自适应卷积注意力与掩码结构协同的显著目标检测

朱 磊 袁金垚* 王文武 蔡小嫚
(武汉科技大学信息科学与工程学院 武汉 430000)

摘 要：显著目标检测(SOD)旨在模仿人类视觉系统注意力机制和认知机制来自动提取场景中的显著物体。虽然现有基于卷积神经网络 (CNN)或Transformer的模型不断刷新该领域方法的性能，但较少研究关注以下两个问题：(1)此领域多数方法常采用逐像素点的密集预测方式以获取像素显著值，然而该方式不符合基于人类视觉系统的场景解析机制，即人眼通常对语义区域进行整体分析而非关注像素级信息；(2)增强上下文信息关联在SOD任务中受到广泛关注，但通过Transformer主干结构获取长程关联特征不一定具有优势。SOD应更关注目标在适当区域内其中心-邻域差异性而非全局长程依赖。针对上述问题，该文提出一种新的显著目标检测模型，将CNN形式的自适应注意力和掩码注意力集成到网络中，以提高显著目标检测的性能。该算法设计了基于掩码感知的解码模块，通过将交叉注意力限制在预测的掩码区域来感知图像特征，有助于网络更好地聚焦于显著目标的整体区域。同时，该文设计了基于卷积注意力的上下文特征增强模块，与Transformer逐层建立长程关系不同，该模块仅捕获最高层特征中的适当上下文关联，避免引入无关的全局信息。该文在4个广泛使用的数据集上进行了实验评估，结果表明，该文提出的方法在不同场景下均取得了显著的性能提升，具有良好的泛化能力和稳定性。

关键词：显著目标检测；卷积神经网络形式的自适应注意力；掩码注意力；特征增强
中图分类号：TN911.7; TP391 **文献标识码**：A **文章编号**：1009-5896(2025)01-0260-11
DOI: 10.11999/JEIT240431

1 引言

人类视觉系统(Human Vision System, HVS)能够高效地从纷杂的输入信息中有选择地处理感兴趣的部分，在视觉信息处理过程中有着非常重要的意义。此研究基础上延伸出的显著目标检测(Salient Object Detection, SOD)任务则指利用视觉注意力原理对图像中最具视觉特征的对象或区域进行提取，对于理解图像中的重要信息起着关键作用[1]。SOD常用于自动关注图像中令人感兴趣的区域，而这些区域通常包含此图像的主体物体，因此可以消除冗余的背景信息，提高如物体跟踪[2]、物体识别[3]、机器人导航[4]等计算机视觉任务的性能。此外，SOD在多模态场景解析任务中也获得研究学者的关注并开拓了包括RGB-D SOD[5], RGB-T SOD[6]和光场SOD[7]等研究领域。

自1998年Itti等人[8]提出基于特征整合理论的可计算注意力模型以来，显著性检测逐步进入计算机视觉研究学者的视野。在此基础上发展起来的SOD则将该理论引入目标级别的分析中。2013年之前，SOD处于无监督的、基于经验线索的方法蓬勃发展的时期，而基于随机森林方法的出现[9]则进一步推动监督学习在该领域的发展。然而，无论上述无监督还是监督学习方法均采用基于经验线索的特征提取，导致难以在复杂背景中获得目标的高可区分性特征描述。2015年以来[10]，由于卷积神经网络(Convolutional Neural Network, CNN)网络具有强大的独特语义特征提取能力，基于CNN的SOD框架已经成为该研究领域的主流方案。例如，Lee等人[11]出将编码的低层距离图(Encoded Low-level Distance map, ELD)与来自深度卷积神经网络模型的高层特征结合，增强了显著性检测的性能。Wang等人[12]提出了深度神经网络局部估计(Deep Neural Network Local estimation, DNN-L)和深度神经网络全局搜索(Deep Neural Network Global search, DNN-G)两阶段的网络架构，通过整合低层显著性特征和高层对象特性，并利用深度神经网络来预测显著目标。而Transformer结构的引入则进一步促进了该领域的发展。由于Transformer结构能够快速建立长程依赖关系，所以与CNN相比，基于Transformer的SOD在定位突出区域方面表现更出色。例如，视觉显著性转换器(Visual Saliency Transformer, VST)[13]首次从序列到序列的视角重新思考了SOD，并提出了基于纯变压器的统一模型，通过引入新的标记上采样方法和多任务解码器，实现了优异的RGB和RGB-D SOD结果。SR-former[14]采用金字塔视觉转换器(Pyramid Vision Transformer, PVT)作为编码器骨干，设计了上下文细化模块(Context Refinement Module, CRM)

收稿日期：2024-05-13；改回日期：2024-09-18；网络出版：2024-09-24
*通信作者： 袁金垚 jyyuan202209@163.com

来引导解码器使用全局上下文信息，并生成局部上下文地图以获得更好的预测细节。

现有基于CNN或Transformer架构的SOD方法大多以逐点预测的方式对每个像素分配显著概率，这与 HVS进行场景解析的机制不一致。HVS通常对语义区域进行整体分析而非关注像素级信息。逐点形式的密集预测若不考虑像素间的依赖关系则可能导致同一语义区域内部像素的显著值分布不均匀、以及突出背景中局部对比度较大的区域。为此，已有的SOD方法通常会设计特征感受野增强方法以充分获取上下文信息[15]，而引入Transformer结构作为主干网络以提取具有全局长程依赖属性的特征似乎能够更好地解决此问题。值得注意的是，虽然Transformer结构在语义分割中(Semantic Segmentation, SS)获得了广泛关注[16,17]，考虑到图像中存在散列在空间位置上的不同类别物体，SS任务需要网络精细地建立像素间长程依赖以实现精确的多类分割。相比之下，SOD任务更关注目标与背景之间的对比度以评估其显著程度。因此，该任务的确需要捕获一定范围内上下文关联信息，但可能不需要全局长程关系。应注意到，相比CNN结构，在小规模数据环境中Transformer结构更难在训练中收敛[18]。

为解决上述问题，本文提出一种基于掩码预测的显著目标检测模型，通过将CNN形式的自适应注意力与掩码注意力集成到网络中，以提高显著目标检测的性能。本文的主要贡献如下：

(1)为解决SOD任务中逐点预测可能导致的同一语义区域内部像素的显著值分布不均匀、以及突出背景中局部对比度较大的区域问题，本文设计了掩码预测结构作为模型的解码器。本方法通过将交叉注意力限制在预测的掩码区域来感知图像特征，有助于网络更好地聚焦于显著目标的整体区域。

