视觉模型的惯例，采用在ImageNet(Krizhevsky等，2017)上预训练的主干网络作为特征编码器。本文使用视觉VMamba-S模型(vision Mamba-small, VMamba-S)作为主干网络来验证MambaCOD的有效性。所有实验均在配备24G显存的RTX 4090 GPU上进行，模型训练80个轮次。采用AdamW优化器，权重衰减设为0.1，初始学习率设为5×10^{-5}，并在40个轮次后衰减为原来的十分之一。所有训练和测试图像的大小都被调整为384×384，批量大小设为4。与先前使用DPT生成深度图的研究工作不同，本文引入Depth Anything V2来生成高质量的深度图。需要说明的是，本文采用Depth Anything V2生成伪深度图，而部分对比方法在原论文中使用DPT生成深度图。为保证公平性，本文在消融实验(表8)中额外验证了本文方法在DPT深度图输入下的性能，结果表明即使在相同深度图条件下，MambaCOD仍优于现有RGB-D方法，比如Samba。

4.3 性能对比

本文在三个公开基准数据集(CAMO、COD10K和NC4K)上，将所提出的MambaCOD模型与11个主流COD模型进行了全面的定量性能对比，其中包括6个基于RGB的COD模型：特征收缩金字塔网络

(feature shrinkage pyramid network, FSPNet)(Huang等,2023)、视觉显著与伪装目标检测网络(visual salient and camouflaged object detection, VSCode)(Luo等,2024)、频域-空间纠缠学习网络(frequency-spatial entanglement learning, FSEL)(Sun等,2024)、多上下文细化网络(multi-context refinement network, MCRNet)(Zhang等,2025)、边缘语义协同网络(edge-semantic collaborative network, ESCNet)(Ye等,2025)、显著性Mamba模型(saliency Mamba, Samba)(He等,2025))和5个基于RGB-D的COD模型:渐进优化与块尺度网络progressively optimized and patch-scale network, PopNet)(Wu等,2023)、深度辅助伪装目标检测网络(depth-aided camouflaged object detection, DaCOD)(Wang等,2023)、RISNet模型(Wang等,2024)、深度感知Segment Anything模型(depth-aware segment anything model, DSAM)(Yu等,2024)、双流适配器(dual stream adapters, DSA)(Liu等,2025)、Samba(He等,2025)。为公平比较,所有模型均采用其官方发布的最优结果或在其默认设置下复现的结果。本文采用四个通用评估指标来衡量模型性能:Sm、Em、wFm和MAE。

1)定量分析

MambaCOD与基于RGB的COD模型对比:如表1所示,在CAMO数据集上,MambaCOD在所有四个指标上均取得了最优性能,Sm、Em、wFm分别达到0.891、0.945和0.859,MAE低至0.037。相较于性能次优的MCRNet和FSEL,本文的方法在MAE指标上降低了7.5%。在更具挑战性的COD10K数据集上,MambaCOD模型同样表现优异。相较于近期提出的Samba,MambaCOD在wFm上提升了2.7%,MAE降低了13.6%。在NC4K数据集上,本文模型取得了与最优模型相当的性能。这些结果表明,本文所提的MambaCOD模型在复杂场景下的伪装目标检测任务中具有显著优势。

MambaCOD与基于RGB-D的COD模型对比:相比于仅使用RGB模态的模型,基于RGB-D的模型额外利用了深度信息来辅助检测。MambaCOD模型在三个数据集上的多数指标上仍超越了所有基于RGB-D的对比模型。例如,在CAMO数据集上,相较于性能最优的RGB-D模型DSA,本文的方法在Sm、wFm上分别提升了2.9%和2.4%,MAE降低了21.3%。这充分证明了本文的MambaCOD模型能够有效挖掘RGB图像和深度图像中的深层互补线索。

2)效率分析

为评估模型的计算效率,本文在表2中对比了不同COD模型的参数量(params)和计算复杂度(floating point operations, FLOPs)。所有模型的输入图像分辨率均统一为384×384。如表2所示,MambaCOD模型参数量仅为58.5M,显著低于大多数对比模型。具体而言,相较于FSPNet,本文的模型参数量减少了78.6%;相较于同为基于状态空间模型的Samba(61.7M),本文的模型参数量仍降低了5.2%。这表明本文的网络设计更加紧凑高效,能够有效控制模型规模,便于在资源受限的边缘设备上部署。在FLOPs方面,MambaCOD模型同样表现出色,仅需47.6G的计算量,在所有对比模型中是最低的。相较于同样高效的Samba(49.6G),本文的FLOPs仍降低了4.0%。

