文章编号：1003-0530(2026)02-0235-14

基于扩散模型的注意力驱动RGB-D显著性目标检测

李恭杨[1,2] 史世翔[2] 李红云*[3]

（1. 泉州职业技术大学环境认知与智能系统实验室，福建泉州 362000；2. 上海大学通信与信息工程学院，上海 200444；3. 泉州职业技术大学联合创新产业学院，福建泉州 362000）

摘 要：显著性目标检测是计算机视觉领域的一个重要研究方向，旨在从复杂的背景中提取出人眼最为关注的区域。传统的RGB图像显著性目标检测方法仅依赖于图像的颜色信息，难以应对复杂场景中的多样性和干扰。为此，RGB-D显著性目标检测在传统RGB图像的基础上，额外引入了深度信息，从而能够更好地感知图像的空间结构，进而提高了显著性目标检测的性能。然而现有RGB-D显著性目标检测方法大多基于卷积神经网络或视觉Transformer，主要依靠判别式学习进行显著性目标检测，即通过对像素级显著性概率进行硬分类实现预测，往往存在模型过度自信的问题，这限制了现有方法在复杂场景下的检测性能。为了应对上述问题，本文提出了一种基于扩散模型的注意力驱动RGB-D显著性目标检测方法，利用扩散模型的渐进式加噪和逐步去噪过程，以生成的方式有效优化了预测结果，减少了模型过度自信导致的错误估计风险，提升了网络在复杂场景下的检测性能。首先，本文采用金字塔形视觉Transformer主干分别对RGB图像和深度图进行四个层级的特征提取；随后，通过提出的双流注意力融合模块实现对对应特征层级的两种跨模态特征的充分融合，接着通过渐进式融合模块对四个不同层级的融合后特征进行融合；最后，把它作为条件信息注入到去噪网络中对扩散模型的输出进行条件约束，并生成预测的显著性图。实验结果表明，所提出的方法在DUT、LFSD、NJU2K、NLPR、SIP、SSD和STERE这七个公开基准数据集上的多个指标均优于现有主流方法，证明了本文提出方法的有效性。

关键词：RGB-D图像；显著性目标检测；扩散模型；注意力机制

中图分类号：TP391 **文献标识码：**A **DOI：**10.12466/xhcl.2026.02.010

引用格式：李恭杨, 史世翔, 李红云. 基于扩散模型的注意力驱动RGB-D显著性目标检测[J]. 信号处理, 2026, 42(2): 235-248. DOI: 10.12466/xhcl.2026.02.010.

Reference format: LI Gongyang, SHI Shixiang, LI Hongyun. Attention-driven RGB-D salient object detection based on diffusion model[J]. Journal of Signal Processing, 2026, 42(2): 235-248. DOI: 10.12466/xhcl.2026.02.010.

Attention-Driven RGB-D Salient Object Detection Based on Diffusion Model

LI Gongyang[1,2] SHI Shixiang[2] LI Hongyun*[3]

（1. Laboratory of Environment Recognition and Intelligent Systems, Quanzhou Vocational and Technical University, Quanzhou, Fujian 362000, China;
2. School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China;
3. Industrial School of Joint Innovation, Quanzhou Vocational and Technical University, Quanzhou, Fujian 362000, China）

收稿日期：2025-05-13；修回日期：2025-07-01

*通信作者：李红云 ynlihy@163.com *Corresponding Author: LI Hongyun, ynlihy@163.com

基金项目：国家自然科学基金(62401350)；上海市科委启明星项目扬帆专项(24YF2713000)；泉州职业技术大学2024年开放课题(LERIS24-02)；泉州市科技计划项目(2025QZC02R)

Foundation Items: The National Natural Science Foundation of China (62401350); Shanghai Sailing Program (24YF2713000); Opening Foundation of Quanzhou Vocational and Technical University in 2024 (LERIS24-02); Quanzhou City Science & Technology Program of China (2025QZC02R)

Abstract: Salient object detection is an important research direction in computer vision, aiming to extract the regions that the human eye pays the most attention to from complex backgrounds. Traditional RGB salient detection methods relied only on the image's color information and had difficulty dealing with the diversity and interference in complex scenes. Therefore, RGB-D salient object detection, based on traditional RGB images, additionally introduced depth information, thereby enabling the perception of the spatial structure of the image better and further improving the performance of salient object detection. However, most existing RGB-D salient object detection methods are based on convolutional neural networks or vision Transformers, mainly relying on discriminative learning for salient object detection, that is, achieving prediction by hard classification of pixel-level saliency probabilities. There is often the problem of model overconfidence, which limits the detection performance of existing methods in complex scenes. To address the above problems, this paper proposes an attention-driven RGB-D salient object detection method based on the diffusion model. By using the progressive noise addition and stepwise denoising processes of the diffusion model, the prediction results were effectively optimized in a generative manner, reducing the risk of incorrect estimation caused by the overconfidence of the model and improving the detection performance of the network in complex scenarios. Firstly, this paper adopted the Pyramid Vision Transformer to achieve four-level feature extraction for RGB images and depth maps. Then, the proposed dual-stream attention fusion module was used to fully fuse the features of the two modes corresponding to the feature level. Subsequently, the fusion features at four different levels were fused through the progressive fusion module to achieve feature fusion. Finally, they were injected into the denoising network as conditional information to impose conditional constraints on the output of the diffusion model and generate the predicted saliency map. The experimental results show that the proposed method outperforms existing mainstream methods in multiple metrics on seven public benchmark datasets, namely DUT, LFSD, NJU2K, NLPR, SIP, SSD, and STERE, which proves the effectiveness of the proposed method.

Key words: RGB-D image; salient object detection; diffusion model; attention mechanism

1 引言

显著性目标检测作为计算机视觉的基础任务，致力于定位图像中吸引人类视觉注意力的核心区域[1-3]，其结果为图像分割[4]、目标识别[5-6]、视觉跟踪[7]等下游任务提供关键预处理信息。传统的基于RGB图像的显著性目标检测方法依赖手工设计特征的方式提取图像中的颜色、纹理等视觉特征，但难以在复杂光照、低对比度或相似背景场景下实现显著性目标检测[8-10]。随着深度相机设备的广泛使用，深度信息作为辅助方式对显著性目标检测性能的提高已经显示出优势，因为深度图像可以提供更多的几何信息，帮助算法更好地理解场景中的空间关系和物体的深度特征，从而提高显著性目标检测的准确性和鲁棒性。

随着深度学习技术的发展[11-12]，研究人员逐渐采用深度学习方法来解决显著性目标检测问题。这些方法主要基于卷积神经网络，采用双流架构分别对RGB图像和深度图像进行特征提取，并通过拼接、卷积、注意力计算等操作实现跨模态特征融合。根据特征融合阶段不同，可以将RGB-D显著性目标检测模型分为输入融合、输出融合和特征级融合三类，其中特征级融合展现出更好的融合效果[13]。例如，FAN等人[14]将骨干网络提取的多层级特征分别融合，随后划分为高层教师特征和低层学生特征，通过两阶段解码器逐步融合多层级信息，提升定位精度和边缘细节。LI等人[15]利用深度到RGB调制和RGB自调制实现相邻尺度特征增强，实现跨模态调制。近年来，视觉Transformer[16]的引入打破了卷积操作感受野的限制，提升了特征长距离建模能力，进一步推动了RGB-D显著性目标检测的发展。WANG等人[17-18]提出的金字塔形视觉Transformer使得完全基于Transformer的网络可以实现与卷积神经网络类似的多尺度特征提取，进而实现Transformer适用于图像分类、分割、检测等多个任务。SUN等人[19]则采用Swin Transformer[20]作为网络主干对特征进行提取，通过卷积、空间注意力、通道注意力等实现跨模态特征融合，进一步提高了检测性能。

然而现有基于卷积神经网络或视觉Transformer的模型都可以归为判别式学习模型，即通过对像素级显著性概率进行硬分类实现预测。这些方法往往忽略了像素之间的潜在关联性，并且过于依赖自信的判别结果，导致在复杂背景和低对比度的场景中容易产生误判。作为一种新型的生成模型，扩散模型在图像生成和医学影像分析等多个领域取得了显著进展[21-22]。其通过前向加噪与反向去噪的马尔可夫过程，能够从高斯噪声中逐步恢复高质量图像或修复缺失部分的图像。这种对生

