联合序列和空间注意力机制的光场显著性检测算法

姜文晖 程一波 方玉明 朱旻炜 左一帆
(江西财经大学信息管理学院 南昌 330013)

摘 要 光场图像包含丰富的空间视差信息和精确的深度信息,利用光场图像的丰富视觉信息实现准确的显著目标检测是重要的研究课题.然而,由于光场图像包含焦堆栈图像序列、全聚焦图像等多幅不同特性的图像数据,面向二维图像的特征提取方法不能有效地融合光场图像中不同序列不同空间的互补信息.针对这一问题,本文提出一种联合序列和空间注意力机制的光场显著性检测模型.针对焦堆栈图像序列,利用RFB模块和特征金字塔结构提取全局信息丰富且细节信息充分的语义特征.同时,提出一种联合序列和空间的自注意力机制,利用多头自注意力操作对焦堆栈图像特征从序列和空间维度联合建模,从而实现对焦堆栈图像序列特征的增强与融合.该机制能够同时建模图像的长距离、空间相关性和特征的内部关联性,从而在不同空间位置反映不同焦堆栈图像的重要性,有利于检测更完整的显著目标.最后,将焦堆栈图像信息和全聚焦图像信息有效融合,以预测最终的显著目标.本文在DUT-LFSD、HFUT-LFSD和LFSD数据集上展开实验,并与28种代表性工作进行对比.结果表明,本文设计的模型效果显著,在多个评价指标上一致地提高了显著目标检测的准确性.定性分析也表明本文提出的方法能够更准确地定位显著目标.

关键词 光场;显著性检测;特征金字塔;注意力机制;特征融合
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.01977

Light Field Saliency Detection Based on Joint Sequence and Spatial Attention Mechanism

JIANG Wen-Hui CHENG Yi-Bo FANG Yu-Ming ZHU Min-Wei ZUO Yi-Fan
(*School of Information Management, Jiangxi University of Finance and Economics, Nanchang* 330013)

Abstract Light field captures rich spatial and 3D layout information of the scenes. Therefore, light field salient object detection has attracted extensive research attentions recently. As light field contains multiple images with different characteristics including focal slices and all-focus images, conventional saliency detection methods based on RGB images fail to explore and integrate semantic information from focal slices, leading to suboptimal results because the relative contribution of different regions in focal slice sequences is ignored. In this paper, we propose a novel light field saliency detection method based on joint sequence and spatial attention mechanism. Firstly, we extract semantic features from focal slice sequence with RFB module and feature pyramid structure. The extracted features not only capture global context, but also retain rich scene details. Secondly, to integrate the salient features of focal slices comprehensively, we propose a joint sequence and spatial self-attention mechanism. Specifically, we introduce self-attention on

收稿日期:2022-09-09;在线出版日期:2023-03-09.本课题得到科技创新2030-"新一代人工智能"重大项目课题(No.2018AAA0100601)、国家自然科学基金项目(No.62132006, No.62161013, No.62162029, No.62271237)、江西省自然科学基金项目(No.20223AEI91002, No.20224BAB212010)、江西省重点研发计划项目(No.20203BBE53033)、江西省教育厅科技项目(No.GJJ2200522)资助。**姜文晖**,博士,讲师,中国计算机学会(CCF)会员(E1553M),主要研究方向为图像内容理解、跨媒体分析.E-mail:jiang1st@bupt.cn. **程一波**,硕士研究生,主要研究方向为图像内容理解. **方玉明(通信作者)**,博士,教授,中国计算机学会(CCF)会员(52103D),主要研究领域为计算机视觉、多媒体信号处理和视觉质量评估.E-mail:leo.fangyuming@foxmail.com. **朱旻炜**,硕士研究生,主要研究方向为计算机视觉. **左一帆**,博士,副教授,中国计算机学会(CCF)会员(E1547M),主要研究领域为图像处理和多媒体信号处理.

semantic features from focal slices within certain spatial context, which builds both spatial relations with long-range dependencies and inter-sequence correlations simultaneously. Such mechanism dynamically integrates semantic information with different importance on different feature, which is beneficial to predict complete salient object. Finally, we propose linear fusion to effectively aggregate information from both focal slices and all-focus image to generate accurate saliency maps. Our proposed method is simple to implement. We conduct comprehensive experiments on DUT-LFSD, HFUT-LFSD and LFSD, which are the most widely used benchmark for light field saliency detection. We compare the performance of our model with 28 state-of-the-art methods. The quantitative evaluation metrics include E-measure, S-measure, F-measure and MAE (mean absolute error). The experimental results demonstrate the superior performance of the proposed model in terms of all evaluation metrics. In addition, we also provide visualization comparisons of our method and representative competitors. The visualization analysis also verifies that our method effectively improves the detection precision of images with tiny objects and complicated backgrounds.

Keywords light field; saliency detection; feature pyramid; attention mechanism; feature aggregation

1 引 言

显著性检测旨在通过计算机模拟人类视觉注意力机制,以自动预测和定位场景的显著视觉信息.该任务既有助于研究人类视觉感知机理,又能服务于其它计算机视觉的高层任务,如目标识别[1]、图像分割[2]、目标跟踪[3]等,因而具有重要的研究价值.

根据输入图像类型的不同,显著性检测任务可以分为 2D(RGB 图像)、3D(RGB-D 图像)、4D(光场图像)三类图像的显著性检测.其中,2D 图像在显著性检测领域研究最多,但是在图像的弱纹理区域、遮挡区域、背景杂乱等场景的检测效果并不理想[4-6].3D 图像在 2D 图像的基础上引入了深度图(Depth Map)[7-11].深度图记录了场景到相机拍摄平面的距离,能够区分不同深度层的物体,从而减少背景的干扰,提高图像在弱纹理区域和遮挡区域的检测效果.然而,深度数据通常质量较低,难以为显著目标检测提供有效信息.光场图像拥有比深度图像更丰富的信息.通过数字重聚焦技术[12],可将光场数据合成为聚焦在不同焦平面的图像序列(即焦堆栈图像),更有利于分离显著目标.融合焦堆栈图像的聚焦区域可以合成一幅全聚焦图像,相比标准的 RGB 图像,其色彩、纹理更为清晰[13-14].鉴于以上优点,基于光场图像的显著性检测吸引了大量研究人员的关注[15-21].

如何有效融合焦堆栈图像序列提供的互补信息,以分离位于不同焦平面的前景与背景,是光场显著性检测的核心问题.主流的研究方法使用卷积神经网络(Convolution Neural Networks,CNN)提取各焦堆栈图像的语义特征,并将特征按预设顺序输入卷积递归网络(ConvLSTM)[22]以预测不同焦堆栈图像的全局权重,最后采用注意力机制融合不同焦堆栈图像的特征[1,15-16].该方案取得了较好的检测效果,但仍存在以下不足.第一,卷积运算的感受野有限,难以建模较大范围的上下文信息,不利于检测小目标和复杂环境下的显著物体.第二,ConvLSTM 网络对输入序列的顺序敏感,而现有光场数据集中焦堆栈图像序列长度不一且排列无序,降低了特征融合的效果.此外,由于递归网络的记忆能力有限,ConvLSTM 容易遗忘最先输入网络的焦堆栈图像,从而弱化该图像对最终结果的影响.第三,现有的全局融合方式忽略了空间位置对显著性预测结果的影响,导致难以完整地检测覆盖较大深度范围的物体.因此,如何有效聚合焦堆栈图像序列的特征信息,在复杂场景下分离位于不同焦平面的目标和背景仍是挑战问题.

针对以上问题,本文提出一种联合序列和空间注意力机制的光场显著性检测模型.受 Transformer 模型的启发,本文对焦堆栈图像序列提取的高层语义特征图,利用多头自注意力机制对焦堆栈图像特征从序列和空间维度联合建模.该模型具有以下优点.首先,在空间维度上,自注意力

机制能够挖掘图像长距离的空间相关性,从而检测更完整的显著目标.其次,在序列维度上,自注意力机制能够更好地构建特征的内部相关性,并且不依赖于输入序列的长度和顺序.最后,本文对焦堆栈图像的序列和空间联合建模,使序列特征融合过程中考虑更大范围的空间上下文信息,增加了空间位置敏感性,从而在不同空间位置反映不同焦堆栈图像的重要性,进而提高了特征融合的有效性.