结合表1和表2的结果可以看出,MambaCOD模型在实现最优测性能的同时,保持了最低的参数量和计算复杂度。此外,相比较使用分割基础模型(Segment Anything Model, SAM)的方法,比如DSAM,Mamba COD使用少得多的参数和计算量取得更好的结果,证明了本文方法的有效性和较高的计算效率。

综上所述,MambaCOD为伪装目标检测任务提供了一种既高效又精准的全新解决方案,在实际应用中具有巨大的潜力。

3)视觉效果分析

为直观展示模型的分割效果,本文选取典型伪装场景样本进行视觉对比,结果如图5所示。

从视觉结果可以看出,本文模型的分割掩码与GT高度一致,能精准还原伪装目标的轮廓与边缘(如蜻蜓的翅脉、鱼类的体态),而FSEL、Samba,DSAM,DaCOD等模型常出现边缘模糊、细节丢失的问题。在高相似性背景(如石缝、沙地)中,本文模型能有效区分伪装目标与背景,避免将背景误检为目标(如第五行样本中,其他模型均出现明显背景噪点,而本文模型输出干净掩码)。对于形态不规则、纹理高度融合的伪装目标(如第一行的动物、第六行的鱼类),本文模型仍能保持完整分割,而PopNet、MCRNet等模型易出现目标断裂、缺失的情况。综上,视觉对比结果进一步验证了本文模型在伪装目标检测任务上的优越性,不仅定量指标领先,定性分

割效果也更精准、完整。

表1　MambaCOD与其他11个COD模型在CAMO，COD10K和NC4K数据集上的定量性能对比

Table 1　Quantitative performance comparison of MambaCOD with 11 other COD models on CAMO, COD10K, and NC4K datasets

模型	出版	CAMO				COD10K				NC4K			
		Sm	Em	wFm	MAE	Sm	Em	wFm	MAE	Sm	Em	wFm	MAE
基于RGB的COD模型													
FSPNet	CVPR23	0.856	0.877	0.738	0.071	0.851	0.901	0.716	0.032	0.897	0.913	0.789	0.044
VSCode	CVPR24	0.838	0.906	0.768	0.060	0.847	0.913	0.744	0.028	0.874	0.924	0.813	0.038
FSEL	ECCV24	0.885	0.942	0.851	0.040	0.873	0.928	0.800	0.021	0.892	0.941	0.853	0.030
MCRNet	IJCV25	0.886	0.942	0.849	0.040	0.874	0.931	0.807	0.021	0.893	0.943	0.853	0.030
ESCNet	ICCV25	0.871	0.934	0.843	0.044	0.873	0.934	0.804	0.021	0.892	0.941	0.859	0.029
Samba	CVPR25	0.883	0.938	0.846	0.042	0.865	0.935	0.791	0.022	0.888	0.939	0.847	0.031
基于RGB-D的COD模型													
PopNet	ICCV23	0.806	0.869	–	0.073	0.827	0.897	–	0.031	0.852	0.908	–	0.043
DaCOD	MM23	0.855	0.911	0.796	0.051	0.840	0.908	0.729	0.028	0.874	0.923	0.814	0.035
RISNet	CVPR24	0.870	0.922	0.827	0.050	0.863	0.931	0.779	0.025	0.882	0.925	0.834	0.037
DSAM	MM24	0.832	0.913	0.821	0.061	0.846	0.921	0.789	0.033	0.871	0.932	0.826	0.040
DSA	ICCV25	0.866	0.943	0.839	0.047	0.871	0.932	0.807	0.023	0.889	0.935	0.847	0.032
Samba	CVPR25	0.871	0.932	0.841	0.045	0.853	0.931	0.793	0.024	0.881	0.934	0.842	0.033
Ours	–	**0.891**	**0.945**	**0.859**	**0.037**	**0.879**	**0.939**	**0.812**	**0.019**	**0.897**	**0.944**	**0.859**	**0.028**