成过程的逐步控制的方式，能够有效降低模型的过度自信，减少误判，尤其在复杂和动态变化的场景中[23]。与传统的判别式方法不同，扩散模型提供了一种更加柔性和稳健的生成框架，可以根据多模态输入信息（如RGB图像和深度图）进行条件生成，从而提升显著性目标检测的性能。

在显著性目标检测任务中，已有初步工作尝试引入扩散模型以提升显著目标的定位质量。ZHANG等人[24]基于预训练的Stable Diffusion[22]，将RGB图像特征作为条件信息注入到扩散模型中，实现了RGB显著性目标检测性能的进一步提升。在后续工作DiMSOD[25]中，ZHANG等人[26]首次将多模态显著性目标检测任务建模为掩码生成过程，同时基于ControlNet实现将多模态特征作为条件信息注入到扩散模型中来指导显著性图生成。然而，现有方法主要利用条件信息来引导扩散模型生成显著性图，没有充分利用扩散特征实现对条件信息的进一步优化。同时，现有模型将输入数据转换到潜在空间进行加噪、去噪处理，可能导致信息损失，降低模型性能。此外，多种预训练模型的引入也导致模型架构较为复杂，未充分利用主干网络的多层次特征，整体算力需求较大。

为解决上述问题，本文提出一种基于扩散模型的注意力驱动显著性目标检测架构，采用双流架构对RGB图像和深度图进行特征提取，并对对应层级特征进行跨模态融合，最终注入到扩散网络中，实现RGB-D显著性目标检测性能的进一步提升。

2　本文方法

2.1　扩散模型

扩散模型是一类基于随机过程的生成式模型，其基本思想是通过一个逐步加入噪声的前向过程，将原始图像逐渐扰动为全高斯噪声，随后再通过一个反向生成过程，从纯噪声逐步还原出高质量图像。前向扩散过程通常被建模为马尔可夫链，即在每一个时间步t，将输入图像x_0添加少量高斯噪声，得到扰动样本x_t。具体而言，该过程可表示为：

$$q(x_t|x_{t-1})=\mathcal{N}\left(x_t;\ \sqrt{1-\beta_t}\,x_{t-1},\beta_t I\right) \tag{1}$$

其中，β_t介于0到1之间，表示指示每个时间步引入的噪声水平的方差表。取$\alpha_t=1-\beta_t$，利用重参数化技巧可由x_0直接加噪得到x_t，具体公式为：

$$q\left(x_t|x_0\right)=\mathcal{N}\left(x_t;\ \sqrt{\bar{\alpha}_t}\,x_0,\left(1-\bar{\alpha}_t\right)I\right) \tag{2}$$

其中，$\bar{\alpha}_t=\Pi_{s=1}^{t}\alpha_t$。

反向过程可视为从纯高斯噪声图像中逐步恢复高质量图像的过程，其表达式如下：

$$p\left(x_{t-1}|x_t\right)=\mathcal{N}\left(x_{t-1};\ u_\theta\left(x_t,t\right),\sigma_t^2 I\right) \tag{3}$$

其中，$\sigma_t^2=(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)\beta_t$，均值$u_\theta\left(x_t,t\right)$由去噪网络通过学习预测得到，并可通过重参数化表示为：

$$u_\theta(x_t,t)=\frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\hat{x}_0 \tag{4}$$

其中，$\hat{x}_0$为模型最终输出结果。

与传统无条件扩散模型不同，显著性目标检测任务需要模型在生成过程中受输入图像内容控制，生成与之结构一致、语义对应的显著图。为此，本文采用条件扩散框架，通过注入融合后的RGB-D图像特征，引导每一步的去噪方向，使生成过程逐步向真实的显著区域分布靠拢，保证预测结果的准确性与可解释性，其表达式如下：

$$p\left(x_{t-1}|x_t,c_r,c_d\right)=\mathcal{N}\left(x_{t-1};\ u_\theta\left(x_t,c_r,c_d,t\right),\sigma_t^2 I\right) \tag{5}$$

其中，c_r、c_d分别表示第t步注入到去噪网络的RGB特征与深度图像特征。

2.2　网络架构

本文采用基于扩散模型的训练机制，将人工标注的真值图作为原始图像x_0，并通过前向扩散过程逐步添加噪声，得到在任意时间步t下的加噪图x_t，如图1上半部分所示。随后采用本文所提出的RGB-D显著性目标检测网络根据加噪的图来训练去噪网络。本文所提出的RGB-D显著性目标检测网络整体架构如图1下半部分所示，包含特征提取网络和去噪网络两部分。该网络首先采用预训练的金字塔形视觉Transformer（Pyramid Vision Transformer，PVT）分别对输入RGB图像与深度图像进行特征提取，为了实现通道对齐，本网络把单通道深度图复制为三通道深度图进行输入，同时在RGB分支中注入扩散噪声掩码图，并在RGB和深度图分支均嵌入时间步信息，来生成适应于扩散过程的条件信息。PVT第一层至PVT第四层分别表示PVT编码器的不同层级，编码器第i层提取的RGB特征和深度图特征分别记为F_i^r和F_i^d，将对应层级的RGB特征和深度图特征经过双流注意力融合模块（Dual Attention Fusion Model，DAFM）进行融合，得到跨模态融合特征F_i。接着，将四个不同尺度的跨模态融合特征通过渐进式融合模块（Progressive Fusion Model，PFM）进行融合，得到语义信息丰富的特征F_4^*。最后，将F_4^*注入到去噪网络中来指导生成显著性图。本文采用的去噪网络结构基于经典扩散模型

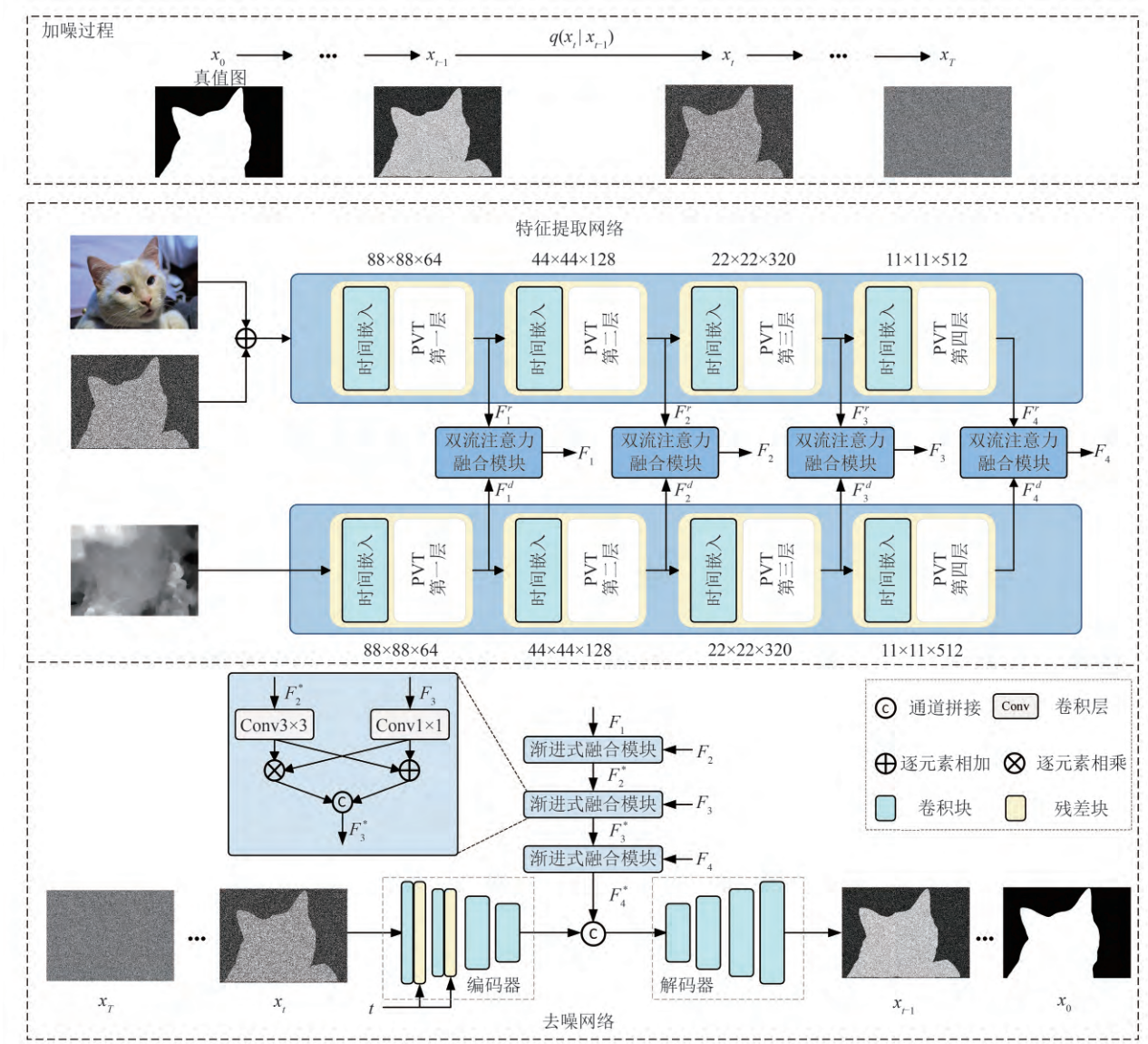

图 1 基于扩散模型的注意力驱动 RGB-D 显著性目标检测框架

Fig. 1 Framework of attention-driven RGB-D salient object detection based on diffusion model

