软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(4):1899−1913 [doi: 10.13328/j.cnki.jos.006833]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

跨模态交互融合与全局感知的 RGB-D 显著性目标检测*

孙福明, 胡锡航, 武景宇, 孙 静, 王法胜

(大连民族大学 信息与通信工程学院, 辽宁 大连 116600)

通信作者: 王法胜, E-mail: wangfasheng@dlnu.edu.cn

摘 要: 近年来, RGB-D 显著性检测方法凭借深度图中丰富的几何结构和空间位置信息, 取得了比 RGB 显著性检测模型更好的性能, 受到学术界的高度关注. 然而, 现有的 RGB-D 检测模型仍面临着持续提升检测性能的需求. 最近兴起的 Transformer 擅长建模全局信息, 而卷积神经网络 (CNN) 擅长提取局部细节. 因此, 如何有效结合 CNN 和 Transformer 两者的优势, 挖掘全局和局部信息, 将有助于提升显著性目标检测的精度. 为此, 提出一种基于跨模态交互融合与全局感知的 RGB-D 显著性目标检测方法, 通过将 Transformer 网络嵌入 U-Net 中, 从而将全局注意力机制与局部卷积结合在一起, 能够更好地对特征进行提取. 首先借助 U-Net 编码-解码结构, 高效地提取多层次互补特征并逐级解码生成显著特征图. 然后, 使用 Transformer 模块学习高级特征间的全局依赖关系增强特征表示, 并针对输入采用渐进上采样融合策略以减少噪声信息的引入. 其次, 为了减轻低质量深度图带来的负面影响, 设计一个跨模态交互融合模块以实现跨模态特征融合. 最后, 5 个基准数据集上的实验结果表明, 所提算法与其他最新的算法相比具有显著优势.

关键词: 显著性目标检测; 跨模态; 全局注意力机制; RGB-D 检测模型

中图法分类号: TP391

中文引用格式: 孙福明, 胡锡航, 武景宇, 孙静, 王法胜. 跨模态交互融合与全局感知的RGB-D显著性目标检测. 软件学报, 2024, 35(4): 1899–1913. http://www.jos.org.cn/1000-9825/6833.htm

英文引用格式: Sun FM, Hu XH, Wu JY, Sun J, Wang FS. RGB-D Salient Object Detection Based on Cross-modal Interactive Fusion and Global Awareness. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 1899–1913 (in Chinese). http://www.jos.org.cn/1000-9825/6833.htm

RGB-D Salient Object Detection Based on Cross-modal Interactive Fusion and Global Awareness

SUN Fu-Ming, HU Xi-Hang, WU Jing-Yu, SUN Jing, WANG Fa-Sheng

(School of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China)

Abstract: In recent years, RGB-D salient detection method has achieved better performance than RGB salient detection model by virtue of its rich geometric structure and spatial position information in depth maps and thus has been highly concerned by the academic community. However, the existing RGB-D detection model still faces the challenge of improving performance continuously. The emerging Transformer is good at modeling global information, while the convolutional neural network (CNN) is good at extracting local details. Therefore, effectively combining the advantages of CNN and Transformer to mine global and local information will help to improve the accuracy of salient object detection. For this purpose, an RGB-D salient object detection method based on cross-modal interactive fusion and global awareness is proposed in this study. The transformer network is embedded into U-Net to better extract features by combining the global attention mechanism with local convolution. First, with the help of the U-Net encoder-decoder structure, this study efficiently extracts multi-level complementary features and decodes them step by step to generate a salient feature map. Then, the Transformer module is used to learn the global dependency between high-level features to enhance the feature representation, and the progressive upsampling fusion strategy is used to process the input and reduce the introduction of noise information. Moreover, to reduce the negative

* 基金项目: 国家自然科学基金 (61976042, 61972068); 兴辽英才计划 (XLYC2007023); 辽宁省高等学校创新人才支持计划 (LR2019020)
收稿时间: 2022-06-29; 修改时间: 2022-09-01, 2022-10-10; 采用时间: 2022-11-01; jos 在线出版时间: 2023-06-14
CNKI 网络首发时间: 2023-06-15

impact of low-quality depth maps, the study also designs a cross-modal interactive fusion module to realize cross-modal feature fusion. Finally, experimental results on five benchmark datasets show that the proposed algorithm has an excellent performance than other latest algorithms.

Key words: salient object detection (SOD); cross-modal; global attention mechanism; RGB-D detection model

显著性目标检测 (salient object detection, SOD) 的目的是找到并分割图像中视觉上最显著的目标[1,2]. 在过去 10 年中, 因其在目标识别[3]、基于内容的图像检索[4]、目标分割[5]、图像编辑[6]、视频分析[7,8]和视觉跟踪[9,10]中的广泛应用而备受关注.

随着 CNN 的发展, RGB 显著性目标检测[11,12]逐渐突破了传统方法[13,14]的性能瓶颈, 取得了良好的效果. 然而, 在某些复杂场景 (例如, 杂乱的背景、多个对象、不同的光照、透明对象等) 中检测效果往往并不理想[15], 主要原因在于缺乏空间位置信息, 这对于显著性目标检测至关重要. 例如, 在某些显著性物体与背景对比度较低时, 只靠 RGB 图像很难区分物体与背景.

近年来, 带有深度信息的 RGB-D 显著性目标检测, 凭借深度图中所含丰富的空间结构、3D 布局以及目标边界等有用信息[16], 在具有挑战性的场景中能够取得出色的性能. 同时, 3D 成像传感器技术的快速发展[17], 降低了深度图像的获取成本, 促进了基于 RGB-D 显著性目标检测的相关研究[18,19], 有效地解决了传统 SOD 存在的问题. 尽管如此, 现有的 RGB-D 显著性目标检测模型仍面临持续提升性能的挑战. 解决这一挑战, 可以采用如下 2 种思路.

(1) 有效融合深度特征与 RGB 特征实现跨模态的信息互补. 现有的融合策略可分为 3 类, 即早期融合、晚期融合和中期融合. 早期融合方法将深度图与原始三通道 RGB 图直接集成为四通道输入[20,21], 这种方式未考虑两种模态的分布差距, 不能有效融合跨模态信息. 晚期融合是使用并行的双流模型生成独立的显著性图, 然后将两个图进行融合得出最终特征. 这种方法忽略了 RGB 图描述图片的颜色和纹理信息及深度图描述不同位置对比度信息的事实. 最近的研究主要集中在中期融合策略上, 该策略利用两个独立的网络分别学习两种模式的中间特征, 然后将融合后的特征反馈给后续的网络或解码器. 例如 Zhu 等人[22]利用一个独立的子网络来提取深度特征, 然后将这些特征直接合并到 RGB 网络中 (如图 1(a) 所示). Fan 等人[23]从通道和空间注意力中挖掘深度信息线索, 然后将深度信息以辅助方式融合进 RGB 特征中 (如图 1(b) 所示). 需要注意的是, 上述方法主要侧重于将每一级的深度特征直接或者增强后作为辅助信息融入 RGB 特征中, 然后使用解码器生成最终的显著性图. 上述融合策略并未实现深度特征与 RGB 特征的双向交互, 导致 SOD 在一些深度特征较差的情况下所取得的检测效果并不理想. 在本文中, 我们通过采用一种双向交互融合的方式, 将融合特征作为 RGB 特征的补充 (如图 1(c) 所示), 以降低低质量深度图带来的负面影响.

