中图法分类号:TP391.41; TP183　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-25

论文引用格式:JI Jiaxin, GUO Xingge, YANG Fazhan, WANG Jiang, ZHAO Peipei, XIAO Tao. Uncertainty-Gated Mamba and Frequency Compensation Network for Saliency Prediction[J/OL]. Journal of Image and Graphics, XXXX:1-25. DOI: 10.11834/jig.260189. (纪嘉歆, 郭星歌, 杨发展, 王江, 赵培培, 肖涛. 不确定性门控Mamba与频域补偿的显著性预测网络[J/OL]. 中国图象图形学报,XXXX:1-25. DOI: 10.11834/jig.260189.)[DOI:10.11834/jig.260189]

不确定性门控Mamba与频域补偿的显著性预测网络

纪嘉歆[1],郭星歌[1],杨发展[1],王江[1],赵培培[1],肖涛[2]

1. 中国矿业大学　信息与控制工程学院,江苏　徐州　221116; 2. 常州海图信息科技股份有限公司,江苏　常州213000

摘　要: **目的**　针对视觉显著性预测中长程建模开销较大、复杂背景响应易随全局传播扩散以及解码上采样导致注视热点不够集中的问题,提出一种不确定性门控Mamba建模增强与动态频域调制相结合的显著性预测网络(Spatio-Spectral Uncertainty-Gated Mamba Network ,S^2UG-Mamba)。**方法**　在编码端,设计不确定性感知状态空间增强模块(Uncertainty-Aware State Space Enhancement Module,UA-SSM),通过双向蛇形交叉扫描Mamba建模捕获长程上下文信息,并结合空间与通道方差统计进行不确定性估计,生成置信度门控,以抑制不可靠区域响应。在解码端,针对连续上采样引起的预测响应扩散问题,提出语义引导的动态频域调制模块(Semantic-Guided Dynamic Frequency Modulation Module,SDFM),利用深层语义先验对频域调制过程进行动态引导,从而提升注视热点区域的响应集中性。**结果**　在SALICON、MIT300等5个公开数据集上的实验结果表明,所提S^2UG-Mamba在多个主流评价指标上均优于现有先进方法。与GSGNet相比,S^2UG-Mamba在LSUN′17竞赛上将KL由0.190降低至0.176,降低7.4%,IG达到0.943,提升4.0%;在MIT300盲测中,CC相对提升2.2%,KL降低9.8%;在MIT1003零样本测试中,NSS、CC和SIM分别提升2.2%、2.0%和5.6%,KL降低5.1%。**结论**　所提方法实现了长程上下文建模、背景噪声抑制和显著结构恢复的协同优化,在保持较高计算效率的同时提升了复杂自然场景下显著性预测的分布一致性、跨域泛化能力和鲁棒性。

关键词: 显著性预测;状态空间模型;Mamba;不确定性感知;频域增强

Uncertainty-Gated Mamba and Frequency Compensation Network for Saliency Prediction

JI Jiaxin[1], GUO Xingge[1], YANG Fazhan[1], WANG Jiang[1], ZHAO Peipei[1], XIAO Tao[2]

1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China; 2. Changzhou Haitu Information Technology Co., Ltd., Changzhou 213000, China

Abstract: **Objective** Visual saliency prediction aims to simulate the human visual attention mechanism by estimating the spatial probability distribution of gaze points in complex scenes. This task is fundamental to various downstream applications, including autonomous driving, robot navigation, and image compression. While Convolutional Neural Networks (CNNs) have long been the backbone of saliency modeling due to their proficiency in local feature extraction and multi-

收稿日期:2026-04-10;**修回日期**:2026-06-07

*通信作者:赵培培 zppcumt@163.com

基金项目:国家重点研发计划项目(项目编号:2022YFC3004703);江苏省研究生科研与实践创新计划(项目编号:KYCX25_2990);中国矿业大学研究生创新计划项目(项目编号:2025WLKXJ114)

Supported by: National Key Research and Development Program of China(2022YFC3004703); Postgraduate Research & Practice Innovation Program of Jiangsu Province(KYCX25_2990); Graduate Innovation Program of China University of Mining and Technology under Grant(2025WLKXJ114)

scale representation, they are inherently limited by their local receptive fields. This restriction often leads to "fixation fragmentation" and background false alarms in cluttered environments. Conversely, Transformer-based models have introduced global dependency modeling via self-attention; however, their quadratic computational complexity becomes a significant bottleneck when processing high-resolution saliency maps. Recently, State Space Models (SSMs), specifically Mamba, have emerged as a promising alternative, offering linear-time complexity while maintaining long-range context. Nevertheless, standard Mamba architectures are designed for 1D sequences, which inherently disrupts the 2D spatial topological continuity of images. Furthermore, during long-range state propagation, background noise is often indiscriminately amplified alongside salient features. Additionally, the continuous bilinear upsampling typically employed in decoding stages causes a "structural over-smoothing" effect, which diminishes the peak energy of salient regions and blurs boundaries. To address these multi-faceted challenges, this study proposes a novel visual saliency prediction network named Spatio-Spectral Uncertainty-Gated Mamba Network (S^2UG-Mamba). The goal is to achieve an optimal synergy between global context modeling, noise suppression, and high-frequency structural restoration within a linear computational framework. **Method** The proposed S^2UG-Mamba follows a "backbone extraction, encoder enhancement, multi-scale decoding, frequency modulation, and prediction" pipeline. We employ an ImageNet-pretrained ConvNeXt-Tiny as the backbone to extract multi-level features. In the encoder enhancement phase, we introduce the Uncertainty-Aware State Space Enhancement Module (UA-SSM). To preserve 2D spatial continuity, UA-SSM utilizes a bidirectional orthogonal serpentine scanning strategy, which unfolds features in both horizontal and vertical directions to capture omnidirectional contextual interactions. To mitigate the propagation of background noise, we design a dual-view uncertainty proxy within UA-SSM. This proxy consists of a spatial variance component (capturing local texture fluctuations in 3x3 windows) and a channel variance component (measuring semantic activation disagreements). These proxies are fused to generate a pixel-level confidence gate G, which dynamically modulates the Mamba input sequence, thereby suppressing responses in high-uncertainty background regions. In the decoding stage, to counteract the smoothing effects of upsampling, we propose the Semantic-Guided Dynamic Frequency Modulation (SDFM) module. Unlike conventional static filters, SDFM employs a joint logical routing mechanism that combines deep semantic priors from the encoder with local content features from the decoder. This routing mechanism dynamically synthesizes a set of learnable complex frequency filters. The feature maps are transformed into the frequency domain via a real Fast Fourier Transform (rFFT) after StarReLU activation, multiplied by the adaptive filters, and then mapped back to the spatial domain using an inverse rFFT. This residual frequency compensation explicitly restores high-frequency structural details and tightens the energy concentration of fixation peaks. The entire network is optimized using a joint loss function of Kullback-Leibler (KL) divergence and Linear Correlation Coefficient (CC). **Result** Extensive experiments were conducted on five benchmark datasets: SALICON, MIT300, MIT1003, CAT2000, and TORONTO. On the SALICON test set (LSUN'17), S^2UG-Mamba achieved state-of-the-art performance, reducing the KL divergence to 0.176 while reaching a CC of 0.918 and an NSS of 2.007. On the challenging MIT300 blind test set, our model yielded a CC of 0.829 and a KL of 0.370, outperforming the advanced GSGNet by 2.2% and 9.8%, respectively. To evaluate cross-domain generalization, we performed zero-shot testing on the MIT1003 dataset using weights trained solely on SALICON; S^2UG-Mamba attained an NSS of 2.386 and a CC of 0.6849, demonstrating robust learning of universal saliency priors. Ablation studies confirmed the efficacy of each component: the introduction of UA-SSM improved the NSS from 1.9627 to 2.0152, while SDFM significantly reduced the KL divergence. The complexity and efficiency analysis demonstrates that our model achieves favorable performance in terms of computational cost and GPU memory consumption. Meanwhile, it maintains comparable parameter scale and inference speed with state-of-the-art methods, fully validating the efficient modeling capability of the Mamba architecture. Qualitative visualizations further illustrated that S^2UG-Mamba effectively suppresses non-salient background clutter in densely textured scenes and produces sharper, more compact salient boundaries compared to existing methods. **Conclusion** This study presents S^2UG-Mamba, a linear-complexity network that synergistically integrates uncertainty-aware state space modeling and semantic-guided frequency modulation for visual saliency prediction. By employing orthogonal scanning and a dual-view uncertainty gating mechanism, the encoder efficiently captures long-range dependencies while adaptively suppressing noise. Simultaneously, the SDFM leverages high-level semantic guidance to perform adaptive frequency filtering, effectively restoring structural details lost during spa-

tial upsampling. The experimental results across multiple benchmarks validate that S²UG-Mamba provides a superior balance between prediction accuracy, cross-domain robustness, and computational efficiency. Future research will explore hardware-friendly parallel 2D scanning algorithms and low-annotation learning paradigms to further enhance the model's practical deployment potential.

Key words: saliency prediction; state space model; Mamba; uncertainty-aware; frequency domain enhancement

论文引用格式：Ji Jiaxin, Guo Xingge, Yang Fazhan, Wang Jiang, Zhao Peipei, Xiao Tao. Uncertainty-Gated Mamba and Frequency Compensation Network for Saliency Prediction[J/OL]. Journal of Image and Graphics, XXXX: 1-XX. DOI: 10.11834/jig.260189.（引用格式：纪嘉歆，郭星歌，杨发展，王江，赵培培，肖涛. 不确定性门控Mamba与频域补偿的显著性预测网络[J/OL]. 中国图象图形学报，XXXX:1-XX. DOI:10.11834/jig.260189.）

0 引　言

人类视觉系统能够在复杂场景中快速分配注意资源，并优先关注信息量较高的区域。视觉显著性预测旨在模拟这一注意机制，预测图像中人眼注视点的空间分布，判断哪些视觉响应真正具有注视价值，并以显著图形式表征视觉注意结果。该任务在图像理解、目标检测、视频分析、人机交互以及自动驾驶等领域具有重要应用价值。

早期视觉显著性研究主要通过整合颜色、亮度、方向等低层视觉线索进行建模。Itti等人(1998)提出经典视觉注意模型，通过多尺度特征融合实现了对显著区域的早期建模；Harel等人(2015)进一步利用图模型增强了局部对比与全局关系建模能力。随着眼动数据集与评价体系的逐步完善，视觉显著性预测由人工设计特征驱动逐渐转向数据驱动，SALICON数据集的提出为深度模型训练提供了大规模数据基础(Jiang等，2015)。Bylinskii等人(2019)对评价指标的系统分析则为不同模型之间的公平比较提供了统一依据。上述工作奠定了显著性预测的基本范式，但低层线索主导的建模方式难以充分解释真实场景中的语义注视行为。

卷积神经网络(convolutional neural networks，CNN)的引入显著提升了显著性预测中的语义表征能力。DeepGaze I模型通过重用预训练网络的深层特征，率先验证了迁移学习在该任务中的有效性(Kümmerer等，2014)；Huang等人(2015)在此基础上进一步微调深度神经网络，有效缩小了视觉特征与人眼注视分布之间的语义鸿沟；DeepFix通过大感受野卷积与位置偏置建模增强了多尺度上下文表达(Kruthiventi等，2017)。随后，SalGAN将生成对抗学习引入显著性预测，以改善预测分布的结构真实性(Pan等，2017)；SAM模型通过卷积长短期记忆网络(long short-term memory，LSTM)与注意机制对显著图进行迭代优化，增强了模型的上下文整合能力(Cornia等，2018)；EMLNet通过多主干特征融合提升了不同层级视觉线索的互补利用效率(Jia等，2020)。这些CNN类方法能够有效提取局部纹理和多尺度特征，但其全局信息主要依赖局部卷积的逐层累积和多级特征融合间接获得。当图像包含密集纹理或多个语义对象时，显著区域响应容易被稀释，非显著区域的局部强激活也可能在融合过程中被保留下来，形成错误预测。

为增强全局信息建模能力，Transformer结构被引入视觉显著性预测任务。自注意力机制能够直接建立远距离位置之间的关联，因此在全局语义建模方面较传统卷积具有更强能力。TranSalNet通过结合卷积特征与Transformer编码器提升显著性预测性能(Lou等，2022)；GSGNet通过融合CNN与Transformer结构实现局部与全局信息的协同建模(Xie等，2024)；UNETRSal将Transformer编码器与U-Net式解码结构结合，在多个公开数据集上取得了有竞争力的结果(Kaibaldiyev等，2026)。然而Transformer虽然扩大了特征交互范围，却并不天然区分参与交互的信息是否可靠。纹理边界、重复图案和噪声响应同样可能进入全局注意力聚合，干扰显著区域的响应分布。

状态空间模型(state space models，SSM)特别是Mamba通过选择性状态空间机制实现了线性复杂度的序列建模，在保留长程依赖建模能力的同时显著降低了计算与显存开销(Gu和Dao，2023)。近期多种视觉Mamba研究表明，状态传播机制需要结合具