DDPM[21]中的 UNet 架构，为与 PVT 提取的特征尺寸对齐，采用四层编解码结构，编码器前两层各包括一个卷积块和一个残差块，其余层为卷积块，其中卷积块用于调整特征尺寸，残差块用于引入时间步信息，并在第四层将去噪网络特征与 PFM 生成的条件特征 F_4^* 进行拼接融合，实现条件信息注入来引导生成对应于输入 RGB-D 图像对的显著性图。而解码器的四层均为卷积块。在推理过程中，去噪网络初始输入为一张服从标准正态分布的全噪声图片，其尺寸与输入图像相同。

2.3 双流注意力融合模块

在 RGB-D 显著性目标检测任务中，如何有效地融合 RGB 图像与深度图像的特征，已成为提升检测性能的一个重要研究问题。虽然传统的特征融合方法（如拼接或加权求和）已在一定程度上提升了性能，但这些方法往往未能充分挖掘 RGB 图像与深度图之间的潜在关系，导致信息的融合不够深刻。为此，本文提出了一种双流注意力特征融合模块，该模块通过引入分组通道注意力和多尺度空间注意力机制，能更精细地调整 RGB 与深度图像的特征交互，从而提高显著性目标检测的性能。

本文提出的双流注意力融合模块（Dual Attention Fusion Model，DAFM）架构如图2所示，整体可分为分组通道协同融合和多尺度空间协同融合两部分，旨在通过两种互补的注意力机制有效提升多模态特征的融合效果，最终增强显著性目标检测的准确性和鲁棒性。分组通道协同融合部分通过引入分组通道注意力机制（Grouped Channel Attention，GCA）实现。传统的通道注意力机制往往对整个特征图的所有通道进行加权，而分组通道机制则将输入特征图的通道划分为若干组，每组通道独立进行注意力学习，突出对显著性区域有贡献的特征，从而增强重要通道的表达能力。这一过程能够有效避免特征中的冗余信息，并使网络能够更精确地聚焦于关键特征，提升特征表达的有效性。具体地，对于PVT的第 i 个编码块提取的RGB特征 F_i^r 和深度特征 F_i^d，首先，使用拼接操作对其进行融合得到特征 F_c；随后，将 F_c 在通道维度上分为4组，分别进行全局最大池化和全局平均池化操作，以提取全局通道信息，将池化后特征相加并应用Sigmoid激活函数得到单组通道协同调制注意力权重 F_{CA}^i；接着，将4组通道注意力权重进行拼接得到最终的通道协同注意力权重 F_{CA}，用于对 F_i^r 和 F_i^d 进行加权调制；最后，将调制后的RGB与深度图特征进行拼接处理得到通道协同融合特征 F_{GCA}，其表达式如下：

$$F_{\mathrm{CA}}^i=\mathrm{CA}(\mathrm{Chunk}_i(\mathrm{Conv}(\mathrm{Concat}(F_i^r,F_i^d)))) \tag{6}$$

$$F_{\mathrm{CA}}=\mathrm{Concat}(F_{\mathrm{GCA}}^1,F_{\mathrm{GCA}}^2,F_{\mathrm{GCA}}^3,F_{\mathrm{GCA}}^4) \tag{7}$$

$$F_{\mathrm{GCA}}=\mathrm{Conv}(\mathrm{Concat}(F_i^r\otimes F_{\mathrm{CA}},F_i^d\otimes F_{\mathrm{CA}})) \tag{8}$$

其中，Concat表示通道拼接操作，Conv表示1×1卷积操作用来调整通道维度，Chunk_i 表示沿通道维度进行切分为4组中的第 i 组特征，$\otimes$ 表示逐元素相乘，CA为通道注意力（Channel Attention，CA）[27]，其表达式为：

$$\mathrm{CA}(f)=\sigma(\mathrm{GAP}(f)+\mathrm{GMP}(f)) \tag{9}$$

其中，GAP表示全局平均池化操作（Global Average Pooling，GAP），GMP表示全局最大池化操作（Global Max Pooling，GMP），σ 表示Sigmoid激活函数。

为了进一步提升跨模态特征的融合效果，本文提出了多尺度空间注意力机制（Multi-Scale Spatial Attention，MSA）。空间注意力机制[28]旨在通过自适应加权不同空间位置的特征，从而强化显著区域的表达。在此基础上，结合多尺度特征信息，通过多尺度卷积核对空间信息进行建模，能够更好地捕捉图像中不同尺寸的显著区域。具体地，对于RGB特征 F_i^r 和深度特征 F_i^d，首先，本模块通过逐元素相乘操作来得到它们的共同特征 F_s；随后，通过通道维度上的全局最大池化和全局平均池化操作来提取全局空间信息，将池化后特征进行拼接卷积操作；接着，通过三个并行的不同卷积核大小的卷积

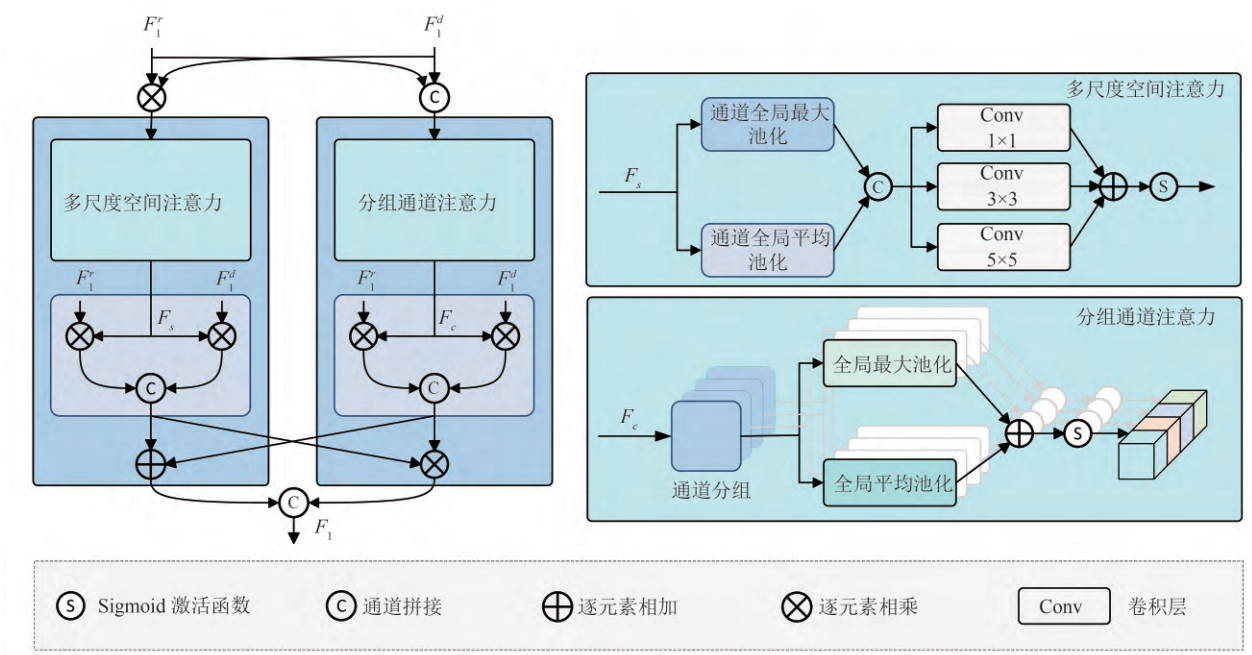

图2　双流注意力融合模块

Fig. 2　Architecture of the proposed DAFM

操作来提取多尺度全局空间信息，进一步强化空间信息的表达，并将三个卷积分支结果进行相加并应用Sigmoid激活函数得到最终的多尺度空间注意力权重F_{SA}，用于对F_i^r和F_i^d进行加权调制；最后，将调制后的RGB与深度图特征进行拼接处理得到空间协同融合特征F_{MSA}，其表达式如下：

$$F_{SA}^i = \mathrm{Conv}_i(\mathrm{SA}(F_i^r \otimes F_i^d)) \tag{10}$$

$$F_{SA} = \sigma(F_{SA}^i \oplus F_{SA}^3 \oplus F_{SA}^5) \tag{11}$$

$$F_{MSA} = \mathrm{Conv}(\mathrm{Concat}(F_i^r \otimes F_{SA}, F_i^d \otimes F_{SA})) \tag{12}$$