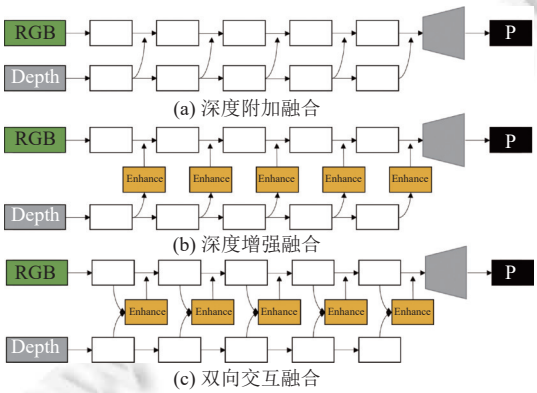


图 1 RGB-D 显著性目标检测跨模态特征融合策略比较

(2) 减轻图像特征表示过程中的图像信息损失. CNN 使用了大量的卷积、池化操作, 这些操作导致图像特征表示会损失大量的图像信息. 对于该问题, 现有的一些方法主要是通过逐层集成特征的方式来进行特征信息的补

充. 一些方法[24−26]以完全连接的方式或启发式的方式组合来自多个层的特征. 然而, 集成过多的特征和不同分辨率之间缺乏平衡容易导致计算成本高、大量噪声和融合困难, 从而干扰自顶向下路径中的后续信息恢复. 此外, Atrus 空间金字塔池模块 (atrous spatial pyramid pooling module, ASPP)[27]和全局上下文模块 (global context module, GCM)[23]用于提取多尺度上下文感知特征并增强单层表示[28−30]. 然而, 现有的 CNN 方法主要通过增大感受野的方式以获取全局信息, 这种操作会导致图像分辨率下降以及大量语义信息丢失. 最近, Transformer[31]架构在计算机视觉领域引起了广泛的关注. 与利用滑动窗口卷积运算而聚焦于图像局部的 CNN 不同, Transformer 在编码层和解码层中多次堆叠自注意力层, 利用自注意力机制可以实现全局上下文建模, 获取长距离的依赖关系, 但在捕捉局部特征方面效果不够理想, 且存在计算成本高的问题.

基于上述分析, 我们提出了一种新的基于跨模态交互融合与全局感知的 RGB-D 显著性目标检测方法, 通过将 Transformer 网络嵌入 U-Net 中, 从而将全局注意力机制与局部卷积结合在一起, 能够更好地对特征进行提取. 该方法采用 U 形结构[32,33]提取 RGB 和深度双流特征, 并利用多级信息重建高分辨率图, 抑制下层的干扰, 减少冗余的细节. 在编码器的早期阶段进行深度特征与 RGB 特征的双向交互融合. 该渐进式的交互方式基于通道注意力机制, 能够从深度特征中充分挖掘丰富的深度信息线索 (几何结构和空间位置信息) 并将其与 RGB 特征相融合, 使得在编码器阶段所提取的双流特征能够互相校正和细化. 同时, 引入 Transformer 模块用于学习高层特征的跨层级间的长距离依赖关系, 更好地利用多级特征, 避免因分辨率差异过大而对特征融合造成干扰, 可以有效地增强特征表示, 同时降低因卷积和池化操作产生的信息损失, 从而改善显著性目标的预测效果.

本文的主要贡献可以概括如下.

(1) 设计了一个 CNN-Transformer 网络架构, 将 Transformer 全局感知特征增强模块嵌入到 U-Net 框架中, 通过 CNN 提取局部特征, 利用 Transformer 学习跨层级的长距离依赖关系以增强特征表示.

(2) 设计了跨模态交互融合模块, 借助注意力机制学习深度图像和 RGB 图像之间的互补信息, 并且将跨模态融合特征作为 RGB 的补充, 以充分利用不同模态的特征信息.

(3) 设计了一个多级融合解码器, 通过不同大小的残差卷积块逐级解码, 并且在解码的过程中融合低级特征, 保留了更多的原始信息.

(4) 采用预训练的 Res2Net 模型作为骨干网络, 进一步提升了检测精度. 在 5 个基准的 RGB-D 显著目标检测数据集上的实验结果表明, 我们的方法获得了最优的性能.

本文第 1 节介绍 RGB-D 显著性目标检测的相关方法和研究现状. 第 2 节介绍本文构建的跨模态交互融合与全局感知的 RGB-D 显著性目标检测方法. 第 3 节通过对比实验验证了所提模型的有效性. 第 4 节总结全文.

1 相关工作

1.1 RGB-D 显著性目标检测

在过去 20 年间, 大量 RGB 显著性目标检测方法[34−39]被提出并且取得了出色的性能. 然而, 在复杂场景下, 它们的检测结果却不够理想, 如低对比度、小目标、复杂背景、多个物体和前景背景相似, 因此需要引入额外的辅助信息来协助完成 SOD 任务. 深度线索因包含丰富的几何结构和空间位置信息, 能够有效地提高复杂场景下的检测性能, 近年来被广泛应用于 SOD 任务.

在 RGB-D SOD 任务中, RGB 特征包含大量外观和纹理信息, 而深度特征则更侧重于三维布局和空间位置信息. 如何将 RGB 特征和深度特征的互补信息进行跨模态融合, 一直是 RGB-D 显著性目标检测任务中的一个重要问题. 针对这一问题, 已经开展了大量的研究工作. Zhao 等人[40]设计了一致性差异聚合结构, 通过多路径融合的方式, 实现跨模态和跨层次融合. Qu 等人[41]使用手工制作的特征向量作为输入来训练基于 CNN 的模型, 与传统方法[42−44]相比, 取得了显著的改进. Chen 等人[45]通过 3D 卷积神经网络在编码器阶段进行预融合和解码器阶段进行深度融合. Chen 等人[46]设计了一个渐进式双流网络, 其中使用跨模态残差函数和互补感知监督来探索跨模型和跨层次互补. Fu 等人[47]将 RGB 和深度输入联合学习, 通过孪生网络挖掘有用的互补特征. Pang 等人[48]通过密集连

接结构生成不同大小感受野的动态过滤器, 实现深度引导融合. Chen 等人[49]引入深度潜能感知对深度图的潜力进行建模, 并在网络后期融合特征, 整合了跨模态互补性.

在本文中, 我们将利用注意力机制, 实现深度特征和 RGB 特征的交互融合; 并且, 为了减轻低质量深度图带来的负面影响, 将跨模态特征作为 RGB 特征的补充.

1.2 Transformer 网络

Transformer 网络由 Vaswani 等人[31]首次提出并应用于机器翻译任务后, 在自然语言处理 (natural language processing, NLP) 领域取得了巨大的成功. 借助于自注意力机制, Transformer 网络能够捕获输入序列元素间的长期依赖关系, 这一特性对于计算机视觉任务也能提供巨大的帮助. 因此, 近年来在计算机视觉领域出现了大量基于 Transformer 模型的相关研究成果, 如目标检测[50,51]、目标跟踪[52]、姿态估计[53]、图像分类[54,55]、语义分割[56,57]等. 其中, ViT[54]将图像分割成一系列平面化的二维块, 然后采用 Transformer 对图像进行分类, 在图像分类任务中取得了巨大的突破. Wang 等人[58]提出了一种适用于密集预测任务的 ViT 金字塔结构. Zhu 等人[59]将 Transformer 引入 SOD 任务, 并首次与深度监督策略相结合, 提出了一种基于 Swin Transformer 的深度监督模型. 此外, 在医学图像分割领域, Chen 等人提出了 TransUNet[60], 以预训练的 ViT 作为骨干网络, 并采取 U-Net[61]网络架构, 取得了良好的结果.

基于卷积运算的 CNN 模型在提取局部特征时更有优势, 而 Transformer 能够更好地捕获远程相关性. 基于它们的特性, 出现了一些 CNN 与 Transformer 的混合结构, 充分发挥两者的优势. MaX-DeepLab[62]采取双路径架构, 引入全局内存路径实现全局交互, 构建了一个用于全景分割的端到端模型. TransFuse[63]提出了一种并行分支架构, 通过 CNN 分支提取空间细节, 通过 Transformer 分支捕获全局依赖关系. Luo 等人[64]提出了一种基于 CNN 和 Transformer 的半监督交叉学习方法, 用一个网络的预测端到端地监督另一个网络. TANet[65]提出了一种非对称网络, 通过 Transformer 主干提取全局信息, 再利用轻量级 CNN 主干提取空间结构信息相结合. TransT[66]通过孪生 CNN 作为主干网络进行特征提取, 并基于自注意力和交叉注意力机制实现特征增强与融合. CoTr[67]通过 CNN 提取特征表示, 并构建一种可变形 Transformer (DeTrans) 获取远程依赖关系.