(2)相比当前流行的Transformer主干框架，本文采用通用卷积框架并采用基于卷积注意力以捕获最高层特征中的适当上下文关联，避免引入无关的全局信息。此外，这样做的好处还有两点：(1)主干网络的选择具有灵活性；(2)在SOD任务这类小规模数据环境中网络更容易收敛。

2　系统架构

本文针对显著性检测任务提出了一种新的框架，整体架构如图1所示。该框架主要包括基于卷积视觉转换器的特征增强模块(CNN Transformer-based Feature Enhancement, CTFE)、掩码感知视觉转换器模块(Mask-Aware Transformer, MAT)、特征融合模块(Feature Fusion Module, FFM)和损失计算模块。不失一般性，本文并未精心设计针对SOD的特征提取模型而选择了一个通用的CNN主干网络，并采用经典的特征金字塔网络(Feature

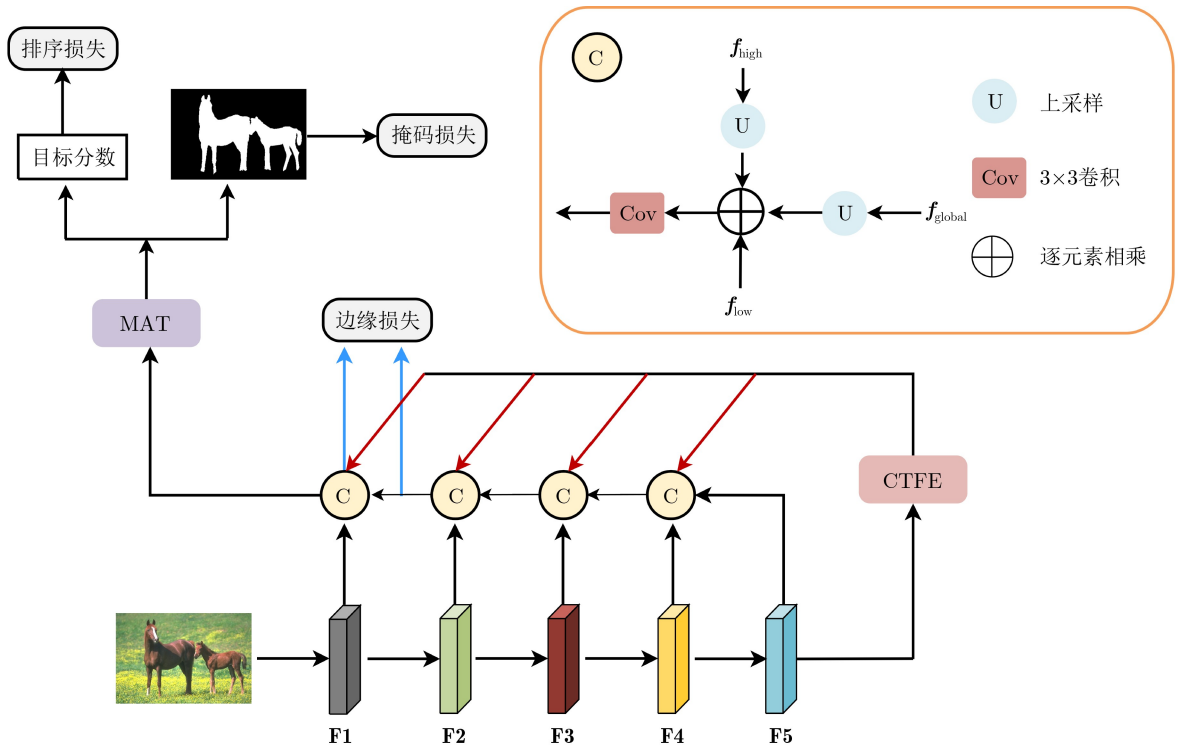


图 1 本文方法总体网络结构框图

Pyramid Network, FPN)以聚合来自主干的多尺度特征。其中本文选取$\mathbf{F}i\{i=1,2,3,4,5\}$的5个特征层作为主干特征，主干特征的大小分别为输入图像大小的1/2, 1/4, 1/8, 1/16和1/32。本文在主干网络自上而下的路径顶部引入了CTFE模块以实现**F5**特征中各像素对于全局上下文的捕获能力。在上采样阶段，本文通过FFM将增强后的**F5**特征与网络中的各尺度次级特征进行融合，并将融合后的特征送入MAT解码模块以预测最终的显著图。在训练阶段，本文采用了多个损失函数来有效地引导不同模块的训练。

2.1 主干网络

本文提出的方法并没有采用对称的U形结构，因此在选择主干网络时具有更大的灵活性。当前流行的CNN架构通常都是基于多尺度设计的，这些网络均可作为本方法的主干网络。本研究选择Res2Net[19]作为主干网络。Res2Net擅长于有效地从图像中提取特征，从而可以为显著目标检测模型提供可靠的基础特征描述。

2.2 基于卷积视觉转换器的特征增强模块(CTFE)

非Transformer结构的主干网络获取的特征难以形成对全局上下文信息的描述。因此，在密集预测任务中使用纯CNN架构大多设计了全局语义提取模块以捕获全局上下文信息。然而Transformer结构中的注意力机制使得其更具优势。受ConvFormer[20]的启发，本文设计了基于CNN注意力的全局语义提取模块，与标准Transformer结构相比，它有助于更好地注意力收敛，并有效避免了在小规模数据环境训练中注意力坍塌的风险。该模块由CNN式的自适应注意力(CNN-style Self-Attention, CSA)和卷积前馈网络(Convolutional Feed-Forward Network, CFFN)两部分组成。对于输入的特征图，CSA通过自适应地生成可缩放卷积来建立适当的依赖性，并在CFFN中通过应用连续卷积来细化每个像素的特征。CTFE的结构图如图2所示，其中，CBR表示卷积、批处理归一化和ReLU的简单组合。

CTFE中长程依赖性的构建依赖于CNN式的自注意力机制，该自注意力通过构建特定的卷积核为每个像素创建自适应感受野。具体来说，对于$\mathbf{F5}\in\mathbb{R}^{c_m\times\frac{H}{32}\times\frac{W}{32}}$中的每个像素$x_{i,j}$，首先通过余弦相似度得到初始卷积核$\boldsymbol{I}_{i,j}\in\mathbb{R}^{\frac{H}{32}\times\frac{W}{32}}$，其计算过程如式(1)、式(2)所示

$$\boldsymbol{Q}=\mathrm{Conv}_{3\times3}(\mathbf{F5}),\boldsymbol{K}=\mathrm{Conv}_{3\times3}(\mathbf{F5}) \tag{1}$$

$$\boldsymbol{I}_{m,n}^{i,j}=\left(\sum_{c=0}^{c_q}\boldsymbol{Q}_{i,j}\boldsymbol{K}_{m,n}\right)\Big/\left(\sqrt{\sum_{c=0}^{c_q}\boldsymbol{Q}_{i,j}^2}\sqrt{\sum_{c=0}^{c_q}\boldsymbol{K}_{m,n}^2}\right) \tag{2}$$