其中，Conv_i代表使用大小为$i\times i$的卷积核进行卷积操作，$\otimes$代表逐元素相乘操作，$\oplus$表示逐元素相加操作，SA代表空间注意力（Spatial Attention，SA）机制，其表达式为：

$$\mathrm{SA}(f) = \mathrm{Concat}(\mathrm{CGAP}(f), \mathrm{CGMP}(f)) \tag{13}$$

其中，CGAP表示通道维度全局平均池化（Channel-wise Global Average Pooling，CGAP）操作，CGMP表示通道维度全局最大池化（Channel-wise Global Max Pooling，CGMP）操作。

在得到通道协同融合特征F_{GCA}与空间协同融合特征F_{MSA}后，本模块将通过逐元素相加和逐元素相乘两种操作对F_{GCA}和F_{MSA}进行进一步处理。首先，使用加法操作将F_{GCA}和F_{MSA}相加，从而结合了两种特征的联合信息；随后，通过乘法操作将F_{GCA}和F_{MSA}相乘，加强了两种特征中共同显著区域的表达；最后，将加法和乘法融合的结果沿通道维度拼接，并通过最终的卷积层处理，生成融合后的特征图F_i，以充分利用两种融合方式所带来的互补信息，提升显著性目标检测的性能。其表达式如下：

$$F_i = \mathrm{Conv}(\mathrm{Concat}(F_{MSA} \oplus F_{GCA}, F_{MSA} \otimes F_{GCA})) \tag{14}$$

2.4 渐进式融合模块

在多模态特征融合任务中，不同尺度的特征包含着丰富的层次信息，如何有效地融合这些信息以提高模型的性能是一个关键挑战。为此，本文提出了一种渐进式融合模块，通过逐步融合不同尺度的特征，以确保各层次信息得到有效整合，从而提升显著性目标检测的准确性和鲁棒性。

如图1下方所示，为得到与去噪网络中间层特征尺寸相同的条件信息，本模块采用自上而下的融合方法，每个阶段结合低分辨率特征和高分辨率特征，在逐步减少分辨率的过程中增强语义信息的表达能力。具体的，对于浅层高分辨率跨模态融合特征，采用卷积核为3×3的卷积层进行处理以便将其匹配到与相邻深层高分辨率跨模态融合特征相同的特征尺寸，对于相邻深层特征则采用卷积核为1×1的卷积层进行通道数调整。接着对调整后的特征进行逐元素相加与相乘操作，同时强调两种特征间的联合信息和共同信息，最后进行拼接-卷积层操作得到融合特征。具体公式表示如下：

$$\begin{cases} F_{i+1}^* = \mathrm{Concat}(\mathrm{Conv}_3(F_i) \oplus \mathrm{Conv}_1(F_{i+1}), \mathrm{Conv}_3(F_i) \otimes \mathrm{Conv}_1(F_{i+1})), & i = 1 \\ F_{i+1}^* = \mathrm{Concat}(\mathrm{Conv}_3(F_i^*) \oplus \mathrm{Conv}_1(F_{i+1}), \mathrm{Conv}_3(F_i) \otimes \mathrm{Conv}_1(F_{i+1})), & 2 \leqslant i \leqslant 3 \end{cases} \tag{15}$$

渐进式融合模块的最终输出特征F_4^*包含了多尺度融合后的丰富语义和空间信息，将其作为条件信息与去噪网络中间层特征进行拼接融合，为扩散模型提供了更为丰富的上下文信息，帮助模型在每个去噪步骤中更精确地恢复显著区域，避免过度模糊或误判，从而提升了生成显著性图的质量和鲁棒性。

2.5 损失函数

为了充分考虑模型预测显著性图与真值图之间的结构差异和分类误差，本文采用了一种基于加权二值交叉熵和加权交并比的复合损失函数，损失函数公式如下：

$$L = L_{BCE}^w + L_{IOU}^w \tag{16}$$

其中，L_{BCE}^w为加权二值交叉熵（Binary Cross Entropy，BCE）损失函数，L_{IOU}^w为加权交并比（Intersection-Over-Union，IOU）损失函数。

3 实验

3.1 数据集

为充分评估所提出网络的性能，本文在7个公共基准RGB-D数据集上进行了评估。DUT[29]包含Lytro相机在现实生活中捕获的1200对RGB-D图像。LFSD[30]包含Lytro相机采集的100对RGB-D图像，它们由多名标注者进行标注。NJU2K[31]包含了1985对复杂场景下的RGB-D图像。NLPR[32]包含Microsoft Kinect相机采集的1000对RGB-D图像，图像中常含有多个显著性目标。SIP[33]包含929张RGB-D图像对，由华为手机拍摄而得，显著目标为图像中的人物。SSD[34]包含80对RGB-D图像，从3D电影的室内外场景采集而得。STERE[35]包含1000对RGB-D图像，从公开网站下载样本图像并通过人工标注而得。

3.2 评价指标

为了全面评估本文所提出的网络使用五个计算机视觉任务常用评价指标：S-measure[36]、F-measure[37]、MAE[38]、Weighted F-measure[39]和E-measure[40]。

S-measure用来评价预测显著性图和真值图之间的结构相似性，计算公式如下：

$$S_a = \alpha \times S_o + (1-\alpha) \times S_r \tag{17}$$

其中，α为平衡参数，值为0.5，S_o为目标结构相似性，S_r为区域结构相似性。

F-measure是精度和召回率的调和平均数，用于综合评估模型的整体性能。其计算公式为：

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{18}$$

其中，β^2为调节参数，一般设为0.3，Precision为精确率，Recall为召回率。

平均绝对误差（Mean Absolute Error，MAE）用于衡量预测显著性图与真实显著性图之间的像素级误差，公式表示如下：

$$\text{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \left| S(i,j) - G(i,j) \right| \tag{19}$$

其中，S为预测显著性图，G为真值图，H和W分别为图像的高度和宽度。

Weighted F-measure是F-measure的一种扩展，旨在通过给精度和召回率赋予不同的权重，强化对特定区域或类别的关注。计算公式如下：

$$F_\beta^w = \frac{(1+\beta^2) \times \text{Precision}^w \times \text{Recall}^w}{\beta^2 \times \text{Precision}^w + \text{Recall}^w} \tag{20}$$

其中，Precision^w和Recall^w分别为加权后的精确率和召回率，w为加权因子，用于调整不同区域或类别在损失计算中的重要性。

E-measure用于计算预测显著性图和真值图的全局统计特性和局部区域像素匹配程度，计算公式如下：

$$E_\xi = \frac{1}{W \times H} \sum_{j=1}^{W} \sum_{i=1}^{H} M_{\text{FM}}\left(\xi_{\text{FM}}(x,y)\right) \tag{21}$$

其中，$\xi_{\text{FM}}(x,y)$表示对齐矩阵，M_{FM}表示增强的一致性矩阵。

3.3 实验设置

本文的模型在PyTorch[41]框架上实现，并使用一块RTX 4090 GPU加速训练过程。与以往方法[42-43]相同，本文采用包括来自NLPR的700张图像对，来自NJU2K的1485张图像对，以及来自DUT的800张图像对作为训练集，采用NLPR、NJU2K和DUT数据集剩余图像对以及LFSD、SIP、SSD、STERE数据集的所有图像对作为测试集。在训练和测试过程中，所有图像对大小均为352×352。为了防止过拟合问题，本文应用了一些数据增强策略，如仿射变换、随机翻转和颜色抖动等。本文采用预训练的PVT作为网络主干来提取RGB和深度图的多尺度特征。为了高效训练，本文使用AdamW优化器进行参数优化，并设置初始学习率为0.001，将批次大小设置为12，总训练轮次为150，使用一块RTX 4090 GPU进行训练大约需要10小时。在测试过程中，扩散模型迭代去噪步数设置为10。