借鉴上述思想, 本文将 CNN 与 Transformer 相结合, 并将 Transformer 全局感知特征增强模块嵌入 U-Net 框架中, 充分结合两种框架的优点, 将检测性能提升到了一个新的水平.

2 本文方法

我们提出一种基于 CNN-Transformer 框架的 RGB-D 显著性目标检测网络, 如后文图 2 所示. 该网络由跨模态融合编码器、全局感知特征增强模块和多级融合解码器组成, 利用 Transformer 更充分地获取图像的全局信息, 提高检测性能. 首先, 双流骨干网络从 RGB 图像和深度图像中分别提取特征. 接着, 利用跨模态交互融合模块 (CIF) 进行跨模态的特征融合, 并将融合特征与 RGB 特征结合作为更高层的输入. 然后, 通过渐进上采样融合将高 3 层的特征转化为相同的尺寸并进行融合, 再将其输入全局感知特征增强模块. 最后, 将得到的增强的高级特征和低两层特征, 输入到多级融合解码器中进行解码, 得到最终的显著预测图.

2.1 跨模态融合编码器

在 SOD 任务中, 不同模态的输入包含不同的信息, RGB 图像包含丰富的色彩信息和纹理信息, 而深度图像则侧重于空间位置信息. 此外, 在训练和测试过程中, 并不能保证深度图的质量, 低质量的深度图容易影响检测结果. 针对上述问题, 我们设计了跨模态交互融合模块 (如图 3 所示), 用于实现跨模态信息的交互融合, 并减轻低质量深度图的负面影响. 图 3 中, 对深度特征和 RGB 特征在通道维度上级联后, 利用两个一维池化操作给特征图嵌入方向信息. 然后进行级联并输入转换层, 经过转换层压缩通道并且编码空间信息, 这里是通过一个卷积层来实现的. 之后将编码后的信息沿 x、y 方向分离, 再通过编码注意力层在各自方向上生成编码注意力图, 并与输入特征图相乘来实现通道注意力感知. 最后输入到空间注意力模块中, 并将输出与输入相乘以获取空间注意力感知. 这一过程可以描述为:

$$F_i^r = f_i \times SA(f_i \times CA_x(trans(p_x(f_i), p_y(f_i))) \times CA_y(trans(p_x(f_i), p_y(f_i)))) \tag{1}$$

其中, $f_i = Cat(f_i^r, f_i^d)$, f_i^r 和 f_i^d 分别表示骨干网提取的颜色特征和深度特征 $i = 1,\ldots,5$, $Cat(\cdot)$ 表示级联操作; p_x 和 p_y 表示水平方向和垂直方向的平均池化操作; $trans(\cdot)$ 表示转换层实现编码信息的嵌入, 其中包括一个卷积层、BN 层和 Sigmoid 层; $CA_x(\cdot)$ 和 $CA_y(\cdot)$ 表示沿 x、y 方向上编码注意力的生成[68], 通过一个包含 Sigmoid 层的卷积层来实现. 通过这种方式, 可以沿一个空间方向捕获远程依赖关系并保留位置信息; $SA(\cdot)$ 表示空间注意力层. 这样, 深度特征和 RGB 特征就能充分结合以增强感兴趣目标的特征表示.

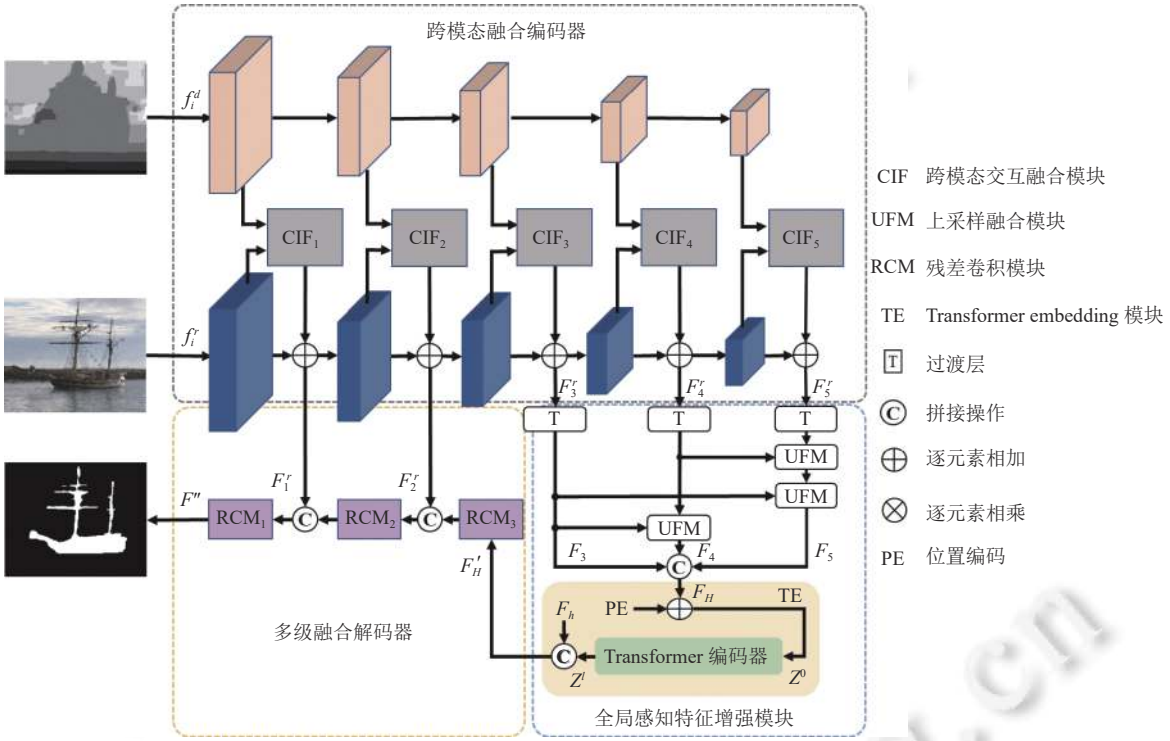


图 2 基于 CNN-Transformer 的模型框架图

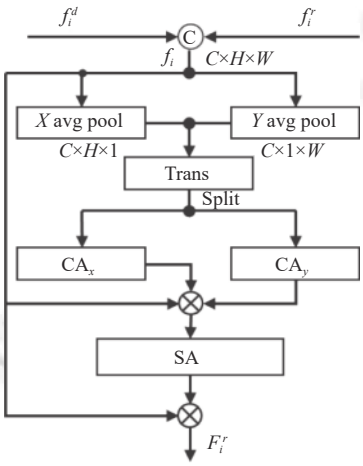


图 3 跨模态交互融合模块 (CIF)

此外, 我们还引入了在 ImageNet 上预训练的 Res2Net 模型作为骨干网络[69]. Res2Net 通过将 3×3 卷积核替换为分层残差连接的卷积核组, 提高了网络的感知能力. 在计算量相近的前提下, 提高了准确率.

2.2 全局感知特征增强模块

Transformer 在获取全局依赖关系方面有着出色的表现, 该特性有助于提取高级语义信息. 基于此我们设计了全局感知特征增强模块, 由尺寸调节模块和 Transformer embedding 模块两部分组成. 尺寸调节模块用于调整多级高级特征到相同的尺寸, 旨在减轻上采样过程中噪声的负面影响. Transformer embedding 模块的作用是获取高级特征间的长距离依赖关系以实现特征增强. 具体来说, 我们首先将第 3–5 层的特征调整到相同的尺寸后进行级联, 再输入到 Transformer embedding 模块来学习跨层级的长距离依赖关系以增强特征表示, 最后将 Transformer embedding 模块的输出与输入进行级联以保留更多原始信息.

2.2.1 尺寸调节模块

Transformer embedding 模块将多层高级特征作为输入, 但是不同层级的高级特征间的尺寸以及通道数是不同的, 因此需要将其调整到相同的大小以便于进行融合.

