Samba: 一种基于Mamba的统一通用显著目标检测框架

何家豪[1] 傅可人[1,3,*] 刘笑宏[2] 赵启军[1,3]
[1]四川大学计算机学院 [2]上海交通大学约翰·霍普克罗夫特计算机科学中心
[3]四川大学视觉合成图形图像技术重点实验室

Abstract

现有显著目标检测*(SOD)*模型主要基于卷积神经网络*(CNNs)*和*Transformer*架构。然而，*CNN*有限的感受野与*Transformer*的二次计算复杂度共同制约了现有模型在注意力目标识别方面的性能。新兴的状态空间模型*Mamba*展现出在平衡全局感受野与计算复杂度方面的潜力。为此，我们提出首个基于纯*Mamba*架构的统一框架——显著*Mamba*模型(***Samba***)，可灵活处理多种*SOD*任务，包括*RGB/RGB-D/RGB-T SOD*、视频*SOD(VSOD)*及*RGB-D VSOD*。具体而言，我们从*SOD*视角重新审视*Mamba*的扫描策略，发现保持显著区域在扫描序列中空间连续性的重要性。基于此，提出显著性引导*Mamba*块*(SGMB)*，通过空间邻域扫描*(SNS)*算法来维持显著区域的空间连续性。同时提出上下文感知上采样方法*(CAU)*，通过建模上下文依赖关系促进层级特征对齐与聚合。实验结果表明，在更低计算成本下，***Samba***在*21*个数据集上的五种*SOD*任务中均超越现有方法，证实了将*Mamba*引入*SOD*领域的优越性。代码已开源：*https://github.com/Jia-hao999/Samba*。

1. 引言

显著目标检测(Salient Object Detection, SOD)作为基础视觉任务，旨在识别并分割场景中最具视觉显著性的目标。该技术在目标跟踪[90]、语义分割[14]、图像增强[51]、自动对焦[28]以及大模型评估[29]等领域具有重要应用价值。

当前最先进的SOD方法主要基于卷积神经网络(CNN)和Transformer架构，涵盖多种SOD任务类型：RGB/RGB-D/RGB-T SOD[4, 23, 40, 62, 63, 72]、视频SOD(VSOD)[21, 80, 85]及RGB-D VSOD[36, 39, 53]。尽管基于CNN的主干网络具有可扩展性和线性计算复杂度优势，但其有限感受野难以建模全局依赖关系。而基于Transformer的主干网络虽通过全局感受野实现更优的视觉建模，但其自注意力机制导致的二次计算复杂度引发效率担忧。尽管已有Swin Transformer[44]和MobileViT[50]等高效Transformer架构被提出，但这些方法往往以牺牲部分全局建模能力为代价换取效率，未能实现二者的最优平衡。

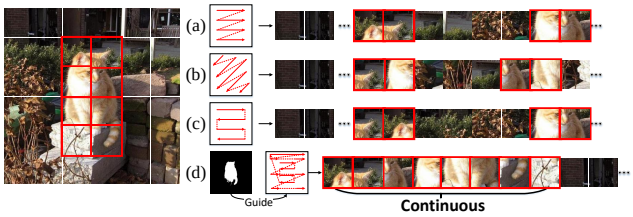

Figure 1. 现有扫描策略和我们的扫描策略的比较。(a) "Z"型序列化扫描[43]。(b) 沿对角线方向的块顺序扫描[60, 84]。(c) 以"S" 型路径的块顺序扫描[76]。(d) 与(a/b/c) 相比，我们的空间邻域扫描(SNS) 能够保留显著块的空间连续性。

近期，一种新兴的状态空间模型(SSM)，Mamba[16]，在平衡全局感受野与计算效率方面展现出巨大潜力。Mamba采用选择性扫描算法建模长程依赖关系，同时保持线性计算复杂度。配合硬件感知算法设计，Mamba在GPU上实现了高效训练。基于此，视觉Mamba主干网络[43, 92]及各类任务专用模型[6, 68, 78, 87, 91]得到快速发展。鉴于Mamba在多项视觉任务中的成功应用及其在SOD领域应用研究的缺乏，我们将探索其在SOD任务中高效全局建模的潜力。