3.4 实验比较

为评估本文提出的模型性能，本文与12个近年提出的RGB-D显著性方法进行了定性和定量比较，包括DSNet[44]、DCFNet[45]、CIRNet[46]、C2DFNet[47]、DIGRNet[48]、CFIDNet[49]、HINet[50]、CAVER[42]、PICRNet[51]、RD3D+[52]、LAFB[53]、MAGNet[54]。评价的显著性图由作者提供或者通过运行源代码生成。

不同显著性目标检测方法在7个测试集上的5个评价指标S-measure（S_a）、Weighted F-measure（F_β^w）、MAE（M）、F-measure（F_β）和E-measure（E_ξ）上的对比结果如表1和表2所示。在DUT、NJU2K、SSD和STERE这四个数据集上，本文所提模型均取得了最优结果。其中DUT、NJU2K和STERE三个数据集包含很多有挑战性的复杂场景，表明本文所提方法较其他先进模型更能适应于复杂背景、相似背景下的场景。在DUT数据集上，S_a、F_β^w和F_β较次优结果分别提高了0.007、0.014和0.011，M较次优结果降低了0.003，实现性能的大幅提升。在LFSD、NLPR、SIP三个数据集上，本文所设计的模型也取得了大部分指标上的最优结果，表明本文方法较强的泛化能力。

为了更全面地评估模型的实际部署能力，本文从参数量、FLOPs和推理速度三个方面对本文方法与当前主流RGB-D显著性目标检测模型进行了系统对比，详见表1。得益于扩散模型逐步推理机制，本文模型在多个数据集上表现出较强的检测性能，但相应地，在参数量、FLOPs方面处于中游水平，在推理速度上有着进一步优化的空间。

图3展示了本文模型与一些先进模型生成的显著性图的视觉对比。其中第1、2行为低对比度场景图，第3、4、5行为多物体场景图，第6、7行为复杂场

表 1 不同方法在数据集 DUT、LFSD 和 NJU2K 下的指标比较

Tab. 1 Metric comparison of different methods under the datasets DUT, LFSD and NJU2K

方法	发表刊物	年份	参数(M)	FLOPs(G)	FPS	DUT				
						S_α	F_β^w	M	F_β	E_ξ
DSNet	TIP	2021	-	-	21.7	0.841	0.774	0.079	0.807	0.857
DCFNet	CVPR	2021	108.5	107.8	57.0	0.836	0.766	0.071	0.812	0.888
CIRNet	TIP	2022	103.1	22.5	42.4	0.932	0.904	0.031	0.923	0.949
C2DFNet	TMM	2022	47.5	22.1	52.0	0.933	0.918	0.026	0.932	0.958
DIGRNet	TMM	2022	201.8	68.2	33.0	0.926	0.898	0.033	0.920	0.946
CFIDNet	NCA	2022	53.9	42.9	21.7	0.916	0.887	0.039	0.903	0.940
HINet	PR	2023	98.9	389.7	16.3	0.884	0.826	0.054	0.854	0.903
CAVER	TIP	2023	55.5	44.3	67.0	0.903	0.874	0.042	0.892	0.932
PICRNet	ACM MM	2023	106.8	121.3	25.5	0.943	0.933	0.021	0.943	0.967
RD3D+	TNNLS	2024	28.9	43.3	20.1	0.936	0.908	0.031	0.923	0.952
LAFB	TCSVT	2024	451.8	137.6	50.0	0.926	0.906	0.032	0.920	0.953
MAGNet	KBS	2024	16.1	9.9	31.3	0.943	0.935	0.021	0.944	0.967
Ours		2025	178.7	61.0	8.9	**0.950**	**0.949**	**0.018**	**0.955**	**0.973**
方法	发表刊物	年份	参数(M)	FLOPs(G)	FPS	LFSD				
						S_α	F_β^w	M	F_β	E_ξ
DSNet	TIP	2021	-	-	21.7	0.868	0.823	0.068	0.848	0.889
DCFNet	CVPR	2021	108.5	107.8	57.0	0.842	0.801	0.075	0.835	0.877
CIRNet	TIP	2022	103.1	22.5	42.4	0.875	0.836	0.068	0.867	0.891
C2DFNet	TMM	2022	47.5	22.1	52.0	0.864	0.831	0.065	0.859	0.896
DIGRNet	TMM	2022	201.8	68.2	33.0	0.873	0.825	0.067	0.851	0.892
CFIDNet	NCA	2022	53.9	42.9	21.7	0.870	0.825	0.070	0.849	0.895
HINet	PR	2023	98.9	389.7	16.3	0.852	0.800	0.076	0.830	0.877
CAVER	TIP	2023	55.5	44.3	67.0	0.873	0.842	0.063	0.864	0.907
PICRNet	ACM MM	2023	106.8	121.3	25.5	0.888	0.862	0.052	**0.885**	**0.918**
RD3D+	TNNLS	2024	28.9	43.3	20.1	0.861	0.805	0.076	0.832	0.876
LAFB	TCSVT	2024	451.8	137.6	50.0	0.864	0.828	0.065	0.856	0.898
MAGNet	KBS	2024	16.1	9.9	31.3	**0.889**	0.859	0.054	0.878	0.917
Ours		2025	178.7	61.0	8.9	0.887	**0.865**	**0.051**	0.880	**0.918**
方法	发表刊物	年份	参数(M)	FLOPs(G)	FPS	NJU2K				
						S_α	F_β^w	M	F_β	E_ξ
DSNet	TIP	2021	-	-	21.7	0.921	0.898	0.034	0.907	0.943
DCFNet	CVPR	2021	108.5	107.8	57.0	0.911	0.893	0.035	0.903	0.944
CIRNet	TIP	2022	103.1	22.5	42.4	0.925	0.895	0.035	0.908	0.940
C2DFNet	TMM	2022	47.5	22.1	52.0	0.907	0.885	0.038	0.898	0.936
DIGRNet	TMM	2022	201.8	68.2	33.0	0.933	0.909	0.028	0.918	0.952
CFIDNet	NCA	2022	53.9	42.9	21.7	0.914	0.886	0.038	0.898	0.937
HINet	PR	2023	98.9	389.7	16.3	0.915	0.881	0.039	0.895	0.933
CAVER	TIP	2023	55.5	44.3	67.0	0.920	0.906	0.031	0.914	0.950
PICRNet	ACM MM	2023	106.8	121.3	25.5	0.927	0.912	0.029	0.919	0.952
RD3D+	TNNLS	2024	28.9	43.3	20.1	0.927	0.898	0.034	0.909	0.943
LAFB	TCSVT	2024	451.8	137.6	50.0	0.916	0.897	0.033	0.906	0.945
MAGNet	KBS	2024	16.1	9.9	31.3	0.928	0.917	0.027	0.923	0.956
Ours		2025	178.7	61.0	8.9	**0.931**	**0.923**	**0.024**	**0.932**	**0.958**

注:最好的结果加粗显示。

表2　不同方法在数据集NLPR、SIP、SSD和STERE下的指标比较

Tab. 2　Metric comparison of different methods under the datasets NLPR, SIP, SSD and STERE