首先, 我们将 F_i^r 通过一个由 3×3 卷积和 ReLU 激活函数组成的变换层 T, 将多级特征的通道数调整到相同的大小. 该过程可以描述为:

$$F_i^{'r} = \sigma(Conv(F_i^r)),\ i = 3,\ldots,5 \tag{2}$$

其中, $Conv(\cdot)$ 是 3×3 卷积操作, $\sigma(\cdot)$ 是 ReLU 激活函数.

为了将特征的尺寸调整到相同的大小, 我们需要对 $F_4^{'r}$ 和 $F_5^{'r}$ 进行上采样操作, 但是直接使用 2 倍和 4 倍上采样会引入一定的噪声信息. 所以, 我们采取渐进上采样融合的策略来处理第 3–5 层的特征. 通过 UFM 模块, 能够有效降低引入的噪声, 并且使高级特征的空间细节更加丰富. 该过程可以表述为:

$$\begin{cases} F_5 = UFM(UFM(F_5^{'r}, F_4^{'r}), F_3^{'r}) \\ F_4 = UFM(F_4^{'r}, F_3^{'r}) \\ F_3 = F_3^{'r} \end{cases} \tag{3}$$

其中, $UFM(\cdot)$ 如图 4 所示. 具体可以表述为:

$$UFM(F_h, F_l) = Cat(Conv(Up(F_h)), F_l) \tag{4}$$

其中, F_h 和 F_l 分别表示较高层和较低层的特征, $Up(\cdot)$ 表示上采样操作. 然后, 将高级特征 $F_H = Cat(F_3, F_4, F_5)$ 作为输入送入 Transformer 编码器.

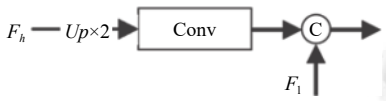


图 4 上采样融合模块示意图 (UFM)

2.2.2 Transformer embedding 模块

在 Transformer embedding 模块中, 首先通过 Patch embedding 层将输入特征图转化为特征嵌入序列, 然后叠加上可训练的位置编码 (position encoder, PE) 以保留位置信息. 接着将特征输入到堆叠的标准 Transformer 编码器[31]中, 利用 Transformer 机制获取高级特征的长距离依赖关系后增强原始特征表示. 最后将输出的特征调整为原始大小. 同时, 为了保留更多的原始信息, 我们将输出特征与原始特征进一步级联生成增强的高级特征 F_H'.

2.3 多级融合解码器

多级融合解码器用于解码增强后的高级特征并融合低级特征, 以生成显著预测. 具体来说, 我们将 Transformer embedding 模块增强后的高级特征与最低两层特征进行结合. 同时, 我们采取多个不同尺度的残差卷积模块 (residual convolution module, RCM) 代替单独的卷积层进行解码. 在 RCM 模块中, 输入特征通过深度可分离卷积 (depth-wise, DW) 层、正则化层 (layer normalization, LN)[70]和逐点卷积 (point-wise, PW) 层进行进一步的过滤. 实

验验证了这种方式可以取得较好的预测结果, 具体过程如下所示:

$$F'' = RCM_1(Cat(RCM_2(Cat(RCM_3(F'_H), F'_2)), F'_1)) \tag{5}$$

最后, 输出的特征 F'' 通过 1×1 卷积操作生成最终的显著图 S_{final}. 在训练过程中, 该显著图由真值图监督生成损失. 在这里, 我们采用像素位置感知损失 (pixel position aware loss) 用于端到端的训练, 总体损失定义为:

$$L = L^s_{\text{ppa}}(S_{\text{final}}, G) \tag{6}$$

其中, G 是真值显著图.

3 实验结果

3.1 数据集和评价指标

我们在 5 个具有挑战性的 RGB-D 数据集上对本文提出的方法进行了评估. DUT[16]包含 Lytro 相机在现实生活场景中捕获的 1 200 张图像; NLPR[21]包括具有单个或多个显著对象的 1 000 张图像; NJU2K[71]包括 2 003 张不同分辨率的立体图像; SIP[72]包含 1 000 幅突出人物的高分辨率图像; DES[73]包含 135 幅由微软 Kinect 采集的室内图像.

为了公平比较, 我们采取与文献 [72,74] 中相同的训练数据集, 包括 NJU2K 数据集的 1 485 幅图像和 NLPR 数据集的 700 幅图像, 合计 2 185 个样本来训练检测算法. NJU2K 和 NLPR 数据集的剩余图像以及 SIP、DUT 和 DES 的整个数据集用于测试. 此外, 在 DUT 数据集上, 我们遵循与文献 [75−77] 中相同的设置, 从 DUT 添加额外的 800 对用于训练, 其余 400 对用于测试.

评估时, 我们采用了 4 个广泛使用的评价指标, 即 E 指标[78]、S 指标[79]、F 指标[80]和平均绝对误差 (mean absolute error, MAE)[81]. E 指标用来衡量局部像素级误差和全局图像级误差; S 指标评估显著图的区域感知和对象感知的空间结构相似性; F 指标是查准率和查全率的加权调和均值, 用来评价系统的整体性能; MAE 测量显著图和真值图之间的每像素绝对差值的平均值. 在实验中, E 指标和 F 指标均采用了自适应的值.

3.2 实施细节

在训练和测试阶段, 输入 RGB 图和深度图像尺寸调整为 256×256. 采用多种增强策略增强所有训练图像, 即随机翻转、旋转和边界剪切. 骨干网络的参数使用 Res2Net-50 网络的预训练参数进行初始化. Transformer 编码器中的超参数设置为: $L = 12$, $D = 768$, $N = 1024$. 其余参数初始化为 PyTorch 默认设置. 我们采用 Adam 优化器[82]训练我们的网络, Batch 为 8, 初始学习率为 1E−5, 学习率为每 60 个 Epoch 除以 10. 我们的模型在具有单个英伟达 GTX 3090 GPU 的机器上进行训练. 该模型在 150 个 Epoch 内收敛, 需要约 15 h.

3.3 与先进方法对比

将本文模型与 CoNet[75]、TriTransNet[76]、SSF[83]、ATSA[84]、AILNet[85]、EBFSP[86]、CDNet[87]、HAINet[88]、RD3D[46]和 DSA2F[89]、DCF[90]这 11 种最新的 RGB-D SOD 模型进行了比较.

3.3.1 定量评估

上文提及的 11 种最新的 RGB-D SOD 模型在 5 个广泛使用的数据集上的定量结果如表 1 所示. 最优和次优结果分别用加粗和下划线表示. 表 1 中的部分统计结果, 为运行作者提供的源代码生成, 包括 (1) TriTransNet 的全部结果; (2) HAINet 在 NLPR、NJU2K 数据集上的结果; (3) ATSA、SSF、AILNet、DSA2F 在 SIP 数据集上的结果; (4) ATSA、CoNet、SSF、AILNet、DSA2F、DCF 在 DES 数据集上的结果. EBFSP、CDNet、RD3D 的实验结果引用自文献 [76], 其余结果均来自原作者发表.

根据 4 个评价指标的结果可以看出, 本文提出的算法在 4 个评价指标上均取得了最好的结果, 相比于现有的最新算法有显著的提升. 具体来说, 在全部 5 个数据集上, 相比于次优方法, 本文的 S 指标平均提高了 0.4%, F 指标平均提高了 0.54%, E 指标平均提高了 0.34%, MAE 值平均提高了 8.9%. 实验结果直观地验证了本文算法在不同数据集及评价指标下的有效性和鲁棒性.