本文提出了一个新的统一模型——显著Mamba(*Samba*)，以灵活处理通用SOD任务。得益于视觉Mamba主干网络的优异性能，越来越多的视觉模型[49, 68, 78, 91]开始利用其提取全局特征。受此启发，我们采用Vmamba [43]作为主干网络，并尝试设计基于Mamba的解码器以生成更精确的结果。通过参考现有Mamba解码器[20, 49, 68, 74, 76, 84]，我们发现了设计Mamba解码器的两个关键问题：

空间连续性问题。在视觉Mamba工作流程中，二维特征图首先被划分为图像块并扫描为一维序列输入SSM。由于SSM专为连续一维因果序列预测设计，当前图像块的预测结果高度依赖前序块(特别是邻近块)。因此，一维序列中连续的显著图像块有助于SSM准确定位完整显著目标，提升特征表达能力。然而，现有扫描策略[43, 60, 76, 84]均忽视了该问题，未能保持显著区域的空间连续性(见图 1(a)(b)(c))，导致SSM难以生成高质量特征。

特征对齐问题。现有Mamba解码器[49, 68, 91]通常

*Corresponding author: Keren Fu *(fkrsuper@scu.edu.cn)*.

采用最近邻插值对低分辨率(高层)特征上采样后与高分辨率(低层)特征融合。该方法存在两个局限：1)缺乏可学习性；2)忽略层级特征间的上下文依赖关系。这些问题导致特征融合错位，最终预测出现偏差。尽管先前工作[42, 64]提出了可学习上采样方法，但仍未建模上采样过程中的层级上下文依赖关系。

针对第一个问题，我们提出显著性引导的Mamba模块(SGMB)，通过空间邻域扫描(SNS)算法生成扫描路径，将二维特征图展开为一维序列时保持显著区域的空间连续性(图 1(d))。这些序列经SSM处理后生成高质量特征。相较于固定扫描策略(图 1(a/b/c))，SNS能动态调整扫描方向以适应不同场景，为扫描策略设计提供新思路。针对第二个问题，我们提出上下文感知的上采样(CAU)方法，通过创新的图像块配对排序机制促进解码阶段的层级特征对齐与聚合。首先，将深浅层特征块配对为子序列以建模层级上下文依赖。这些配对子序列经拼接后输入SSM，利用其强大的因果预测能力，使深层特征逐步学习浅层的特征分布，最后扩展至与浅层特征相同尺寸进行融合。为灵活处理多模态SOD任务，我们还提出多模态融合Mamba(MFM)模块探索多模态信息的交互与整合。

本文主要贡献包括：

- 首次将状态空间模型引入SOD领域，提出首个基于纯Mamba架构的统一框架，可灵活处理多种SOD任务。
- 提出显著性引导的Mamba模块(SGMB)，通过空间邻域扫描(SNS)算法保持显著区域空间连续性，增强特征表达能力。
- 提出上下文感知上采样(CAU)方法，通过建模上下文依赖促进层级特征对齐与聚合。
- *Samba*在21个数据集上的五类SOD任务中取得了最先进的结果，验证了其有效性及Mamba在SOD领域的应用潜力。

2. 相关工作

2.1. 基于深度学习的SOD方法

RGB SOD. 早期SOD研究聚焦于RGB模态，开发了边界增强[31, 83]、特征优化[73, 86]及注意力机制[72, 86]等方法。近期基于Transformer的方法[40, 48, 93]凭借强大的全局建模能力成为主流，但在复杂背景、低对比度等挑战性场景中仍存在局限。

RGB-D与RGB-T SOD. 为应对挑战性场景，研究者引入深度[3, 12, 23, 24, 32, 63, 88]或热成像[2, 4, 25, 65, 81, 82]辅助显著检测。例如Fu等人[12, 13]采用孪生网络提取RGB与深度信息的共享特征，提升复杂场景检测精度；Tu等人[65]提出双解码器架构融合RGB与热成像的多模态交互。

VSOD. 动态视频场景因运动模式多样性显著增加检测难度，因此有效利用时序上下文信息对VSOD方法至关重要。部分研究[19, 21, 80, 85]通过视频帧间交互提取运动线索，另一些方法[26, 35, 41, 59]则预计算相邻帧光流图以获取互补运动信息。

RGB-D VSOD. 文献[46]验证了深度信息融入VSOD模型的有效性，由此催生RGB-D VSOD这一新兴研究方向。随着RGB-D视频采集技术的普及，相关数据集[36, 39, 53]相继发布。同时，DCTNet+[53]和ATFNet[39]等RGB-D VSOD方法展现出令人鼓舞的检测性能。