本文在 LSFD[13]、HFUT-LFSD[14] 和 DUT-LFSD[15] 三个公开数据集上进行实验,并与 28 种先进的显著性检测模型进行比较.结果表明,本文设计的显著性检测模型在 MAE[23](Mean Absolute Error)、E-measure[24] (Enhanced-alignment measure)、S-measure[25] (Structure measure)、F-measure[26] 多个评价指标上一致地优于其它对比方法.可视化分析表明,本文提出的方法可以有效提高较小目标和复杂背景下的显著性检测精度.

2　相关工作

2.1　二维图像的显著性检测

早期的研究方法主要基于显著目标与背景对比度高、背景简单、光源单一等假设,设计具有颜色和纹理对比度的人工特征[27-32],或引入空间位置先验等信息度量图像的显著性[33-34].近年来,大量学者开始研究基于深度神经网络的显著性检测方法.

Vig 等人[4]较早地使用深度神经网络提取多层特征并输入线性分类器实现显著性检测.随后,大量学者挖掘和利用卷积神经网络不同特征图的性质以提高显著性检测的准确性.例如,Wu 等人[5]对特征在显著性检测任务上的效率进行分析,发现中间层的特征既保留了人眼可识别的底层信息,又具有高层语义信息,并提出了一种级联解码器(CPD)框架,只融合较深层的特征预测相对精确的显著图,放弃底层特征以加快网络预测效率.Wang 等人[35]提出了渐进式特征抛光网络(PFPN),通过对多层级特征渐进式优化,提升特征的质量,以预测高质量的显著图.为进一步融合不同尺度的特征图,Zhang 等人[36]提出一种注意力机制引导的上下文特征融合网络(ACFFNet)以调整不同通道的重要性,实现更鲁棒的显著性检测特征表达.为进一步优化特征,Wei 等人[37]提出了级联反馈解码器(F^3Net),级联多个相同的解码单元;每级解码器都生成一幅显著图;生成上一级显著图的特征反馈至下一级解码器的输入,以实现特征的逐级精化.另一方面,Hu 等人[38]提出了基于空间衰减上下文的显著性检测算法,通过在特征图中自适应地传播和聚合可变衰减的图像上下文特征用于预测显著图.为进一步抑制背景噪声对显著图预测的影响,Lu 等人[33]在神经网络模型中嵌入中心先验知识.Jian 等学者[39]利用视频帧的空间位置信息过滤背景的干扰,从而实现视频序列的显著性检测.

针对显著目标边缘检测不准确的问题,部分研究者利用目标的边缘作为辅助信息,提高显著目标边界检测的准确性.例如,Zhao 等人[40]提出边缘指导模型(EGNet)建模显著目标的准确边界,协助多层级显著性检测得到最后的显著图.相似地,Qin 等人[41]提出边界感知模型(BASNet),利用残差优化模块和混合损失函数对边界部分训练,得到的显著图具有更精确边界.近期,Fang 等人[42]提出基于不确定性感知的显著目标检测模型,该模型通过构建外轮廓和内轮廓像素处理子模块,针对性地处理目标边缘像素,有效提升了轮廓预测的准确性.

尽管二维图像的显著性检测技术取得了重要进步,但由于二维图像在纹理相似、背景暗光、场景复杂等情况下,前景和背景难以区分,导致显著目标检测仍不够准确.

2.2　光场图像的显著性检测

光场图像提供更丰富的深度信息,有利于分离复杂场景下的显著物体.因此,基于光场图像的显著目标检测是当前的研究热点.

Li 等人[13]提出首个由室内和室外场景组成的光场显著性数据库 LFSD.为了聚合各类光场数据,通过前背景线索将不同焦堆栈图像特征加权融合以预测图像的显著性.随后,Zhang 等人[14]提出加权稀疏编码显著性方法,对多种不同的特征进行加权稀疏编码预测图像的显著性,再将多组结果融合预测最终的显著图.Piao 等人[43]基于超像素的深度、位置、颜色等光场数据特征构建图模型预测显著目标,既融合了不同特征的互补性,又强调了显著图的空间一致性,从而较大程度地提高了光场显著性检测的效果.

此后,更多研究者构建深度神经网络实现光场显著性检测.例如,Li 等人[44]提出联合聚焦的方法,在前-背景相似或杂乱背景的场景中均匀地突出显著区域,同时更好地抑制背景区域,但该方法未建模

焦堆栈图像特征之间的联系.为解决该问题,Zhang等人[21]首次利用3D卷积网络提取焦堆栈图像的序列特征.Wang等人[1]先通过卷积神经网络提取各焦堆栈图像的语义特征,再利用ConvLSTM和注意力机制自适应地融合提取的特征序列,从而实现更有效的焦堆栈图像特征表达.然而,该方法仅对焦堆栈图像序列的最高层语义特征进行融合.作为改进,Zhang等人[15]则利用ConvLSTM对焦堆栈图像序列提取的多尺度语义特征逐一融合,以全面建模焦堆栈图像之间的内在关联.Piao等人[16]使用不同的注意权重融合不同堆栈图像的特征,这些注意权值由ConvLSTM通过多个时间步计算.作者还采用知识蒸馏来提高不同焦堆栈图像分支的特征表示能力.另一方面,Zhang等人[21]则对光场数据中的子孔径图像间的空间关联性建模,有效提高了光场显著性检测的准确性.针对光场显著性检测数据标注困难的问题,Feng等人[45]提出利用注意力模型预测的噪声标签指导光场显著性检测模型的学习.值得注意的是,该工作也采用ConvLSTM对焦堆栈图像的特征进行融合.

综上所述,ConvLSTM在光场数据特征融合中起到举足轻重的作用.但是,ConvLSTM在光场显著性检测中具有以下局限.第一,卷积操作的感受野较小,难以对图像较大距离的上下文进行建模;第二,LSTM结构难以记忆最先输入的焦堆栈图像,从而容易忽略其对显著性检测结果的影响;第三,全局融合方式缺乏空间敏感性,致使网络难以检测覆盖较大深度范围的物体.本文提出使用联合序列和空间的自注意力机制对焦堆栈图像的特征进行融合和增强,该模块不依赖于输入序列的长度和顺序,且融合过程中能考虑更大范围的空间上下文信息,同时对空间位置敏感.

3 模型设计

本文模型的整体结构如图1所示.模型由三部分组成,分别是焦堆栈图像预测模块、全聚焦图像预测模块和融合模块.其中,全聚焦图像预测模块针对全聚焦图像,建模不同空间区域的颜色对比度、纹理、形状等信息预测图像的显著目标;焦堆栈图像预测模块挖掘不同焦堆栈图像之间的深度差异性,实现显著目标与复杂背景的分离.焦堆栈图像预测模块为本模型的核心.为更好地建模焦堆栈图像序列中不同焦平面成像之间的序列关联,本文提出一种联合序列和空间的自注意力机制实现不同焦堆栈图像特征的融合与增强.最后,融合模块利用深度与颜色的互补性,将以上两个模块的预测结果有效融合,形成最终的显著图.