注：黑色字体表示最优结果。“–”表示对比方法论文没有提供该数据。

表2　不同COD模型的计算效率对比

Table 2　Comparison of computational efficiency of different COD models

模型	FSPNet	VSCode	FSEL	Samba	MCRNet	ESCNet	PopNet	DaCOD	RISNet	DSAM	Ours
参数(M)	274	74.7	67.2	61.7	66.2	98	–	267	87	324	**58.5**
FLOPs(G)	283	59.7	109.5	49.6	143.1	129	228.8	234	63	342	**47.6**

注：黑色字体表示最优结果。“–”表示对比方法论文没有提供该数据。

4.4　消融实验

为验证本文所提出各核心模块的有效性，本节设计并开展了一系列消融实验。所有实验在COD10K数据集上进行，采用与主实验相同的训练设置和评估指标。

1）MambaCOD中组件的消融实验

为深入MambaCOD中M3FM和DCM-Conv模块的作用，本文设计了逐步添加各模块的消融实验。基线模型采用RGB-D双分支编码器，但仅通过简单拼接操作融合多模态特征和3×3 DWConv替代M3FM和DCM-Conv模块。在此基础上，依次引入M3FM和DCM-Conv模块。消融实验结果如表3所示。

2）多模态状态空间融合模块的消融实验

为深入分析多模态状态空间融合模块(M3FM)各组件的作用，本文设计了逐步添加各组件的消融实验。基线模型采用RGB-D双分支编码器，但仅通过简单拼接操作融合多模态特征。在此基础上，依次引入跨模态交叉交互、SS2D模块和Bi-SS2D模块。

如表4所示，引入跨模态交叉交互后，模型性能

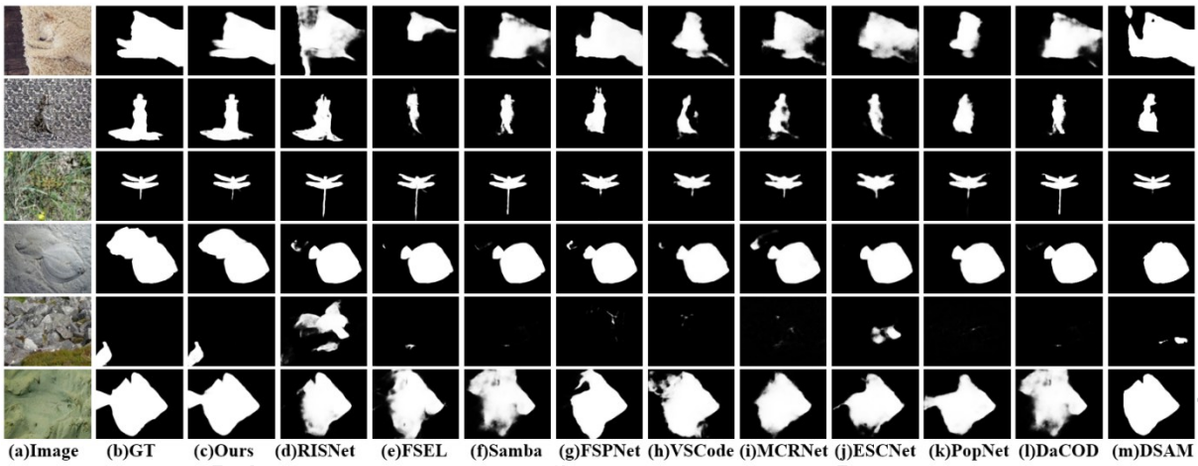

图5　不同模型的视觉效果对比：(a)图片；(b)GT；(c)本文方法结果；(d)RISNet方法结果；(e)FSEL方法结果；(f)Samba方法结果；(g)FSPNet方法结果；(h)VSCode方法结果；(i)MCRNet方法结果；(j)ESCNet方法结果；(k)PopNet方法结果；(l)DaCOD方法结果；(m)DSAM方法结果

Fig. 5　Comparison of visual effects of different models：(a) Image；(b)GT；(c)Ours；(d)RISNet；(e)FSEL；(f)Samba；(g)FSPNet；(h)VSCode；(i)MCRNet；(j)ESCNet；(k)PopNet；(l)DaCOD；(m)DSAM