表 1 先进算法及本文提出的算法在 5 个 RGB-D 数据集上的定量指标

数据集	指标	ATSA	CoNet	SSF	AILNet	EBFSP	CDNet	HAINet	RD3D	DSA2F	DCF	TriTransNet	Ours
DUT	S↑	0.918	0.918	0.915	0.926	0.858	0.880	0.910	0.931	0.921	0.924	0.928	**0.933**
	F_β↑	0.920	0.908	0.915	0.917	0.842	0.874	0.920	0.924	0.926	0.926	0.924	**0.927**
	E_ξ↑	0.948	0.941	0.946	0.951	0.890	0.918	0.944	0.949	0.950	0.952	0.952	**0.958**
	MAE↓	0.032	0.034	0.033	0.031	0.067	0.048	0.038	0.031	0.030	0.030	0.031	**0.028**
NLPR	S↑	0.907	0.907	0.914	0.912	0.909	0.902	0.924	0.930	0.918	0.922	0.921	**0.931**
	F_β↑	0.876	0.848	0.875	0.857	0.887	0.848	0.891	0.892	0.897	0.893	0.891	**0.901**
	E_ξ↑	0.945	0.936	0.949	0.935	0.940	0.935	0.956	0.958	0.950	0.956	0.955	**0.962**
	MAE↓	0.028	0.031	0.026	0.029	0.028	0.032	0.024	0.022	0.024	0.023	0.025	**0.020**
NJU2K	S↑	0.901	0.894	0.899	0.898	0.907	0.885	0.912	0.916	0.903	0.918	0.916	**0.924**
	F_β↑	0.893	0.872	0.886	0.876	0.895	0.866	0.898	0.901	0.901	0.897	0.903	**0.916**
	E_ξ↑	0.921	0.912	0.913	0.912	0.908	0.911	0.921	0.918	**0.922**	**0.922**	0.912	**0.922**
	MAE↓	0.040	0.047	0.043	0.045	0.038	0.048	0.038	0.036	0.039	0.038	0.035	**0.031**
SIP	S↑	0.887	0.858	0.868	0.889	0.877	0.823	0.880	0.885	0.862	0.880	0.886	**0.899**
	F_β↑	0.873	0.842	0.851	0.866	0.863	0.805	0.892	0.874	0.865	0.877	0.892	**0.895**
	E_ξ↑	0.915	0.909	0.911	0.914	0.911	0.880	0.922	0.920	0.908	0.920	0.924	**0.930**
	MAE↓	0.049	0.063	0.056	0.050	0.052	0.076	0.053	0.048	0.057	0.051	0.043	**0.040**
DES	S↑	0.923	0.911	0.905	0.922	0.937	0.875	0.935	0.935	0.903	0.923	0.942	**0.948**
	F_β↑	0.897	0.861	0.876	0.881	0.913	0.839	0.924	0.917	0.901	0.912	0.927	**0.933**
	E_ξ↑	0.961	0.945	0.948	0.952	0.974	0.921	0.973	0.975	0.923	0.963	0.981	**0.982**
	MAE↓	0.021	0.027	0.025	0.023	0.018	0.034	0.018	0.019	0.039	0.021	0.016	**0.014**

3.3.2 定性评估

为了进行定性评估, 我们将本文算法的结果与一些具有代表性的最新算法进行了可视化的比较. 其中包含了一些具有代表性的困难场景, 如前景和背景相似 (行 1 行、第 2 行)、复杂场景 (第 3 行、第 4 行)、低质量深度图 (第 5 行、第 6 行)、多目标 (第 7 行、第 8 行) 和小目标 (第 9 行、第 10 行) 的情况, 比较结果如后文图 5 所示. 从图 5 中可以看出, 本文的模型能够更精确的定位和分割显著目标, 并且在困难场景下仍能保证优秀的检测性能. 这些实验进一步验证了该模型的有效性和鲁棒性.

3.4 消融实验

为了验证每个模块的有效性, 我们在 DUT、NLPR、NJU2K 这 3 个数据集上进行了消融实验, 从不同方面验证各个模块在本文算法中的有效性. 实验结果如表 2–表 5 所示, 从结果中我们可以看出, 全局感知特征增强模块对检测结果的影响最大, 跨模态交互融合模块的影响次之, 两者相比于基线模型都获得了约 3% 的性能提升, 而多级融合解码器带来的性能提升相对较小, 只有不到 1%. 这印证了跨模态特征融合以及全局特征感知在显著性目标检测任务中的重要性. 详细的消融实验结果将在以下几节中给出.

3.4.1 跨模态交互融合模块的有效性

为了验证跨模态交互融合模块的有效性, 本文进行了 4 个实验. 第 1 个实验, 基线模型去掉了跨模态交互融合模块 (CIF), 直接在编码器中将深度特征与 RGB 特征相加. 第 2 个实验, 使用 JL-DCF[47]中提出的跨模态融合模块 (CM). 第 3 个实验, 使用 CBAM[91]提出的通道-空间注意力模块. 第 4 个实验, 使用本文提出的跨模态交互融合模块 (CIF). 为了公平起见, 这些模块都采取与本文方法相同的连接方式.

根据表 2 的结果, 我们可以观察到, 基线模型加入 CIF 后, 在 3 个数据集的 4 个评价指标上的结果都有明显的提升. 并且, 相比于其他 2 种特征融合模块, 本文的 CIF 取得了最佳的实验结果. 这说明本文提出的跨模态交互融合模块, 借助一维编码注意力机制, 能够在不增加计算负担的前提下, 获取各自维度上更长距离的注意力信息再进行结合, 从而有效地实现 RGB 特征和深度特征的跨模态融合. 同时将输出的融合特征附加到 RGB 特征上作为补充, 以减轻低质量深度图带来的负面影响.

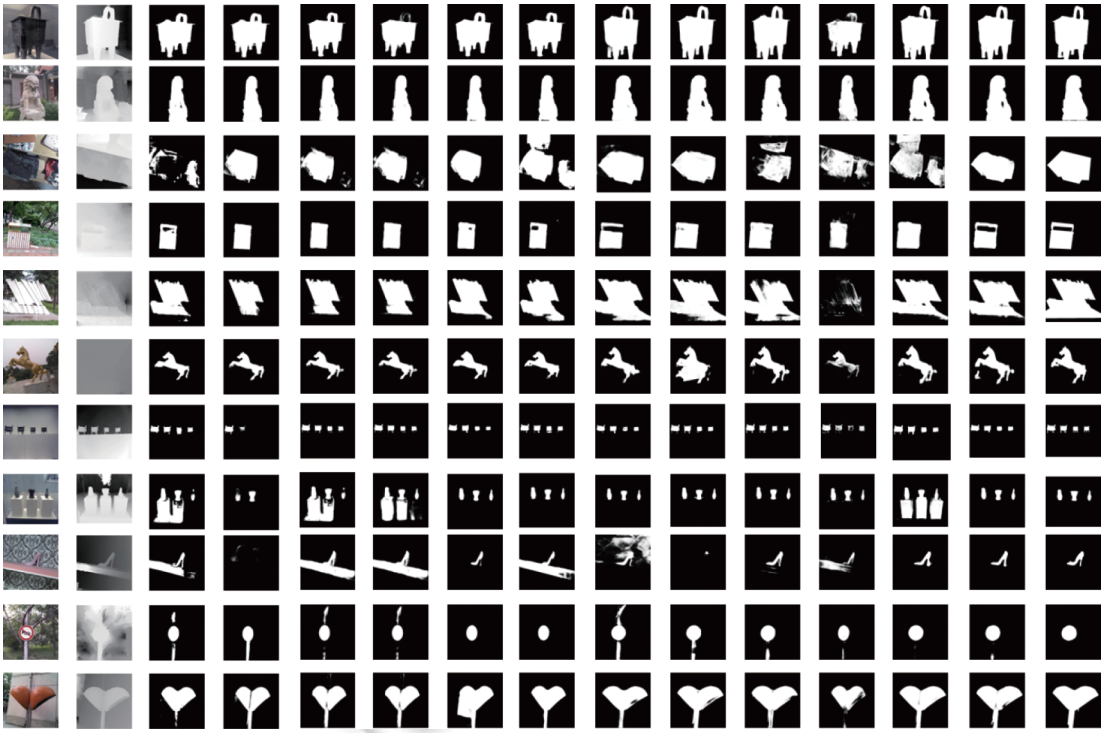


图 5　本文算法与前沿 RGB-D 显著性模型的定性比较

表 2　跨模态交互融合模块的消融实验结果

Variant	Candidate	DUT				NLPR				NJU2K			
		S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓
(a)	Baseline	0.898	0.889	0.927	0.044	0.903	0.864	0.941	0.031	0.896	0.888	0.915	0.045
(b)	(a)+CM	0.907	0.896	0.940	0.040	0.879	0.828	0.924	0.037	0.872	0.855	0.889	0.059
(c)	(a)+CBAM	0.930	0.925	0.954	0.030	0.921	0.891	0.948	0.027	0.924	0.912	0.917	0.032
(d)	(a)+CIF	0.937	0.933	0.960	0.026	0.935	0.905	0.965	0.019	0.923	0.913	0.922	0.032