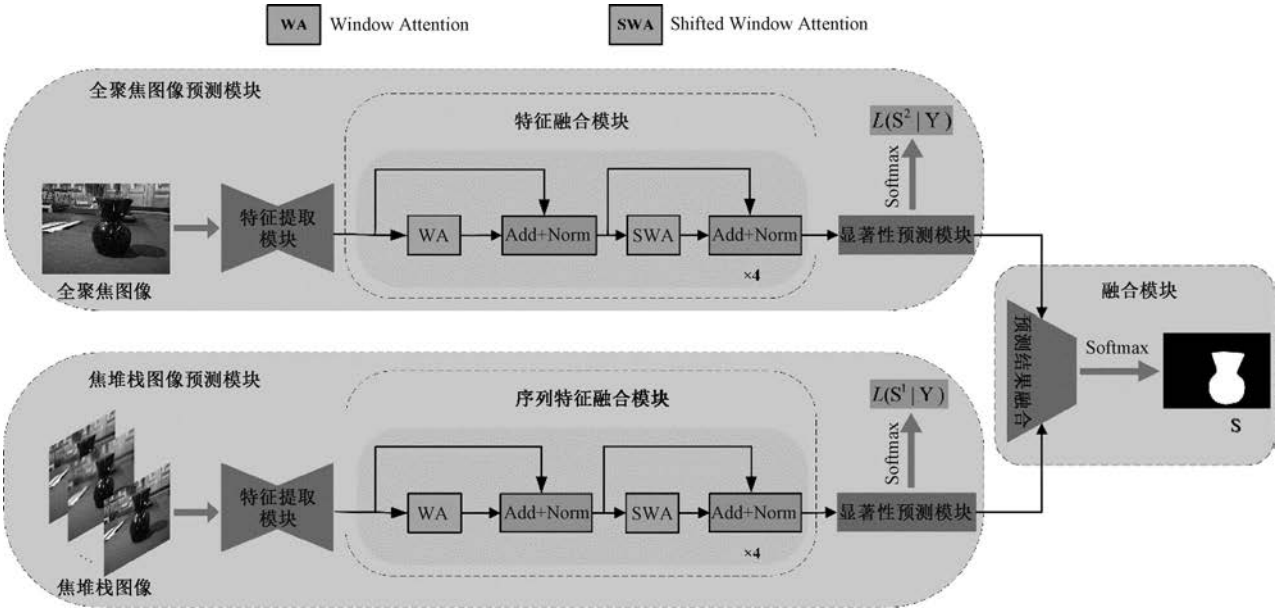

图1 模型整体结构图

3.1 焦堆栈图像预测模块

焦堆栈图像由一组焦距不同的图像堆叠组成.显著目标通常处于近邻的深度面.通过比较不同焦堆栈图像的聚焦区域,可以更有效地挖掘背景信息,从而完整地定位图像中的显著目标.

为充分利用焦堆栈图像的聚焦度信息,首先通

过特征提取模块,分别提取每幅焦堆栈图像的高分辨率语义特征图.随后,通过序列特征融合模块,以融合焦堆栈图像序列特征的互补信息,获得更精确的显著区域特征,提高光场图像显著性检测的准确性.

3.1.1　特征提取模块

本文面向显著性检测任务,基于 VGG19 网络[46]提取焦堆栈图像的高分辨语义特征.特征提取模块如图 2 所示.对于每张焦堆栈图像,通过 VGG19 网络提取图像的多尺度特征图.其中,高分辨率的特征图包含更多图像细节,而低分辨率的特征图提取了丰富的高层语义.为进一步利用多尺度特征图的丰富信息,本文使用特征金字塔结构(FPN)[47]融合多尺度特征图,使模块在得到高级语义信息的同时保留更多细节.同时,引入 RFB 模块[5]扩大网络的感受野,以建模更丰富的上下文信息.具体地,以第 i 幅焦堆栈图像为例,先将 VGG19 最后三层特征图分别输入 RFB 模块,输出结果按尺度由小到大分别记为 $\boldsymbol{V}_3$、$\boldsymbol{V}_4$、$\boldsymbol{V}_5$.随后从 $\boldsymbol{V}_3$ 层起逐级经过上采样和卷积操作与大尺度特征相加,最后将三层特征增强后的输出在通道维度拼接得到第 i 幅焦堆栈图像的特征图 $\boldsymbol{F}_i$.$\boldsymbol{F}_i \in R^{C\times W\times H}$,其中 C 表示特征维度,W 和 H 分别表示特征图的宽和高.

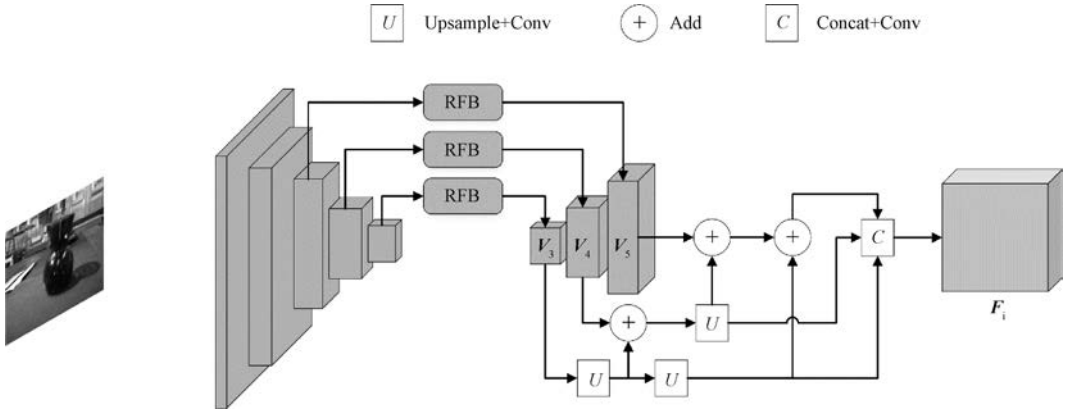

图 2　特征提取模块结构图

3.1.2　序列特征融合模块

为充分融合焦堆栈特征,本文提出一种联合序列和空间的自注意力机制融合焦堆栈图像的特征.该融合方式不依赖于输入序列的长度和顺序,且融合过程中能考虑更大范围的空间上下文信息.如图 1 所示,在序列特征融合模块,将窗口划分特征图计算注意力(Window Attention,WA)和滑动窗口划分特征图计算注意力(Shift Window Attention,SWA)作为一个注意力基本单元.本文重复多次基本运算单元实现焦堆栈图像特征的融合与增强.窗口划分特征图计算注意力(WA)的过程如下.

首先,为建立焦堆栈图像序列特征图 $\boldsymbol{F}=\{\boldsymbol{F}_1,\boldsymbol{F}_2,\ldots,\boldsymbol{F}_N\}$ 的上下文信息,以 $m\times m$ 为大小将特征图 $\boldsymbol{F}_i$ 划分为 M 个子图 $\{\boldsymbol{B}_{i,1},\boldsymbol{B}_{i,2},\ldots,\boldsymbol{B}_{i,M}\}$,其中 $\boldsymbol{B}_{i,j}\in R^{C\times m\times m}$ 是大小为 $m\times m$ 的特征子图,i 表示堆栈的序列,j 表示子图序号.

其次,引入多头自注意力机制对图像序列处于相同空位置子图的视觉特征进行自注意计算.如图 3 所示,将特征子图扁平化排列(flatten)后,得到子图的视觉特征编码 $\boldsymbol{X}_{i,j}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{m\times m}]$.焦堆栈图像序列第 j 个子图的特征编码为 $\ddot{\boldsymbol{X}}_j=[\boldsymbol{X}_{1,j},\boldsymbol{X}_{2,j},\ldots,\boldsymbol{X}_{N,j}]$,将 $\ddot{\boldsymbol{X}}_j$ 分别通过线性映射形成查询矩阵、键矩阵和值矩阵,即

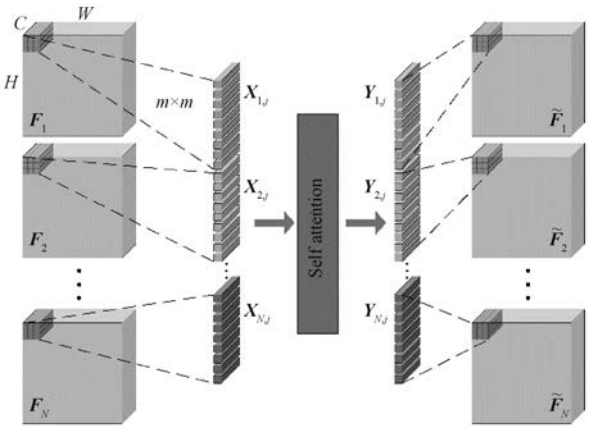

图 3　联合序列和空间的自注意力机制流程图

$$\begin{aligned}\boldsymbol{Q}_j &= \boldsymbol{W}^q \ddot{\boldsymbol{X}}_j \\ \boldsymbol{K}_j &= \boldsymbol{W}^k \ddot{\boldsymbol{X}}_j \\ \boldsymbol{V}_j &= \boldsymbol{W}^v \ddot{\boldsymbol{X}}_j\end{aligned} \tag{1}$$

其中，$\boldsymbol{Q}_j \in R^{(N\times m\times m)\times C}$ 为查询矩阵，$\boldsymbol{K}_j \in R^{(N\times m\times m)\times C}$ 为键矩阵，$\boldsymbol{V}_j \in R^{(N\times m\times m)\times C}$ 为值矩阵，$\boldsymbol{W}^q$、$\boldsymbol{W}^k$ 和 $\boldsymbol{W}^v$ 代表查询矩阵、键矩阵和值矩阵的映射矩阵.