其中，c_m和c_q表示特征通道数，$\mathrm{Conv}_{n\times n}(\cdot)$为$n\times n$卷积操作符。$c_q$对应于ViT中$\boldsymbol{Q}$, $\boldsymbol{K}$和$\boldsymbol{V}$的嵌入维度。其将3×3邻域中相邻像素的特征合并到$x_{i,j}$。$\boldsymbol{I}_{\mathrm{m},n}^{i,j}$对应于注意力分数。然后，本文通过引入了一个可学习的高斯距离图$\boldsymbol{M}$[20]来动态地确定$x_{i,j}$的自定义的卷积核的大小。$\boldsymbol{M}^{i,j}$计算过程如式(3)所示

$$\boldsymbol{M}_{m,n}^{i,j}=\exp\left(-\frac{(i-m)^2\left(\frac{2^d}{H}\right)^2+(j-n)^2\left(\frac{2^d}{W}\right)^2}{2(\theta\times\alpha)^2}\right) \tag{3}$$

其中，$\theta\in(0,1)$是一个可学习的网络参数，用于控制卷积核$\boldsymbol{A}$的感受野，它与感受野成正比。而α是一个超参数，用于控制感受野的趋势。接着，$\boldsymbol{A}^{i,j}$可由$\boldsymbol{A}^{i,j}=\boldsymbol{I}^{i,j}\times\boldsymbol{M}^{i,j}$计算得出。这样，每一个像素$x_{i,j}$都有一个大小可扩展的卷积核变量$\boldsymbol{A}^{i,j}$。通过将变量$\boldsymbol{A}$与变量$\boldsymbol{V}$相乘，CSA可以构建自适应的远程依赖关系，其中变量$\boldsymbol{V}$的表达式类似于式(1)。最后，利用CBR来整合从远程依赖关系中学习到的特征。

CFFN则由两个CBR组件构成，可对CSA生成的特征精细化处理。整个CTFE模块的过程可由式(4)来表示

$$\begin{aligned}\boldsymbol{X}'=\boldsymbol{X}+\mathrm{CSA}(\boldsymbol{A})+\sum_{i=1}^{N}\mathrm{CFFN}(\boldsymbol{X}+\mathrm{CSA}(\boldsymbol{X})\\+\sum_{j=1}^{i-1}\mathrm{CFFN}(\boldsymbol{X}_j))\end{aligned} \tag{4}$$

2.3 特征融合模块(FFM)

特征融合模块的目的是聚合骨干网络中不同尺

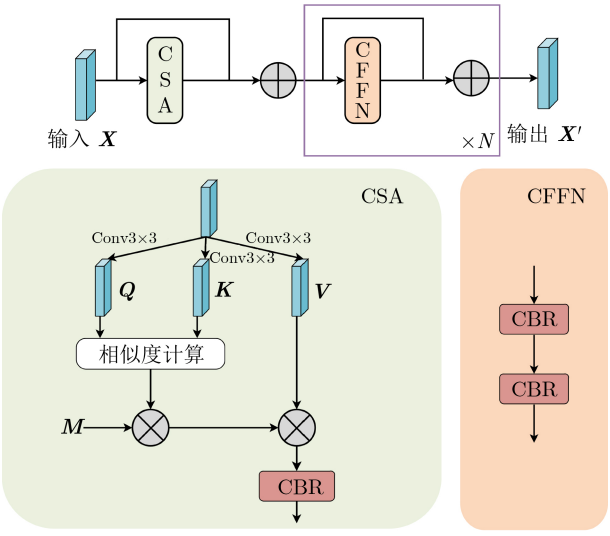


图 2 基于卷积视觉转换器的特征增强模块(CTFE)

度上提取的特征，实现低级细节信息和高级语义信息之间的交互。在所提出的网络中，每个尺度的特征融合包括3种类型的特征：来自当前尺度的特征$\boldsymbol{f}_{\text{high}}$、来自下一尺度的特征$\boldsymbol{f}_{\text{low}}$和全局语义特征$\boldsymbol{f}_{\text{global}}$。本文设计了以下两种特征融合方法，如图3。

在Attention-Fusion中，本文将来自CTFE模块的全局语义特征$\boldsymbol{f}_{\text{global}}$经过转置卷积和双线性插值操作进行尺度调整，以匹配低级特征$\boldsymbol{f}_{\text{low}}$的大小。然后，利用全局语义特征作为注意力权重，筛选出低尺度特征中有用的信息。最后，将过滤后的低级特征与高级特征$\boldsymbol{f}_{\text{high}}$进行相加(第1次应该是与来自layer5的特征进行相加)，以增强模型的特征表示，其计算过程如式(5)和式(6)所示

$$\boldsymbol{f}_{\text{att}} = \text{BL}(\text{T}-\text{Conv}(\boldsymbol{f}_{\text{global}})) \tag{5}$$

$$\boldsymbol{f}_{\text{fusion}} = \boldsymbol{f}_{\text{low}} \times \text{CA}(\boldsymbol{f}_{\text{att}}) + \boldsymbol{f}_{\text{high}} \tag{6}$$

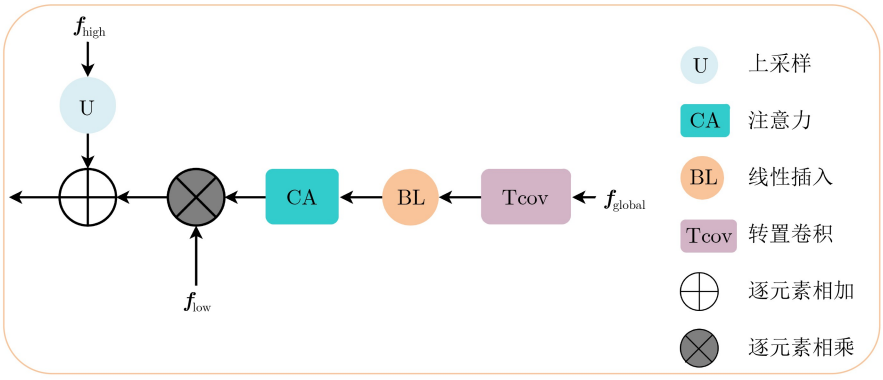
在Simple-Fusion中，本文将来自CTFE模块的全局语义特征$\boldsymbol{f}_{\text{global}}$和高级特征$\boldsymbol{f}_{\text{high}}$进行上采样，与来自主干的低级特征$\boldsymbol{f}_{\text{low}}$相加，之后通过一个3×3的卷积，得到融合特征。在实验过程中本文发现方法Simple-Fusion的评价指标优于方法Attention-Fusion，所以本文最终采用方法Simple-Fusion。