2.2. 视觉Mamba架构

受Mamba在语言建模中成功的启发，Zhu等人[92]将其引入视觉领域，提出首个视觉Mamba主干网络(Vim)，通过双向状态空间模型(SSM)实现全局上下文建模。Liu等人[43]设计视觉状态空间块，构建新型视觉Mamba主干(Vmamba)，在语义分割[68, 91]和目标检测[6, 87]等任务中展现优异性能。针对SSM选择性扫描方向对有效感受野的影响，Zhao等人[84]在Vmamba四方向扫描基础上增加四个对角方向(见图 1(b))，实现多方向大空间特征提取。Yang等人[76]提出PlainMamba，采用连续二维扫描策略(见图 1(c))，确保扫描序列保持空间连续性。

3. 方法

3.1. 预备知识

状态空间模型(SSMs) [17, 18, 61]是基于线性时不变（LTI）系统的序列建模方法，旨在捕获长程依赖关系。该系统通过隐藏状态$h(t) \in \mathbb{R}^N$将输入序列$x(t) \in \mathbb{R}$映射至输出序列$y(t) \in \mathbb{R}$，其常微分方程（ODEs）表示为：

$$y(t) = \mathbf{C}h(t) + \mathbf{D}x(t), h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad (1)$$

其中$\mathbf{A} \in \mathbb{R}^{N\times N}$，$\mathbf{B} \in \mathbb{R}^{N\times 1}$，$\mathbf{C} \in \mathbb{R}^{1\times N}$和$\mathbf{D} \in \mathbb{R}^1$为系统参数。$h'(t)$表示$h(t)$的时间导数.

将连续系统离散化是SSM 集成至深度学习框架的关键步骤。常用零阶保持法（ZOH）[17]进行离散化：

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}, \quad (2)$$

其中$\Delta \in \mathbb{R}^D$为预设时间尺度参数。

离散化后的SSM 可表示为递归形式，将输入序列$\{x_1, x_2, ..., x_k\}$映射至输出序列$\{y_1, y_2, ..., y_k\}$：

$$y_k = \mathbf{C}h_k + \mathbf{D}x_k, h_k = \overline{\mathbf{A}}h_{k-1} + \overline{\mathbf{B}}x_k. \quad (3)$$

传统SSM 虽能以线性复杂度建模长程依赖，但其时不变性限制了动态上下文的捕捉。Mamba [16] 通过输入依赖的选择机制（S6）放松时不变约束，使系统参数$\mathbf{B}$, $\mathbf{C}$, Δ随输入动态调整，从而增强模型对长序列复杂交互的建模能力。此外，Mamba 提出并行扫描算法，支持GPU 高效训练与推理。

3.2. 总体框架

本节我们展示如图 2 所示的*Samba* 模型架构，其适用于通用SOD 任务。输入（“Input”）涵盖多种SOD 任务类型（如RGB-T SOD 等未列出的扩展任务）。编码器（“Encoder”）采用含四层视觉状态空间（VSS）[43] 的孪生主干网络，提取多级特征。转换