随后，通过计算查询矩阵与键矩阵之间的相似性预测注意力权重矩阵. 较大的权重表示对应的值向量与查询的相关性更大. 结合权重矩阵和值矩阵，对不同的值向量加权融合，得到增强后的向量表示：

$$\ddot{\boldsymbol{Y}}_j = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_j (\boldsymbol{K}_j)^T}{sqrt(C)}\right)\boldsymbol{V}_j \tag{2}$$

其中，$\ddot{\boldsymbol{Y}}_j \in R^{(N\times m\times m)\times C}$. 当所有子图完成自注意力计算后，得到增强后的焦堆栈图像特征表示 $\ddot{\boldsymbol{Y}} = [\ddot{\boldsymbol{Y}}_1, \ddot{\boldsymbol{Y}}_2, \ldots, \ddot{\boldsymbol{Y}}_M]$. 最后，如图 3 所示，将 $\ddot{\boldsymbol{Y}}$ 按元素在原始特征序列的位置还原为特征图 $\widetilde{\boldsymbol{F}} = \{\widetilde{\boldsymbol{F}}_1, \widetilde{\boldsymbol{F}}_2, \ldots, \widetilde{\boldsymbol{F}}_N\}$，其中 $\widetilde{\boldsymbol{F}}_i \in R^{C\times W\times H}$.

由于窗口划分导致不同窗口之间的特征缺乏信息交互，限制了特征的表达能力. 本文引入了滑动窗口操作，通过重新划分窗口以实现跨窗口信息交互. 如图 4 所示，滑动窗口操作将窗口划分的边界向右下方向移动 $m/2$ 步长. 对位于特征图边缘的子图，本文参考 swin transformer[48] 的做法，将一侧不完整的子图平移至另一侧对不完整的子图补全. 其中相同色块代表同一窗口.

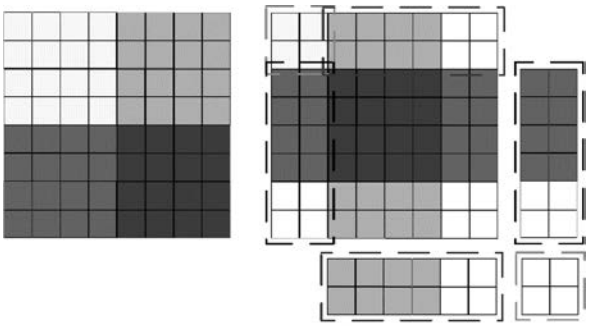

图 4 注意力模块窗口滑动划分示意图

3.1.3 显著性预测模块

如图 5 所示，将特征增强后的堆栈图像序列的特征图在通道维度上拼接.

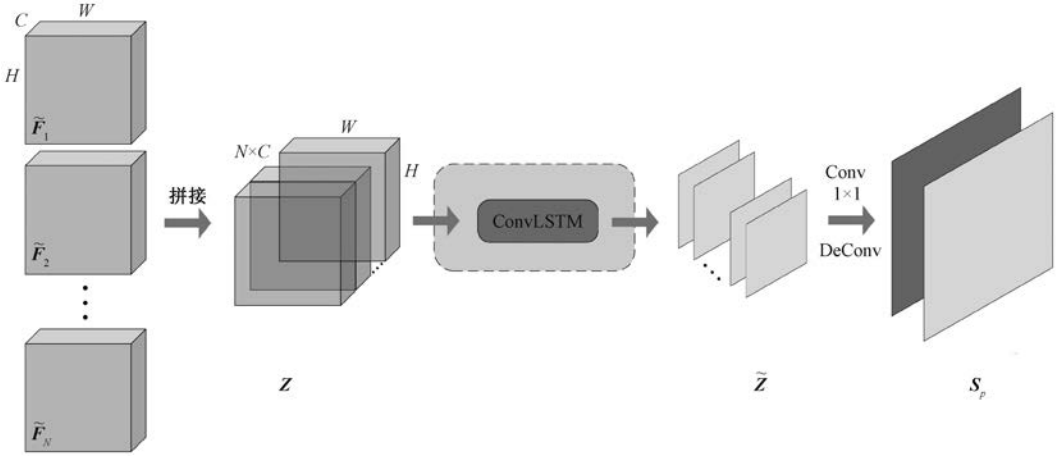

图 5 显著性预测模块结构图(其中 ConvLSTM 结构(虚线部分)仅在焦堆栈图像预测模块中使用)

$$\boldsymbol{Z} = concat(\widetilde{\boldsymbol{F}}_1, \widetilde{\boldsymbol{F}}_2, \ldots, \widetilde{\boldsymbol{F}}_N) \tag{3}$$

将拼接后的特征 Z 输入 ConvLSTM 模块[1]，进一步融合堆栈序列的全局信息，得到显著性预测特征 $\widetilde{\boldsymbol{Z}}$. 随后，使用 1×1 的卷积网络将 $\widetilde{\boldsymbol{Z}}$ 的特征通道降为 2，并使用反卷积网络[21]进行上采样，得到与输入图像相同分辨率的显著性检测结果 $\boldsymbol{S}_p^1 \in R^{2\times W\times H}$. 其中 $\boldsymbol{S}_p^1$ 中两个通道分别表示将图像预测为显著目标和背景的置信度.

3.2 全聚焦图像预测模块

全聚焦图像每个像素都是清晰的，因此本文利用全聚焦图像建模图像区域之间的对比度信息.

如图 1 所示，与焦堆栈图像预测模块相似，全聚焦图像预测模块主要由特征提取模块和特征融合模块组成. 其中，特征提取模块与焦堆栈图像特征提取模块相同，通过特征金字塔实现高分辨率的特征编码以保留输出显著图的细节信息，同时使用 RFB 模块加强特征的全局性. 随后，通过特征融合模块提取图像特征区域之间的对比度信息，用于预测最终的

显著图.特征融合模块基于空间注意力机制实现,可以视为联合序列和空间的自注意力机制在处理 N 为 1 的序列时的特例.与焦堆栈图像的预测不同,全聚焦图像预测模块不需要使用 ConvLSTM 模块(图 5 虚线部分),而是直接将注意力模块增强后的特征送入 1×1 的卷积网络将通道降为 2,再用反卷积网络上采样得到显著性检测结果 $\boldsymbol{S}_p^2$.

3.3　融合模块

本文将焦堆栈图像预测模块与全聚焦图像预测模块预测的结果融合,以充分利用两类图像数据的互补信息.记焦堆栈图像预测模块与全聚焦图像预测模块输出的显著性检测结果分别为 $\boldsymbol{S}_p^1$ 与 $\boldsymbol{S}_p^2$,本文将两个模块的预测结果线性融合:

$$\widetilde{\boldsymbol{S}}_p=\alpha\boldsymbol{S}_p^1+(1-\alpha)\boldsymbol{S}_p^2 \tag{4}$$

其中 $\alpha\in[0,1]$.最后,对 $\widetilde{\boldsymbol{S}}_p$ 在通道维度使用 $softmax$ 概率化,得到融合后的显著图.

3.4　损失函数

在训练焦堆栈图像预测模块时,首先,将焦堆栈图像预测模块和全聚焦图像预测模块的输出 $\boldsymbol{S}_p^1$ 和 $\boldsymbol{S}_p^2$ 在通道维度上使用 $softmax$ 函数概率化,得到显著图 $\boldsymbol{S}^1$ 和 $\boldsymbol{S}^2$.随后,使用交叉熵损失函数和 IoU 损失函数分别指导 $\boldsymbol{S}^1$ 和 $\boldsymbol{S}^2$ 的训练.具体地,本文对焦堆栈图像预测模块和全聚焦图像预测损失函数为:

$$Loss=\sum_{k=1}^{2}\left[L_{CE}(\boldsymbol{S}^k\mid\boldsymbol{Y})+L_{IoU}(\boldsymbol{S}^k\mid\boldsymbol{Y})\right] \tag{5}$$

其中:

$$L_{CE}(\boldsymbol{S}\mid\boldsymbol{Y})=-\frac{1}{n}\sum_{i=1}^{n}y_i log(\boldsymbol{S}_i) \tag{6}$$

$$L_{IoU}(\boldsymbol{S}\mid\boldsymbol{Y})=1-\frac{\sum_{i=1}^{n}y_i*\boldsymbol{S}_i}{\sum_{i=1}^{n}(y_i+\boldsymbol{S}_i-y_i*\boldsymbol{S}_i)} \tag{7}$$

其中 $\boldsymbol{S}_i^1$ 和 $\boldsymbol{S}_i^2$ 分别是 $\boldsymbol{S}^1$ 和 $\boldsymbol{S}^2$ 第 i 个像素的显著值,$\boldsymbol{Y}=\{y_i\}_{i=1}^n$ 是真实显著图,n 为显著图的像素总数.