2.4　掩码感知视觉转换器模块(MAT)

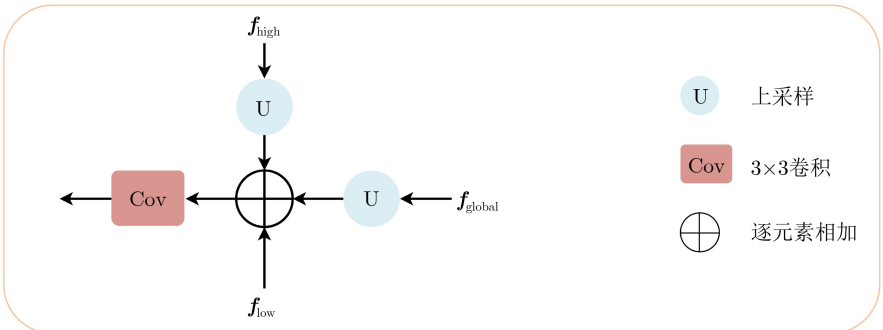
传统SOD方法的解码部分通过逐像素预测将显著概率单独分配给每个像素。而本文采用语义区域预测将显著性值分配给图像中具有相同语义属性的像素集。主干网络输出的顶层特征**F5**经过CTFE模块后，与来自主干网络的不同尺度的特征进行融合，融合后的特征进一步送入MAT，其结构图如图4所示。

掩码预测架构[21]最初用于语义/实例分割任务，将交叉注意力替换为掩码注意力，且只在每个查询的预测掩码的前景区域内进行关注。相比之下，MAT模块在以下两方面进行了创新：(1)多尺度特征融合：MAT模块的输入是融合后的多尺度特征，而非简单的特征金字塔，提升了特征的语义一致性和模型的鲁棒性。(2)目标分数输出：MAT模块输出每个掩码的目标分数，而非分类预测，使模型能够更精确地评估显著目标区域的置信度。

MAT模块有两组不同的输出：掩码预测和目标分数。该模块包括一个像素解码器和一个transformer解码器。融合后的特征图$\boldsymbol{I}$送入到像素解码器得到解码特征$g(\boldsymbol{I})$。同时，融合特征也被传递到transformer解码器中，在这个过程中我们将特征图作为键特征(Keys)和值特征(Values)，并将可学习的嵌



(a) Attention-Fusion



(b) Simple-Fusion

图 3 两种特征融合方法对比

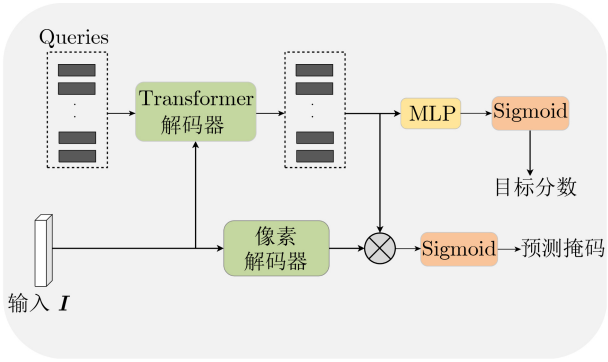


图 4 掩码感知视觉转换器模块

入作为查询特征(Queries)，输出得到n_q个逐掩码嵌入$\boldsymbol{q}_i$。之后通过在解码特征和逐掩码嵌入之间应用矩阵乘法并对结果应用一个逐元素sigmoid函数$\sigma(\cdot)$来得到最终的预测**Mask**，计算过程如式(7)所示

$$\mathbf{Mask}=\{\mathbf{Mask}_i|\mathbf{Mask}_i=\sigma(g(\boldsymbol{I})\boldsymbol{q}_i),i=1,2,\cdots,n_q\} \tag{7}$$

其中，$\boldsymbol{q}_i$表示第i次查询(即掩码嵌入)。对于每个掩码$\mathbf{Mask}_i$，通过将对应的逐掩模嵌入$\boldsymbol{q}_i$送到具有两个隐藏层的MLP和sigmoid函数来得到目标分数$o_i\in[0,1]$。

2.5 实验设置

在训练阶段，本文使用了3个目标函数用于优化网络，包括：掩码损失、排序损失和边缘损失，分别用L_{mask}，L_{rank}和L_{edge}表示。

(1)掩码损失：本文采用Dice损失作为掩码损失，通过最小化预测掩码与真值的差异来训练网络。计算方法如式(8)所示

$$L_{\text{mask}}=\text{Dice}(\boldsymbol{M}_i,\boldsymbol{g})=1-\frac{2\cdot\boldsymbol{g}\cdot\boldsymbol{M}_i}{\|\boldsymbol{g}\|+\|\boldsymbol{M}_i\|} \tag{8}$$

其中，$\|\cdot\|$表示L_1范数，$\boldsymbol{M}_i$和$\boldsymbol{g}$分别表示预测掩码和真值。

(2)排序损失：本文引入了排序损失作为辅助损失函数。具体来说，根据Dice损失值对每个样本进行从大到小排列得到索引，按照索引值选择对应的目标分数o_i。然后计算目标分数之间的差异，并将差异小于零的部分的绝对值相加，作为排序损失。该过程可通过构建目标分数与其转置的上三角矩阵来实现，排序损失能够约束模型在SOD任务中对于不同区域的优先级差异，进而提升目标检测任务的性能，计算方法如式(9)所示

$$L_{\text{rank}}=\sum_{i=1}^{n-1}\sum_{j>1}^{n}\max(0,\left|o_{s_i}-o_{s_j}\right|) \tag{9}$$