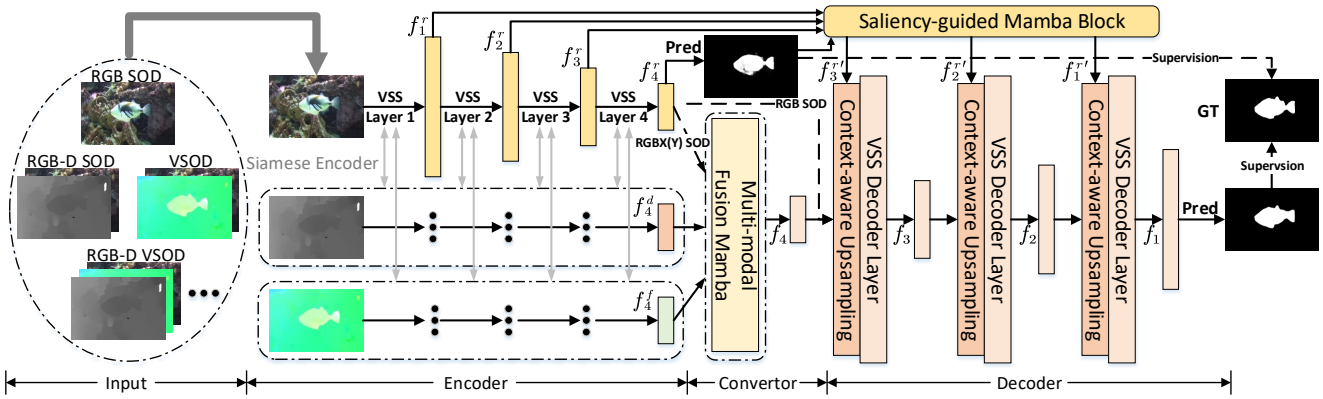

Figure 2. *Samba* 模型的总体架构图。

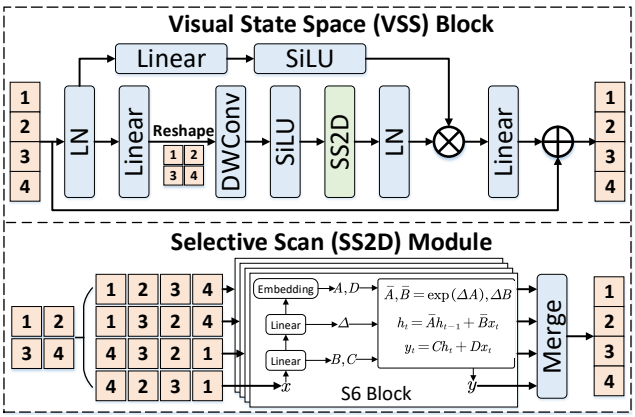

Figure 3. 视觉状态空间模块（VSS）和选择性扫描模块（SS2D）的结构图。

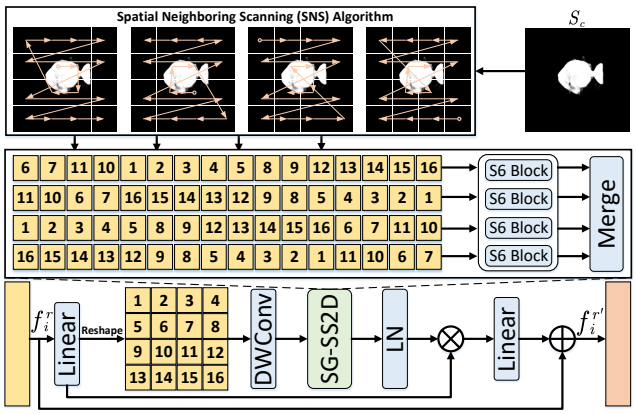

Figure 4. 显著性引导的Mamba模块的结构图。

器（“Convertor”）通过多模态融合Mamba（MFM）整合多源信息。解码器（“Decoder”）包含两大核心设计：1）显著性引导的Mamba模块（SGMB），采用新颖扫描策略保持显著区域的空间连续性；2）上下文感知上采样（CAU），通过建模层级上下文依赖促进特征对齐。下文我们将分别描述各组件的细节。

3.3. 编码器

为了处理通用SOD任务，我们基于视觉状态空间(VSS)层实现孪生编码器架构。该编码器首先将输入图像划分为图像块，随后通过四级VSS层(每级包含多个VSS块和下采样操作)提取多级特征f_i^m，其中$m \in [r, d, f, t]$分别表示RGB、深度、光流及热成像模态，$i \in [1, 2, 3, 4]$对应层级索引。VSS块及其核心组件选择性扫描模块(SS2D)的详细结构见图 3。

VSS. 输入特征首先通过层归一化(LN)处理，随后分流为两个信息分支。主分支依次执行线性投影(Linear)、维度重塑(Reshape)、深度可分离卷积(DWConv)和SiLU激活函数[7]，随后经SS2D模块建模全局依赖关系，最后通过另一LN层。旁路分支仅包含线性投影和SiLU激活。两分支特征经逐元素相乘后通过线性层输出，最终通过残差连接与原始输入相加。