表3　MambaCOD组件的消融实验

Table 3　Ablation experiment of MambaCOD component

型号	M3FM	DCM-Conv	Sm	Em	wFm	MAE
A0			0.841	0.907	0.763	0.029
A1	✔		0.863	0.921	0.790	0.024
A2		✔	0.859	0.922	0.788	0.025
MambaCOD	✔	✔	**0.879**	**0.939**	**0.812**	**0.019**

注：黑色字体表示最优结果。“✔”表示模型配置时使用了该模块。

即获得初步提升，Sm达到0.866，M降至0.022。进一步添加SS2D模块，通过全局上下文建模增强了特征表示能力，wFm提升至0.802。完整配置下（包含Bi-SS2D模块），模型在所有指标上达到最优，相较于仅使用交叉交互的版本。这一消融结果表明，M3FM模块中的各组件均对最终性能有正向贡献，其中Bi-SS2D的双向交互机制能够有效促进跨模态信息的深度融合。

图6展示了RGB-D伪装目标检测任务中，单模态与融合模态的特征可视化对比结果，从左到右依次为RGB图像、Depth图像、真值（GT）、RGB特征、Depth特征、融合特征。通过观察可以发现，单模态特征存在固有缺陷，其中RGB模态易受目标与背景颜色相似性干扰，无法有效区分伪装目标与背景；Depth模态仅能提供空间结构信息，缺乏外观语义，特征完整性不足。而融合模态通过多模态信息互补，有效整合了RGB的外观语义与Depth的空间结构，不仅精准定位了目标的完整区域，抑制了背景噪声，还提升了特征的判别性与鲁棒性，验证了多模态融合在伪装目标检测任务中的有效性与优越性。

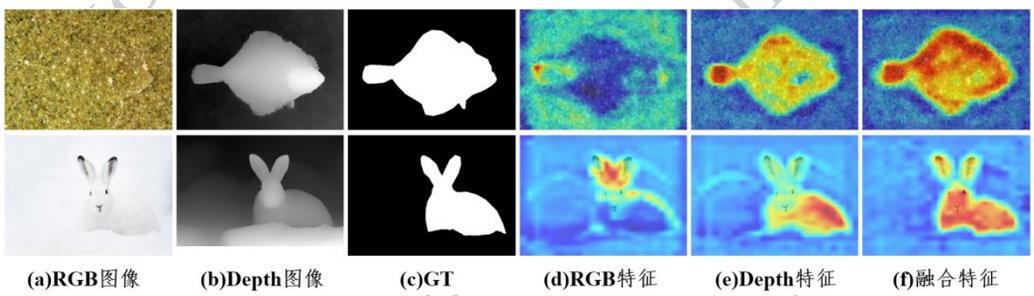

图6　特征的可视化效果对比：(a)RGB图像；(b)深度图像；(c)GT；(d)RGB特征；(e)Depth特征；(f)融合特征

Fig. 6　Comparison of feature visualization：(a)RGBimage；(b)depth map；(c)GT；(d)RGB feature；(e)Depth feature；(f)fusion feature

表4　M3FM模块各组件的消融实验

Table 4　Ablation experiment of each component of the M3FM module

型号	交叉交互	SS2D	Bi-SS2D	Sm	Em	wFm	MAE
B0				0.859	0.922	0.788	0.025
B1	✔			0.866	0.928	0.796	0.022
B2	✔	✔		0.872	0.933	0.802	0.021
Ours	✔	✔	✔	**0.879**	**0.939**	**0.812**	**0.019**

注：黑色字体表示最优结果。“✔”表示模型配置时使用了该模块。

3）双向上下文混合卷积模块的消融实验

为验证所提出的双向上下文混合卷积模块（DCM-Conv）的内部设计的有效性，本文以3×3 DWConv替代DCM-Conv作为极限模型C0，在此基础上逐步加入DCM-Conv的垂直分支，水平分支和通道混合，并分别记为C1和C2。对比结果如表5所示。