体任务重新约束,如非因果扫描、傅里叶空间建模和时序运动补偿等策略可以分别被用于改善二维结构建模、频率相关性建模和长程时空信息传播(肖杰等,2025;Li等,2025;Sun等,2025)。SUM将Mamba与U-Net架构结合,用于多类型图像的显著性预测(Hosseini等,2025);SalM²通过轻量化Mamba结构实现驾驶场景下的实时注意力预测(Zhao等,2024)。相关工作说明Mamba在显著性预测中具有较好的应用潜力,但显著性预测需要的不只是低成本长程依赖,还需要在传播过程中保持二维结构信息,并区分不同区域响应的可信度。现有视觉Mamba方法通常将二维特征按预设路径展平为一维序列,虽然保持了较高效率,却也将扫描顺序引入空间建模;当局部纹理波动较大时,低可信响应可能被写入隐状态,并沿序列传递到后续位置。

这种传播风险与显著性预测中的不确定性密切相关,现有显著性预测方法大多将该任务建模为确定性回归问题,即直接从输入图像学习到单一的注视概率分布。然而,人眼注视分布本身具有一定的不确定性来源(Gal和Ghahramani,2016;Kendall和Gal,2017):局部纹理干扰、标注噪声、场景歧义以及不同观察者之间的注视差异,都会造成空间响应的不稳定。在长程状态传播框架下,这类不稳定响应一旦被编码为状态信息,便可能影响后续位置的特征更新。这种传播方式容易使干扰响应与真实显著线索同时被强化。因而引入不确定性估计与门控调制有助于提高全局状态传播的鲁棒性。

除编码阶段的全局建模外,解码阶段的特征恢复同样会直接影响显著图质量。显著性预测通常采用逐级上采样与多尺度融合生成最终分布,在这一过程中,高频结构信息容易被逐步平滑,进而导致显著区域响应不够集中、分布形态不够紧凑。已有研究表明,频域表示能够较自然地区分低频语义信息与高频结构信息,为解码阶段的结构补偿提供了可行途径。快速傅里叶卷积和GFNet模型等均验证了频域建模在视觉任务中的有效性(Chi等,2020;Rao等,2021);TFGNet说明频率引导机制能够改善显著性相关任务中的结构恢复能力(Wang等,2024)。但显著性预测中的高频成分并非都应增强。目标轮廓和局部关键结构有助于形成集中响应,而纹理边缘和重复图案则可能引入干扰。故频域补偿需要借助高层语义先验进行选择性调制,使结构恢复服务于注视分布,而不是简单锐化图像细节。

基于上述分析,本文提出一种融合不确定性门控Mamba建模增强与动态频域调制相结合的视觉显著性预测网络S²UG-Mamba。核心思路是在状态传播阶段引入抗噪约束,并在语义先验引导下进行频域结构补偿,从而同时缓解低可信响应累积和上采样造成的热点扩散。具体而言,编码端设计不确定性感知状态空间增强模块(UA-SSM),通过双向蛇形交叉扫描建模水平与垂直方向的长程上下文;同时,利用空间—通道双视点方差统计生成像素级门控,对细粒度特征传播进行连续调制,降低低可信响应对隐状态更新的影响。解码端设计语义引导的动态频域调制模块(SDFM),高层语义特征与局部内容特征联合路由可学习复数滤波器基,对解码特征进行样本自适应频域重加权,选择性补偿与注视结构相关的频率成分。通过编码端抑制低可信响应传播、解码端选择性恢复注视结构,所提出方法在保持复杂度优势的同时,提高了显著性预测的响应集中性和结构一致性。

1 理论与方法

1.1 S²UG-Mamba网络架构概述

如图1所示,网络整体采用特征提取-编码增强-多尺度解码-频域调制-显著图预测的框架。首先,输入图像经主干网络提取四级特征,并通过侧边映射层统一到相同嵌入维度,获得兼具高层语义与浅层空间细节的多尺度表示。在编码增强阶段,本文在中间层特征上引入不确定性感知状态空间增强模块UA-SSM,对复杂场景中的上下文信息进行建模与筛选。该模块通过正交蛇形扫描双向Mamba建立长程空间依赖,结合双视点不确定性生成置信度门控,以抑制复杂背景中的不可靠响应。

在解码阶段,网络采用自顶向下的逐级融合方式,对多尺度特征进行上采样恢复和跨层整合,逐步形成兼具语义信息与空间细节的解码特征。在解码末端进一步引入语义引导的频域校正模块SDFM,利用解码融合特征与编码阶段提供的语义引导信息,对频域响应进行自适应补偿,从而增强显著区域的结构表达。

最终,经过频域矫正的特征通过输出卷积头与Sigmoid激活函数生成单通道显著性概率分布图。

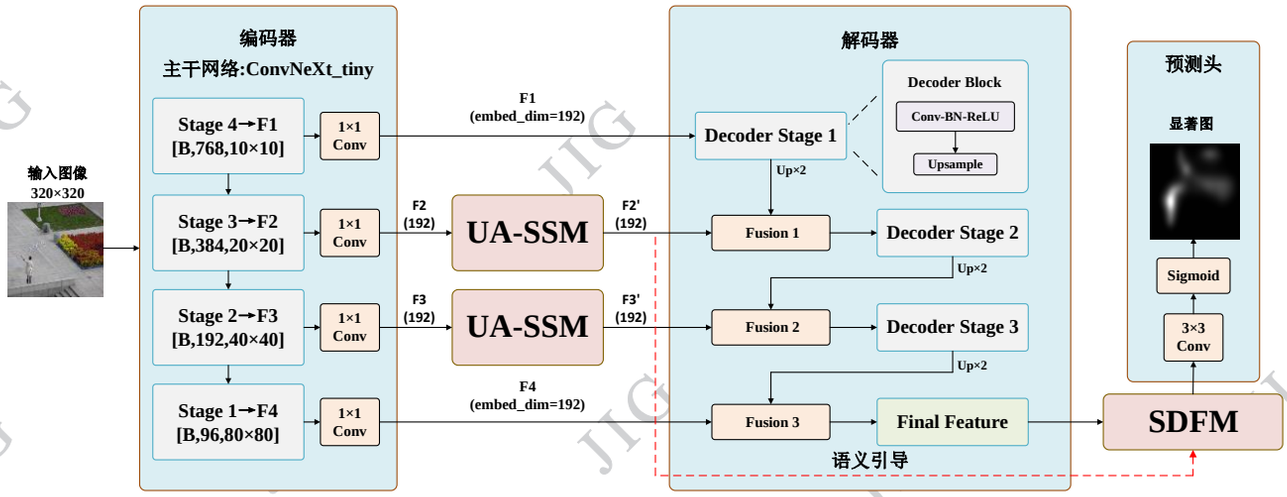


图1　S²UG-Mamba整体架构

Fig. 1　Overall architecture of S²UG-Mamba.

1.2　不确定性感知状态空间增强模块(UA-SSM)

现有显著性预测方法如TranSalNet、GSGNet等通过Transformer实现全局上下文建模。然而，全局建模并不意味着所有响应都应被同等传播，复杂背景中的纹理边缘、重复结构和语义歧义区域也可能产生较强激活，若缺少可靠性约束，容易在全局交互中形成非注视区域误响应。SUM等Mamba显著性预测工作验证了状态空间模型在高效长程建模中的潜力，但也说明视觉SSM需要针对图像结构重新设计传播方式。基于此，本文提出UA-SSM，如图2所示，模块主要包含三个组成部分：通过水平和垂直两个方向的蛇形扫描，将二维特征展开为一维序列，并利用双向Mamba建模长程上下文关系；根据空间方差和通道方差生成像素级置信度门控，用于刻画不同空间位置特征传播的可靠性；将粗粒度全局上下文分支与细粒度门控增强分支进行融合，从而在保持全局建模能力的同时抑制复杂背景中的不可靠响应。

1.2.1　正交蛇形扫描双向Mamba建模

图2(a)展示了本文采用的正交蛇形扫描双向Mamba建模。其作用是将二维特征分别沿水平和垂直方向蛇形展开为一维序列，并利用双向Mamba对序列进行长程上下文建模。

记输入的任意特征为：

$$\boldsymbol{Z} \in \mathbf{R}^{C\times H\times W} \tag{1}$$

式中，C、H、W分别表示通道数、高度和宽度。该模块的输出记为：

$$\mathcal{C}_{mamba}(\boldsymbol{Z}) \in \mathbf{R}^{C\times H\times W} \tag{2}$$

标准Mamba源于选择性状态空间模型，其基本思想是通过状态递推实现长程序列建模。对于一维输入序列中的第t个token，离散状态空间模型可表示为：

$$\boldsymbol{h}_t = \bar{\boldsymbol{A}}\boldsymbol{h}_{t-1} + \bar{\boldsymbol{B}}\boldsymbol{z}_t \tag{3}$$

$$\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{h}_t + \boldsymbol{D}\boldsymbol{z}_t \tag{4}$$

式中，$\boldsymbol{z}_t$表示序列输入，$\boldsymbol{h}_t$表示隐状态，$\boldsymbol{y}_t$表示序列输出，$\bar{\boldsymbol{A}}$、$\bar{\boldsymbol{B}}$、$\boldsymbol{C}$、$\boldsymbol{D}$分别表示离散状态转移矩阵、输入映射矩阵、输出映射矩阵和残差映射矩阵。Mamba在此基础上引入输入相关的选择性机制，使状态更新能够根据当前序列内容自适应调整，因此适合用于长程视觉上下文建模。

然而，视觉特征$\boldsymbol{Z}$是二维空间结构，不能直接作为一维序列输入Mamba。若采用普通光栅扫描，图像在换行位置会产生较大的空间跳变，使一维序列中的相邻token不一定对应二维图像中的相邻位置。为缓解这一问题，本文采用水平和垂直两个方向的蛇形扫描方式。

具体而言，模型在水平与垂直两个正交方向上同步实施蛇形展开。以图2(a)水平分支(Horizontal Serpentine Flattening)为例，定义二维空间坐标为(h,w)，其中$0\leqslant h<H$且$0\leqslant w<W$。将其映射到一维时间步$t\in[1,HW]$的映射函数Φ_H定义为：

$$t=\Phi_H(h,w)=\begin{cases}h\cdot W+w, & \text{if } h \text{ is even}\\ h\cdot W+(W-1-w), & \text{if } h \text{ is odd}\end{cases} \tag{5}$$

该映射使偶数行从左到右展开，奇数行从右到左展开，从而减小换行位置带来的空间不连续。经

过水平蛇形扫描后，二维特征 $\boldsymbol{Z}$ 被展开为一维序列：

$$\boldsymbol{Z}_{seq}^{H}=\mathcal{S}_{H}(\boldsymbol{Z})\in\mathbf{R}^{HW\times C} \tag{6}$$

式中，$\mathcal{S}_{H}(\cdot)$ 表示上述水平蛇形扫描操作。类似地，垂直蛇形扫描通过对特征的高度和宽度维度进行转置后执行蛇形展开得到：

$$\boldsymbol{Z}_{seq}^{V}=\mathcal{S}_{V}(\boldsymbol{Z})\in\mathbf{R}^{HW\times C} \tag{7}$$

通过水平和垂直两个方向的序列化，模型能够分别捕获横向和纵向的空间上下文关系。

得到一维序列后，本文对每个方向均采用双向Mamba建模。以水平序列为例，前向Mamba沿扫描方向建模上下文，后向Mamba沿相反方向建模上下文：

$$\vec{\boldsymbol{Y}}_{seq}^{H}=Mamba_{fwd}^{H}(\boldsymbol{Z}_{seq}^{H}) \tag{8}$$

$$\overleftarrow{\boldsymbol{Y}}_{seq}^{H}=Reverse\left(Mamba_{bwd}^{H}\left(Reverse(\boldsymbol{Z}_{seq}^{H})\right)\right) \tag{9}$$

式中，$Mamba_{fwd}^{H}$ 和 $Mamba_{bwd}^{H}$ 分别表示水平分支中的前向Mamba和后向Mamba建模，$Reverse(\cdot)$ 表示沿序列维度翻转。$\vec{\boldsymbol{Y}}_{seq}^{H}$ 和 $\overleftarrow{\boldsymbol{Y}}_{seq}^{H}$ 分别表示经过前向和后向状态传播后得到的增强特征序列。

水平分支的双向输出相加后，再通过水平逆蛇形映射恢复为增强后的空间特征 $\boldsymbol{Y}^{H}\in\mathbf{R}^{C\times H\times W}$：

$$\boldsymbol{Y}^{H}=\mathcal{S}_{H}^{-1}\left(\vec{\boldsymbol{Y}}_{seq}^{H}+\overleftarrow{\boldsymbol{Y}}_{seq}^{H}\right) \tag{10}$$

垂直分支采用相同方式进行双向Mamba建模得到 $\boldsymbol{Y}^{V}$。

最后，将水平增强特征 $\boldsymbol{Y}^{H}$ 和垂直增强特征 $\boldsymbol{Y}^{V}$ 在通道维度进行拼接，并通过线性映射恢复为 C 通道，再与输入特征 $\boldsymbol{Z}$ 进行残差融合：

$$\mathcal{C}_{mamba}(\boldsymbol{Z})=\boldsymbol{Z}+\lambda\cdot Linear\left(Concat(\boldsymbol{Y}^{H},\boldsymbol{Y}^{V})\right) \tag{11}$$

式中，$Concat$ 表示沿通道维度的拼接操作，$Linear$ 为降维投影层将 $2C$ 通道还原为 C 通道，λ 为可学习的标量系数，初始化为0.5，用于动态平衡原始特征 $\boldsymbol{Z}$ 与全局上下文。通过这一处理，UA-SSM在保持线性复杂度的同时，实现了对二维空域信息的充分建模。