方法	NLPR					SIP				
	S_a	F_β^w	M	F_β	E_ξ	S_a	F_β^w	M	F_β	E_ξ
DSNet	0.926	0.881	0.024	0.897	0.950	0.876	0.832	0.052	0.863	0.910
DCFNet	0.924	0.883	0.022	0.897	0.957	0.876	0.838	0.052	0.874	0.915
CIRNet	0.933	0.884	0.023	0.901	0.952	0.888	0.840	0.052	0.875	0.911
C2DFNet	0.928	0.888	0.022	0.903	0.955	0.872	0.830	0.053	0.864	0.912
DIGRNet	0.935	0.889	0.023	0.905	0.955	0.885	0.841	0.052	0.879	0.913
CFIDNet	0.922	0.876	0.026	0.892	0.948	0.864	0.816	0.060	0.856	0.899
HINet	0.922	0.871	0.026	0.888	0.945	0.856	0.796	0.066	0.840	0.886
CAVER	0.928	0.895	0.020	0.906	0.961	0.893	0.864	0.042	0.889	0.930
PICRNet	0.935	0.906	0.019	0.917	**0.965**	0.899	0.874	0.041	0.899	0.933
RD3D+	0.933	0.883	0.022	0.899	0.953	0.892	0.850	0.046	0.881	0.917
LAFB	0.925	0.886	0.024	0.899	0.954	0.877	0.840	0.052	0.876	0.917
MAGNet	**0.939**	0.908	**0.018**	0.916	0.964	**0.908**	0.888	0.037	0.911	**0.942**
Ours	0.933	**0.913**	**0.018**	**0.920**	0.963	0.904	**0.895**	**0.035**	**0.915**	**0.942**
方法	SSD					STERE				
	S_a	F_β^w	M	F_β	E_ξ	S_a	F_β^w	M	F_β	E_ξ
DSNet	0.885	0.829	0.045	0.859	0.907	0.915	0.876	0.036	0.894	0.940
DCFNet	0.865	0.804	0.050	0.835	0.902	0.902	0.867	0.039	0.886	0.939
CIRNet	0.878	0.807	0.049	0.840	0.897	0.916	0.866	0.039	0.890	0.932
C2DFNet	0.872	0.816	0.048	0.846	0.910	0.902	0.862	0.038	0.881	0.936
DIGRNet	0.866	0.797	0.052	0.830	0.889	0.916	0.870	0.037	0.891	0.940
CFIDNet	0.879	0.819	0.050	0.850	0.914	0.901	0.861	0.043	0.882	0.933
HINet	0.865	0.799	0.049	0.836	0.900	0.892	0.831	0.049	0.859	0.918
CAVER	0.878	0.824	0.041	0.849	0.919	0.913	0.882	0.033	0.896	0.947
PICRNet	0.874	0.822	0.048	0.845	0.916	0.920	0.892	0.031	0.906	0.951
RD3D+	0.882	0.812	0.044	0.841	0.900	0.914	0.860	0.039	0.881	0.932
LAFB	0.882	0.833	0.042	0.862	0.915	0.899	0.862	0.040	0.885	0.935
MAGNet	0.885	0.836	0.043	0.856	0.923	0.922	0.892	0.031	0.903	0.951
Ours	**0.895**	**0.865**	**0.032**	**0.882**	**0.941**	**0.927**	**0.909**	**0.026**	**0.920**	**0.957**

注：最好的结果加粗显示。

景图，第8、9行为相似背景图，其余为小目标场景图。可以看出本文模型生成的显著性图在低对比度、多物体、复杂场景、相似背景、小目标等场景中均表现出良好的检测性能。其中第6、7行复杂场景图中，本文模型生成的显著性图优于所有对比方法，进一步证明了扩散模型在复杂场景下优于判别式模型的检测能力。综合来看，本文模型生成的显著性图更接近真值图，优于其他先进模型。

3.5　消融实验

本文以PVT为网络主干，采用RGB图像与深度图相加融合并将四个层级的融合结果相加注入去噪网络的模型为基线模型，以评估所提出的GCA、MSA模块和PFM的有效性。所有评估的模型均在相同的训练集和初始参数下进行训练，并在DUT、SIP、SSD数据集上测试指标。

（1）GCA和MSA模块的有效性。

本文通过设定对照组与多种参数组合对所提出的GCA和MSA模块进行了系统化的实验评估，实验结果展示于表3中。其中Base+CA_1、Base+CA_2分别表示在基线模型基础上增加通道注意力分

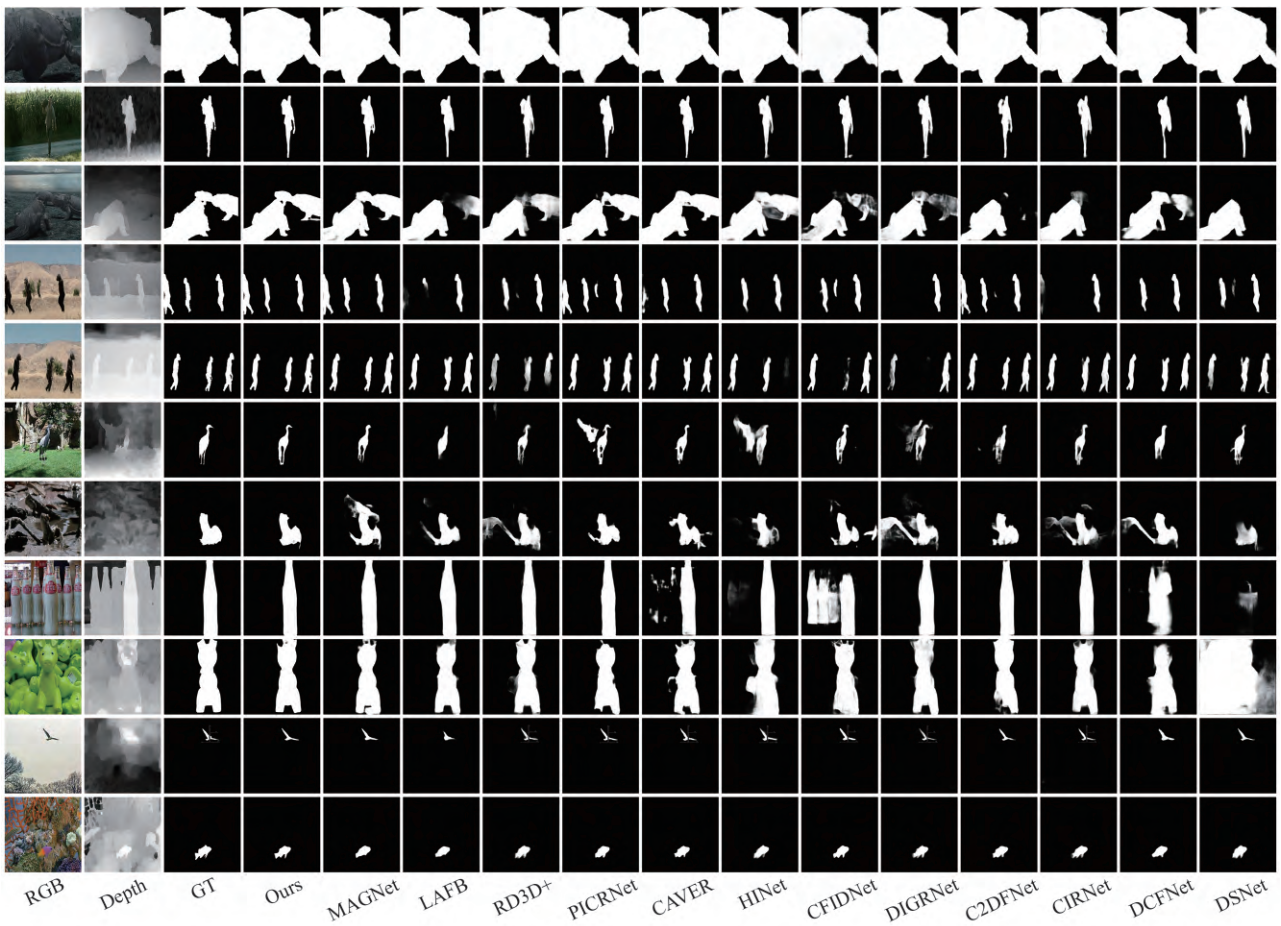

图 3 与其他先进 RGB-D 模型的视觉对比

Fig. 3 Visual comparison with state-of-the-art RGB-D models

支和通道划分两组注意力分支，Base＋GCA 表示基线模型基础上增加本文提出的分组通道注意力分支。对于多尺度空间注意力分支，本文对 MSA 中卷积核大小分别为 1×1、3×3、5×5 的分支进行了比较，分别以 Base＋SA_1、Base＋SA_3、Base＋SA_5 表示，Base＋MSA 表示基线模型基础上增加本文提出的多尺度空间注意力分支。将 Base＋GCA 实验结果与 Base＋CA_1、Base＋CA_2 比较可验证 GCA 模块的有效性，将 Base＋MSA 实验结果与 Base＋SA_1、Base＋SA_3、Base＋SA_5 比较可验证 MSA 模块的有效性。