表5揭示了DCM-Conv各设计组件的有效性。首先，引入垂直分支（C1）后，模型性能相较于C0显著提升，Sm从0.863提升至0.869，wFm从0.790提升至0.799。其次，加入水平分支（C2）后，性能进一步提升。最后，引入通道混合模块（Ours）后，模型达到最优性能。这一结果表明，DCM-Conv的三个组件协同作用，共同构建了高效的多尺度上下文特征提取模块。

4）双向上下文混合卷积模块与其他卷积的对比

表5　DCM-Conv模块的消融实验

Table 5　Ablation experiment of DCM-Conv module

型号	垂直分支	水平分支	通道混合	Sm	Em	wFm	MAE
C0				0.863	0.921	0.790	0.024
C1	✔			0.869	0.929	0.799	0.022
C2	✔	✔		0.871	0.931	0.806	0.020
MambaCOD	✔	✔	✔	**0.879**	**0.939**	**0.812**	**0.019**

注：黑色字体表示最优结果。“✔”表示模型配置时使用了该模块。

为验证所提出的DCM-Conv的优越性，本文将其他广泛使用的卷积模块用于替代DCM-Conv，包括ASPP和DenseASPP模块，并在相同实验设置下对比性能。同时统计各模块的参数量和计算复杂度（FLOPs），结果如表6所示。

从表6可以看出，DCM-Conv在性能和效率之间取得了最佳平衡。具体而言，采用ASPP模块（D1）的模型参数量达74.1M，FLOPs为52.86G，但检测性能有限，Sm仅为0.852，MAE高达0.026。这表明ASPP虽然通过并行空洞卷积分支捕捉多尺度信息，但空洞卷积的网格效应可能导致局部信息丢失，且参数量较大。DenseASPP模块通过密集连接增强了特征复用能力，但同时也带来了更高的计算开销。相比之下，本文提出的DCM-Conv模块在参数量（58.5M）和FLOPs（47.5G）均为最低的情况下，取得了最优的检测性能。

表6　不同卷积的消融实验

Table 6　Ablation experiments with different convolutions

型号	卷积模块	参数（M）	FLOPs（G）	Sm	Em	wFm	MAE
D1	ASPP	74.1	52.86	0.852	0.932	0.778	0.026
D2	DenseASPP	87.4	59.1	0.865	0.927	0.803	0.021
Ours	DCM-Conv	**58.5**	**47.5**	**0.879**	**0.939**	**0.812**	**0.019**

注：黑色字体表示最优结果。

5）不同主干网络的对比

为验证主干网络对模型性能的影响，本文在不同设置下对比了MambaCOD与Samba模型的性能。主干网络包括VMamba-S和Swin Transformer。需要说明的是，该实验对比结果统一使用DPT生成的深度图像以保证对比的公平性。实验结果如表7所示。

从表7可以得出以下结论：在不同主干网络和深度图组合下，MambaCOD均优于同配置的Samba模型。这验证了MambaCOD所提出的M3FM和DCM-Conv模块相较于Samba的基线设计具有普适性的性能优势。最后，采用Swin Transformer作为主

干网络时，Samba和MambaCOD模型依旧能够达到与VMamba相似的性能，表明MambaCOD的性能并不依赖于先进的主干网络。

表7　主干网络的对比

Table 7　Comparison between different backbone network

型号	模型	VMamba	SwinT	Sm	Em	wFm	MAE
E1	Samba	✔		0.853	0.931	0.793	0.024
E2	Samba		✔	0.851	0.932	0.794	0.025
E3	MambaCOD	✔		0.866	**0.934**	**0.804**	**0.022**
E4	MambaCOD		✔	**0.868**	0.932	0.803	**0.022**

注：黑色字体表示最优结果。“✔”表示模型配置时使用了该模块。

6）对深度图生成器依赖性的讨论

为验证本文方法是否严重依赖于特定的深度图生成器（如Depth Anything V2），本文在表8中对比了使用DPT与Depth Anything V2（DA v2）两种深度图生成器时的模型性能，并同时与同期SOTA模型进行同条件对比。