其中，o_{s_i}表示索引为s_i的样本的目标分数。

(3)边缘损失：考虑到边缘损失能够约束模型对边缘的预测，即使在输入图像中存在一些噪声或者不规则边缘，模型也能正确地分割出物体。本文利用交并比IOU和2元交叉熵BCE来计算边缘损失，将这些信息作为额外的监督信号以约束网络预测结果的边缘锐利程度和与人工标注的一致性。特别地，对于(i,j)处的像素，其边缘信息可由式(10)得到

$$E_{i,j}^k=\max(\boldsymbol{P}_{i,j}^k)-\min(\boldsymbol{P}_{i,j}^k) \tag{10}$$

其中，$\boldsymbol{P}_{i,j}^k$表示零填充后的预测和真值中以(i,j)为中心$k\times k$大小的局部区域，$\max(\cdot)$表示取$k\times k$个像素中的最大值，$\min(\cdot)$则表示取$k\times k$个像素中的最小值，本文将k设置为3。然后，计算预测边缘和真值边缘之间的BCE损失以及两者之间的IOU。最后，结合BCE损失与IOU的差异，得到每个样本的边缘损失。本文在两个不同尺度的特征上(大小分别为输入图像1/2和1/4)应用边缘损失，计算方法如式(11)–式(13)所示

$$L_{\text{edge1}}=1-\text{IOU}_1+\text{BCE}_1 \tag{11}$$

$$L_{\text{edge2}}=1-\text{IOU}_2+\text{BCE}_2 \tag{12}$$

$$L_{\text{edge}}=L_{\text{edge1}}+L_{\text{edge2}} \tag{13}$$

其中，IOU_1和IOU_2分别表示两个不同尺度特征上预测边缘和真值边缘之间的IOU，同理，BCE_1和BCE_2分别表示两个不同尺度特征上预测边缘和真值边缘之间的BCE损失。总体而言，本文的最终目标函数为

$$L=\lambda_1L_{\text{mask}}+\lambda_2L_{\text{rank}}+L_{\text{edge}} \tag{14}$$

其中，λ_1和λ_2是加权因子，在本文实验中都设置为1.0。

3 实验结果与分析

3.1 实验设置

为验证方法的性能，本文利用Pytorch深度学习框架实现所提出算法，并在单块RTX3090 GPU上进行训练。实验中本文将RGB图像统一缩放至384×384。初始学习率设置为6e–6，批量大小为4，共训练60个epoch。同时，使用水平翻转和随机裁剪作为数据增强方式。参数优化器采用Adam方法，其中权重衰减设置为0.000 1。本文将提出方法在4个广泛使用的数据集上进行实验，包括SOD, DUTS-TE, DUT-OMRON和ECSSD。这些数据集涵盖了不同领域的语义分割任务，具有不同的场景和物体类别，能够全面评估模型的性能和泛化能力。

本文采用2个评估指标对模型结果进行评估，其中包括平均绝对值误差(Mean Absolute Error,

MAE)和Max F-measure指标。具体来说，MAE是一种常用于评估预测模型准确度的指标，用于衡量预测值与真实值之间的平均差异。F-measure(F_β)是精确率(Precision)和召回率(Recall)的调和平均值，而Max F-measure($F_\beta^{\max}$)则指在不同的阈值下F-measure的最大值。

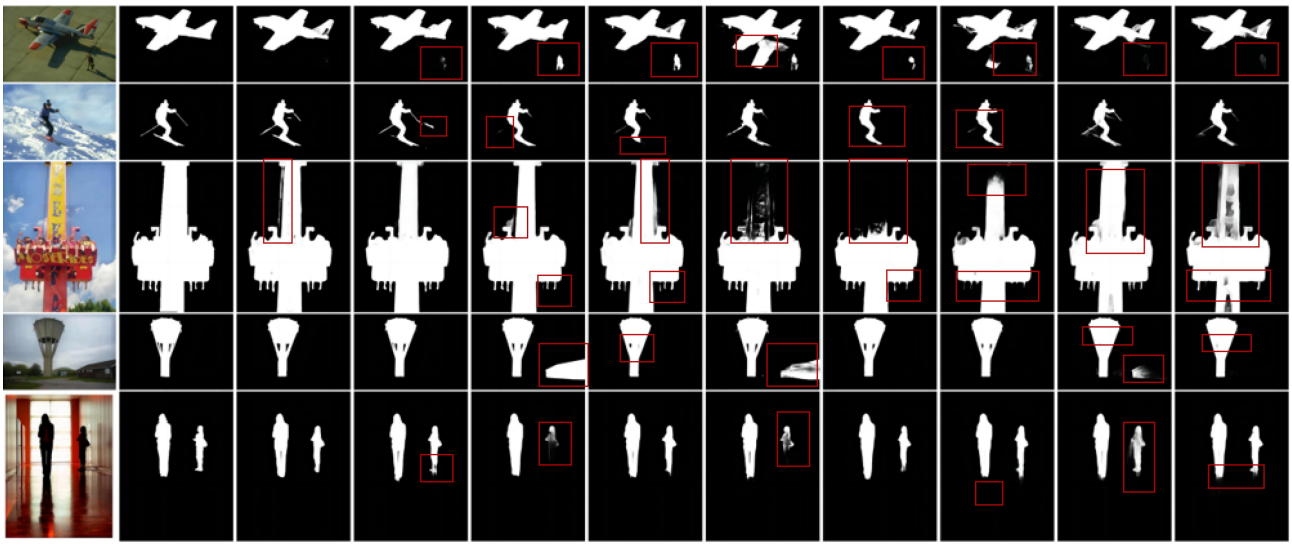
3.2 实验对比

本节将提出方法与8种代表性的方法进行了比较，包括EGNet[22], PoolNet[23], MINet[24], AADFNet[15], SACNet[25], ICON[26], MENet[27], VSCode[28]。表1为本文模型与这8种深度学习模型在MAE, Max F-measure()评价指标下的对比结果，其中，标注为红色字体的数值为最佳指标，标注为蓝色字体的数值为次优指标。可以观察到，本文提出的方法在4个数据库上的指标评估中取得了综合最佳的性能。在比较不同数据集时，本文算法在ECSSD数据集上表现最佳。此外，本文算法在SOD数据集上的MAE和最大F-measure方面达到了最佳性能。值得注意的是，本文方法得到的MAE值明显低于其他方法。本文方法与其他几种方法的定性评价结果如图5所示。从左至右分别为RGB图像、人工标注结果、本文提出方法结果和其他参与对比方法的结果。

根据对评估结果图的分析，可以看出，其他方法往往包括非显著区域或者忽略了一些细节部分，而本文方法得到的显著图与人工标注的结果更接近。特别是对于具有挑战性的情况，如(1)精细结构(见第1～3行)，(2)多个目标(见第5行)，(3)具有线状部分的物体(见第2行)，本文方法仍然可以得到比其他方法更合理的显著性图，显示了它的鲁棒性。进一步分析可发现，第1张图和第4张图的RGB信息中含有干扰部分，MENet和SACNet等方法在对该图进行分析时表现最差，预测结果中出现了非显著区域，表明其抗干扰能力不足。相比之下，本