同时，为进一步验证本文所提出的 GCA 和 MSA 模块的先进性，本小节在相同网络框架下分别将其替换为近年先进的注意力模块矩形自校准注意力模块（Rectangular Self-Calibration Attention Module，RSCAM）[55] 和重排加权空间注意力模块（Shuffle Weighted Spatial Attention Module，SWSAM）[56]，并以 Ours－GCA＋RSCAM 和 Ours－MSA＋SWSAM 表示，Ours 表示本文所提出的完整模型。实验结果表明本文所提出的 GCA 和 MSA 相比当前先进的注意力模块具有一定的性能优势。

（2）PFM 的有效性。

本文对所提出的 PFM 的效果进行了系统化的实验评估，实验结果展示于表 4 中。其中 Base 表示基线模型，Base＋PFM 表示在基线模型基础上增加 PFM 模块，Base＋DAFM＋MF、Base＋DAFM＋CF、Base＋DAFM＋AF 表示基线模型基础上保留 DAFM 并以相乘、拼接、相加操作对四个阶段的 RGB-D 融合特征进行融合注入到去噪网络的结果。Ours 表示本文所提出的完整模型。由 Base 与 Base＋PFM 结果对比可得加入 PFM 后基线模型效果得到了显著提升，由 Ours 与 Base＋DAFM＋MF、Base＋DAFM＋CF、Base＋DAFM＋AF 结果对比可得 PFM 模块比传统的相乘、拼接、相加操作对多阶段特征的融合性能更好。同时，所提出的结合了 DAFM 和 PFM 的完整模型取得了最好的实验效果。

表3　GCA和MSA模块消融结果

Tab. 3　Ablation results of GCA and MSA modules

模型	DUT				SIP				SSD			
	S_a	F_β^w	M	E_ξ	S_a	F_β^w	M	E_ξ	S_a	F_β^w	M	E_ξ
Base + CA_1	0.941	0.939	0.023	0.969	0.902	0.890	0.036	0.942	0.881	0.850	0.042	0.928
Base + CA_2	0.946	0.943	**0.021**	0.972	0.901	0.890	0.036	0.943	0.882	0.845	0.044	0.923
Base + GCA	**0.947**	**0.946**	**0.021**	**0.973**	**0.906**	**0.894**	**0.035**	**0.945**	**0.890**	**0.853**	**0.040**	**0.937**
Base + SA_1	0.945	0.944	**0.019**	**0.973**	**0.893**	0.877	0.041	0.935	0.872	0.832	0.043	0.925
Base + SA_3	0.945	0.943	0.019	**0.973**	0.892	0.876	**0.039**	0.936	0.873	0.830	0.044	0.925
Base + SA_5	0.943	0.941	0.020	0.972	0.891	0.877	0.040	0.934	0.874	0.835	**0.041**	0.923
Base + MSA	**0.947**	**0.946**	**0.019**	**0.973**	**0.893**	**0.883**	**0.039**	**0.937**	**0.881**	**0.847**	**0.041**	**0.931**
Ours – GCA + RSCAM	0.942	0.940	0.023	0.968	0.903	**0.901**	0.036	0.943	0.876	0.850	0.037	0.926
Ours – MSA + SWSAM	0.945	0.944	0.021	0.971	0.902	0.898	0.036	0.943	0.876	0.839	0.044	0.919
Ours	**0.950**	**0.949**	**0.018**	**0.976**	**0.904**	0.895	**0.035**	**0.945**	**0.895**	**0.865**	**0.032**	**0.943**

表4　PFM消融结果

Tab. 4　Ablation results of PFM

模型	DUT				SIP				SSD			
	S_a	F_β^w	M	E_ξ	S_a	F_β^w	M	E_ξ	S_a	F_β^w	M	E_ξ
Base	0.930	0.925	0.027	0.959	0.886	0.870	0.046	0.928	0.860	0.820	0.050	0.922
Base + PFM	0.932	0.925	0.025	0.962	0.890	0.874	0.042	0.932	0.871	0.828	0.049	0.922
Base + DAFM + MF	0.948	0.944	0.019	0.973	0.900	0.888	0.036	0.944	0.894	**0.865**	**0.031**	0.942
Base + DAFM + CF	0.941	0.939	0.023	0.969	0.902	0.890	0.036	0.942	0.881	0.850	0.042	0.928
Base + DAFM + AF	0.938	0.936	0.022	0.968	0.899	0.887	0.037	0.942	0.880	0.839	0.039	0.930
Ours	**0.950**	**0.949**	**0.018**	**0.976**	**0.904**	**0.895**	**0.035**	**0.945**	**0.895**	**0.865**	0.032	**0.943**

3.6　本文方法的不足

值得关注的是，虽然本文方法在多个主流公开RGB-D显著性检测数据集上表现出较强的性能优势，但仍存在一定的资源开销问题。从参数规模和计算复杂度的角度来看，本文模型在实现高质量显著性预测的同时，也引入了较多的训练参数与计算负担。此外，由于扩散模型本身依赖于多步去噪采样以逐步生成预测结果，其推理过程相较于传统基于CNN或Transformer结构的显著性目标检测模型，具有一定的延迟，尤其在对实时性要求较高的应用场景中存在一定局限性。未来，我们将面向提高基于扩散模型的显著性目标检测模型展开研究，致力于研究兼顾检测精度和计算高效性的模型。

4　结论

本文提出了一种基于扩散模型的注意力驱动RGB-D显著性目标检测方法，利用扩散模型的渐进式加噪和逐步去噪过程，以生成的方式优化了预测结果，降低了模型过度自信导致错误估计的风险。具体地，本文通过并行的PVT主干网络分别提取RGB图像和深度图的多尺度特征，并通过设计的双流注意力融合模块进行高效全面的特征融合，接着将多阶段融合特征通过渐进式融合模块进行融合并注入到扩散模型去噪网络中实现条件信息注入以指导显著性图的生成。实验结果表明本文所提模型在多个数据集上的测试结果均优于现有的先进模型。未来，将研究更好的条件信息注入方式，来适应于扩散模型的逐步去噪预测特性，进一步提升模型性能。

参考文献

[1] CHEN Hao, LI Youfu. Three-stream attention-aware network for RGB-D salient object detection[J]. IEEE Transactions on Image Processing, 2019.

[2] WANG Jie, SONG Kechen, BAO Yanqi, et al. CGFNet: Cross-guided fusion network for RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems

for Video Technology, 2022, 32(5): 2949-2961.
[3] LI Junxia, PAN Zefeng, LIU Qingshan, et al. Stacked U-shape network with channel-wise attention for salient object detection[J]. IEEE Transactions on Multimedia, 2020, 23: 1397-1409.
[4] YU Zeng, ZHUGE Yunzhi, LU Huchuan, et al. Joint learning of saliency detection and weakly supervised semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea. IEEE, 2019: 7222-7232.
[5] QIN Xuebin, ZHANG Zichen, HUANG Chenyang, et al. BASNet: Boundary-aware salient object detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 7471-7481.
[6] 周晓飞, 郭舒瑶, 温洪发, 等. 基于深度学习的RGBD图像协同显著目标检测[J]. 信号处理, 2022, 38(6): 1213-1221.
ZHOU Xiaofei, GUO Shuyao, WEN Hongfa, et al. Deep learning-based co-salient object detection on RGBD images[J]. Journal of Signal Processing, 2022, 38(6): 1213-1221. (in Chinese)
[7] LEE M S, SHIN W, HAN S W. TRACER: Extreme attention guided salient object tracing network (student abstract)[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(11): 12993-12994.
[8] CONG Runmin, LEI Jianjun, ZHANG Changqing, et al. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion[J]. IEEE Signal Processing Letters, 2016, 23(6): 819-823.
[9] REN Jianqiang, GONG Xiaojin, LU Yu, et al. Exploiting global priors for RGB-D saliency detection[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Boston, MA, USA. IEEE, 2015: 25-32.
[10] GUO Jingfan, REN Tongwei, BEI Jia. Salient object detection for RGB-D image via saliency evolution[C]// 2016 IEEE International Conference on Multimedia and Expo (ICME). Seattle, WA, USA. IEEE, 2016: 1-6.
[11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations (ICLR 2015). Computational and Biological Learning Society, 2015.
[12] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 770-778.
[13] LI Gongyang, LIU Zhi, LING Haibin. ICNet: Information conversion network for RGB-D based salient object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 4873-4884.
[14] FAN Dengping, ZHAI Yingjie, BORJI A, et al. BBS-net: RGB-D salient object detection with a bifurcated backbone strategy network[C]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 275-292.
[15] LI Gongyang, LIU Zhi, YE Linwei, et al. Cross-modal weighting network for RGB-D salient object detection [C]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 665-681.
[16] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]// Proceedings of the International Conference on Learning Representations, 2020: 611-632.
[17] WANG Wenhai, XIE Enze, LI Xiang, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada. IEEE, 2021: 548-558.
[18] WANG Wenhai, XIE Enze, LI Xiang, et al. PVT v2: Improved baselines with pyramid vision transformer[J]. Computational Visual Media, 2022, 8(3): 415-424.
[19] SUN Fuming, REN Peng, YIN Bowen, et al. CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection[J]. IEEE Transactions on Multimedia, 2023, 26: 2249-2262.
[20] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada. IEEE, 2021: 9992-10002.
[21] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.
[22] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA. IEEE, 2022: 10674-10685.
[23] LAI Zeqiang, DUAN Yuchen, DAI Jifeng, et al. Denoising diffusion semantic segmentation with mask prior modeling[EB/OL]. 2023: 2306.01721. https://arxiv.org/abs/2306.01721v2.
[24] ZHANG Shuo, HUANG Jiaming, CHEN Shizhe, et al. SOD-diffusion: Salient object detection via diffusion-based image generators[J].Computer Graphics Forum,

2024，43(7)：e15251.