表 3　全局感知特征增强模块的消融实验结果

Variant	Candidate	DUT				NLPR				NJU2K			
		S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓
(a)	Baseline	0.894	0.888	0.934	0.046	0.872	0.828	0.925	0.037	0.868	0.859	0.902	0.060
(b)	(a)+TE	0.922	0.924	0.950	0.034	0.919	0.891	0.956	0.025	0.889	0.884	0.913	0.054
(c)	(a)+TE+UFM	0.937	0.933	0.960	0.026	0.935	0.905	0.965	0.019	0.923	0.913	0.922	0.032

表 4　多级融合解码器的消融实验结果

Variant	Candidate	DUT				NLPR				NJU2K			
		S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓
(a)	Baseline	0.933	0.923	0.955	0.030	0.925	0.891	0.956	0.023	0.922	0.911	0.917	0.031
(b)	(a)+MFD′	0.931	0.928	0.953	0.030	0.925	0.898	0.960	0.024	0.921	0.911	0.918	0.032
(c)	(a)+MFD	0.937	0.933	0.960	0.026	0.935	0.905	0.965	0.019	0.923	0.913	0.922	0.032

表 5 多级融合解码器层数的消融实验结果

层数	DUT				NLPR				NJU2K			
	S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓	S↑	F_β↑	E_ξ↑	MAE↓
2	0.934	0.932	0.960	0.028	0.920	0.888	0.953	0.025	0.914	0.902	0.923	0.039
3	0.937	0.933	0.960	0.026	0.935	0.905	0.965	0.019	0.923	0.913	0.922	0.032
4	0.922	0.918	0.943	0.035	0.914	0.883	0.952	0.028	0.901	0.888	0.910	0.046

3.4.2 全局感知特征增强模块的有效性

在这一部分, 我们去掉全局感知特征增强模块作为基线模型. 第 2 个实验, 将全局感知特征增强模块的渐进上采样融合模块 (UFM) 替换为 2×、4×上采样. 第 3 个实验, 使用本文提出的全局感知特征增强模块.

根据表 3 的结果可以看出, 我们的全局感知特征增强模块能够有效增强融合特征, 改善了检测结果. 这得益于 Transformer 中的自注意力机制, 通过自注意力机制, 能够对图像信息进行全局交互, 从而获得更大范围尺度上的高级语义信息, 这对定位显著目标起到决定性的影响. 并且, 从实验 (b) 和实验 (c) 的对比可以看出, 相比于直接上采样, UFM 采取逐级融合邻层特征再上采样的方式, 以邻层特征互为指导, 可以减轻噪声的负面影响, 获得更好的检测结果.

3.4.3 多级融合解码器的有效性

为了验证多级融合解码器的有效性, 我们将多级融合解码器替换为单层卷积构成的解码器作为基线模型, 并对比了多级融合解码器 (MFD) 和未融合低级特征的多级解码器 (MFD′) 的性能差距. 根据表 4 的结果, 我们可以看出, 相比于单层卷积解码器, 我们的多级融合解码器借助残差卷积块, 能进一步提取并保留有效的显著信息, 减轻低级特征中噪声的干扰. 同时, 可以看出融合低级特征能够显著提升检测效果, 这是由于低级特征中包含了大量边缘信息, 在提取高级特征的过程中往往会丢失, 通过这种方式能够得到有效补充, 以实现显著目标的精确分割. 最后, 我们对残差卷积模块 (RCM) 的层数进行消融分析. 我们从表 5 的结果可以看出, 当层数为 3 时具有最好的检测结果.

4 结 论

我们针对 RGB-D 显著目标检测如何更好地挖掘局部和全局信息的问题, 从 CNN 和 Transformer 各自的优势及局限性出发将 Transformer 与 U-Net 框架相结合, 设计了一个新的 RGB-D 显著目标检测框架. 我们利用跨模态交互融合模块对深度特征和 RGB 特征进行互补融合, 并利用 Transformer 全局感知特征增强模块学习不同层级高级特征间的长距离依赖关系以增强特征表示. 此外, 设计了多级融合解码器以实现显著特征图的精确生成. 在 5 个数据集上的实验结果表明, 该方法与其他最新算法相比较将性能提升到了一个新的水平. 但 Transformer 中自注意力机制的计算量会随着数据量呈平方复杂度增长, 这限制了模型的扩展. 后续研究工作中, 我们将扩展本文模型, 尝试优化自注意力中的二次运算, 同时针对边缘细节细化问题进行进一步的研究, 并将其推广到 RGB-T 显著性目标检测任务中.

References:

[1] Borji A, Cheng MM, Jiang HZ, Li J. Salient object detection: A benchmark. IEEE Trans. on Image Processing, 2015, 24(12): 5706–5722. [doi: 10.1109/TIP.2015.2487833]

[2] Wang WG, Lai QX, Fu HZ, Shen JB, Ling HB, Yang RG. Salient object detection in the deep learning era: An in-depth survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3239–3259. [doi: 10.1109/TPAMI.2021.3051099]

[3] Cheng MM, Zhang ZM, Lin WY, Torr P. BING: Binarized normed gradients for objectness estimation at 300 fps. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Columbus: IEEE, 2014. 3286–3293. [doi: 10.1109/CVPR.2014.414]

[4] Cheng MM, Hou QB, Zhang SH, Rosin PL. Intelligent visual media processing: When graphics meets vision. Journal of Computer Science and Technology, 2017, 32(1): 110–121. [doi: 10.1007/s11390-017-1681-7]

[5] Wang WG, Shen JB, Yang RG, Porikli F. Saliency-aware video object segmentation. IEEE Trans. on Pattern Analysis and Machine

Intelligence, 2018, 40(1): 20–33. [doi: 10.1109/TPAMI.2017.2662005]

[6] Cheng MM, Zhang FL, Mitra NJ, Huang XL, Hu SM. RepFinder: Finding approximately repeated scene elements for image editing. ACM Trans. on Graphics, 2010, 29(4): 83. [doi: 10.1145/1778765.1778820]

[7] Fan DP, Wang WG, Cheng MM, Shen JB. Shifting more attention to video salient object detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 8546–8556. [doi: 10.1109/CVPR.2019.00875]

[8] Yan PX, Li GB, Xie Y, Li Z, Wang C, Chen TS, Lin L. Semi-supervised video salient object detection using pseudo-labels. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 7283–7292. [doi: 10.1109/ICCV.2019.00738]

[9] Wang YB, Wang FS, Wang C, Sun FM, He JJ. Learning saliency-aware correlation filters for visual tracking. The Computer Journal, 2022, 65(7): 1846–1859. [doi: 10.1093/comjnl/bxab026]

[10] Zhou ZK, Pei WJ, Li X, Wang HP, Zheng F, He ZY. Saliency-associated object tracking. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 9846–9855. [doi: 10.1109/ICCV48922.2021.00972]

[11] Liu JJ, Hou QB, Cheng MM, Feng JS, Jiang JM. A simple pooling-based design for real-time salient object detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 3912–3921. [doi: 10.1109/CVPR.2019.00404]

[12] Wang LZ, Wang LJ, Lu HC, Zhang PP, Ruan X. Salient object detection with recurrent fully convolutional networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1734–1746. [doi: 10.1109/TPAMI.2018.2846598]

[13] Cheng MM, Mitra NJ, Huang XL, Torr PHS, Hu SM. Global contrast based salient region detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015, 37(3): 569–582. [doi: 10.1109/TPAMI.2014.2345401]

[14] Zhang DW, Meng DY, Han JW. Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(5): 865–878. [doi: 10.1109/TPAMI.2016.2567393]