4　实验结果与分析

4.1　数据集和评估指标

本文利用 LFSD[13]、HFUT-LFSD[14] 和 DUT-LFSD[15] 三个广泛使用的光场数据集对显著目标检测的性能进行评价.其中 LFSD 数据集[13] 是最早提出的光场显著性检测数据集,包含 100 组光场图像.HFUT-LFSD 数据集[14] 则包含 255 组光场图像,其中 100 组为训练数据.DUT-LFSD 数据集[15] 是光场显著性领域最大的数据集,包含了 1462 组光场图像(其中 1000 组为训练数据,462 组为测试数据),场景较为丰富.其中,1 组光场图像由 1 张真实显著图、1 张全聚焦图像和多张焦堆栈图像组成.参照光场显著性检测的一般设置[16,21,49-50],本文使用 DUT-LFSD 训练集和 HFUT-LFSD 训练集进行模型训练,并分别在 DUT-LFSD 的测试集和 LSFD 全集测试.为评估显著性检测的质量,我们使用 E-measure[24]、S-measure[25]、F-measure[26] 和 MAE[23] 等标准的评估指标验证模型的预测效果.

4.2　实施细节

对焦堆栈图像预测模块,本文使用 SGD 优化算法进行训练[16],权值衰减系数为 0.0005,动量为 0.99,初始学习率为 10e-8,迭代次数设置为 500000,批处理大小(batch size)为 1.对于全聚焦图像预测模块,本文参照显著性检测工作的常用方法[16,28,35],采用 Adam 优化器进行训练.初始学习率设为 0.0001,训练周期(epoch)设为 45.学习率在第 15 个训练周期之后开始衰减,每完成 1 个训练周期学习率衰减为当前的 0.99,批处理大小为 10.

考虑到公开数据集的规模相对较小,本文在训练中采用光场显著性模型常用的数据增强方式[51],即对图像随机裁剪、旋转和镜像翻转.另外,本文参照常用的数据增强方法对图像进行色度和对比度变化以进行数据增强.

4.3　消融实验与分析

4.3.1　联合序列和空间的自注意力机制的有效性

为验证联合序列和空间的自注意力机制的有效性,针对焦堆栈数据设计了 4 种不同的特征融合结构与本文提出的方法进行对比.

第一种结构对每组堆栈特征仅在堆栈维度采用自注意力机制进行特征融合,不引入空间上下文信息.该结构等价于将联合序列和空间的自注意力机制的空间窗口 m 设为 1.第二种结构对每张焦堆栈图像的特征独立地采用空间自注意力机制进行特征增强,不进行序列维度的特征融合.第三种结构先采用序列注意力机制融合堆栈序列特征,再采用空间注意力机制融合空间上下文信息.第四种结构先采用空间注意力机制融合空间上下文信息,再采用序列注意力机制融合堆栈序列特征.不同结构性能对比结果如表 1 所示.其中基准模型不采用任何额外的注意力融合策略.

表 1 DUT-LFSD 数据集上不同的焦堆栈序列融合方式对显著性检测性能的影响

方法	DUT-LFSD				LFSD			
	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑
基准	0.0362	0.9161	0.9432	0.9119	0.0810	0.8627	0.8721	0.8293
序列注意力	0.0354	0.9169	0.9445	0.9137	0.0797	0.8663	0.8758	0.8284
空间注意力	0.0355	0.9179	0.9442	0.9140	0.0803	0.8644	0.8754	0.8282
序列注意力＋空间注意力	0.0349	0.9182	0.9462	0.9146	0.0795	0.8693	0.8779	0.8346
空间注意力＋序列注意力	0.0346	0.9174	0.9457	0.9153	0.0790	0.8689	0.8794	0.8338
本文方法	**0.0338**	**0.9186**	**0.9472**	**0.9170**	**0.0781**	**0.8726**	**0.8918**	**0.8465**

由表 1 可见，序列注意力机制和空间注意力机制都一定程度提高了显著性预测的结果，说明通过自注意力机制建模焦堆栈图像的序列特征关联和空间特征关联均可以提高模型对显著性目标检测的准确性. 将这两种注意力机制以不同顺序串联加入基准模型，显著性检测性能相比基准模型有更大提升，说明结合序列和空间注意力可以有效提高模型性能. 但是独立地建模序列和空间维度的上下文不能充分挖掘特征之间的相关性. 本文提出的联合序列和空间的自注意力机制的模型结果在所有指标上都比其他对比结构显著提升，验证了联合序列和空间的自注意力机制的有效性.

4.3.2 空间窗口大小的影响

为进一步探究焦堆栈图像预测模块中联合序列和空间的自注意力机制的有效性，本文探究了空间窗口大小 m 对显著性检测性能的影响. 从表 2 可以看出，当窗口由小到大变化时，在 DUT-LFSD 测试集和 LFSD 数据集上显著性检测的性能先提高，后降低，取 8 时性能达到了最佳. 这是由于联合序列和空间的自注意力机制中的空间窗口大小的选择会影响融合特征的空间邻域视野以及参加自注意力计算的值向量的数量. 选择窗口过小会使得模块在融合堆栈特征时考虑的空间上下文信息有限，影响模块有效性. 而过大的窗口引入了不相关的背景噪声，降低模型融合堆栈特征的能力. 表 3 展示了窗口大小变化对全聚焦图像预测模块性能的影响. 相似地，8×8 的窗口大小下模型取得最佳性能. 因此，本文将焦堆栈图像预测模块和全聚焦图像预测模块中 m 均设为 8.

表 2 联合序列和空间的自注意力机制中 m 的变化对 DUT-LFSD 和 LFSD 数据集性能影响

m	DUT-LFSD				LFSD			
	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑
4	0.0352	0.9170	**0.9489**	0.9141	0.0797	0.8741	0.8783	0.8324
8	**0.0338**	**0.9186**	0.9472	**0.9170**	**0.0781**	0.8726	**0.8918**	**0.8465**
16	0.0355	0.9160	0.9473	0.9102	0.0790	**0.8747**	0.8794	0.8338

表 3 空间自注意力机制中 m 的变化对 DUT-LFSD 和 LFSD 数据集性能影响

m	DUT-LFSD				LFSD			
	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑
4	0.0387	0.9099	0.9443	0.9051	0.0840	0.8687	0.8713	0.8223
8	**0.0376**	**0.9143**	0.9456	**0.9091**	**0.0821**	0.8717	**0.8743**	**0.8269**
16	0.0379	0.9096	**0.9457**	0.9060	0.0825	**0.8719**	0.8737	0.8260
32	0.0383	0.9086	0.9449	0.9042	0.0832	0.8711	0.8724	0.8241

4.3.3 模型融合的有效性

为分析融合模块的有效性，设计了 2 种不同的融合结构与本文提出的线性融合方法进行对比. 第一种结构使用卷积操作，先将焦堆栈图像预测模块与全聚焦图像预测模块的显著性检测结果 $\boldsymbol{S}_p^1$ 和 $\boldsymbol{S}_p^2$ 从通道维度拼接，然后使用卷积操作降维，重新得到显著性检测结果 $\widetilde{\boldsymbol{S}}_p$. 第二种结构采用门机制[52]对显著性检测结果 $\boldsymbol{S}_p^1$ 和 $\boldsymbol{S}_p^2$ 融合.

实验结果如表 4 所示，本文采用的线性融合得到的结果最优，融合模型在数据集 DUT-LFSD 上的显著性检测结果在 MAE 指标上达到了 0.0304. 卷积融合和基于门机制的融合方法均一定程度利用了焦堆栈图像和全聚焦图像的互补信息，并一定程度提高了显著性预测的结果，但由于 DUT-LFSD 数据集的数据量有限，其效果不如线性融合.