1.2.2 双视点不确定性估计

Mamba的长程状态传播在提升全局建模能力的同时，也可能将背景噪声一并扩散。为此，本文引入轻量的不确定性估计机制，对特征传播进行约束。参考贝叶斯深度学习中的划分方式，不确定性主要来源于输入扰动和语义不确定性。考虑到显著性预测对计算效率的要求，本文不采用显式概率建模，而是使用简单统计量作为不确定性的近似描述。

如图2(b)，为降低计算复杂度，输入特征 $\boldsymbol{X}$ 首先经1×1卷积降维至 $C/4$ 通道，记为 $\boldsymbol{F}_{red}$，计算以下两个不确定性代理分量：

具体而言，利用(3×3)局部窗口 Ω 内的空间方差描述区域响应的波动程度，将其作为输入噪声的代理，反映纹理复杂度：

$$\boldsymbol{U}_{spa}=\sqrt{\max\left(\mathbf{E}_{\Omega}[\boldsymbol{F}_{red}^{2}]-(\mathbf{E}_{\Omega}[\boldsymbol{F}_{red}])^{2},10^{-6}\right)} \tag{12}$$

式中，$E_{\Omega}[\cdot]$ 表示在该局部窗口内求数学期望，即局部均值统计。

同时，利用通道维度上的响应分歧衡量语义一致性，将通道方差作为模型认知不稳定性的代理，后者反映语义分歧。通道方差 $\boldsymbol{U}_{sem}$ 估计定义如下：

$$\boldsymbol{U}_{sem}=\frac{1}{C/4}\sum_{c=1}^{C/4}(\boldsymbol{F}_{red,c}-\bar{\boldsymbol{F}}_{red})^{2} \tag{13}$$

式中，$\bar{\boldsymbol{F}}_{red}$ 表示该局部特征在通道维度上的平均响应值。为实现这两种异构不确定性的深度交互，本文将维度扩展对齐后的 $\boldsymbol{U}_{spa}$ 与 $\boldsymbol{U}_{sem}$ 沿通道维度拼接，得到维度为 $C/2$ 的联合特征。随后，利用分组数等于 $C/4$ 的3×3分组卷积进行成对融合，确保每一个通道的认知不确定性都能与其对应的空间不确定性进行独立交互。最后，经3×3卷积 H_{gate} 与Sigmoid非线性激活，映射为单通道的像素级置信度门控矩阵 $\mathbf{G}$。

$$\boldsymbol{G}=\sigma\left(\mathbf{H}_{gate}\left(Concat(\boldsymbol{U}_{spa},\boldsymbol{U}_{sem})\right)\right),\quad \boldsymbol{G}\in[0,1]^{1\times H\times W} \tag{14}$$

上述方法并非严格的概率不确定性估计，而是一种面向工程实现的近似策略。其作用在于为状态传播提供简单有效的可靠性约束，从而在不增加明显计算开销的情况下，抑制噪声区域的特征扩散。

1.2.3 粗细粒度特征融合

为平衡全局上下文与局部细节，如图2(c)，UA-SSM采用粗细粒度双路并行增强结构。粗粒度分支对输入特征降采样后利用 $\mathcal{C}mamba$ 模块进行全局交互；细粒度分支在原分辨率下利用门控矩阵 对特征进行加权整流，再进行全局交互，以抑制高不确定性区域的噪声传播。

第一路为粗粒度全局分支(Coarse-Grained Global Branch)：对输入特征 $\boldsymbol{X}$ 进行自适应平均池化，压缩为低分辨率紧凑表征 $\boldsymbol{X}_{coarse}$。随后，利用Mamba模块进行全局交互：

$$\boldsymbol{X}_{global} = \mathcal{C}^{c}_{mamba}(\boldsymbol{X}_{coarse}) \tag{15}$$

$$\boldsymbol{X}_{c\_up} = UpSample\left(\boldsymbol{X}_{global}\right) \tag{16}$$

式中$\boldsymbol{X}_{global}$表示经过Mamba模块进行全局交互后得到的低频全局上下文特征，$UpSample$表示将特征图恢复至原始分辨率的上采样操作，$\boldsymbol{X}_{c_up}$为上采样后得到的粗粒度增强特征。该过程能够快速聚合场景的整体语义信息，典型如物体间的共现关系，且由于分辨率较低，自然过滤了大部分高频噪声。

第二路为细粒度不确定性整流分支（Fine-Grained Uncertainty Rectification Branch），在原始分辨率下运行，保留并增强显著目标的边缘与纹理细节。高分辨率特征包含丰富的细节信息，但同时也携带大量背景噪声，因此本文利用前文生成的像素级置信度门控$\boldsymbol{G}$对该分支进行特征整流（Feature Rectification），构建软阈值掩码，整流公式为：

$$\boldsymbol{X}_{masked} = \boldsymbol{X} \odot (0.5 + 0.5 \cdot \boldsymbol{G}) \tag{17}$$

式中，$\odot$为逐元素相乘操作，$\boldsymbol{X}_{masked}$表示经过门控加权后的整流特征。引入0.5的偏移量以避免高不确定性区域的特征被完全抑制，保留少量的信息用于后续建模，防止特征过度稀疏导致模型收敛困难。该调制方式不是二值截断，而是连续软权重调制。当$\boldsymbol{G}\to 1$时，可靠区域特征基本完整保留；当$\boldsymbol{G}\to 0$时，不可靠区域特征仍保留50%的响应，从而避免特征被完全抑制导致梯度传播困难。门控并不直接作用于Mamba的隐状态，而是先调制进入Mamba的输入特征，进而间接影响后续状态传播。

随后，将整流后的特征送入细粒度Mamba模块进行处理，得到上下文特征$\boldsymbol{X}_{fine}$：

$$\boldsymbol{X}_{fine} = \mathcal{C}^{f}_{mamba}\left(\boldsymbol{X}_{masked}\right) \tag{18}$$

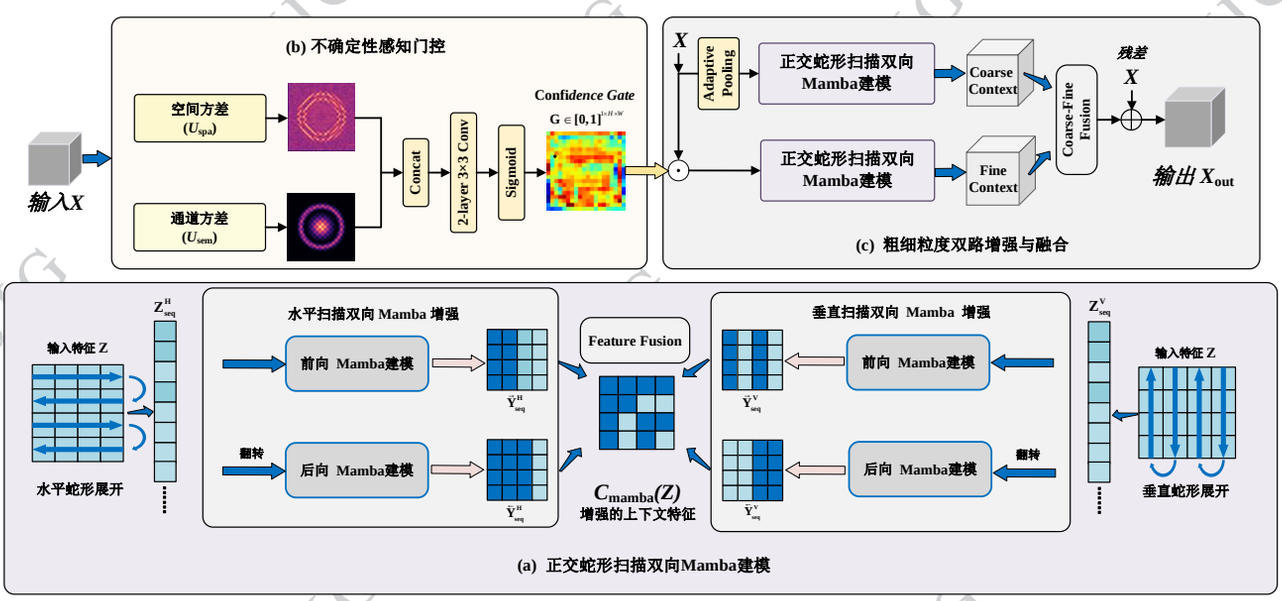
式中的$\mathcal{C}^{c}_{mamba}$和$\mathcal{C}^{f}_{mamba}$表示两个结构相同但参数独立的正交蛇形扫描双向Mamba块，分别用于粗粒度分支和细粒度分支。

最后，本文设计自适应加权融合策略，将粗粒度分支的全局上下文特征、细粒度分支的细节增强特征进行融合，同时为进一步提高输出特征的稳定性，再次利用置信度门控$\boldsymbol{G}$对细粒度分支的输出进行加权约束，最终$\boldsymbol{X}_{out}$输出特征的计算公式为：

$$\boldsymbol{X}_{out} = \boldsymbol{X} + Conv_{fuse}\left(\boldsymbol{X}_{c\_up} + \boldsymbol{X}_{fine} \odot \boldsymbol{G}\right) \tag{19}$$

式中，$Convfuse$为1个3×3卷积层，用于融合多分支特征并调整通道数，保证输出特征与输入特征的维度一致，便于残差连接。

综上，UA-SSM的双路架构，通过粗粒度分支保障了对整体场景的全局理解能力，同时通过细粒度分支保留了显著目标的局部结构与细节信息，有效缓解了全局建模与背景噪声抑制之间的矛盾。



(a)正交蛇形扫描双向Mamba建模；(b)不确定性感知门控；(c)粗细粒度双路增强与融合。

((a) Orthogonal serpentine scanning with bidirectional Mamba modeling; (b) uncertainty-aware gating; (c) coarse-fine dual-path enhancement.)

图2　UA-SSM模块结构图

Fig. 2　Architecture of the UA-SSM module.

1.3 语义引导的动态频域调制模块（SDFM）

解码阶段决定了显著性响应能否从高层语义恢复为紧凑、连续且具有清晰峰值结构的概率分布。常规逐级上采样虽然能够恢复空间分辨率，但容易削弱注视热点、边界过渡和局部峰值等结构成分，使显著图出现预测响应扩散和注视点偏移问题。FourierMamba等工作表明，频率学习能够增强退化结构恢复能力。但在显著性预测中，高频信息并非全部有益，真实显著结构与背景纹理干扰都可能包含高频响应，简单增强反而可能放大噪声。为此，本文提出SDFM。如图3所示，SDFM以解码端融合特征作为待调制对象，并引入UA-SSM输出的深层语义特征作为跨层先验，通过语义—内容联合路由生成样本自适应的复数频域滤波器，从而在频域内补偿上采样带来的结构响应衰减。

1.3.1 语义—内容联合路由生成

为了在频域内更有针对性地调制与显著区域相关的结构响应，并抑制背景纹理干扰，SDFM摒弃了全局统一滤波方法，引入基于语义先验的动态滤波器组路由机制：高频调制过程同时受到局部内容需求与高层语义条件的联合约束。设解码端融合特征为$\boldsymbol{Fdec}$，UA-SSM输出的深层语义特征为$\boldsymbol{Fsem}$，SDFM分别由二者生成内容路由和语义路由：

$$\boldsymbol{R}_d = \sigma(MLP_d(GAP(\boldsymbol{F}_{dec}))) \tag{20}$$

$$\boldsymbol{R}_s = \sigma(MLP_s(GAP(\boldsymbol{F}_{sem}))) \tag{21}$$

式中，$\boldsymbol{Rd}, \boldsymbol{Rs} \in R^{B\times K\times Cm}$，$B$表示批大小，$K$表示复数频域滤波器基数量，$Cm$表示通道映射后的中间通道数。通过全局平均池化（global average pooling，GAP）操作，空间特征被压缩为全局描述向量，随后经多层感知机（multi-layer perceptron，MLP）映射到滤波器基与通道联合路由空间，从而将解码内容信息和语义先验转化为可用于频域滤波器组合的路由权重：

$$\alpha_{b,k,c} = \frac{\boldsymbol{R}_{d,b,k,c}\boldsymbol{R}_{s,b,k,c}}{\sum_{j=1}^{K}\boldsymbol{R}_{d,b,j,c}\boldsymbol{R}_{s,b,j,c} + \epsilon}. \tag{22}$$

式中，b表示样本索引，k表示第k个滤波器基，c表示中间通道，K为滤波器基数量，ϵ为数值稳定项。内容路由刻画当前解码特征对不同频谱重塑模式的需求，语义路由提供高层语义约束，二者乘性耦合后决定各复数频域滤波器基的组合权重。因此，语义先验并非直接生成显著概率图，而是通过调节滤波器组的路由权重间接控制频域调制过程。与统一的全局频域滤波相比，这种联合路由方式能够在保持实现简洁的同时，使频域调制过程同时受到当前解码内容与高层语义先验的约束，从而提高频域增强的针对性。综合模型容量与计算开销，本文将滤波器基数量设定为4。

1.3.2 自适应频谱调制与残差重构

在计算出路由权重后，SDFM将K个可学习的复数频域滤波器基$\boldsymbol{W}_j$进行线性聚合，生成当前输入图像特有的自适应滤波器：

$$\boldsymbol{W}_{b,c}^{dyn}(u,v,c) = \sum_{k=1}^{K}\alpha_{b,k,c}\boldsymbol{W}_k(u,v) \tag{23}$$