首先，DA v2生成的伪深度图质量显著优于DPT。因此，无论对于MambaCOD还是其他的RGB-D COD方法，采用DA v2时均能获得比DPT更优的检测性能，这是深度图质量改善带来的自然提升。其次，更为关键的是，在相同深度图输入条件下，本文提出的MambaCOD始终优于其他的RGB-D方法。这表明，MambaCOD的性能优势主要来源于其设计的M3FM跨模态融合模块与DCM-Conv多尺度解码器，而非对某一特定深度图生成器的依赖。

综上所述，本文方法并不严重依赖于Depth Anything V2。采用DA v2是为了获得更高质量的几何先验以进一步提升性能，但即便在质量较低的深度图（如DPT生成）输入下，MambaCOD仍能显著超越同类效果最优的方法，充分体现了模型自身的结构优越性。

表8　深度图的对比

Table 8　Comparison between different depth map

模型	DPT	DA v2	Sm	Em	wFm	MAE
DaCOD	✔		0.840	0.908	0.729	0.028
RISNet	✔		0.863	0.931	0.779	0.025
Samba	✔		0.853	0.931	0.793	0.024
MambaCOD	✔		**0.866**	**0.934**	**0.804**	**0.022**
DaCOD		✔	0.867	0.922	0.769	0.025
RISNet		✔	0.871	0.939	0.804	0.022
Samba		✔	0.872	**0.942**	0.809	0.021
MambaCOD		✔	**0.879**	0.939	**0.812**	**0.019**

注：黑色字体表示最优结果。“✔”表示模型配置时使用了该模块。

5　总　结

本文针对伪装目标检测任务中单一RGB模态信息受限、CNN感受野不足以及Transformer计算复杂度高等问题，提出了一种状态空间模型引导多模态融合的RGB-D伪装目标检测方法MambaCOD。该方法通过引入深度几何先验、设计高效的多模态融合机制和轻量化的解码器模块，实现了检测精度与计算效率的良好平衡。其中，提出的基于状态空间模型的多模态融合模块M3FM，通过跨模态交叉交互和双向扫描机制，实现了RGB与深度特征的高效互惠融合，在保持线性计算复杂度的同时充分挖掘跨模态互补信息；设计的双向混合上下文卷积模块DCM-Conv，通过“分而治之”的策略在扩大感受野的同时有效控制参数量，并构建多尺度特征解码器实现伪装目标的精细化分割。在三个公开基准数据集上进行的大量实验表明，本文方法在各项评估指标上均超越11种主流方法，达到先进的检测性能。消融实验进一步验证了各核心模块的有效性。

参考文献（References）

Zhang Y and Lv C. 2024. TinySegformer: A lightweight visual segmentation model for real-time agricultural pest detection. Computers and Electronics in Agriculture, 218: 108740 [DOI: 10.1016/j.compag.2024.108740]

Xiao B, Hu J W, Li W S, Pun C M and Bi X L. 2024. CTNet: Contrastive transformer network for polyp segmentation. IEEE Transactions on Cybernetics, 54(9): 5040-5053 [DOI: 10.1109/TCYB.2024.3368154]

Li J, Xu W K, Deng L M, Xiao Y, Han Z Z and Zheng H Y. 2023. Deep learning for visual recognition and detection of aquatic animals: A review. Reviews in Aquaculture, 15(2): 409-433 [DOI: 10.1111/raq.12726]

Lv Y Q, Zhang J, Dai Y C, Li A X, Barnes N and Fan D P. 2023.

Toward deeper understanding of camouflaged object detection. IEEE Transactions on Circuits and Systems for Video Technology, 33(7): 3462-3476 [DOI: 10.1109/TCSVT.2023.3234578]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of the 33th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows// Proceedings of the 18th IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 10012-10022. [DOI: 10.1109/ICCV48922.2021.00986]

Fan D P, Ji G P, Sun G L, Cheng M M, Shen J B and Shao L. 2020. Camouflaged object detection//Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 2777-2787 [DOI: 10.1109/CVPR42600.2020.00281]

Yang L H, Kang B Y, Huang Z L, Zhao Z, Xu X G, Feng J S, et al. 2024. Depth anything v2//Proceedings of the 38th Advances in Neural Information Processing Systems, 37: 21875-21911 [DOI: 10.52202/079017-0688]