最后，本文研究了融合权重 α 的变化对融合结果的影响. 如图 6 所示，当 α 为 0 时，模型退化为仅使用全聚焦图像进行显著性预测；当 α 增大时，

MAE 与 S-measure 性能逐步提高,当 α 为 0.5 时,线性融合效果最好.随着 α 进一步增大,全聚焦图像的重要性不断降低,显著性融合的效果也不断下降.因此,在本文的实验中,α 均取 0.5.

表 4　三种焦堆栈图像预测模块与全聚焦图像预测模块融合方法在 DUT-LFSD 和 LFSD 数据集的对比结果

方法	DUT-LFSD				LFSD			
	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑
焦堆栈图像预测模块	0.0338	0.9186	0.9472	0.9170	0.0781	0.8726	0.8918	0.8465
全聚焦图像预测模块	0.0376	0.9143	0.9456	0.9091	0.0821	0.8717	0.8743	0.8269
卷积融合	0.0320	0.9200	0.9520	0.9205	0.0736	0.8800	0.8966	0.8526
门机制融合	0.0326	0.9196	0.9508	0.9193	0.0743	0.8735	0.8921	0.8458
线性融合(本文)	**0.0304**	**0.9365**	**0.9553**	**0.9245**	**0.0720**	**0.8831**	**0.8973**	**0.8536**

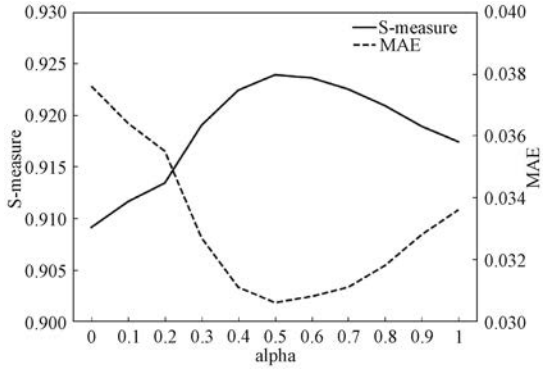

图 6　α 变化对 DUT-LFSD 数据集性能影响

4.3.4　多尺度特征融合结构的有效性

本文在特征提取模块中使用特征金字塔结构(FPN)[47]融合多尺度特征图,其核心是通过加法运算融合多尺度语义特征.本文与 CPD 网络[5]的多尺度特征融合结构进行了比较.具体地,采用 CPD 网络的乘法运算融合多尺度语义特征,其它模块保持不变.如表 5 所示,对焦堆栈图像预测模块和全聚焦图像预测模块,采用 FPN 结构实现多尺度特征融合都高于 CPD 结构,验证了本文设计的多尺度特征融合的有效性.

表 5　多尺度特征融合策略在 DUT-LFSD 和 LFSD 数据集的对比结果

方法	DUT-LFSD				LFSD			
	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑
FPN+焦堆栈	**0.0338**	**0.9186**	**0.9472**	**0.9170**	**0.0781**	**0.8726**	**0.8918**	**0.8465**
CPD+焦堆栈	0.0342	0.9118	0.9465	0.9129	0.0790	0.8747	0.8794	0.8338
FPN+全聚焦	**0.0376**	**0.9143**	**0.9456**	**0.9091**	**0.0821**	**0.8717**	**0.8743**	**0.8269**
CPD+全聚焦	0.0379	0.9144	0.9449	0.9073	0.0828	0.8708	0.8732	0.8252

4.4　对比实验与分析

我们与 28 种先进的显著性检测方法进行比较.其中,基于光场图像的显著性检测方法包括 PANet[49]、LPNL[45]、DGENet[53]、WSLF[54]、SANet[21]、ERNet[16]、DCN[49]、LFNet[50]、DLFS[18]和 LFS[13];基于 3D 图像的显著性检测方法包括 FSLNet[11]、TMFNet[55]、ASIF-Net[10]、HAINet[56]、D3Net[9]、HDFNet[57]、JLDCF[58]、S2MA[59]和 PDNet[7];基于 RGB 图像的显著性检测方法包括 ICON[60]、UDNet[42]、F^3Net[37]、MINet[61]、EGNet[40]、PoolNet[62]、BASNet[41]、PiCANet[63]和 R^3Net[64].

由表 6 可见,对比基于光场图像的显著性检测模型,在 DUT-LFSD 数据集上,本文提出的方法在所有指标都取得了当前最高性能.其中 MAE 为 0.030,比当前最好的模型 SANet 高出 0.002,说明本文模型的显著性检测结果在像素级别的预测上更加精确.在 S-measure、E-measure 和 S-measure 指标也比其它最好的结果有一致提升,进一步证明该方法的鲁棒性.在 LFSD 测试集上,本文提出的方法优势更加明显.例如,在 F-measure 指标上,比最好的 ERNet 提高 0.028,其它指标也显著高于光场显著性检测的其它方法,说明本文提出的方法具有更强的泛化能力.

4.5　可视化分析

为了更直观地展示本文的模型和其他代表性模型的结果差异,本文在图 7 中展示了各方法在 DUT-LFSD 测试集上的显著性检测结果.

从对比结果中可见,本文提出的方法在前背景对比度差异较小、复杂背景和显著对象较小等情况下,更准确地预测了显著目标.从 3、6、7 等样例的对比结果可见,当前景与背景的颜色对比度较小时,BASNet、ERNet、MINet、PiCANet 不能很好地抑制背景干扰,而本文的模型可以准确地从复杂的背景中找出显著目标.从 8、9 等样例的对比结果可见,当显著对象和背景都比较复杂时,BASNet、HAINet、HDINet 和 HDRNet 检测的显著对象不够完整,而

表 6 与主流方法在 DUT-LFSD 和 LFSD 数据集性能比较

类型	方法	时间	DUT-LFSD				LFSD			
			MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑	MAE ↓	F-measure ↑	E-measure ↑	S-measure ↑
4D	PANet[49]	2023	0.042	0.892	0.941	0.897	0.080	0.853	0.882	0.842
	LPNL[45]	2022	0.091	0.813	—	—	0.111	0.804	—	—
	DGENet[53]	2022	0.040	0.881	0.944	0.897	0.075	0.839	0.890	0.847
	WSLF[54]	2022	0.043	0.884	0.937	0.889	0.080	0.835	0.880	0.831
	SANet[21]	2021	0.032	0.920	0.954	0.918	0.074	0.844	0.889	0.841
	ERNet[16]	2020	0.040	0.889	0.943	0.899	0.080	0.855	0.895	0.849
	DCN[49]	2020	—	—	—	—	0.116	—	—	0.810
	LFNet[50]	2020	0.055	0.833	0.878	0.913	0.092	0.805	0.882	0.820
	DLFS[18]	2019	0.076	0.801	0.891	0.841	0.147	0.715	0.806	0.737
	LFS[13]	2017	0.240	0.484	0.728	0.563	0.208	0.740	0.771	0.680
	本文		**0.030**	**0.937**	**0.955**	**0.925**	**0.072**	**0.883**	**0.897**	**0.854**
3D	FSLNet[11]	2022	—	—	—	—	0.074	0.861	0.894	0.862
	TMFNet[55]	2022	—	—	—	—	0.084	0.846	0.865	0.849
	ASIF-Net[10]	2021	—	—	—	—	0.089	0.858	—	0.814
	HAINet[56]	2021	0.042	0.897	0.932	0.902	—	—	—	—
	D3Net[9]	2021	0.039	0.911	0.947	0.906	0.086	0.821	0.877	0.827
	HDFNet[57]	2020	0.040	0.926	0.938	0.905	0.076	**0.883**	0.891	0.854
	JLDCF[58]	2020	0.047	0.881	0.926	0.896	0.078	0.854	0.882	0.854
	S2MA[59]	2020	0.048	0.874	0.920	0.891	0.094	0.820	0.873	0.837
	PDNet[7]	2020	0.111	0.763	0.864	0.803	0.116	0.780	0.849	0.786
2D	UDNet[42]	2023	0.035	0.933	0.944	0.922	0.077	0.865	0.883	0.843
	ICON[60]	2022	0.046	0.915	0.924	0.904	0.083	0.858	0.861	0.844
	F^3Net[37]	2020	0.054	0.884	0.912	0.888	0.104	0.801	0.818	0.810
	MINet[61]	2020	0.061	0.861	0.897	0.866	0.094	0.799	0.821	0.815
	EGNet[40]	2019	0.053	0.870	0.914	0.886	0.085	0.828	0.850	0.838
	PoolNet[62]	2019	0.051	0.868	0.919	0.889	0.088	0.813	0.857	0.813
	BASNet[41]	2019	0.042	0.897	0.927	0.899	0.082	0.834	0.874	0.832
	PiCANet[63]	2018	0.073	0.852	0.899	0.859	0.124	0.764	0.836	0.791
	R^3Net[64]	2018	0.097	0.801	0.828	0.815	0.117	0.811	0.786	0.798