SS2D. 二维输入特征沿四个方向扫描展开为四个一维序列，然后各序列经S6块[16]处理以捕获长程依赖，最后重排序后沿原方向合并序列信息。

3.4. 转换器

为实现单模态SOD(RGB SOD)向双模态(RGB-D/T SOD、VSOD)及三模态SOD(RGB-D VSOD)的灵活扩展，我们设计多模态融合Mamba(MFM)作为转换器并插入编码器-解码器之间。处理RGB SOD任务时，转换器保持空置状态，然后f_4^r直接输入解码器。双模态转换器，$f_4^r \in \mathbb{R}^{H_4 \times W_4 \times C_4}$和$f_4^x \in \mathbb{R}^{H_4 \times W_4 \times C_4}$，其中$x \in [d, f, t]$，分别经线性投影和深度可分离卷积处理，随后展平为$\mathbb{R}^{L \times C_4}$，其中$L = H_4 \times W_4$，然后沿着$L$维度拼接。为探索多模态信息交互，使用S6块处理拼接序列，最终分割为两个输出求和后经线性投影。该过程可形式化为：

$$\begin{aligned} \bar{f}_4^r &= DWConv\left(Linear\left(f_4^r\right)\right), \\ \bar{f}_4^x &= DWConv\left(Linear\left(f_4^x\right)\right), \\ \tilde{f}_4^r, \tilde{f}_4^x &= Split\left(S6\left(Cat\left(\bar{f}_4^r, \bar{f}_4^x\right)\right)\right), \\ f_4 &= Linear\left(\tilde{f}_4^r + \tilde{f}_4^x\right). \end{aligned} \tag{4}$$

Algorithm 1 SNS的核心步骤

输入：形状为(h, w)的二维粗糙显著图S_c
输出：保存所有显著图像块索引的一维数组I_s

初始化S_c当前的扫描行为第一行$(cur = 1)$, 当前行cur的扫描方向为从左到右$(dir = l \rightarrow r)$，一维数组I_s为空$(I_s = \varnothing)$.
while $cur <= h$ **do**
    根据扫描方向dir扫描当前行所有显著图像块，然后将它们的索引添加到I_s.
    计算I_s中最后一个显著图像块和下一行中最左边$(dist_{left})$和最右边$(dist_{right})$的显著图像块之间的距离
    if $dist_{left} <= dist_{right}$ **then**
        将当前行的扫描方向设置为从左到右$(l \rightarrow r)$.
    else
        将当前行的扫描方向设置为从右到左$(r \rightarrow l)$.
    end if
    跳转到下一行$(cur = cur + 1)$.
end while
返回: I_s

Figure 5. SNS生成的显著区域的扫描路径。

在两模态转换器的基础上，我们可以很自然地将其扩展到三模态转换器。

3.5. 解码器

如第 1节所述，设计基于Mamba的SOD解码器仍需解决两个关键问题。为此，我们提出显著性引导的Mamba模块(Saliency-guided Mamba Block, SGMB)与上下文感知上采样(Context-aware Upsampling, CAU)方法。

3.5.1. 显著性引导的**Mamba**模块

如图 2所示，提取的RGB特征f_i^r, 其中$i \in [1, 2, 3]$,与粗粒度显著图S_c输入SGMB以增强RGB特征。如第 1节所述，保持一维序列中显著区域的空间连续性对SSM预测至关重要。为此，我们设计空间邻域扫描(SNS)算法，在将二维特征图展平为一维序列时保持显著区域空间连续性（见图 4）。具体地，我们将显著区域空间连续性问题转化为显著区域的最短路径遍历问题，即扫描过程中下一个待扫描显著区域应与当前区域空间相邻。为降低计算复杂度并保持最短遍历路径，SNS通过近似最短路径扫描每个显著区域。SNS核心流程详见Algorithm 1。

输入为S_c，其中每个显著块被分配一个表示其在特征图中位置的索引。输出为列表I_s，存储所有显著块的索引，反映显著区域的扫描路径。SNS 从第一行开始，逐行扫描所有显著块。由于同一行内的显著块几乎连续，我们可以近似认为同一行内相邻的两个块