Fan D P, Ji G P, Zhou T, Chen G, Fu H Z, Shen J B, et al. 2020. Pranet: Parallel reverse attention network for polyp segmentation// Proceedings of the 23th the International Conference on Medical Image Computing and Computer-assisted Intervention. Cham: Springer International Publishing: 263-273 [DOI: 10.1007/978-3-030-59710-8_26]

Hu X B, Wang S, Qin X B, Dai H, Ren W Q, Luo D H, et al. 2023. High-resolution iterative feedback network for camouflaged object detection//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington DC: AAAI Press: 37 (1): 881-889 [DOI: 10.1609/aaai.v37i1.12847]

Yang F, Zhai Q, Li X, Huang R, Lou A, Cheng H, et al. 2021. Uncertainty-guided transformer reasoning for camouflaged object detection//Proceedings of the 18th IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 4146-4155 [DOI: 10.1109/ICCV48922.2021.00419]

Ranftl R, Lasinger K, Hafner D, Schindler K and Koltun V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3): 1623-1637[DOI: 10.1109/TPAMI.2020.3019967]

Ranftl R, Bochkovskiy A and Koltun V. 2021. Vision transformers for dense prediction//Proceedings of the 18th IEEE/CVF International Conference on Computer Vision. Chengdu: IEEE: 12179-12188 [DOI: 10.1109/ICCV51070.2021.01220]

Wang Q W, Yang J Y, Yu X S, Wang F Y, Chen P and Zheng F. 2023. Depth-aided camouflaged object detection//Proceedings of the 31st ACM International Conference on Multimedia. Montreal: ACM: 3297-3306 [DOI: 10.1145/3581783.3592145]

Liu X, Qi L, Song Y X and Wen Q. 2024. Depth awakens: A depth-perceptual attention fusion network for RGB-D camouflaged object detection. Image and Vision Computing, 143: 104924 [DOI: 10.1016/j.imavis.2024.104924]

Gu A and Dao T. 2024. Mamba: Linear-time sequence modeling with selective state spaces//Proceedings of the 12th International Conference on Learning Representations. Vienna: ICLR: 1-14

Zhu L h, Liao B C, Zhang Q, Wang X L, Liu W Y and Wang X G. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model[EB/OL]. [2024-01-18]. https://arxiv.org/pdf/2401.09417.pdf

Liu Y, Tian Y J, Zhao Y Z, Yu H T, Xie L X , Wang Y W. 2024. Vmamba: Visual state space model//Proceedings of the 38th Advances in Neural Information Processing Systems, 37: 103031-103063

Ma J, Li F F, Wang B. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation[EB/OL]. [2024-01-09]. https://arxiv.org/pdf/2401.04722.pdf

Wang Z Y, Zheng J Q, Zhang Y C, Cui G and Li L. 2024. Mamba-unet: Unet-like pure visual mamba for medical image segmentation [EB/OL].[2024-02-09]. https://arxiv.org/pdf/2402.05079.pdf

Xing Z H, Ye T, Yang Y J, Cai D, Gai B W, Wu X J, et al. 2025. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. IEEE Transactions on Medical Imaging, 45(1): 4-15 [DOI: 10.1109/TMI.2025.3589797]

Le T N, Nguyen T V, Nie Z, Tran M T and Sugimoto A. 2019. Anabranch network for camouflaged object segmentation. Computer Vision and Image Understanding, 184: 45-56 [DOI: 10.1016/j.cviu.2019.04.006]

Lv Y Q, Zhang J, Dai Y C, Li A X, Liu B W, Barnes N, et al. 2021. Simultaneously localize, segment and rank the camouflaged objects//Proceedings of the 38th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 11591-11601 [DOI: 10.1109/CVPR46437.2021.01143]

Zhou X F, Wu Z C and Cong R M. 2024. Decoupling and integration network for camouflaged object detection. IEEE Transactions on Multimedia, 26: 7114-7129 [DOI: 10.1109/TMM.2024.3360710]

Cong R M, Sun M Y, Zhang S Y, Zhou X F, Zhang W and Zhao Y. 2023. Frequency perception network for camouflaged object detection//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM: 1179-1189 [DOI: 10.1145/3581783.3612016]

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Huang Z, Dai H, Xiang T Z, Wang S, Chen H X, Qin J, et al. 2023. Feature shrinkage pyramid for camouflaged object detection with transformers//Proceedings of the 40th IEEE/CVF Conference on