表 1　所有参与评价方法在4个数据集上的Max F-measure, MAE测度的定量评价结果

方法(年份)	速度 (fps)	SOD		ECSSD		DUTS-TE		DUT-OMRON	
		MAE ↓	$F_\beta^{\max}$ ↑	MAE ↓	$F_\beta^{\max}$ ↑	MAE ↓	$F_\beta^{\max}$ ↑	MAE ↓	$F_\beta^{\max}$ ↑
EGNet(2019)	30.5	0.0969	0.8778	0.0374	0.9474	0.0386	0.8880	0.0528	0.8155
PoolNet(2019)	32.0	0.1000	0.8690	0.0390	0.9440	0.0400	0.8860	0.0560	0.8300
MINet(2020)	86.1	0.0920	0.8680	0.0342	0.9475	0.0373	0.8833	0.0559	0.8098
AADFNet(2020)	15.0	0.0903	0.8677	0.0280	0.9543	0.0314	0.8993	0.0488	0.8143
SACNet(2021)	11.2	0.0934	0.8804	0.0309	0.9512	0.0339	0.8944	0.0523	0.8287
ICON(2022)	58.5	0.0841	0.8790	0.0318	0.9503	0.0370	0.8917	0.0569	0.8254
MENet(2023)	45.0	0.0874	0.8780	0.0307	0.9549	0.0281	0.9123	0.0380	0.8337
VSCode(2024)	39.8	0.0602	0.8817	0.0245	0.9560	0.0262	0.9150	0.0473	0.8315
本文	46.0	0.0567	0.8872	0.0230	0.9508	0.0243	0.8966	0.0352	0.8290



(a) RGB　(b) GT　(c) 本文　(d) VSCode　(e) MENet　(f) ICON　(g) SACNet　(h) AADNet　(i) MINet　(j) PoolNet　(k) EGNet

图 5 本文方法与其他几种方法的定性评价结果

文所提CTFE模块采用基于卷积的注意力来捕获最高层特征中适当的上下文关联，避免引入无关的全局信息。针对第2张图中的滑雪板和雪杖边缘部分，MENet, ICON, AADNet 和 MINet 等方法均存在类似的模糊残缺问题，预测效果不理想。而本文方法通过利用IOU和BCE计算边缘损失，很好地改善了这一问题。

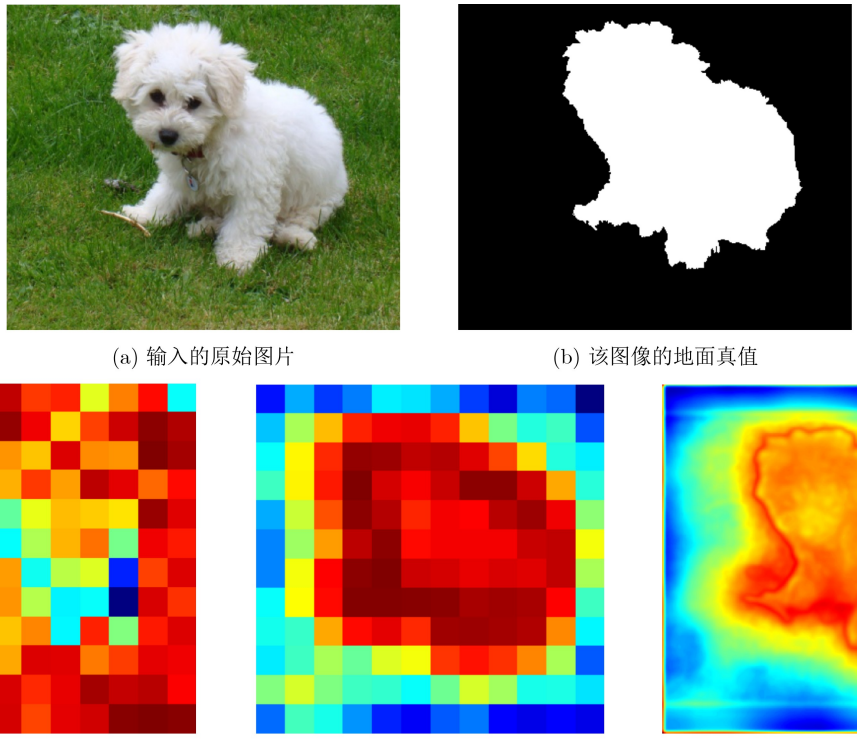
为了验证本文提出的CTFE模块和MAT模块在显著目标检测任务中的有效性，通过特征可视化方法生成了相应的热图，如图6所示。其中，图6(a)为输入的原始图片，图6(b)为该图像的地面真值，图6(c)–图6(e)分别为主干网络输出的特征图可视化结果、经过CTFE模块后的特征可视化结果和经过MAT模块后的特征可视化结果。从图6(d)可以看出，目标的激活值显著高于背景区域，体现出目标与背景之间更强的对比度。这说明CTFE模块使网络更加关注目标区域内的中心-邻域差异性。其次，图6(e)中的热图展示了目标整体激活的效果，显著目标与背景之间的界限更清晰。这意味着MAT模块能够有效地捕捉目标的全局特征，帮助模型更全面地理解显著目标的整体语义信息，而不仅仅是单一像素的显著性。

3.3 消融实验

为了证明本文提出的模型中每个模块的有效性，针对模型结构及优化策略中的边缘监督部分进行了消融实验，即CTFE模块和MAT模块、Canny边缘损失方法和利用IOU, BCE的边缘损失方法。这些消融实验在SOD数据集上进行，并使用与比较定量实验中相同的评价指标进行评估。消融实验中的实验设置与比较定量实验中的设置相同。实验结果如表2中的实验a～e所示。可以看出，添加CTFE和MAT模块后，评价指标均优于基线方法，该评价结果充分说明了提出模型的有效性。除此之外，当采用利用IOU, BCE的边缘损失方法时，评价指标优于Canny边缘损失，所以本文采用前者作为我们的监督策略之一。

前节提到了两种不同的特征融合模块。本文通过消融实验对这两种特征融合方法进行了定量评估，结果如表2中的实验f和g所示。可以观察到，当采用融合方法Simple-Fusion时，MAE指标下降了0.008，而最大F-measure值则增加了0.011。因此本研究采用了方法Simple-Fusion。

此外，本文还研究了不同损失函数的比重对检测效果的影响。根据表3中的数据可以看出，掩码损失的比重较大时(例如在设置1:1:1和1:0.5:0.5中)，评价指标表现较好。这表明，掩码损失对模型性能具有显著的正面影响，有助于模型更好地捕捉目标的整体结构。相反，当排序损失或边缘损失的比重较大时，MAE值增加，F分数降低。这表明，虽然边缘损失和排序损失可以帮助提高检测中的细节识