[25] ZHANG Shuo，HUANG Jiaming，TANG Wenbing，et al. DiMSOD：A diffusion-based framework for multi-modal salient object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence，2025，39(10)：10103-10111.

[26] ZHANG Lvmin，RAO Anyi，AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]//2023 IEEE/CVF International Conference on Computer Vision（ICCV）. Paris，France. IEEE，2023：3813-3824.

[27] HU Jie，SHEN Li，SUN Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Salt Lake City，UT，USA. IEEE，2018：7132-7141.

[28] FU Jun，LIU Jing，TIAN Haijie，et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Long Beach，CA，USA. IEEE，2019：3141-3149.

[29] PIAO Yongri，JI Wei，LI Jingjing，et al. Depth-induced multi-scale recurrent attention network for saliency detection[C]//2019 IEEE/CVF International Conference on Computer Vision（ICCV）. Seoul，Korea. IEEE，2019：7253-7262.

[30] LI Nianyi，YE Jinwei，JI Yu，et al. Saliency detection on light field[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(8)：1605-1616.

[31] JU Ran，GE Ling，GENG Wenjing，et al. Depth saliency based on anisotropic center-surround difference[C]//2014 IEEE International Conference on Image Processing（ICIP）. Paris，France. IEEE，2014：1115-1119.

[32] PENG Houwen，LI Bing，XIONG Weihua，et al. RGBD salient object detection：A benchmark and algorithms[C]//Computer Vision - ECCV 2014. Cham：Springer International Publishing，2014：92-109.

[33] FAN Dengping，LIN Zheng，ZHANG Zhao，et al. Rethinking RGB-D salient object detection：Models，data sets，and large-scale benchmarks[J]. IEEE Transactions on Neural Networks and Learning Systems，2021，32(5)：2075-2089.

[34] LI Ge，ZHU Chunbiao. A three-pathway psychobiological framework of salient object detection using stereoscopic technology[C]//2017 IEEE International Conference on Computer Vision Workshops（ICCVW）. Venice，Italy. IEEE，2017：3008-3014.

[35] NIU Yuzhen，GENG Yujie，LI Xueqing，et al. Leveraging stereopsis for saliency analysis[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence，RI，USA. IEEE，2012：454-461.

[36] FAN Dengping，CHENG Mingming，LIU Yun，et al. Structure-measure：A new way to evaluate foreground maps[C]//2017 IEEE International Conference on Computer Vision（ICCV）. Venice，Italy. IEEE，2017：4558-4567.

[37] ACHANTA R，HEMAMI S，ESTRADA F，et al. Frequency-tuned salient region detection[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami，FL，USA. IEEE，2009：1597-1604.

[38] PERAZZI F，KRÄHENBÜHL P，PRITCH Y，et al. Saliency filters：Contrast based filtering for salient region detection[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence，RI，USA. IEEE，2012：733-740.

[39] MARGOLIN R，ZELNIK-MANOR L，TAL A. How to evaluate foreground maps[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus，OH，USA. IEEE，2014：248-255.

[40] FAN Dengping，GONG Cheng，CAO Yang，et al. Enhanced-alignment measure for binary foreground map evaluation[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm，Sweden. International Joint Conferences on Artificial Intelligence Organization，2018：698-704.

[41] PASZKE A，LERER A，KILLEEN T，et al. PyTorch：An imperative style，high-performance deep learning library[C]//Advances in Neural Information Processing Systems 32，Volume 11 of 20：32nd Conference on Neural Information Processing Systems（NeurIPS 2019）. Vancouver(CA).8-14 December 2019.2020.

[42] PANG Youwei，ZHAO Xiaoqi，ZHANG Lihe，et al. CAVER：Cross-modal view-mixed transformer for bi-modal salient object detection[J]. IEEE Transactions on Image Processing，2023，32：892-904.

[43] HU Xihang，SUN Fuming，SUN Jing，et al. Cross-modal fusion and progressive decoding network for RGB-D salient object detection[J]. International Journal of Computer Vision，2024，132(8)：3067-3085.

[44] WEN Hongfa，YAN Chenggang，ZHOU Xiaofei，et al. Dynamic selective network for RGB-D salient object detection[J]. IEEE Transactions on Image Processing，2021，30：9179-9192.

[45] JI Wei，LI Jingjing，YU Shuang，et al. Calibrated RGB-D salient object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Nashville，TN，USA. IEEE，2021：9466-9476.

[46] CONG Runmin，LIN Qinwei，ZHANG Chen，et al. CIR-net：Cross-modality interaction and refinement for RGB-D salient object detection[J]. IEEE Transactions

on Image Processing, 2022, 31: 6800-6815.
[47] ZHANG Miao, YAO Shunyu, HU Beiqi, et al. C2DFNet: Criss-cross dynamic filter network for RGB-D salient object detection[J]. IEEE Transactions on Multimedia, 2022, 25: 5142-5154.
[48] CHENG Xiaolong, ZHENG Xuan, PEI Jialun, et al. Depth-induced gap-reducing network for RGB-D salient object detection: An interaction, guidance and refinement approach[J]. IEEE Transactions on Multimedia, 2022, 25: 4253-4266.
[49] CHEN Tianyou, HU Xiaoguang, XIAO Jin, et al. CFIDNet: Cascaded feature interaction decoder for RGB-D salient object detection[J]. Neural Computing and Applications, 2022, 34(10): 7547-7563.
[50] BI Hongbo, WU Ranwan, LIU Ziqi, et al. Cross-modal hierarchical interaction network for RGB-D salient object detection[J]. Pattern Recognition, 2023, 136: 109194.
[51] CONG Runmin, LIU Hongyu, ZHANG Chen, et al. Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa ON Canada. ACM, 2023: 406-416.
[52] CHEN Qian, ZHANG Zhenxi, LU Yanye, et al. 3-D convolutional neural networks for RGB-D salient object detection and beyond[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 4309-4323.
[53] WANG Kunpeng, TU Zhengzheng, LI Chenglong, et al. Learning adaptive fusion bank for multi-modal salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 7344-7358.
[54] ZHONG Mingyu, SUN Jing, REN Peng, et al. MAGNet: Multi-scale Awareness and Global fusion Network for RGB-D salient object detection[J]. Knowledge-Based Systems, 2024, 299: 112126.
[55] NI Zhenliang, CHEN Xinghao, ZHAI Yingjie, et al. Context-guided spatial feature reconstruction for efficient semantic segmentation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 239-255.
[56] LI Gongyang, BAI Zhen, LIU Zhi, et al. Salient object detection in optical remote sensing images driven by transformer[J]. IEEE Transactions on Image Processing, 2023, 32: 5257-5269.

作者简介

李恭杨 男，1993 年生，浙江台州人。上海大学副教授，主要研究方向为多模态图像处理、显著性检测、对象分割与缺陷检测。
E-mail：ligongyang@shu.edu.cn

史世翔 男，2003 年生，江苏宿迁人。上海大学硕士研究生，主要研究方向为多模态图像处理、显著性检测。
E-mail：shishixiang@shu.edu.cn

李红云 女，1982 年生，河南濮阳人。泉州职业技术大学副教授，主要研究方向为计算机应用技术、视频图像复原、图像/视频分割以及图像增强。
E-mail：ynlihy@163.com

（责任编辑：刘建新）