Computer Vision and Pattern Recognition. Vancouver: IEEE/CVF: 5557-5566 [DOI: 10.1109/CVPR52729.2023.00538]

Luo Z Y, Liu N, Zhao W B, Yang X G, Zhang D W, Fan D P, et al. 2024. Vscode: General visual salient and camouflaged object detection with 2d prompt learning//Proceedings of the 41st IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE/CVF: 17169-17180 [DOI: 10.1109/CVPR52733.2024.01625]

Sun Y G, Xu C Y, Yang J, Xuan H Y, and Luo L. 2024. Frequency-spatial entanglement learning for camouflaged object detection//Proceedings of the 18th European Conference on Computer Vision. Cham: Springer: 343-360 [DOI: 10.1007/978-3-031-72664-3_20]

Zhang D W, Cheng L B, Liu Y, Wang X G, and Han J W. 2025. Mamba capsule routing towards part-whole relational camouflaged object detection. International Journal of Computer Vision, 133(10): 7201-7221 [DOI: 10.1007/s11263-025-02530-3]

Ye S, Chen X, Zhang Y, Lin X M, Cao L J. 2025. Escnet: Edge-semantic collaborative network for camouflaged object detection//Proceedings of the 41st IEEE/CVF International Conference on Computer Vision. Honolulu: IEEE: 20053-20063

He J H, Fu K R, Liu X H, and Zhao Q J. 2025. Samba: A unified mamba-based framework for general salient object detection//Proceedings of the 41st IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 25314-25324 [DOI: 10.1109/CVPR52734.2025.02357]

Wu Z W, Paudel D P, Fan D P, Wang J J, Demonceaux C, Timofte R, et al. 2023. Source-free depth for object pop-out//Proceedings of the 39th IEEE/CVF International Conference on Computer Vision. Paris: IEEE: 1032-1042 [DOI: 10.1109/ICCV51070.2023.00101]

Wang L Q, Yang J Y, Zhang Y F, Wang F Y and Zheng F. 2024. Depth-aware concealed crop detection in dense agricultural scenes//Proceedings of the 40th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 17201-17211

Yu Z N, Zhang X Q, Zhao L, Bin Y and Xiao G B. 2024. Exploring deeper! segment anything model with depth perception for camouflaged object detection//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM: 4322-4330 [DOI: 10.1145/3664647.3681119]

Liu J M, Kong L H and Chen G H. 2025. Improving SAM for Camouflaged Object Detection via Dual Stream Adapters//Proceedings of the 41st IEEE/CVF International Conference on Computer Vision. IEEE: 21906-21916

Lai Jie, Peng Ruihui, Sun Dianxing, Huang Jie. 2024. Detection of camouflage targets based on attention mechanism and multi-detection layer structure. Journal of Image and Graphics, 29(01): 0134-0146 (赖杰, 彭锐晖, 孙殿星, 黄杰. 2024. 融合注意力机制与多检测层结构的伪装目标检测. 中国图象图形学报, 29(01):0134-0146).[DOI: 10.11834/jig.221189]

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

Ye Xinyue, Zhu Lei, Wang Wenwu, Fu Yun. 2024. RGB_D salient object detection algorithm based on complementary information interaction. Journal of Image and Graphics, 29(05): 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(05):1252-1264) DOI: 10.11834/jig.230583.[DOI: 10.11834/jig.230583.]

作者简介

黄荣梅,女,助教,主要研究方向为计算机视觉。E-mail:1683551760@qq.com

洪如霞,通讯作者,女,教授,主要研究方向为主要研究方向为深度学习和目标检测。E-mail:lm199703295@163.com

余宏,男,副教授,主要研究领域为人工智能和显著性目标检测。E-mail:154945907@qq.com

张永选,男,副教授,主要研究方向为计算机视觉。E-mail:zyx126com@126.com

谢彩云,女,讲师,主要研究方向为深度学习和图像处理。E-mail:469323169@qq.com

陈颖,女,教授,主要研究方向为计算机视觉。E-mail:24394409@qq.com

戴靓婕,女,讲师,主要研究方向为计算机视觉和目标检测。E-mail:11620891@qq.com