式中，(u,v)为空间频率坐标，$\boldsymbol{W}k(u,v)$为第k个复数滤波器基。随后，解码特征先经过通道映射和$StarReLU(x) = s\cdot ReLU(x)^2 + b$激活，得到中间特征$\boldsymbol{F}_m$：

$$\boldsymbol{F}_m = StarReLU(Conv_{in}(\boldsymbol{F}_{dec})) \tag{24}$$

相比普通ReLU，StarReLU在非负响应区域提供更平滑的幅值调节，有助于降低频域变换前特征硬截断带来的谱响应不稳定。

之后再对$\boldsymbol{Fm}$执行二维实值快速傅里叶变换得到频谱表示：

$$\hat{\boldsymbol{F}}_m = \mathcal{F}(\boldsymbol{F}_m) \tag{25}$$

式中，$\mathcal{F}(\cdot)$表示二维实值快速傅里叶变换rFFT2。由于输入特征为实数信号，rFFT2可利用共轭对称性减少冗余频域计算。

在频域中，利用动态滤波器进行逐点复数调制，得到调制后的更新频谱$\hat{\boldsymbol{F}}_{mod}$：

$$\hat{\boldsymbol{F}}_{mod}^{(b)}(u,v,c) = \hat{\boldsymbol{F}}_m^{(b)}(u,v,c)\cdot\boldsymbol{W}_{b,c}^{dyn}(u,v,c) \tag{26}$$

最后通过逆傅里叶变换回到空间域，并以残差形式注入原解码特征：

$$\boldsymbol{F}_{out} = \boldsymbol{F}_{dec} + Conv_{out}\left(\mathcal{F}^{-1}(\hat{\boldsymbol{F}}_{\mathrm{mod}})\right) \tag{27}$$

式中，$Conv_{out}$为降维投影卷积，$\mathcal{F}^{-1}(\cdot)$表示二维逆实值快速傅里叶变换即irFFT2，$\boldsymbol{F}_{out}$为最终经过残差频域补偿的解码特征。

残差连接使频域分支主要承担结构补偿作用，而不直接替代原始解码特征。该过程在空间域中表现为注视热点更加集中、边界过渡更加清晰以及背景扩散响应减弱，从而缓解连续上采样造成的局部结构平滑问题。

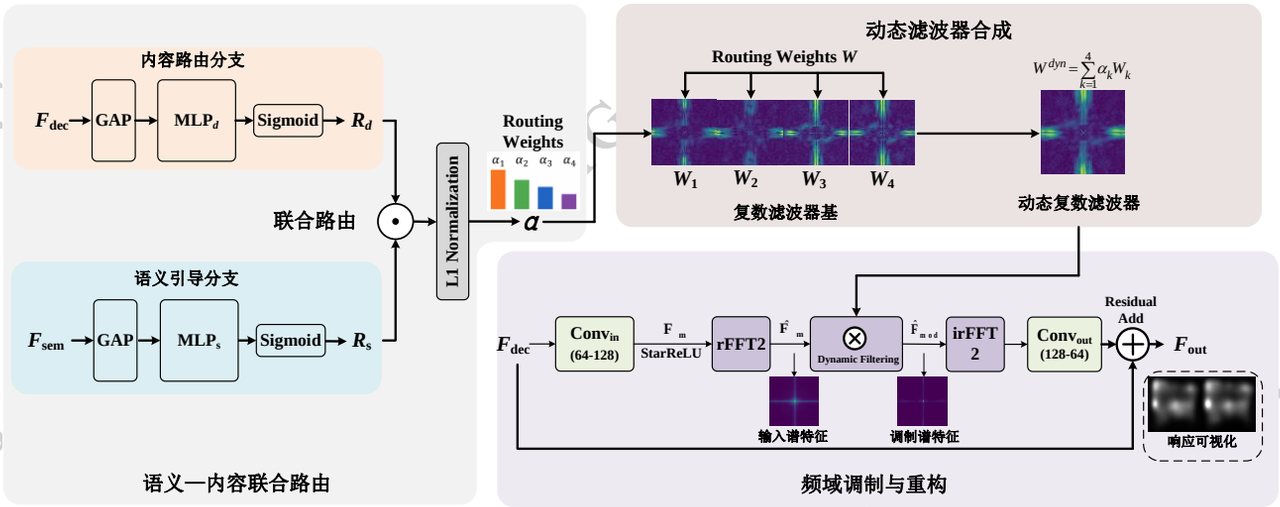


图3　SDFM模块结构图

Fig. 3　Architecture of the SDFM module.

1.4　损失函数与联合监督策略

显著性预测可建模为概率分布回归。训练时将预测图与真值图归一化为概率分布，并联合使用KL与CC损失约束分布一致性和空间相关性。

在实现过程中，预测显著图S与真值图T首先被归一化为概率分布形式，保证KL散度计算的合理性：

$$S \leftarrow \frac{S}{\sum_i S_i}, T \leftarrow \frac{T}{\sum_i T_i} \tag{28}$$

KL散度损失定义为：

$$L_{KL}(S,T) = \frac{1}{B}\sum_{b=1}^{B}\sum_i T_{b,i}\log\left(\varepsilon + \frac{T_{b,i}}{S_{b,i}+\varepsilon}\right) \tag{29}$$

式中B为批大小，i为像素索引，ε为数值稳定项。损失能够有效约束预测显著图与真实注视分布在整体形态与概率质量分配上的一致性。

然而，仅依赖分布距离约束可能导致预测结果在局部结构上出现偏移。为进一步提升显著图与真值图之间的全局空间一致性，本文引入线性相关系数损失Lcc，用于刻画预测分布与真实分布之间的整体相关程度。综合两类损失，本文采用最小化KL散度、最大化CC的联合训练目标，主损失函数定义为：

$$L_{main} = L_{KL}(S,T) - L_{CC}(S,T) \tag{30}$$

此外，为缓解深层网络训练过程中的梯度衰减问题、加速模型收敛，本文在解码器的中间尺度输出处引入辅助监督分支，通过生成中间显著图并计算辅助损失回传梯度。最终，总训练目标由主输出损失与辅助监督损失加权组成：

$$L = L_{main} + \gamma L_{aux} \tag{31}$$

式中$Laux$表示辅助输出损失，γ为权重系数。该策略可有效稳定训练过程，同时提升多尺度显著结构的预测一致性。

2　实验与结果分析

2.1　实验设置

2.1.1　数据集设置

本文选取5个公开数据集进行训练与评估：

1）SALICON：目前规模最大的显著性预测数据集，包含10000幅训练图像、5000幅验证图像和5000幅真实注视点分布保密的测试图像。其注视数据通过鼠标轨迹模拟眼动数据生成，常作为模型预训练的基础。LSUN′17挑战赛由SALICON测试集衍生而来，参与挑战赛时需自行在竞赛网站上提交推理结果。

2）MIT300：包含300幅具有高度挑战性的自然场景图像，该数据集的真实注视点分布同样不对外公开，需将模型预测的显著图提交至托管团队进行盲测，是目前视觉显著性预测领域最具影响力且使用最广泛的基准测试集之一。在本文测试中，模型先在MIT1003数据集上进行微调，随后提交至该平台进行统一评测。

3）MIT1003：包含1003幅来自Flickr和Google的自然图像，配备高精度的眼动仪记录数据。

4）CAT2000：包含20个不同类别的2000幅图像

（如卡通、素描、户外场景等），适合用于评估模型的跨域泛化能力。

5）TORONTO：包含120幅城市街景图像，用于检验模型在规则结构场景下的表现。

2.1.2 评价指标

采用显著性预测任务中通用的评价指标，分为两类：

1）基于分布的指标：含线性相关系数（correlation coefficient，CC）、相似度（similarity，SIM）与Kullback-Leibler散度（Kullback-Leibler divergence，KL）。CC、SIM指标核心衡量预测与真值分布的空间结构相关性、重叠程度，数值越高性能越优。KL散度以信息论方法刻画两类分布的差异，为分布形态合理性检验的核心指标，适配频域约束、概率分布校正机制的贡献度分析，数值越低性能越优。

2）基于注视点的指标：含标准化扫描路径显著性（normalized scanpath saliency，NSS）、Judd版曲线下面积（area under curve-Judd，AUC-J）、Borji版曲线下面积（area under curve-Borji，AUC-B）与混洗曲线下面积（shuffled area under curve，sAUC），以及相对于基准先验在真实注视点处带来的信息增益（Information Gain，IG）。此类指标数值越高性能越优。NSS反映预测显著图在真实注视点处的归一化响应强度，可衡量模型对注视点的直接命中能力。AUC系列指标将显著图视为二分类器得分图，综合衡量阈值变化下的性能表现。不同变体的负样本构造逻辑存在差异，可从多维度评估模型对注视点的区分能力。

2.1.3 对比方法

为全面评估所提出方法的性能，本文选取共33种方法进行对比，对比方法涵盖：

1）经典显著性模型：Itti、GBVS、BMS、CovSal等，这类方法主要基于人工设计特征进行显著性建模；

2）CNN-based方法：如eDN、MLNet、SalGAN、SAM、EMLNet、MSINet等，通过卷积网络建模多尺度特征；

3）Transformer或混合模型：如TranSalNet、GSGNet、UNETRSal等，利用自注意力机制增强全局建模能力；

4）Mamba建模方法，如SUM。

这些方法覆盖了显著性预测从传统方法到深度学习再到Transformer、Mamba结构的发展路径，能够全面验证本文方法的性能优势。

2.1.4 实验细节

本文提出的S²UG-Mamba基于PyTorch深度学习框架构建，并在单张NVIDIA GeForce RTX 4090D GPU上完成所有的训练与推理任务。输入图像在送入网络前统一调整为320 × 320分辨率，并依据ImageNet数据集的均值与方差进行归一化处理。为了加速收敛并利用大规模视觉特征，模型的主干网络ConvNeXt-Tiny使用ImageNet-1K预训练权重进行初始化，而UA-SSM、SDFM及解码器等新引入模块则采用随机初始化策略。

针对训练过程中的超参数设置，本文采用了AdamW优化器进行参数更新，并将权重衰减系数设定为0.05以防止过拟合。考虑到主干网络已具备良好的特征提取能力，而新设计的状态空间与频域模块需从头学习，故实施分层学习率策略：主干网络的初始学习率设定为1×10^{-4}，而新引入模块的初始学习率设定为1×10^{-3}，即主干学习率的10倍。训练过程共持续15个Epoch，批大小Batch Size设定为16。学习率调度采用多步衰减策略，在第1、6、11个Epoch结束时，学习率分别衰减为当前值的0.1倍，最小学习率限制为1×10^{-8}。

此外，在面向数据量较小的MIT1003等数据集进行微调时，冻结主干网络的前两个阶段（Stem和Stage 1），仅更新高层参数，以避免在数据稀疏场景下发生灾难性遗忘。

2.2 定量对比实验

2.2.1 主流竞赛基准性能对比

表1展示了本文方法与经典方法和最新主流方法在LSUN′17竞赛上的定量对比结果。其中加粗字体为最优值。S²UG-Mamba在IG、sAUC、CC、SIM和KL指标上取得已公开的最优结果。在NSS和AUC上，本文方法也较为优秀。与GSGNet相比，S²UG-Mamba在将KL由0.190降低至0.176，降低7.4%，IG达到0.943，提升4.0%。综合来看，S²UG-Mamba在复杂场景下能够较好地平衡注视点定位准确性与分布一致性。

在MIT300数据集上的定量结果如表2所示，具体而言，在基于分布的指标方面，本文方法在CC（0.829）和KL（0.370）等指标上均取得了最高分，优于公认的先进模型GSGNet。相比之下，本文方法在CC和KL上分别提升了2.2%和9.8%。在基于注视

点的指标方面，本文方法在AUC指标上处于第一梯队，而在NSS和sAUC指标上略低于当前排名第一的模型。与传统经典方法GBVS相比，本文方法在AUC、sAUC、NSS、CC、SIM和KL指标上分别提升了约9.2%、20.3%、99.4%、73.1%、42.6%和58.3%。

表1　LSUN′17竞赛上的性能表现

Table1　Performance comparison of saliency prediction models on LSUN′17 competition

		基于注视点的指标				基于分布的指标		
模型	发表情况	IG ↑	AUC ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑	KL ↓
SAM-Res	TIP2018	0.538	0.865	0.741	1.990	0.899	0.793	0.610
GazeGAN	TIP2019	0.720	0.864	0.736	1.899	0.879	0.773	0.376
DINet	TMM2019	0.195	0.862	0.739	1.959	0.902	0.795	0.864
SimpleNet	IROS2020	0.880	0.869	0.743	1.960	0.907	0.793	0.201
UNISAL	ECCV2020	-	0.864	0.739	1.952	0.879	0.775	-
EMLNet	IVC2020	0.736	0.866	0.746	2.050	0.886	0.780	0.520
MSINet	NN2020	0.793	0.865	0.736	1.931	0.889	0.784	0.307
SalED	IVC2021	0.909	0.868	0.745	1.984	0.910	0.801	0.190
FBNet	MVA2021	0.343	0.843	0.706	1.687	0.785	0.694	0.708
ACNet	NC2021	0.856	0.866	0.739	1.948	0.896	0.786	0.228
TranSalNet	IJON2022	-	0.868	0.747	2.014	0.907	0.803	0.373
SalFBNet	IVC2022	0.839	0.868	0.740	1.952	0.892	0.772	0.236
TempSAL	CVPR2023	0.896	0.869	0.745	1.967	0.911	0.800	0.195
SalDA	TCDS2024	0.577	0.851	0.714	1.727	0.821	0.693	0.369
GSGNet	KBS2024	0.907	0.870	0.746	1.988	0.912	0.800	0.190
AugSal	ECCV2024	-	0.870	0.744	1.973	0.914	0.805	0.191
UNETRSal	ACIVS2025	0.821	0.870	0.745	**2.058**	0.914	0.808	0.316
SUM	WACV2025	-	**0.876**	-	1.981	0.909	0.804	0.192
SimpleSalNet	JVCIR2026	-	0.875	-	1.978	0.906	0.798	0.197
本文	-	**0.943**	0.870	**0.749**	2.007	**0.918**	**0.814**	**0.176**