(a) 输入的原始图片 (b) 该图像的地面真值

(c) 主干网络输出的特征图可视化结果 (d) 经过CTFE模块后的特征可视化结果 (e) 经过MAT模块后的特征可视化结果

图 6 特征可视化结果图

表 2　不同模块的定量消融实验结果

实验	方法	SOD	
		MAE ↓	$F_\beta^{\max}$ ↑
a	Baseline	0.109 1	0.869 6
b	Baseline+CTFE	0.102 0	0.875 5
c	Baseline+CTFE+MAT	0.056 7	0.887 2
d	Baseline+CTFE+MAT+ Canny Loss	0.058 0	0.885 3
e	Baseline+CTFE+MAT+ IOU_BCE Loss	0.056 7	0.887 2
f	Attention-Fusion	0.064 7	0.876 1
g	Simple-Fusion	0.056 7	0.887 2

表 3　不同损失比重的实验结果

损失比重			SOD	
L_{mask}	L_{rank}	L_{edge}	MAE ↓	$F_\beta^{\max}$ ↑
1	0.5	0.5	0.060 0	0.883 3
0.5	1	0.5	0.058 9	0.873 5
0.5	0.5	1	0.073 5	0.871 4
1	1	1	0.056 7	0.887 2

别能力，但过高的比重可能对整体检测效果产生负面影响。因此，应该合理地调整各损失的比重。在实验中采用了1:1:1的设置。

4　结论

本文提出一种基于掩码预测的显著目标检测模型，通过将CNN形式的自适应注意力与掩码注意力集成到网络中，以提高显著目标检测的性能。为解决SOD任务中逐点预测可能导致的同一语义区域内部像素的显著值分布不均匀、以及突出背景中局部对比度较大的区域问题，本文设计了掩码预测结构作为模型的解码器。该方法通过将交叉注意力限制在预测的掩码区域来感知图像特征，有助于网络更好地聚焦于显著目标的整体区域。进一步地，相比当前流行的Transformer主干框架，本文采用通用卷积框架并采用基于卷积注意力以捕获最高层特征中的适当上下文关联，避免引入无关的全局信息。在4个广泛使用公开数据集上的定量和定性的实验结果表明，提出方法相比较参与对比的方法取得了更优秀的性能。同时，本文通过消融实验进一步分析了所提出方法的网络结构对于网络整体性能的贡献。

参 考 文 献

[1] ZHOU Huajun, XIE Xiaohua, LAI Jianhuang, *et al*. Interactive two-stream decoder for accurate and fast saliency detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 9138–9147. doi: 10.1109/CVPR42600.2020.00916.

[2] LIANG Pengpeng, PANG Yu, LIAO Chunyuan, *et al*. Adaptive objectness for object tracking[J]. *IEEE Signal Processing Letters*, 2016, 23(7): 949–953. doi: 10.1109/LSP.2016.2556706.

[3] RUTISHAUSER U, WALTHER D, KOCH C, *et al*. Is bottom-up attention useful for object recognition?[C]. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, USA, 2004: II-II. doi: 10.1109/CVPR.2004.1315142.

[4] ZHANG Jing, FAN Dengping, DAI Yuchao, *et al*. RGB-D saliency detection via cascaded mutual information minimization[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 4318–4327. doi: 10.1109/ICCV48922.2021.00430.

[5] LI Aixuan, MAO Yuxin, ZHANG Jing, *et al*. Mutual information regularization for weakly-supervised RGB-D salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(1): 397–410. doi: 10.1109/TCSVT.2023.3285249.

[6] LIAO Guibiao, GAO Wei, LI Ge, *et al*. Cross-collaborative fusion-encoder network for robust RGB-thermal salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(11): 7646–7661. doi: 10.1109/TCSVT.2022.3184840.

[7] CHEN Yilei, Li Gongyang, AN Ping, *et al*. Light field salient object detection with sparse views via complementary and discriminative interaction network[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(2): 1070–1085. doi: 10.1109/TCSVT.2023.3290600.

[8] ITTI L, KOCH C, and NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(11): 1254–1259. doi: 10.1109/34.730558.

[9] JIANG Huaizu, WANG Jingdong, YUAN Zejian, *et al*. Salient object detection: A discriminative regional feature integration approach[C]. 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 2083–2090. doi: 10.1109/CVPR.2013.271.

[10] LI Guanbin and YU Yizhou. Visual saliency based on multiscale deep features[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 5455–5463. doi: 10.1109/CVPR.2015.7299184.

[11] LEE G, TAI Y W, and KIM J. Deep saliency with encoded low level distance map and high level features[C]. 2016

IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 660–668. doi: 10.1109/CVPR.2016.78.

[12] WANG Linzhao, WANG Lijun, LU Huchuan, *et al*. Salient object detection with recurrent fully convolutional networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(7): 1734–1746. doi: 10.1109/TPAMI.2018.2846598.

[13] LIU Nian, ZHANG Ni, WAN Kaiyuan, *et al*. Visual saliency transformer[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 4702–4712. doi: 10.1109/ICCV48922.2021.00468.

[14] YUN Yike and LIN Weisi. SelfReformer: Self-refined network with transformer for salient object detection[J]. arXiv: 2205.11283, 2022.

[15] ZHU Lei, CHEN Jiaxing, HU Xiaowei, *et al*. Aggregating attentional dilated features for salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 30(10): 3358–3371. doi: 10.1109/TCSVT.2019.2941017.

[16] XIE Enze, WANG Wenhai, YU Zhiding, *et al*. SegFormer: Simple and efficient design for semantic segmentation with transformers[C]. The 35th International Conference on Neural Information Processing Systems, 2021: 924.

[17] WANG Libo, LI Rui, ZHANG Ce, *et al*. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, 190: 196–214. doi: 10.1016/j.isprsjprs.2022.06.008.

[18] ZHOU Daquan, KANG Bingyi, JIN Xiaojie, *et al*. DeepViT: Towards deeper vision transformer[J]. arXiv: 2103.11886, 2021.

[19] GAO Shanghua, CHENG Mingming, ZHAO Kai, *et al*. Res2Net: A new multi-scale backbone architecture[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(2): 652–662. doi: 10.1109/TPAMI.2019.2938758.

[20] LIN Xian, YAN Zengqiang, DENG Xianbo, *et al*. ConvFormer: Plug-and-play CNN-style transformers for improving medical image segmentation[C]. The 26th International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, Canada, 2023: 642–651. doi: 10.1007/978-3-031-43901-8_61.

[21] CHENG Bowen, MISRA I, SCHWING A G, *et al*. Masked-attention mask transformer for universal image segmentation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 1280–1289. doi: 10.1109/CVPR52688.2022.00135.

[22] ZHAO Jiaxing, LIU Jiangjiang, FAN Dengping, *et al*. EGNet: Edge guidance network for salient object detection[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 8778–8787. doi: 10.1109/ICCV.2019.00887.