[15] Chen H, Li YF. Three-stream attention-aware network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2019, 28(6): 2825–2835. [doi: 10.1109/TIP.2019.2891104]

[16] Piao YR, Ji W, Li JJ, Zhang M, Lu HC. Depth-induced multi-scale recurrent attention network for saliency detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 7253–7262. [doi: 10.1109/ICCV.2019.00735]

[17] Giancola S, Valenti M, Sala R. State-of-the-art devices Comparison. In: Giancola S, Valenti M, Sala R, eds. A Survey on 3D Cameras: Metrological Comparison of Time-of-flight, Structured-light and Active Stereoscopy Technologies. Cham: Springer, 2018. 29–39. [doi: 10.1007/978-3-319-91761-0_3]

[18] Li NY, Ye JW, Ji Y, Ling HB, Yu JY. Saliency detection on light field. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Columbus: IEEE, 2014. 2806–2813. [doi: 10.1109/CVPR.2014.359]

[19] Zhao JX, Cao Y, Fan DP, Cheng MM, Li XY, Zhang L. Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 3922–3931. [doi: 10.1109/CVPR.2019.00405]

[20] Cong RM, Lei JJ, Fu HZ, Huang QM, Cao XC, Ling N. HSCS: Hierarchical sparsity based co-saliency detection for RGBD images. IEEE Trans. on Multimedia, 2019, 21(7): 1660–1671. [doi: 10.1109/TMM.2018.2884481]

[21] Peng HW, Li B, Xiong WH, Hu WM, Ji RR. RGBD salient object detection: A benchmark and algorithms. In: Proc. of the 13th European Conf. on Computer Vision (ECCV). Zurich: Springer, 2014. 92–109. [doi: 10.1007/978-3-319-10578-9_7]

[22] Zhu CB, Cai X, Huang K, Li TH, Li G. PDNet: Prior-model guided depth-enhanced network for salient object detection. In: Proc. of the 2019 IEEE Int'l Conf. on Multimedia and Expo (ICME). Shanghai: IEEE, 2019. 199–204. [doi: 10.1109/ICME.2019.00042]

[23] Fan DP, Zhai YJ, Borji A, Yang JF, Shao L. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 275–292. [doi: 10.1007/978-3-030-58610-2_17]

[24] Zhang PP, Wang D, Lu HC, Wang HY, Ruan X. Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 202–211. [doi: 10.1109/ICCV.2017.31]

[25] Hou QB, Cheng MM, Hu XW, Borji A, Tu ZW, Torr PHS. Deeply supervised salient object detection with short connections. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5300–5309. [doi: 10.1109/CVPR.2017.563]

[26] Wang TT, Zhang LH, Wang S, Lu HC, Yang G, Ruan X, Borji A. Detect globally, refine locally: A novel approach to saliency detection. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 3127–3135. [doi: 10.1109/CVPR.2018.00330]

[27] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets,

atrous convolution, and fully connected CRFs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

[28] Li CY, Cong RM, Piao YR, Xu QQ, Loy CC. RGB-D salient object detection with cross-modality modulation and selection. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 225–241. [doi: 10.1007/978-3-030-58598-3_14]

[29] Deng ZJ, Hu XW, Zhu L, Xu XM, Qin J, Han GQ, Heng PA. R^3Net: Recurrent residual refinement network for saliency detection. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence (IJCAI). Stockholm: AAAI, 2018. 684–690.

[30] Wang TT, Borji A, Zhang LH, Zhang PP, Lu HC. A stagewise refinement model for detecting salient objects in images. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 4039–4048. [doi: 10.1109/ICCV.2017.433]

[31] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Neural Information Processing Systems (NIPS). Long Beach: Curran Associates Inc., 2017. 6000–6010.

[32] Qin XB, Zhang ZC, Huang CY, Dehghan M, Zaiane OR, Jagersand M. U^2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, 2020, 106: 107404. [doi: 10.1016/j.patcog.2020.107404]

[33] Tang ZQ, Peng X, Geng SJ, Wu LF, Zhang ST, Metaxas D. Quantized densely connected U-nets for efficient landmark localization. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 348–364. [doi: 10.1007/978-3-030-01219-9_21]

[34] Li GB, Yu YZ. Visual saliency based on multiscale deep features. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 5455–5463. [doi: 10.1109/CVPR.2015.7299184]

[35] Li GB, Yu YZ. Visual saliency detection based on multiscale deep CNN features. IEEE Trans. on Image Processing, 2016, 25(11): 5012–5024. [doi: 10.1109/TIP.2016.2602079]

[36] Wang LJ, Lu HC, Ruan X, Yang MH. Deep networks for saliency detection via local estimation and global search. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 3183–3192. [doi: 10.1109/CVPR.2015.7298938]

[37] Zhao R, Ouyang WL, Li HS, Wang XG. Saliency detection by multi-context deep learning. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1265–1274. [doi: 10.1109/CVPR.2015.7298731]

[38] Li X, Zhao LM, Wei LN, Yang MH, Wu F, Zhuang YT, Ling HB, Wang JD. DeepSaliency: Multi-task deep neural network model for salient object detection. IEEE Trans. on Image Processing, 2016, 25(8): 3919–3930. [doi: 10.1109/TIP.2016.2579306]

[39] Wang WG, Shen JB, Jia YD. Review of visual attention detection. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 416–439 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]

[40] Zhao XQ, Pang YW, Zhang LH, Lu HC, Ruan X. Self-supervised pretraining for RGB-D salient object detection. In: Proc. of the 36th AAAI Conf. Artificial Intelligence (AAAI). AAAI, 2022. 3463–3471. [doi: 10.1609/aaai.v36i3.20257]

[41] Qu LQ, He SF, Zhang JW, Tian JD, Tang YD, Yang QX. RGBD salient object detection via deep fusion. IEEE Trans. on Image Processing, 2017, 26(5): 2274–2285. [doi: 10.1109/TIP.2017.2682981]

[42] Cong RM, Lei JJ, Fu HZ, Huang QM, Cao XC, Hou CP. Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation. IEEE Trans. on Image Processing, 2018, 27(2): 568–579. [doi: 10.1109/TIP.2017.2763819]

[43] Feng D, Barnes N, You SD, McCarthy C. Local background enclosure for RGB-D salient object detection. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 2343–2350. [doi: 10.1109/CVPR.2016.257]

[44] Ren JQ, Gong XJ, Yu L, Zhou WH, Yang MY. Exploiting global priors for RGB-D saliency detection. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition workshops (CVPR). Boston: IEEE, 2015: 25–32. [doi: 10.1109/CVPRW.2015.7301391]

[45] Chen Q, Liu Z, Zhang Y, Fu KR, Zhao QJ, Du HW. RGB-D salient object detection via 3D convolutional neural networks. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence (AAAI). AAAI, 2021. 1063–1071. [doi: 10.1609/aaai.v35i2.16191]

[46] Chen H, Li YF. Progressively complementarity-aware fusion network for RGB-D salient object detection. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 3051–3060. [doi: 10.1109/CVPR.2018.00322]

[47] Fu KR, Fan DP, Ji GP, Zhao QJ, Shen JB, Zhu C. Siamese network for RGB-D salient object detection and beyond. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5541–5559. [doi: 10.1109/TPAMI.2021.3073689]

[48] Pang YW, Zhang LH, Zhao XQ, Lu HC. Hierarchical dynamic filtering network for RGB-D salient object detection. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 235–252. [doi: 10.1007/978-3-030-58595-2_15]

[49] Chen ZY, Cong RM, Xu QQ, Huang QM. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 7012–7024. [doi: 10.1109/TIP.2020.3028289]

[50] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[51] Zhu XZ, Su WJ, Lu LW, Li B, Wang XG, Dai JF. Deformable DETR: Deformable Transformers for end-to-end object detection. In: Proc.

of the 9th Int'l Conf. on Learning Representations (ICLR). OpenReview.net, 2021.