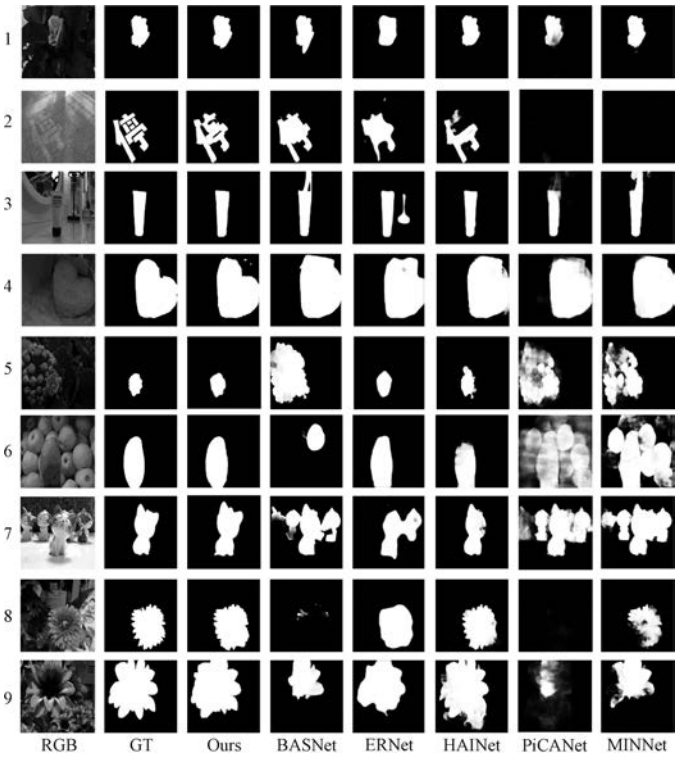

图 7 本文与其他代表性模型预测结果可视化对比

本文提出的方法可以完整地突出显著目标.由样例2的可视化结果可见,当显著目标与背景结合紧密,难以通过深度信息分离时,PiCANet、MINNet不能定位显著目标,BASNet、ERNet、HAINet无法准确区分显著目标与背景,而本文的方法可以在定位显著目标的同时将结合紧密的目标与背景分离.从样例1、5可以看出,当显著目标较小时,BASNet、ERNet、HAINet、PiCANet、MINNet都存在预测显著对象不准确或者预测显著对象不完整等问题,而本文的方法可以准确检测小目标.这些预测结果进一步验证了本文提出模型的有效性和鲁棒性.

5　结　论

本文提出一种以焦堆栈图像预测模块与全聚焦图像预测模块组成的双流结构网络实现显著目标检测.焦堆栈图像预测模块通过在特征提取模块中加入RFB和特征金字塔结构抽取全局信息丰富且保留更多细节的图像特征.同时,提出基于联合序列和空间的自注意力机制的序列特征模块,充分提取和融合不同焦堆栈图像的显著特征.最后将焦堆栈图像预测模块与全聚焦图像预测模块的预测结果经过线性融合得到最终的显著性检测结果.本文在光场数据集DUT-LFSD和LFSD上展开实验,并与28种代表性工作进行对比.结果表明,本文设计的模型效果显著,在多个评价指标上一致地提高了显著目标检测的准确性.可视化分析也表明本文提出的方法能够预测更加准确的显著目标.

参　考　文　献

[1] Wang T, Piao Y, Li X, et al. Deep learning for light field saliency detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 8838-8848

[2] Peng Q, Cheung Y M. Automatic video object segmentation based on visual and motion saliency. IEEE Transactions on Multimedia, 2019, 21(12): 3083-3094

[3] Zhang Wei Jun, Zhong Sheng, Xu Wen Hui, Wu Ying. Translated title of the contribution: Correlation filter based visual tracking integrating saliency and motion cues. ACTA AUTOMATICA SINICA, 2021, 47(7): 1572-1588 (in Chinese)
(张伟俊, 钟胜, 徐文辉. 融合显著性与运动信息的相关滤波跟踪算法. 自动化学报, 2021, 47(7): 1572-1588)

[4] Vig E, Dorr M, Cox D. Large-scale optimization of hierarchical features for saliency prediction in natural images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2798-2805

[5] Wu Z, Su L, Huang Q. Cascaded partial decoder for fast and accurate salient object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3907-3916

[6] Luo HuiLan, Yuan Pu, Tong Kang. Review of the methods for salient object detection based on deep learning. Acta Electronica Sinica, 2021, 49(7): 1417-1427. (in Chinese)
(罗会兰, 袁璞, 童康. 基于深度学习的显著性目标检测方法综述. 电子学报, 2021, 49(7): 1417)

[7] Zhu C, Cai X, Huang K, et al. PDNet: Prior-model guided depth-enhanced network for salient object detection//Proceedings of the IEEE International Conference on Multimedia and Expo. Shanghai, China, 2019: 199-204

[8] Chen H, Li Y. Progressively complementarity-aware fusion network for RGB-D salient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 3051-3060

[9] Fan D P, Lin Z, Zhang Z, et al. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(5): 2075-2089

[10] Li C, Cong R, Kwong S, et al. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. IEEE Transactions on Cybernetics, 2021, 51(1): 88-100

[11] Fu K, He J, Yang X. Few-shot learning-based RGB-D salient object detection: A case study. Neurocomputing, 2022, 512: 142-152

[12] Ji Xinxin, Piao Yongri, Zhang Miao, Jia Lingyao, Li Peihua. Robust depth estimation via light field focal stacks. Chinese Journal of Computers, 2022, 45(6): 1226-1240 (in Chinese)
(吉新新, 朴永日, 张淼, 贾令尧, 李培华. 基于光场焦点堆栈的鲁棒深度估计. 计算机学报, 2022. 45(6): 1226-1240)

[13] Li N, Ye J, Ji Y, et al. Saliency detection on light field//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA,2014: 2806-2813

[14] Zhang J, Wang M, Lin L, et al. Saliency detection on light field: A multi-cue approach. ACM Transactions on Multimedia Computing, Communications, and Applications, 2017, 13(3): 1-22

[15] Zhang M, Li J, Wei J, et al. Memory-oriented decoder for light field salient object detection. //Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, USA. 2019: 898-908

[16] Piao Y, Rong Z, Zhang M, et al. Exploit and replace: An asymmetrical two-stream architecture for versatile light field

saliency detection//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 11865-11873

[17] Liu Y M, Zhang J, Zhang X D, Sun R, Gao J. Review of saliency detection on light fields. Journal of Image and Graphics, 2020, 25(12): 2465-2483. (in Chinese)
(刘亚美, 张骏, 张旭东, 孙锐, 高隽. 光场显著性检测研究综述. 中国图象图形学报, 2020. 25(12): 2465-2483)

[18] Piao Y, Rong Z, Zhang M, et al. Deep light-field-driven saliency detection from a single view//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 904-911

[19] Zhang J, Wang M, Gao J, et al. Saliency detection with a deeper investigation of light field//Proceedings of the 24th International Conference on Artificial Intelligence. Tokyo, Japan, 2015: 2212-221

[20] Zhang M, Xu S, Piao Y, et al. Exploring spatial correlation for light field saliency detection: Expansion from a single view. IEEE Transactions on Image Processing, 2022, 31: 6152-6163

[21] Zhang Y, Chen G, Chen Q, et al. Learning synergistic attention for light field salient object detection//Proceedings of the 32nd British Machine Vision Conference. 2021

[22] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 802-810

[23] Perazzi F, Krähenbühl P, Pritch Y, et al. Saliency filters: Contrast based filtering for salient region detection//2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 733-740

[24] Fan D P, Gong C, Cao Y, et al. Enhanced-alignment measure for binary foreground map evaluation//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 698-704

[25] Fan D P, Cheng M M, Liu Y, et al. StructurE-measure: A new way to evaluate foreground maps//Proceedings of the IEEE International Conference on Computer Vision. Honolulu, USA, 2017: 4548-4557