2.2.2　跨数据集泛化能力评估

为了评估模型的泛化性能，S^2UG-Mamba在MIT1003数据集上进行了零样本测试，即仅使用SALICON训练的模型直接进行评估。如表3所示，NSS达到2.386，CC达到0.685，均优于TranSalNet和GSGNet等先进模型。这表明模型有效地学习到了具有普适性的显著性特征，而非仅仅拟合特定数据集的分布。

表4和表5给出了模型在数据集CAT2000与TORONTO上的跨数据集测试对比结果，包含仅使用SALICON训练的模型，以及在此基础上引入少量MIT1003样本进行微调后的结果。红色加粗和加粗字体分别为最优和次优结果。在不进行任何目标域训练的情况下，模型在两个数据集上相较于GSGNet、TranSalNet等公认先进方法在NSS、CC与SIM等指标上均表现出稳定提升，同时KL明显降低。这表明在类别更加丰富、注视模式更加分散的场景中，模型仍能够较好地保持预测分布的整体结构，而不会因域偏移而出现显著退化。不过各方法之间的差距相对较小，个别分布型指标与最优方法接近或略有差异，但在注视点型指标上仍保持小幅优势或持平。这种分化现象反映了不同数据集在中心偏置强度、目标尺寸分布及背景复杂度方面的差异，也说明单一指标难以全面刻画跨域性能。当引入少量

表2　MIT300竞赛上的性能表现

Table2　Performance comparison of saliency prediction models on MIT300 competition

		基于注视点的指标			基于分布的指标		
模型	发表情况	AUC ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑	KL ↓
Itti	TPAMI1998	0.5434	0.5357	0.4081	0.1307	0.3378	1.4964
GBVS	NIPS2006	0.806	0.629	1.245	0.479	0.484	0.887
CAS	TPAMI2011	0.758	0.640	1.018	0.384	0.431	1.072
BMS	ICCV2013	0.771	0.691	1.151	0.413	0.445	1.023
CovSal	JoV2013	0.811	0.589	1.336	0.500	0.505	1.722
LDS	TNNLS2016	0.810	0.602	1.364	0.517	0.522	1.063
eDN	CVPR2014	0.817	0.618	1.140	0.452	0.411	1.137
MLNet	ICPR2016	0.838	0.739	1.974	0.663	0.581	0.800
SalGAN	CVPR2017	0.849	0.735	1.862	0.674	0.593	0.757
DVA	TIP2018	0.843	0.725	1.930	0.663	0.584	0.629
CASNet II	CVPR2018	0.855	0.739	1.985	0.705	0.580	0.585
SAM-VGG	TIP2018	0.847	0.730	1.955	0.663	0.598	1.274
SAM-Res	TIP2018	0.852	0.739	2.062	0.689	0.611	1.171
GazeGAN	TIP2019	0.860	0.731	2.211	0.757	0.649	1.339
EML-Net	IVC2020	0.876	0.746	**2.487**	0.789	0.675	0.843
MSINet	NN2020	0.873	0.779	2.305	0.780	0.670	0.670
UNISAL	ECCV2020	0.877	0.784	2.369	0.785	0.675	0.415
Cheng	IVC2021	0.870	0.740	2.350	0.790	0.680	0.880
SalED	IVC2021	0.880	0.720	2.400	0.810	0.690	0.400
TranSalNet	IJON2022	0.873	0.746	2.413	0.807	0.689	1.014
SalDA	TCDS2024	0.840	0.710	1.690	0.620	0.550	0.740
GSGNet	KBS2024	0.878	**0.788**	2.423	0.811	0.690	0.410
本文	-	**0.880**	0.757	2.483	**0.829**	**0.690**	**0.370**

MIT1003样本进行微调后，模型在两个数据集上的性能均出现显著提升。值得注意的是，微调带来的增益在场景类别更多、注视分布更加分散的数据集中更为明显，这说明模型并非仅依赖大规模数据集的统计先验，而是可以利用少量真实眼动数据对目标域分布进行快速对齐。

2.2.3　稳定性与鲁棒性评估

为避免单次数据划分带来的偶然性影响，在MIT1003与CAT2000数据集上进一步采用五折交叉验证策略进行评估，结果分别列于表6与表7。S^2UG-Mamba在五次不同划分下的平均NSS、CC与SIM均稳定高于GSGNet模型，同时KL保持更低水平。各项指标的方差均处于较低量级，与表中方法相比在不同指标上呈现一定取舍差异，未出现明显性能波动。这表明模型性能的提升在统计意义上具有稳定性。现有方法在中心偏置建模、类别敏感性及评价指标侧重方面具有不同设计取向，而S^2UG-Mamba更倾向于在注视点命中能力与整体分布一致性之间取得平衡。

2.2.4　复杂度与效率分析

为进一步评估所提出双向蛇形扫描策略以及Coarse-Fine双分支处理的计算开销以及模型复杂度与效率，构建了两个变体：变体1为单分支Mamba变体Single-Fine Path，不包含UA-SSM中的不确定性门

表3　MIT1003数据集上零样本测试

Table 3　Zero-Shot Testing Results on the MIT1003 Dataset

	基于注视点的指标			基于分布的指标		
模型	AUC-J ↑	AUC-B ↑	NSS ↑	CC ↑	SIM ↑	KL ↓
AIM	0.6025	0.5987	0.3809	0.1220	0.2430	1.9053
GBVS	0.8232	0.8133	1.3656	0.4174	0.3627	1.2971
LDS	0.7793	0.7128	1.0185	0.3239	0.3540	1.8726
DVA	0.8702	0.8066	2.3145	0.6435	0.4993	0.9467
MLNet	0.8532	0.7723	2.2167	0.5921	0.4900	1.3441
EML-NET	0.8814	0.8289	2.4453	0.6757	0.5521	1.4397
ACNet-R	0.8828	0.8072	2.4558	0.6749	0.5401	0.8907
ACNet-V	0.8859	0.8637	2.2596	0.6502	0.5022	0.8638
TranSalNet-R	0.8855	0.8646	2.2706	0.6516	0.5058	1.0318
TranSalNet-D	0.8876	0.8666	2.2838	0.6543	0.5045	1.0151
GSGNet	0.8901	**0.8702**	2.3354	0.6714	0.5131	0.7932
本文	**0.8911**	0.8685	**2.3859**	**0.6849**	**0.5420**	**0.7527**

表4　CAT2000跨数据集测试

Table 4　Cross-Dataset Testing Results on CAT2000

	基于注视点的指标				基于分布的指标		
模型	AUC-J ↑	AUC-B ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑	KL ↓
AIM	0.7243	0.7211	0.6054	0.8824	0.3395	0.4277	1.1702
GBVS	0.8005	0.7899	0.6242	1.2457	0.4864	0.4982	0.8615
LDS	0.8320	0.7970	0.6319	1.5506	0.6054	0.5673	0.8326
EML-NET	0.8308	0.7756	0.6611	1.6470	0.6076	0.5822	1.8280
ACNet-R	0.8484	0.7767	0.6813	**1.9163**	**0.6864**	0.6065	0.8284
ACNet-V	0.8364	0.8157	0.6773	1.6210	0.6142	0.5850	0.7515
SAM-VGG	0.8389	0.7516	0.6664	1.7643	0.6281	0.5764	1.1117
SAM-Res	0.8432	0.7619	0.6717	1.8257	0.6550	0.5910	1.1576
TranSalNet-R	0.8420	0.8263	0.6868	1.6313	0.6182	0.5763	0.8144
TranSalNet-D	0.8456	**0.8307**	**0.6920**	1.6422	0.6238	0.5757	0.8211
GSGNet	0.8460	0.8270	**0.6911**	1.6994	0.6429	0.5935	0.5955
本文	**0.8494**	**0.8274**	0.6592	1.7500	0.6603	**0.6101**	**0.5752**
本文 (w/ MIT)	**0.8606**	0.7878	0.6436	**2.0787**	**0.7556**	**0.6530**	**0.4874**

控、coarse/fine双分支增强结构，扫描方式仍为蛇形双向扫描，主要用于反映减少Mamba分支后的效率上限；变体2 Four-Scan则在保持骨干网络、解码器、coarse/fine双分支不确定性引导结构和语义引导频域校正模块不变的情况下，仅将本文的双向蛇形扫描替换为传统四向扫描，即把二维特征图按四个方

表5　TORONTO跨数据集测试

Table 5　Cross-Dataset Testing Results on TORONTO

模型	基于注视点的指标			基于分布的指标			
	AUC-J ↑	AUC-B ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑	KL ↓
AIM	0.7446	0.7369	0.6822	0.8957	0.3347	0.3750	1.3374
GBVS	0.8239	0.8123	0.6993	1.4564	0.5690	0.4877	0.8521
LDS	0.8333	0.7812	0.6956	1.6538	0.6350	0.5640	1.1442
EML-NET	0.8462	0.7703	0.7245	2.0184	0.7291	0.6125	2.6097
ACNet-R	0.8632	0.7687	0.7265	**2.0431**	0.7386	0.6164	0.8038
ACNet-V	0.8619	0.8256	0.7597	1.9755	0.7390	0.6151	0.6712
SAM-Vgg	0.8536	0.7295	0.6957	1.9375	0.6875	0.5881	2.0306
SAM-Res	0.8548	0.7330	0.6989	1.9699	0.6913	0.5926	1.9968
TranSalNet-R	0.8607	0.8211	0.7592	1.9725	0.7318	0.6183	1.3881
TranSalNet-D	0.8602	0.8183	0.7570	1.9837	0.7353	0.6175	1.3607
GSGNet	**0.8642**	0.8287	0.7648	2.0234	0.7564	0.6255	**0.5321**
本文	0.8608	**0.8243**	**0.7643**	2.0262	**0.7580**	**0.6334**	0.5389
本文(w/MIT)	0.8716	0.7801	0.7390	2.2291	0.8093	0.6596	0.4809

向展开成一维序列，再分别做Mamba建模，因此可用于公平比较扫描策略本身的开销。与此同时，选取 Retina-VGG、TranSalNet-R、TranSalNet-D、TempSAL和GSGNet等代表性显著性预测方法进行对比。图4给出了不同方法在参数量、浮点运算量(floating point operations，FLOPs)、显存占用和推理速度方面的比较结果。

如图4(a)所示，在320×320分辨率下，本文方法的参数量和FLOPs与两个变体基本一致，说明所提出蛇形扫描和双分支的处理并未显著引入参数或计算量。图4(b)指出随着输入分辨率升高，本文方法的FLOPs始终与四向扫描变体接近，并明显低于GSGNet等方法，表明其具有较好的计算扩展性。

从显存占用看，图4(c)中本文方法与变体2几乎重合，略微高于变体1，说明蛇形扫描和双路处理不会造成额外显存负担。推理速度方面，图4(d)显示本文方法相比两个变体有一定差距，主要原因是双路处理和蛇形展开与逆映射引入了额外的特征重排与访存开销，但在高分辨率下该差距明显缩小。

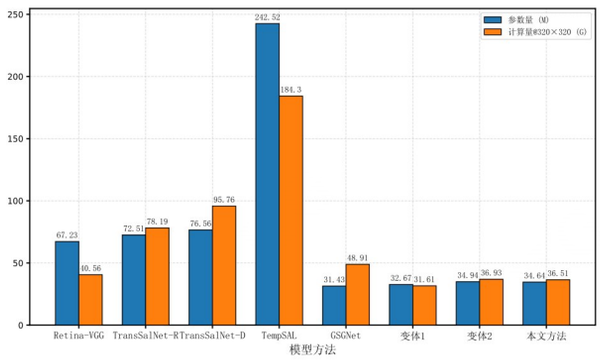
总体来看，本文方法在参数量、FLOPs和显存占用上与两个变体保持同一量级，仅在推理速度上存在一定劣势。相比GSGNet等方法，本文方法在复杂度与计算效率方面更具优势。

2.3 消融实验

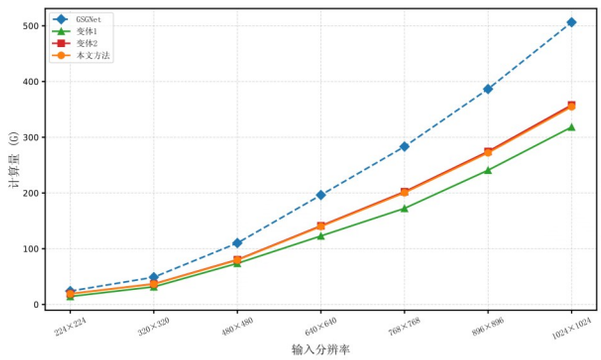
2.3.1 核心组件有效性验证

为了深入验证所提S²UG-Mamba中各个关键组件的有效性及其对显著性预测性能的具体贡献，本文在SALICON验证集上进行了严格的逐步消融实验。实验的基础模型(Baseline)仅采用预训练的ConvNeXt-Tiny作为特征提取主干，并搭配基础的多尺度特征拼接解码器。表8详细展示了逐步引入UA-SSM和SDFM后的量化性能变化。