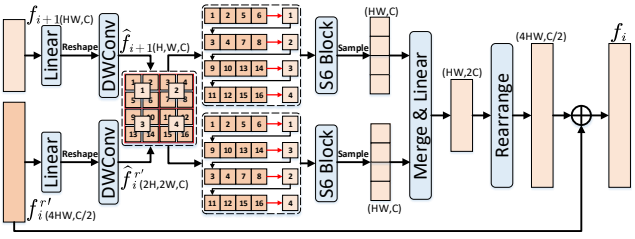

Figure 6. 上下文感知的上采样方法（CAU）的结构图。

彼此距离最近。因此，每行内的显著块可按从左到右或从右到左的方向扫描。扫描完当前行后，算法移动到下一行。为保持空间连续性并最小化计算量，我们比较当前行最后一个显著块与下一行最左和最右显著块的距离，选择距离较小的块作为下一行的起始块。重复上述步骤直至扫描完所有行，生成最终的I_s。获得I_s后，我们将所有非显著块的索引按顺序存储到列表I_{ns}，然后将其拼接在I_s之后，生成二维特征图的完整扫描路径。为增强SNS 的鲁棒性，我们通过改变方向生成三种扫描路径变体：1)将I_s拼接到I_{ns}; 2)反转I_s和I_{ns}并将I_{ns}拼接到I_s; 3)反转I_s和I_{ns}并将I_s拼接到I_{ns}后。这些扫描路径随后应用于RGB 特征，将其展平为一维序列。SGMB 的剩余步骤与VSS 模块类似。最终，输入的RGB 特征f_i^r f_i^r被增强为高质量特征$f_i^{r\prime}$.

SNS 的讨论**. (1)** 与图1 中的“S” 型路径相比，我们的SNS 需要额外计算来确定每行的扫描方向以保持空间连续性。然而在某些情况下，SNS 可能生成与“S” 型相似的显著区域扫描路径。因此，为验证SNS 的必要性，我们统计了显著区域扫描路径与“S” 型不同的图像数量，其占所有图像的比例约为∼31%。此外，图5 提供了一些可视化示例。**(2)** 值得注意的是，从另一角度看，所提出的SNS 可视为一种通过根据先验图改变块顺序来向处理后的特征注入先验图信息的新方法。这种方法可能为未来基于Mamba 架构的设计提供思路。

3.5.2. 上下文感知的上采样方法

先前的上采样方法缺乏可学习性，且无法对层级特征间的上下文依赖关系进行建模，从而导致特征融合时出现不对齐问题。为解决这一问题，我们提出了一种可学习的上下文感知上采样（CAU）方法，并将其集成到解码器中。

图∼6 展示了CAU 的结构示意图。首先，对两个输入特征$f_{i+1} \in \mathbb{R}^{HW \times C}$和$f_i^{r\prime} \in \mathbb{R}^{4HW \times C/2}$, 其中$i \in [1, 2, 3]$，分别进行线性变换、重塑和深度可分离卷积操作，得到$\hat{f}_{i+1} \in \mathbb{R}^{H \times W \times C}$ and $\hat{f}_i^{r\prime} \in \mathbb{R}^{2H \times 2W \times C}$. 为使上采样后的$\hat{f}_{i+1}$ 和$\hat{f}_i^{r\prime}$对齐，我们提出利用$\hat{f}_i^{r\prime}$的信息来指导上采样过程。由于层级特征间的块具有邻域相关性，$\hat{f}_{i+1}$中的每个块可与$\hat{f}_i^{r\prime}$中四个最相关的块建立关联。基于此，我们可以对$\hat{f}_{i+1}$和$\hat{f}_i^{r\prime}$之间的上下文依赖关系进行建模。具体而言，我们首先使用2×2的窗口对$\hat{f}_i^{r\prime}$的块进行分组。然后我们从$\hat{f}_{i+1}$逐个采样块，并将其与$\hat{f}_i^{r\prime}$的块组依次配对。接着，将配对后的子序列拼接成一个长序列，并输入到一个S6 模块中。

Table 1. 我们的*Samba* 与其他先进RGB 显著目标检测（SOD）方法在五个基准数据集上的定量对比结果如下："-" 表示结果不可用，"↑" 表示数值越大越好。最优的两个结果分别用红色 和蓝色 标注。