[23] LIU Jiangjiang, HOU Qibin, CHENG Mingming, *et al*. A simple pooling-based design for real-time salient object detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 3912–3921. doi: 10.1109/CVPR.2019.00404.

[24] PANG Youwei, ZHAO Xiaoqi, ZHANG Lihe, *et al*. Multi-scale interactive network for salient object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 9410–9419. doi: 10.1109/CVPR42600.2020.00943.

[25] HU Xiaowei, FU C, ZHU Lei, *et al*. SAC-Net: Spatial attenuation context for salient object detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, 31(3): 1079–1090. doi: 10.1109/TCSVT.2020.2995220.

[26] ZHUGE Mingchen, FAN Dengping, LIU Nian, *et al*. Salient object detection via integrity learning[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3738–3752. doi: 10.1109/TPAMI.2022.3179526.

[27] WANG Yi, WANG Ruili, FAN Xin, *et al*. Pixels, regions, and objects: Multiple enhancement for salient object detection[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 10031–10040. doi: 10.1109/CVPR527292023.00967.

[28] LUO Ziyang, LIU Nian, ZHAO Wangbo, *et al*. VSCode: General visual salient and camouflaged object detection with 2D prompt learning[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 17169–17180. doi: 10.1109/CVPR52733.2024.01625.

朱 磊：男，副教授，研究方向为目标检测与识别、语义分割、场景解析等.

袁金垚：女，硕士生，研究方向为深度学习、语义分割.

王文武：男，副教授，研究方向为目标检测与识别、语义分割、场景解析等.

蔡小嫚：女，硕士生，研究方向为深度学习、语义分割.

责任编辑：余 蓉

Saliency Object Detection Utilizing Adaptive Convolutional Attention and Mask Structure

ZHU Lei　YUAN Jinyao　WANG Wenwu　CAI Xiaoman

(*School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan* 430000, *China*)

Abstract:

Objective　Salient Object Detection (SOD) aims to replicate the human visual system's attentional processes by identifying visually prominent objects within a scene. Recent advancements in Convolutional Neural Networks (CNNs) and Transformer-based models have improved performance; however, several limitations remain: (1) Most existing models depend on pixel-wise dense predictions, diverging from the human visual system's focus on region-level analysis, which can result in inconsistent saliency distribution within semantic regions. (2) The common application of Transformers to capture global dependencies may not be ideal for SOD, as the task prioritizes center-surround contrasts in local areas rather than global long-range correlations. This study proposes an innovative SOD model that integrates CNN-style adaptive attention and mask-aware mechanisms to enhance contextual feature representation and overall performance.

Methods　The proposed model architecture comprises a feature extraction backbone, contextual enhancement modules, and a mask-aware decoding structure. A CNN backbone, specifically Res2Net, is employed for extracting multi-scale features from input images. These features are processed hierarchically to preserve both spatial detail and semantic richness. Additionally, this framework utilizes a top-down pathway with feature pyramids to enhance multi-scale representations. High-level features are further refined through specialized modules to improve saliency prediction. Central to this architecture is the ConvoluTional attention-based contextual Feature Enhancement (CTFE) module. By using adaptive convolutional attention, this module effectively captures meaningful contextual associations without relying on global dependencies, as seen in Transformer-based methods. The CTFE focuses on modeling center-surround contrasts within relevant regions, avoiding unnecessary computational overhead. Features refined by the CTFE module are integrated with lower-level features through the Feature Fusion Module (FFM). Two fusion strategiesAttention-Fusion and Simple-Fusion—were evaluated to identify the most effective method for merging hierarchical features. The decoding process is managed by the Mask-Aware Transformer (MAT) module, which predicts salient regions by restricting attention to mask-defined areas. This strategy ensures that the decoding process prioritizes regions relevant to saliency, enhancing semantic consistency while reducing noise from irrelevant background information. The MAT module's ability to generate both masks and object confidence scores makes it particularly suited for complex scenes. Multiple loss functions guide the training process: Mask loss, computed using Dice loss, ensures that predicted masks closely align with ground truth. Ranking loss prioritizes the significance of salient regions, while edge loss sharpens boundaries to clearly distinguish salient objects from their background. These objectives are optimized jointly using the Adam optimizer with a dynamically adjusted learning rate.

Results and Discussions　Experiments were conducted using the PyTorch framework on an RTX 3090 GPU, with training configurations optimized for SOD datasets. The input resolution was set to 384×384 pixels, and data augmentation techniques, such as horizontal flipping and random cropping, were applied. The learning rate was initialized at 6e–6 and adjusted dynamically, with the Adam optimizer employed to minimize the combined loss functions. Experimental evaluations were performed on four widely used datasets: SOD, DUTS-TE, DUT-OMRON, and ECSSD.

The proposed model demonstrated exceptional performance across all datasets, showing significant improvements in Mean Absolute Error (MAE) and maximum F-measure metrics. For instance, on the DUTS-TE dataset, the model achieved an MAE of 0.023 and a maximum F-measure of 0.9508, exceeding competing methods such as MENet and VSCode.

Visual comparisons indicate that the proposed method generates saliency maps that closely align with the ground truth, effectively addressing challenging scenarios including fine structures, multiple objects, and complex backgrounds. In contrast, other methods often incorporate irrelevant regions or fail to accurately capture object details.

Ablation experiments validated the effectiveness of crucial components. For example, the incorporation of the CTFE module resulted in a reduction of MAE from 0.109 to 0.102. Additionally, the Simple-Fusion strategy outperformed the Attention-Fusion approach, yielding a lower MAE and a higher maximum F-measure score. The integration of IOU and BCE-based edge loss further enhanced boundary sharpness, demonstrating superior performance compared to Canny-based edge loss.

Heatmaps illustrate the contributions of the CTFE and MAT modules in emphasizing salient regions while preserving semantic consistency. The CTFE effectively accentuates center-surround contrasts, while the MAT captures global object-level semantics. These visualizations highlight the model's ability to focus on critical areas while minimizing background noise.

Conclusions This study presents a novel SOD framework that integrates CNN-style adaptive attention with mask-aware decoding mechanisms. The proposed model addresses the limitations of existing approaches by enhancing semantic consistency and contextual representation while avoiding excessive dependence on global variables. Comprehensive evaluations demonstrate its robustness, generalization capability, and significant performance enhancements across multiple benchmarks. Future research will investigate further optimization of the architecture and its application to multimodal SOD tasks, including RGB-D and RGB-T saliency detection.

Key words: Saliency Object Detection (SOD); Convolutional Neural Network (CNN)-style adaptive attention; Mask attention; Feature fusion