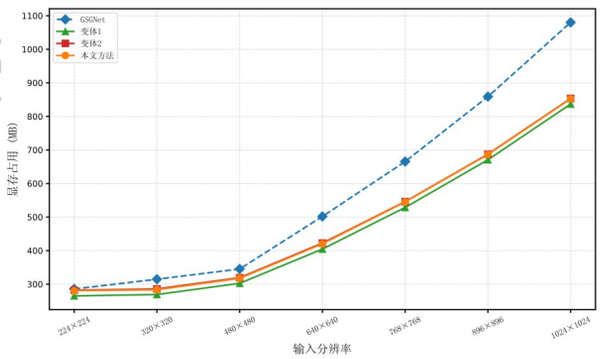
如表8所示，当在基线模型中单独加入UA-SSM后(表中写为UA-M)，模型在所有评价指标上均获得了显著提升。其中，反映预测分布一致性的KL散度从0.1908明显下降至0.1872，同时对注视点极为



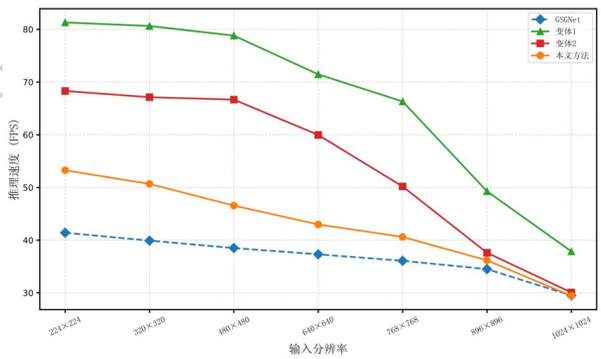
(a) 参数量与计算量对比



(b) 计算量随输入分辨率的变化趋势



(c) 峰值显存随输入分辨率的缩放趋势



(d) 推理速度随输入分辨率的变化趋势

((a) Comparison of parameters and FLOPs; (b) Variation of inference FLOPs with input resolution; (c) Trend of GPU memory usage with varying input resolutions; (d) Trend of FPS with varying input resolutions)

图4 不同模型及变体的复杂度与推理效率对比分析

Fig. 4 Comparative analysis of complexity and inference efficiency among different models and their variants

表6　MIT1003自然场景五折测试

Table 6　Five-Fold Cross-Validation Results on the MIT1003 Natural Scene Dataset

	基于注视点的指标				基于分布的指标	
模型	AUC-J ↑	AUC-B ↑	NSS ↑	CC ↑	SIM ↑	KL ↓
ACNet-V	0.9005	0.8403	2.7260	0.7573	0.61 18	0.9303
(Variance)	0.000007	0.000040	0.004864	0.000053	0.000027	0.002303
ACNet-R	0.9019	0.8443	2.7407	0.7594	0.6099	0.8495
(Variance)	0.000006	0.000038	0.004437	0.000041	0.000016	0.001187
TranSalNet-R	0.9022	0.8333	2.7514	0.7535	0.6136	0.8531
(Variance)	0.000010	0.000046	0.003863	0.000024	0.000016	0.000078
TranSalNet-D	0.9056	0.8347	2.8329	0.7715	0.6288	0.8502
(Variance)	0.000007	0.000045	0.004060	0.000010	0.000016	0.001338
GSGNet	**0.9073**	0.8429	2.8416	0.7833	0.6155	0.5721
(Variance)	0.000008	0.000052	0.003889	0.000029	0.000010	0.000075
本文	0.9066	**0.8456**	**2.9225**	**0.7991**	**0.6531**	**0.5508**
方差波动	0.00001	0.000012	0.005138	0.000024	0.000018	0.000142

表7　CAT2000多类别场景五折测试

Table 7　Five-Fold Cross-Validation Results on the CAT2000 Multi-Category Dataset

	基于注视点的指标				基于分布的指标		
模型	AUC-J ↑	AUC-B ↑	sAUC ↑	NSS ↑	CC ↑	SIM ↑	KL ↓
ACNet-V	0.8818	0.8146	0.6259	2.4057	0.8848	0.7527	0.4992
(Variance)	0.000003	0.000002	0.000007	0.000248	0.000015	0.000010	0.000144
ACNet-R	0.8811	0.8325	0.6371	2.3678	0.8749	0.7403	0.4678
(Variance)	0.000002	0.000005	0.000013	0.000326	0.000013	0.000012	0.000219
TranSalNet-R	0.8806	0.8063	0.6242	2.4093	0.8726	0.7444	0.4959
(Variance)	0.000002	0.000002	0.000005	0.000311	0.000008	0.000004	0.000118
TranSalNet-D	0.8823	0.8024	0.6263	2.4492	0.8799	0.7507	0.5010
(Variance)	0.000002	0.000005	0.000006	0.000297	0.000012	0.000006	0.000264
GSGNet	0.8845	**0.8148**	0.6306	2.4472	0.8931	0.7509	0.2642
(Variance)	0.000002	0.000004	0.000005	0.000243	0.000005	0.000002	0.000016
本文	**0.8846**	0.8108	**0.6321**	**2.4791**	**0.8994**	**0.7675**	**0.2581**
方差波动	0.000002	0.000004	0.000007	0.000234	0.000007	0.000005	0.000023

敏感的NSS指标从1.9627提升至2.0152，抗中心偏移的sAUC指标也有所提高。这表明基于不确定性感知的交叉扫描建模机制有助于减弱复杂背景对特征建模的不利影响。

另一方面，若在基线中仅单独引入SDFM模块，模型的CC（0.9179）和SIM（0.8115）等基于分布的相似度指标同样获得了明显改善，且其KL散度（0.1858）的降低幅度甚至略优于单独使用UA-SSM。这表明在频域视角下进行特征的高频细节补偿与自适应滤波，能够有效帮助模型恢复和保留图像的精细空间结构信息，有效缓解了常规空域下采样操作带来的显著边缘模糊问题。

当同时集成UA-SSM与SDFM模块构成完整的S^2UG-Mamba时，模型在全部七项评估指标上均达到了最优水平。其中，KL散度进一步降至最小值0.1839，NSS达到2.0235峰值，且体现鲁棒性的AUC-B也升至0.8502。这不仅确立了两个模块各自的结构优越性，进一步说明了二者之间具有较好的协同作用：UA-SSM输出的增强特征为SDFM提供了高质量的语义先验，使频域调制过程能够更准确地定位需要补偿的结构区域。二者在空域全局抗噪与频域局部细节重构上形成了互补效应，共同构建了兼顾效率与精度的显著性特征表征体系。

2.3.2　UA-SSM模块内部结构设计分析

表9展示了UA-SSM模块内部不同门控与序列建模策略的消融实验结果。首先，上述的单分支不确定性蛇形双向Mamba增强处理（表中Single-Fine Path）由于其采用单一的全局处理方式，在特征增强时不可避免地丢失了注视点信息，各项指标表现欠佳。其次，无门控（表中w/o Gating）的Coarse-Fine分支蛇形双向Mamba增强虽然在部分基于注视点的绝对指标上取得了较高数值，但由于其不能滤除无效的背景响应，引入了大量背景噪声，导致基于分布的指标表现欠佳。而仅将该模块Mamba扫描方式换为传统四向Mamba，其在推理速度上有所优势，但受限于扫描机制无法充分捕捉二维图像复杂的空间结构上下文，使得特征分布的拟合能力遭遇瓶颈。随机门控（表中Random Gating）虽然直接将计算密度降低，却导致了NSS与KL等核心指标的全面退化，说明简单粗暴的特征稀疏化会割裂空间信息的连续性，破坏Mamba序列建模的有效过程，进一步证明了不确定性估计的有效性。本文方法在保持计算效率的同时获得最优的AUC-B、sAUC、CC、SIM和KL。结果表明，所提出的扫描方式在捕捉局部连续

性上有一定优势，基于不确定性的门控能够有效提升注视分布的拟合能力。

2.3.3 SDFM模块设计策略分析

表10详细给出了SDFM内部关键结构的消融实验结果。对比变体(a)基础ReLU且无引导和(b)StarReLU且无引导可以发现，采用StarReLU后模型在多项指标上均有改善，说明更平滑的非线性映射有助于频域特征建模。凭借其更平滑的非线性映射能力，StarReLU在维持复杂频域特征数值稳定性、保留关键结构信息方面具有一定优势。在此基础上，变体(c)完整SDFM模块进一步引入了来自Mamba编码器的空间语义特征作为调制机制。与无引导的变体(b)相比，引入语义指导后，各类指标达到了最优水平，证明融合高层语义信息生成语义指导向量，与内容路由向量联合计算滤波器权重，能够有效赋予频域滤波器动态的空间感知能力。综上所述，SDFM模块最终实现了预测精度与分布一致性的全面提升。

表8 核心组件消融实验结果

Table 8 Ablation study on key components

组件			基于注视点的指标				基于分布的指标		
ConvNeXt_t	UA-M	SDFM	AUC-J↑	AUC-B↑	sAUC↑	NSS↑	CC↑	SIM↑	KL↓
√	–	–	0.8774	0.8449	0.7502	1.9627	0.9126	0.8049	0.1908
√	√	–	0.8782	0.8494	0.7575	2.0152	0.9172	0.8108	0.1872
√	–	√	0.8776	0.8473	0.7566	2.0087	0.9179	0.8115	0.1858
√	√	√	**0.8786**	**0.8502**	**0.7582**	**2.0235**	**0.9192**	**0.8126**	**0.1839**

表9 UA-SSM不同策略消融实验结果

Table 9 Ablation Study of Different Gating Strategies in UA-SSM

	基于注视点的指标				基于分布的指标		
策略变体	AUC-J↑	AUC-B↑	sAUC↑	NSS↑	CC↑	SIM↑	KL↓
单分支增强	0.8779	0.8424	0.7523	1.9984	0.9173	0.8099	0.1871
无门控	**0.8789**	0.8469	0.7564	**2.0272**	0.9185	0.8124	0.1857
传统四向Mamba	0.8783	0.8475	0.7567	2.0186	0.9185	0.8119	0.1853
随机门控	0.8782	0.8456	0.7558	2.0106	0.9177	0.8109	0.1868
本文模型	0.8786	**0.8502**	**0.7582**	2.0235	**0.9192**	**0.8126**	**0.1839**

表10 SDFM模块内部结构消融实验结果

Table 10 Ablation Study on the Internal Structure of SDFM

			基于注视点的指标				基于分布的指标		
变体	激活函数	语义引导	AUC-J↑	AUC-B↑	sAUC↑	NSS↑	CC↑	SIM↑	KL↓
(a)	ReLU	–	0.8774	0.8449	0.7502	1.9627	0.9126	0.8049	0.1908
(b)	StarReLU	–	0.8782	0.8494	0.7575	2.0152	0.9172	0.8108	0.1872
(c)	StarReLU	√	**0.8786**	**0.8502**	**0.7582**	**2.0235**	**0.9192**	**0.8126**	**0.1839**

2.4 定性对比与可视化分析

本节从定性对比与模块机理可视化两个角度，进一步评估S²UG-Mamba在复杂自然场景中的预测能力。

图5给出了本文方法与经典方法和最新先进方法在不同挑战场景下的定性对比。在复杂杂乱背景与高密度纹理场景下，对比模型在非显著的背景区域如草地、杂乱室内存在不同程度的误响应现象。本文方法生成的显著图在背景区域的响应相对较弱，表明模型对背景噪声具有一定的抑制作用。在多尺度小目标与精细轮廓场景中，受上采样平滑效应影响，部分模型对小目标的定位不够准确。而本文方法生成的预测分布在目标边缘处表现出更好的紧凑性，与真值的轮廓一致性较高。在严重偏心分布与远距离语义关联场景中，CNN架构模型受感受野限制，在边缘目标的捕获上存在偏差。本文方法利用Mamba的长程建模特性，在非中心区域依然能维持相对稳定的定位表现。综合结果来看，本文方法在不同复杂度的场景下均能产生与人类注视分布基本一致的预测结果，表现出较好的鲁棒性。

为观察不确定性门控对背景误响应的抑制作用，本文对比了去除门控与加入门控后的预测结果，如图6所示。在缺少不确定性门控时，模型的显著响应更容易向复杂纹理背景或非注视区域扩散，如原图中球场标线区域、人的腿部、设备边缘等。相比之下，加入不确定性门控后，预测响应更加集中于真实注视区域，非注视背景区域的虚假响应明显减弱。这说明门控能够根据局部可靠性对Mamba状态传播过程进行软约束，从而降低复杂背景噪声在长程建模过程中的累积传播。

为进一步直观验证语义调制频域补偿模块SDFM在恢复高频边缘和结构细节方面的有效性，图7展示了使用模块前后的可视化对比结果。观察结果可以发现，在未引入频域约束的情况下，模型虽然能够定位到显著性目标的位置，但在目标轮廓和精细结构处存在较为明显的边缘模糊与响应弥散现象。相比之下，在引入SDFM模块后，本文模型预测的显著图在多目标和复杂场景下呈现出更高的边缘锐度，显著区域的能量分布更加紧凑，与人类真实注视分布形态更为吻合。