[52] Yan B, Peng HW, Fu JL, Wang D, Lu HC. Learning spatio-temporal Transformer for visual tracking. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 10428–10437. [doi: 10.1109/ICCV48922.2021.01028]

[53] Stoffl L, Vidal M, Mathis A. End-to-end trainable multi-instance pose estimation with transformers. arXiv:2103.12115, 2021.

[54] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations (ICLR). OpenReview.net, 2021.

[55] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

[56] Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 7242–7252. [doi: 10.1109/ICCV48922.2021.00717]

[57] Xie EZ, Wang WH, Yu ZD, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Proc. of the 35th Neural Information Processing Systems (NIPS). 2021. 12077–12090.

[58] Wang WH, Xie EZ, Li X, Fan DP, Song KT, Liang D, Lu T, Luo P, Shao L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 548–558. [doi: 10.1109/ICCV48922.2021.00061]

[59] Zhu HQ, Sun X, Li YX, Ma K, Zhou SK, Zheng YF. DFTR: Depth-supervised fusion Transformer for salient object detection. arXiv:2203.06429, 2022.

[60] Chen JN, Lu YY, Yu QH, Luo XD, Adeli E, Wang Y, Lu L, Yuille AL, Zhou YY. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306, 2021.

[61] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI). Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[62] Wang HY, Zhu YK, Adam H, Yuille A, Chen LC. MaX-DeepLab: End-to-end panoptic segmentation with mask Transformers. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 5459–5470. [doi: 10.1109/CVPR46437.2021.00542]

[63] Zhang YD, Liu HY, Hu Q. TransFuse: Fusing Transformers and CNNs for medical image segmentation. In: Proc. of the 24th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI). Strasbourg: Springer, 2021. 14–24. [doi: 10.1007/978-3-030-87193-2_2]

[64] Luo XD, Hu MH, Song T, Wang GT, Zhang ST. Semi-supervised medical image segmentation via cross teaching between CNN and Transformer. In: Proc. of the 2022 Int'l Conf. on Medical Imaging with Deep Learning. Zurich: PMLR, 2022. 820–833.

[65] Liu C, Yang G, Wang S, Wang HX, Zhang YH, Wang YT. TANet: Transformer-based asymmetric network for RGB-D salient object detection. arXiv:2207.01172, 2022.

[66] Chen X, Yan B, Zhu JW, Wang D, Yang XY, Lu HC. Transformer tracking. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 8122–8131. [doi: 10.1109/CVPR46437.2021.00803]

[67] Xie YT, Zhang JP, Shen CH, Xia Y. CoTr: Efficiently bridging CNN and Transformer for 3D medical image segmentation. In: Proc. of the 24th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI). Strasbourg: Springer, 2021. 171–180. [doi: 10.1007/978-3-030-87199-4_16]

[68] Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 13708–13717. [doi: 10.1109/CVPR46437.2021.01350]

[69] Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P. Res2Net: A new multi-scale backbone architecture. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652–662. [doi: 10.1109/TPAMI.2019.2938758]

[70] Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450, 2016.

[71] Ju R, Ge L, Geng WJ, Ren TW, Wu GS. Depth saliency based on anisotropic center-surround difference. In: Proc. of the 2014 IEEE Int'l Conf. on Image Processing (ICIP). Paris: IEEE, 2014. 1115–1119. [doi: 10.1109/ICIP.2014.7025222]

[72] Fan DP, Lin Z, Zhang Z, Zhu ML, Cheng MM. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(5): 2075–2089. [doi: 10.1109/TNNLS.2020.2996406]

[73] Cheng YP, Fu HZ, Wei XX, Xiao JJ, Cao XC. Depth enhanced saliency detection method. In: Proc. of the 2014 Int'l Conf. on Internet Multimedia Computing and Service (ICIMCS). Xiamen: ACM, 2014. 23–27. [doi: 10.1145/2632856.2632866]

[74] Chen SH, Fu Y. Progressively guided alternate refinement network for RGB-D salient object detection. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 520–538. [doi: 10.1007/978-3-030-58598-3_31]

[75] Ji W, Li JJ, Zhang M, Piao YR, Lu HC. Accurate RGB-D salient object detection via collaborative learning. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 52–69. [doi: 10.1007/978-3-030-58523-5_4]

[76] Liu ZY, Wang Y, Tu ZZ, Xiao Y, Tang B. TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In: Proc. of the 29th ACM Int'l Conf. on Multimedia (ACM). ACM, 2021. 4481–4490. [doi: 10.1145/3474085.3475601]

[77] Zhao XQ, Zhang LH, Pang YW, Lu HC, Zhang L. A single stream network for robust and real-time RGB-D salient object detection. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 646–662. [doi: 10.1007/978-3-030-58542-6_39]

[78] Fan DP, Gong C, Cao Y, Ren B, Cheng MM, Borji A. Enhanced-alignment measure for binary foreground map evaluation. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence (IJCAI). Stockholm: AAAI, 2018. 698–704.

[79] Fan DP, Cheng MM, Liu Y, Li T, Borji A. Structure-measure: A new way to evaluate foreground maps. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 4558–4567. [doi: 10.1109/ICCV.2017.487]

[80] Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned salient region detection. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Miami: IEEE, 2009. 1597–1604. [doi: 10.1109/CVPR.2009.5206596]

[81] Perazzi F, Krähenbühl P, Pritch Y, Hornung A. Saliency filters: Contrast based filtering for salient region detection. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Providence: IEEE, 2012. 733–740. [doi: 10.1109/CVPR.2012.6247743]

[82] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations (ICLR). San Diego: ICLR, 2015. 1–13.

[83] Zhang M, Ren WS, Piao YR, Rong ZK, Lu HC. Select, supplement and focus for RGB-D saliency detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 3469–3478. [doi: 10.1109/CVPR42600.2020.00353]

[84] Zhang M, Fei SX, Liu J, Xu S, Piao YR, Lu HC. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 374–390. [doi: 10.1007/978-3-030-58604-1_23]

[85] Wu JY, Sun FM, Xu R, Meng J, Wang FS. Aggregate interactive learning for RGB-D salient object detection. Expert Systems with Applications, 2022, 195: 116614. [doi: 10.1016/j.eswa.2022.116614]

[86] Huang NC, Yang Y, Zhang DW, Zhang Q, Han JG. Employing bilinear fusion and saliency prior information for RGB-D salient object detection. IEEE Trans. on Multimedia, 2022, 24: 1651–1664. [doi: 10.1109/TMM.2021.3069297]

[87] Jin WD, Xu J, Han Q, Zhang Y, Cheng MM. CDNet: Complementary depth network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 3376–3390. [doi: 10.1109/TIP.2021.3060167]

[88] Li GY, Liu Z, Chen MY, Bai Z, Lin WS, Ling HB. Hierarchical alternate interaction network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 3528–3542. [doi: 10.1109/TIP.2021.3062689]

[89] Sun P, Zhang WH, Wang HY, Li SY, Li X. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 1407–1417. [doi: 10.1109/CVPR46437.2021.00146]

[90] Ji W, Li JJ, Yu S, Zhang M, Piao YR, Yao SY, Bi Q, Ma K, Zheng YF, Lu HC, Cheng L. Calibrated RGB-D salient object detection. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 9466–9476. [doi: 10.1109/CVPR46437.2021.00935]

[91] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

附中文参考文献:

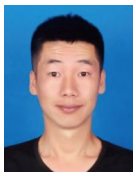
[39] 王文冠, 沈建冰, 贾云得. 视觉注意力检测综述. 软件学报, 2019, 30(2): 416–439. http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]

孙福明(1972－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为人工智能, 计算机视觉, 多媒体技术.

孙静(1992－), 女, 博士, 讲师, 主要研究领域为人工智能, 计算机视觉, 多媒体技术.

胡锡航(1997－), 男, 硕士生, 主要研究领域为人工智能, 计算机视觉, 多媒体技术.

王法胜(1983－), 男, 博士, 教授, CCF 高级会员, 主要研究领域为人工智能, 计算机视觉, 多媒体技术.



武景宇(1994－), 男, 硕士生, 主要研究领域为人工智能, 计算机视觉, 多媒体技术.