[26] Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami Beach, USA, 2009: 1597-1604

[27] Liu Feng, Shen Tongsheng, Han Yanli, and Ma Xinxing. Saliency detection via background aware and color contrast. Journal of Computer-Aided Design and Computer Graphics, 2016, 28(10): 1705-1712. (in Chinese)
(刘峰, 沈同圣, 韩艳丽, 马新星. 融合背景感知和颜色对比的显著性检测方法. 计算机辅助设计与图形学学报, 2016, 28(10): 1705-1712)

[28] Lin HuaFeng, Li Jing, Liu Guo-Dong, Liang DaChuan, Li DongMin. Saliency detection method using adaptive background template and spatial prior. ACTA AUTOMATICA SINICA, 2017, 43(10): 1736-1748. (in Chinese)
(林华锋, 李静, 刘国栋, 梁大川, 李东民. 基于自适应背景模板与空间先验的显著性物体检测方法. 自动化学报, 2017, 43(10): 1736-1748)

[29] Zhang DongMing, Jin GuoQing, Dai Feng, et al. Salient object detection based on deep fusion of hand-crafted features. Chinese Journal of Computers, 2019, 42(9): 2076-2086. (in Chinese)
(张冬明, 靳国庆, 代锋等. 基于深度融合的显著性目标检测算法. 计算机学报, 2019, 42(9): 2076-2086)

[30] Cheng M M, Mitra N J, Huang X, et al. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(3): 569-582

[31] Zhu W, Liang S, Wei Y, et al. Saliency optimization from robust background detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2814-2821

[32] Jian M, Zhang W, Yu H, et al. Saliency detection based on directional patches extraction and principal local color contrast. Journal of Visual Communication and Image Representation, 2018, 57: 1-11

[33] Lu X, Jian M, Wang X, et al. Visual saliency detection via combining center prior and U-Net. Multimedia Systems, 2022, 28(5): 1689-1698

[34] Jian M, Wang J, Yu H, et al. Visual saliency detection by integrating spatial position prior of object with background cues. Expert Systems with Applications, 2021, 168: 114219

[35] Wang B, Chen Q, Zhou M, et al. Progressive feature polishing network for salient object detection//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 12128-12135

[36] Zhang J, Shi Y, Zhang Q, et al. Attention guided contextual feature fusion network for salient object detection. Image and Vision Computing, 2022, 117: 104337

[37] Wei J, Wang S, Huang Q. F^3Net: fusion, feedback and focus for salient object detection//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 12321-12328

[38] Hu X, Fu C W, Zhu L, et al. SAC-Net: Spatial attenuation context for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(3): 1079-1090

[39] Jian M, Wang J, Yu H, et al. Integrating object proposal with attention networks for video saliency detection. Information Sciences, 2021, 576: 819-830

[40] Zhao J X, Liu J J, Fan D P, et al. EGNet: Edge guidance

network for salient object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 8779-8788

[41] Qin X, Zhang Z, Huang C, et al. Basnet: Boundary-aware salient object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7479-7489

[42] Fang Y, Zhang H, Yan J, et al. UDNet: Uncertainty-aware deep network for salient object detection. Pattern Recognition, 2023, 134: 109099

[43] Piao Y, Li X, Zhang M, et al. Saliency detection via depth-induced cellular automata on light field. IEEE Transactions on Image Processing, 2020, 29: 1879-1889

[44] Li S, Deng H P, Zhu L, Zhang L. 2020. Saliency detection on a light field via the focusness and propagation mechanism. Journal of Image and Graphics, 25(12): 2578-2586. (in Chinese)
(李爽, 邓慧萍, 朱磊, 张龙. 2020. 联合聚焦度和传播机制的光场图像显著性检测. 中国图象图形学报, 25(12): 2578-2586)

[45] Feng M, Liu K, Zhang L, et al. Learning from pixel-level noisy label: A new perspective for light field saliency detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA, 2022: 1756-1766

[46] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 7-9

[47] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117-2125

[48] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 10012-10022

[49] Piao Y, Jiang Y, Zhang M, et al. PANet: Patch-aware network for light field salient object detection. IEEE Transactions on Cybernetics, 2023, 53(1): 379-391

[50] Zhang M, Ji W, Piao Y, et al. LFNet: Light field fusion network for salient object detection. IEEE Transactions on Image Processing, 2020, 29: 6276-6287

[51] Zhang J, Liu Y, Zhang S, et al. Light field saliency detection with deep convolutional networks. IEEE Transactions on Image Processing, 2020, 29: 4421-4434

[52] Jiang W H, Zhan K, Cheng Y B, Xia X, Fang Y M. 2022. The integrated mechanism of hierarchical decoders and dynamic fusion for image captioning. Journal of Image and Graphics, 27(9): 2775-2787. (in Chinese)
(姜文晖, 占锟, 程一波, 夏雪, 方玉明. 2022. 结合多层级解码器和动态融合机制的图像描述. 中国图象图形学报, 27(9): 2775-2787)

[53] Liang Y, Qin G, Sun M, et al. Dual guidance enhanced network for light field salient object detection. Image and Vision Computing, 2022, 118: 104352

[54] Liang Z, Wang P, Xu K, et al. Weakly-supervised salient object detection on light fields. IEEE Transactions on Image Processing, 2022, 31: 6295-6305

[55] Zhou W, Pan S, Lei J, et al. TMFNet: Three-input multilevel fusion network for detecting salient objects in RGB-D images. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(3): 593-601

[56] Li G, Liu Z, Chen M, et al. Hierarchical alternate interaction network for RGB-D salient object detection. IEEE Transactions on Image Processing, 2021, 30: 3528-3542

[57] Pang Y, Zhang L, Zhao X, et al. Hierarchical dynamic filtering network for rgb-d salient object detection// Proceedings of the European Conference on Computer Vision. Glasgow, USA, 2020: 235-252

[58] Fu K, Fan D P, Ji G P, et al. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3052-3062

[59] Liu N, Zhang N, Han J. Learning proceedings of the selective self-mutual attention for RGB-D saliency detection// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13756-13765

[60] Zhuge M, Fan D P, Liu N, et al. Salient object detection via integrity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3738-3752

[61] Pang Y, Zhao X, Zhang L, et al. Multi-scale interactive network for salient object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9413-9422

[62] Liu J J, Hou Q, Cheng M M, et al. A simple pooling-based design for real-time salient object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3917-3926

[63] Liu N, Han J, Yang M H. Picanet: Learning pixel-wise contextual attention for saliency detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3089-3098

[64] Deng Z, Hu X, Zhu L, et al. R3net: Recurrent residual refinement network for saliency detection//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Menlo Park, USA, 2018: 684-690

JIANG Wen-Hui, Ph. D., lecturer. His research interests include image content understanding and cross media analysis.

CHENG Yi-Bo, master student. His research interest is image content understanding.

FANG Yu-Ming, Ph. D., professor. His current research interests include computer vision, multimedia signal processing and visual quality assessment.

ZHU Min-Wei, master student. His research interest is computer vision.

ZUO Yi-Fan, Ph. D., associate professor. His current research interests include image processing and multimedia signal processing.

Background

Salient object detection aims to accurately identify and segment objects that most attract human's visual attention. It plays an important role in various tasks, such as object recognition, semantic segmentation and visual tracking.

Light field contains multiple images with different characteristics including focal slices and all-focus images, offering advantages in robustness to challenging scenes. Most existing light field saliency detection methods have achieved great success by exploiting unique light field data-focus information in focal slices. However, they process light field data in a slice-wise way, leading to suboptimal results because the relative contribution of different regions in focal slices is ignored.

In this paper, we propose a novel light field saliency detection method based on joint sequence and spatial attention mechanism. Our main idea is to utilize self-attention mechanism to build sequential and spatial relationship for better feature representation.

Overall, our method consists of three modules: semantic feature extraction, focal slices features aggregation and saliency prediction fusion. "Semantic feature extraction" extracts semantic features from focal slice sequence with RFB module and feature pyramid structure. The extracted features not only capture global context, but also retain rich scene details. "Focal slices features aggregation" builds both spatial relations with long-range dependencies and inter-sequence correlations simultaneously for focal slices features. "Saliency prediction fusion" effectively aggregates information from both focal slices and all-focus image to generate accurate saliency maps.

The experimental results show that our method predicts more accurate salient regions in various complex scenes compared with most other methods.