为进一步定量验证图7中观察到的响应集中性变化，本文在SALICON验证集上比较了无SDFM变体与完整模型，并引入响应扩散半径（response diffusion radius，RDR）、中心偏差距离（centroid error distance，CED）和显著区域增益占比（salient gain ratio，SGR）三个辅助指标。RDR衡量预测响应相对于自身质心的扩散程度，CED衡量预测质心与真实注视质心之间的距离，SGR衡量SDFM带来的正向增益中有多少位于真实高响应区域。三者定义如下：

$$\mathrm{RDR}=\frac{\sqrt{\sum_i P(i)\|p_i-\mu_{pred}\|_2^2}}{\sqrt{H^2+W^2}} \tag{32}$$

$$\mathrm{CED}=\frac{\|\mu_{pred}-\mu_{gt}\|_2}{\sqrt{H^2+W^2}} \tag{33}$$

$$\mathrm{SGR}=\frac{\sum_i D^+(i)M_{gt}(i)}{\sum_i D^+(i)+\epsilon} \tag{34}$$

式中的中间变量定义如下：

经过归一化处理后的预测显著性概率分布：

$$P(i)=\frac{S(i)}{(\sum_i S(i)+\epsilon)} \tag{35}$$

预测显著图的概率分布质心：

$$\mu_{pred}=\sum_i P(i)p_i \tag{36}$$

局部正向概率响应增益：

$$D^+(i)=\max(P_{w/SDFM}(i)-P_{w/oSDFM}(i),0) \tag{37}$$

式中，i为特征图的像素索引，p_i表示像素索引i对应的二维空间坐标向量，H和W分别表示预测图的高度与宽度，μ_{gt}为真实注视图的概率质心，Mgt为真实注视图top 20%高响应区域掩码，ϵ为数值稳定项。

表11　SDFM对响应集中性的影响分析

Table 11　Effect of SDFM on response concentration

模型	RDR↓	CED↓	SGR↑
w/o SDFM	0.1885	0.0235	—
w/ SDFM	0.1867	0.0219	0.6580
改善幅度	0.95%↓	6.81%↓	—

从图7差异图分布和表11可以看出，加入SDFM后，约65.8%的正向增益位于真实高响应区域内，而在非显著背景区域则几乎没有额外的能量激活。预测响应扩散得到一定缓解，且响应中心更接近真实注视中心。

3　结　论

针对复杂自然场景下视觉显著性预测中长程依赖建模不足、噪声易随全局扩散以及解码过程细节

丢失的难题，本文提出了一种融合空频双域调制的线性复杂度网络S²UG-Mamba。区别于传统依赖局部感受野或全局统一注意力的范式，本文构建了不确定性门控引导增强和语义频域补偿的双重约束机制。在编码端，通过设计正交蛇形扫描双向Mamba与空间-通道双视点不确定性估计门控，在恢复二维图像拓扑的同时，有效阻断了复杂背景噪声随长程状态传播的无序扩散；在解码端，突破了常规空域上采样的平滑局限，利用深层语义先验联合路由频域滤波器，显式恢复了显著目标的精细结构与高频定位线索。多域公共数据集的综合评估证实，该架构在分布一致性与注视点命中率上取得了优于现有先进模型的性能，且交叉验证与零样本测试证明了其统计稳定性与跨域泛化能力。此外，复杂度分析表明模型实现了性能与计算效率的良好平衡。

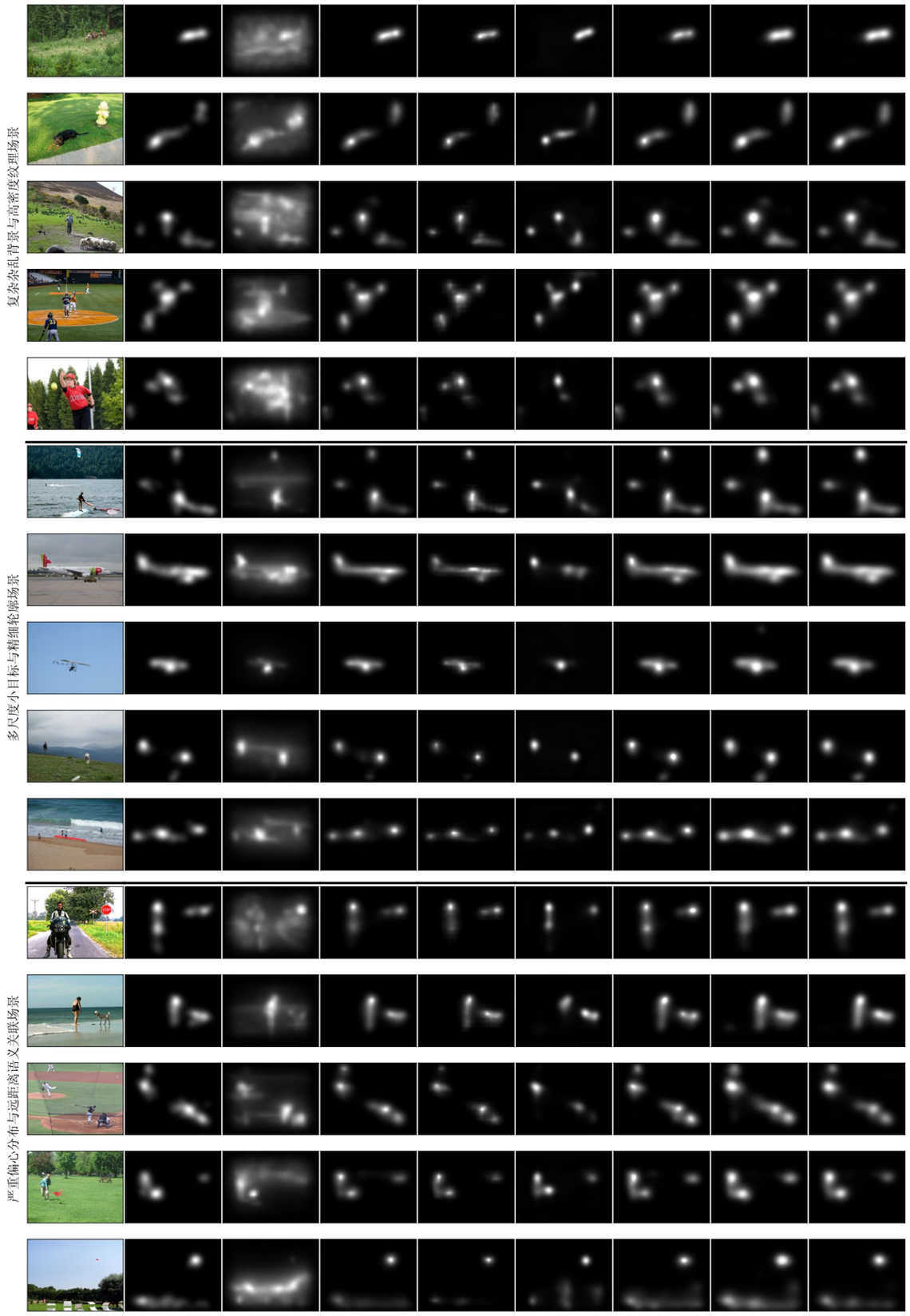
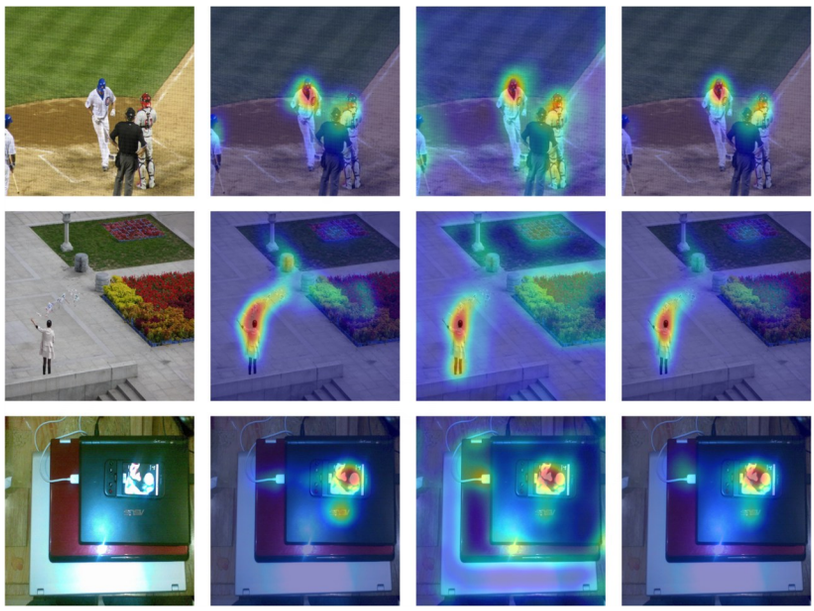


图5　不同方法在典型场景下的显著性预测结果定性对比

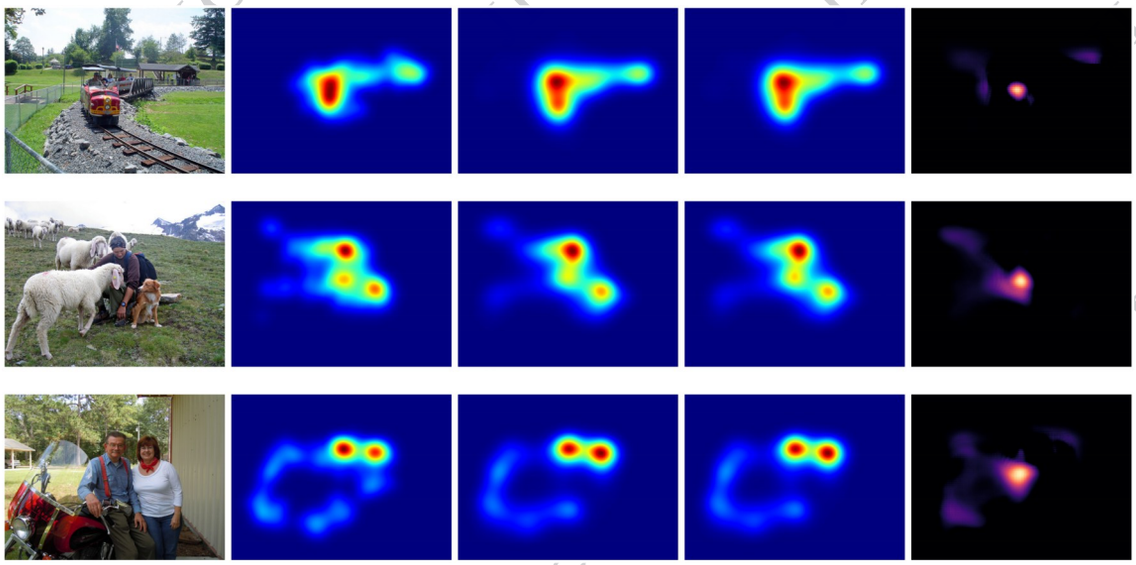
Figure 5　Qualitative comparison of saliency prediction results of different methods in typical scenes.



原图 真值图 无门控 本文模型

图 6　不确定性门控的可视化分析

Fig. 6　Visualization analysis of uncertainty-guided gating.



原图 真值图 未引入频域约束 本文模型 差异图

图 7　语义调制频域恢复的可视化分析

Figure 7　Visualization analysis of semantic-modulated frequency-domain recovery.

参考文献(References)

Aydemir B, Bhattacharjee D, Zhang T, Salzmann M and Süsstrunk S. 2024. Data augmentation via latent diffusion for saliency prediction//Proceedings of the European Conference on Computer Vision. Cham: Springer: 360-377 [DOI: 10.1007/978-3-031-73229-4_21]

Aydemir B, Hoffstetter L, Zhang T, Salzmann M and Süsstrunk S. 2023. TempSAL: uncovering temporal information for deep saliency prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 6461-6470 [DOI: 10.1109/CVPR52729.2023.00625]

Borji A and Itti L. 2015. CAT2000: a large scale fixation dataset for boosting saliency research[EB/OL]. [2026-03-28]. https://arxiv.org/abs/1505.03581

Bruce N D B and Tsotsos J K. 2006. Saliency based on information maximization//Advances in Neural Information Processing Systems.

Cambridge: MIT Press: 155-162

Bylinskii Z, Judd T, Oliva A, Torralba A and Durand F. 2019. What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3): 740-757 [DOI: 10.1109/TPAMI.2018.2815601]

Che Z, Borji A, Zhai G, Min X, Guo G and Le Callet P. 2020. How is gaze influenced by image transformations? dataset and model. IEEE Transactions on Image Processing, 29: 2287-2300 [DOI: 10.1109/TIP.2019.2945857]

Cheng D, Liu R, Li J, Liang S, Kou Q and Zhao K. 2021. Activity guided multi-scales collaboration based on scaled-con for saliency prediction. Image and Vision Computing, 114: 104250 [DOI: 10.1016/j.imavis.2021.104250]

Chi L, Jiang B and Mu Y. 2020. Fast Fourier convolution//Advances in Neural Information Processing Systems. Red Hook: Curran Associates: 4479-4488

Cornia M, Baraldi L, Serra G and Cucchiara R. 2016. A deep multi-level network for saliency prediction//Proceedings of the 23rd International Conference on Pattern Recognition. Washington, D.C.: IEEE: 3488-3493 [DOI: 10.1109/ICPR.2016.7900174]

Cornia M, Baraldi L, Serra G and Cucchiara R. 2018. Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Transactions on Image Processing, 27(10): 5142-5154 [DOI: 10.1109/TIP.2018.2851672]

Ding G, Imamoglu N, Caglayan A, Murakawa M and Nakamura R. 2021. FBNet: feedback-recursive CNN for saliency detection//Proceedings of the 17th International Conference on Machine Vision Applications. Nagoya: MVA: 1-5 [DOI: 10.23919/MVA51890.2021.9511371]

Ding G, Imamoglu N, Caglayan A, Murakawa M and Nakamura R. 2022. SalFBNet: learning pseudo-saliency distribution via feedback convolutional networks. Image and Vision Computing, 120: 104395 [DOI: 10.1016/j.imavis.2022.104395]

Droste R, Jiao J and Noble J A. 2020. UNISAL: unified image and video saliency modeling//Proceedings of the European Conference on Computer Vision. Cham: Springer: 419-435 [DOI: 10.1016/j.imavis.2022.104395]

Erdem E and Erdem A. 2013. Visual saliency estimation by nonlinearly integrating features using region covariances. Journal of Vision, 13(4): 32 [DOI: 10.1167/13.4.32]

Fan S, Shen Z, Jiang M, Koenig B L, Xu J and Kankanhalli M S. 2018. Emotional attention: a study of image sentiment and visual attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, D.C.: IEEE: 7521-7531 [DOI: 10.1109/CVPR.2018.00785]

Fang S, Li J, Tian Y, Huang T and Chen X. 2017. Learning discriminative subspaces on random contrasts for image saliency analysis. IEEE Transactions on Neural Networks and Learning Systems, 28(5): 1095-1108 [DOI: 10.1109/TNNLS.2016.2517002]

Fu R G, Li B and Gao Y H. 2016. Content-based image retrieval based on CNN and SVM//Proceedings of the 2nd IEEE International Conference on Computer and Communications. Chengdu: IEEE: 638-642 [DOI: 10.1109/CompComm.2016.7924779]

Gal Y and Ghahramani Z. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning//Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR: 1050-1059

Goferman S, Zelnik-Manor L and Tal A. 2012. Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(10): 1915-1926 [DOI: 10.1109/TPAMI.2011.272]

Gu A and Dao T. 2023. Mamba: linear-time sequence modeling with selective state spaces[EB/OL]. [2026-03-28]. https://arxiv.org/abs/2312.00752

Guo J C, Yue H H, Zhang Y, Liu D, Liu X W and Zheng S D. 2022. The analysis of image enhancement on salient object detection. Journal of Image and Graphics, 27(7): 2129-2147 (郭继昌, 岳惠惠, 张怡, 刘迪, 刘晓雯, 郑司达. 2022. 图像增强对显著性目标检测的影响研究. 中国图象图形学报, 27(7): 2129-2147) [DOI: 10.11834/jig.200735]

Harel J, Koch C and Perona P. 2006. Graph-based visual saliency//Advances in Neural Information Processing Systems. Cambridge: MIT Press: 545-552

He W and Pan C. 2022. The salient object detection based on attention-guided network. Journal of Image and Graphics, 27(4): 1176-1190 (何伟, 潘晨. 2022. 注意力引导网络的显著性目标检测. 中国图象图形学报, 27(4): 1176-1190) [DOI: 10.11834/jig.200658]

Hinton G E, Srivastava N and Krizhevsky A. 2012. Improving neural networks by preventing co-adaptation of feature detectors[EB/OL]. [2018-05-22]. https://arxiv.org/pdf/1207.0580.pdf

Hosseini A, Kazerouni A, Akhavan S, Brudno M and Taati B. 2025. SUM: Saliency Unification Through Mamba for Visual Attention Modeling//2025 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 1597-1607 [DOI: 10.1109/WACV61041.2025.00163]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Huang X, Shen C, Boix X and Zhao Q. 2015. SALICON: reducing the semantic gap in saliency prediction by adapting deep neural networks//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 262-270 [DOI: 10.1109/ICCV.2015.40]

Itti L, Koch C and Niebur E. 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern

Analysis and Machine Intelligence, 20(11): 1254-1259 [DOI: 10.1109/34.730558]

Jia S and Bruce N D B. 2020. EML-NET: an expandable multi-layer network for saliency prediction. Image and Vision Computing, 95: 103887 [DOI: 10.1016/j.imavis.2020.103887]

Jin S, Qiao D, Ge Y F, Zhang C Y, Chen J Y and Zou Y F. 2026. Perceptually diverse visual saliency prediction with global context attention. Journal of Visual Communication and Image Representation, 117: 104776[DOI:10.1016/J.JVCIR.2026.104776]

Judd T, Durand F and Torralba A. 2012. A benchmark of computational models of saliency to predict human fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(10): 1978-1993 [DOI:10.1109/TPAMI.2012.214]

Judd T, Ehinger K, Durand F and Torralba A. 2009. Learning to predict where humans look//Proceedings of the IEEE 12th International Conference on Computer Vision. Kyoto, Japan: IEEE: 2106-2113 [DOI:10.1109/ICCV.2009.5459237]

Kaibaldiyev A, Pantin J, Lechervy A, Maurel F, Chahir Y and Dias G. 2025. UNETRSal: saliency prediction with hybrid transformer-based architecture//Advanced Concepts for Intelligent Vision Systems. Cham: Springer: 389-400 [DOI: 10.1007/978-3-032-07343-3_31]

Kendall A and Gal Y. 2017. What uncertainties do we need in Bayesian deep learning for computer vision?[EB/OL]. [2026-04-08]. https://arxiv. org/abs/1703.04977 [DOI: 10.48550/arXiv. 1703.04977]

Kong L, Hu X M, Wang D, Liu Y F, Zhang Y and Chen L. 2022. Eye fixation prediction combining with multiple attention mechanism. Journal of Image and Graphics, 27(12): 3503-3515 (孔力,胡学敏,汪顶,刘艳芳,张龑,陈龙. 2022. 融合多重注意力机制的人眼注视点预测. 中国图象图形学报,27(12): 3503-3515) [DOI: 10.11834/jig.210590]

Kroner A, Senden M, Driessens K and Goebel R. 2020. Contextual encoder-decoder network for visual saliency prediction. Neural Networks, 129: 261-270 [DOI: 10.1016/j.neunet.2020.05.004]

Kruthiventi S S S, Ayush K and Babu R V. 2017. DeepFix: a fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing, 26(9): 4446-4456 [DOI: 10.1109/TIP.2017.2710620]

Kümmerer M, Theis L and Bethge M. 2014. Deep Gaze I: boosting saliency prediction with feature maps trained on ImageNet [EB/OL]. [2026-03-28]. https://arxiv.org/abs/1411.1045

Kümmerer M, Wallis T S A and Bethge M. 2017. DeepGaze II: reading fixations from deep features trained on object recognition [EB/OL]. [2026-03-28]. https://arxiv.org/abs/1610.01563

Li D, Liu Y-D, Fu X-Y, Huang J, Xu S-Y, Zhu Q, et al. 2025. FourierMamba: Fourier Learning Integration with State Space Models for Image Deraining//Proceedings of the 42nd International Conference on Machine Learning. Vancouver: PMLR, 267:35342-35365.

Li P, Xing X, Xu X, Cai B and Cheng J. 2021. Attention-aware concentrated network for saliency prediction. Neurocomputing, 429: 199-214 [DOI: 10.1016/j.neucom.2020.10.083]

Liu Z, Mao H, Wu C Y, Feichtenhofer C, Darrell T and Xie S. 2022. A ConvNet for the 2020s//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 11976-11986[DOI: 10.1109/CVPR52688.2022.01167]

Lou J, Lin H, Marshall D, Saupe D and Liu H. 2022. TranSalNet: towards perceptually relevant visual saliency prediction. Neurocomputing, 494: 455-467 [DOI: 10.1016/j.neucom.2022.04.080]

Melgani F and Bruzzone L. 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing, 42(8): 1778-1790 [DOI: 10.1109/TGRS.2004.831865]

Pan J, Canton Ferrer C, McGuinness K, O'Connor N E, Torres J, Sayrol E, et al. 2018. SalGAN: visual saliency prediction with generative adversarial networks[EB/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4781-4790[DOI:10.1109/CVPR.2018.00500]

Rao Y M, Zhao W L, Zhu Z, Lu J W and Zhou J. 2021. Global filter networks for image classification [EB/OL]//Advances in Neural Information Processing Systems. Virtual, Canada: MIT Press: [DOI:10.48550/arXiv.2107.00645]

Reddy N, Jain S, Yarlagadda P and Gandhi V. 2020. Tidying deep saliency prediction architectures//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas: IEEE: 10241-10247 [DOI: 10.1109/IROS45743.2020.9341574]

Song C C, Hu J W and Jin Q W. 2026. High-quality infrared and visible image fusion combining frequency information and atmospheric scattering model. Journal of Image and Graphics, 31(2): 573-588 (宋程程,胡辑伟,靳淇文. 2026. 融合频率信息与大气散射模型的高质量红外与可见光图像融合. 中国图象图形学报,31(2): 573-588) [DOI: 10.11834/jig.250159]

Sun Z-J, Xu S-Y, Liu K, Tian R-Z, Fu X-Y and Zha Z-J. 2025. EVDM: Event-based Real-world Video Deblurring with Mamba//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu: IEEE/CVF:13793-13803.

Tang Y, Gao P and Wang Z. 2024. SalDA: DeepConvNet greets attention for visual saliency prediction. IEEE Transactions on Cognitive and Developmental Systems, 16(1): 319-331 [DOI: 10.1109/TCDS.2023.3283286]

Tang L F, Zhang H, Xu H and Ma J Y. 2023. Deep learning-based image fusion: a survey. Journal of Image and Graphics, 28(1): 3-36 (唐霖峰,张浩,徐涵,马佳义. 2023. 基于深度学习的图像融合方法综述. 中国图象图形学报,28(1): 3-36) [DOI: 10.11834/jig.220422]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention Is All You Need[EB/OL]//Advances in Neural Information Processing Systems. Long Beach, USA: MIT Press: 5998-6008[DOI:10.48550/arXiv.1706.03762]

Vig E, Dorr M and Cox D. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 2798-2805[DOI: 10.1109/CVPR.2014.358]

Wang W and Shen J. 2018. Deep visual attention prediction. IEEE Transactions on Image Processing, 27(5): 2368-2378 [DOI: 10.1109/TIP.2017.2787612]

Wang Y, Wang R, Liu J, Xu R, Wang T, Hou F, et al. 2024. TFG-Net: frequency-guided saliency detection for complex scenes. Applied Soft Computing, 160: 111707 [DOI: 10.1016/j.asoc.2024.111707]

Wang Z Q, Zhang Y S, Yu Y, Min J and Tian H. 2022. Review of deep learning based salient object detection. Journal of Image and Graphics, 27(7): 2112-2128 (王自全,张永生,于英,闵杰,田浩. 2022. 深度学习背景下视觉显著性物体检测综述. 中国图象图形学报,27(7): 2112-2128)[DOI: 10.11834/jig.200649]

Wang Z, Liu Z, Wei W and Duan H. 2021. SalED: saliency prediction with a pithy encoder-decoder architecture sensing local and global information. Image and Vision Computing, 109: 104254 [DOI: 10.1016/j.imavis.2021.104254]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the European Conference on Computer Vision. Cham: Springer: 3-19 [DOI: 10.1109/ICAIET65052.2025.11211214]

Wu X, Li X, Song Y, Zhang R, Zhang Y and Fan J. 2023. Pyramid pooling Transformer for scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4974-4987 [DOI: 10.1109/TPAMI.2022.3181827]

Xiao J, Fan Z-H, Li D, Fu X-Y and Zha Z-J. 2025. Non-causal selective state space model for image restoration. Journal of Image and Graphics, 30(10):3173-3186 (肖杰,范子豪,李东,傅雪阳,查正军. 2025. 面向图像复原的非因果选择性状态空间模型. 中国图象图形学报,30(10):3173-3186).[DOI:10.11834/jig.240517]

Xie J, Liu Z, Li G, Lu X and Chen T. 2024. Global semantic-guided network for saliency prediction. Knowledge-Based Systems, 284: 110394 [DOI: 10.1016/j.knosys.2023.110394]

Yang F, Qian J, Guo X and Liang S. 2026. BioSalNet: biologically inspired saliency prediction. Expert Systems With Applications, 304: 130757 [DOI: 10.1016/j.eswa.2026.130757]

Yang S, Lin G, Jiang Q and Lin W. 2020. A dilated inception network for visual saliency prediction. IEEE Transactions on Multimedia, 22(8): 2163-2176 [DOI: 10.1109/TMM.2019.2947352]

Zhang J and Sclaroff S. 2013. Saliency detection: a boolean map approach//Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE: 153-160 [DOI: 10.1109/ICCV.2013.26]

Zhao X, Li Y, Wang Z, Chen H, Liu T, Sun J, et al. 2024. SalM2: lightweight Mamba-based saliency prediction for driving scenes. IEEE Transactions on Intelligent Transportation Systems, 25(10): 9521-9533 [DOI: 10.1109/TITS.2024.3392817]

作者简介

纪嘉歆,男,硕士研究生,主要研究方向为图像处理和计算机视觉。E-mail:18852183269@163.com

赵培培,通信作者,女,副教授,主要研究方向为计算机视觉、矿山物联网。E-mail:zppcumt@163.com

郭星歌,男,副教授,主要研究方向为人工智能、计算机视觉、移动通信。E-mail:guoxingge@163.com

杨发展,男,博士研究生,主要研究方向为图像处理和计算机视觉。E-mail:fazhanyang@cumt.edu.cn

王江,男,硕士研究生,主要研究方向为计算机视觉。E-mail:1981033498@qq.com

肖涛,男,工程师,主要研究方向为计算机视觉。E-mail:taoxiaocumt@163.com