Method	Params (M)	MACs (G)	DUTS [71]			DUT-O [77]			HKU-IS [34]			PASCAL-S [38]			ECSSD [75]		
			$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
CNN-based																	
GateNet-R [86]	128.63	162.22	0.891	0.874	0.932	0.840	0.782	0.878	0.921	0.926	0.959	0.863	0.836	0.886	0.924	0.935	0.955
CSF-R2 [15]	36.53	18.96	0.890	0.869	0.929	0.838	0.775	0.869	-	-	-	0.863	0.839	0.885	0.931	0.942	0.960
EDN [73]	42.85	20.41	0.892	0.893	0.933	0.849	0.821	0.884	0.924	0.940	0.963	0.864	0.879	0.907	0.927	0.950	0.957
ICON-R [93]	33.09	20.91	0.890	0.876	0.931	0.845	0.799	0.884	0.920	0.931	0.960	0.862	0.844	0.888	0.928	0.943	0.960
MENet [72]	27.83	94.62	0.905	0.895	0.943	0.850	0.792	0.879	0.927	0.939	0.965	0.871	0.848	0.892	0.927	0.938	0.956
Transformer-based																	
EBM [79]	118.96	53.38	0.909	0.900	0.949	0.858	0.817	0.900	0.930	0.943	0.971	0.877	0.856	0.899	0.941	0.954	0.972
ICON-S [93]	94.30	52.59	0.917	0.911	0.960	0.869	0.830	0.906	0.936	0.947	0.974	0.885	0.860	0.903	0.941	0.954	0.971
BBRF [48]	74.40	48.60	0.908	0.905	0.951	0.855	0.820	0.898	0.935	0.946	0.936	0.871	0.884	0.925	0.939	0.957	0.972
VST-S ++ [40]	74.90	32.73	0.909	0.897	0.947	0.859	0.813	0.890	0.932	0.941	0.969	0.880	0.859	0.901	0.939	0.951	0.969
VSCode-S [47]	74.72	93.76	0.926	0.922	0.960	0.877	0.840	0.912	0.940	0.951	0.974	0.887	0.864	0.904	0.949	0.959	0.974
Samba	49.59	46.68	0.932	0.930	0.966	0.889	0.859	0.922	0.945	0.956	0.978	0.892	0.896	0.931	0.953	0.965	0.978

Table 2. 我们的*Samba* 与其他先进RGB-D 显著目标检测（SOD）方法在五个基准数据集上的定量对比结果如下（"↑" 表示数值越大越好，最优和次优结果分别用红色 和蓝色 标注）。

Method	Params (M)	MACs (G)	NJUD [30]			NLPR [56]			SIP [9]			STERE [54]			DUTLF-D [58]		
			$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
CNN-based																	
BBSNet [11]	49.77	31.20	0.921	0.919	0.949	0.931	0.918	0.961	0.879	0.884	0.922	0.908	0.903	0.942	0.882	0.870	0.912
JL-DCF [12]	143.52	211.06	0.877	0.892	0.941	0.931	0.918	0.965	0.885	0.894	0.931	0.900	0.895	0.942	0.894	0.891	0.927
SP-Net [88]	67.88	175.29	0.925	0.928	0.957	0.927	0.919	0.962	0.894	0.904	0.933	0.907	0.906	0.949	0.895	0.899	0.933
DCF [27]	53.92	108.60	0.904	0.905	0.943	0.922	0.910	0.957	0.874	0.886	0.922	0.906	0.904	0.948	0.925	0.930	0.956
SPSN [32]	-	-	0.918	0.921	0.952	0.923	0.912	0.960	0.892	0.900	0.936	0.907	0.902	0.945	-	-	-
Transformer-based																	
SwinNet-B [45]	199.18	122.20	0.920	0.924	0.956	0.941	0.936	0.974	0.911	0.927	0.950	0.919	0.918	0.956	0.918	0.920	0.949
CATNet [63]	262.73	172.06	0.932	0.937	0.960	0.938	0.934	0.971	0.910	0.928	0.951	0.920	0.922	0.958	0.952	0.958	0.975
VST-S ++ [40]	143.15	45.41	0.928	0.928	0.957	0.935	0.925	0.964	0.904	0.918	0.946	0.921	0.916	0.954	0.945	0.950	0.969
CPNet [23]	216.50	129.34	0.935	0.941	0.963	0.940	0.936	0.971	0.907	0.927	0.946	0.920	0.922	0.960	0.951	0.959	0.974
VSCode-S [47]	74.72	93.76	0.944	0.949	0.970	0.941	0.932	0.968	0.924	0.942	0.958	0.931	0.928	0.958	0.960	0.967	0.980
Samba	54.94	71.64	0.949	0.956	0.975	0.947	0.941	0.976	0.931	0.949	0.966	0.935	0.933	0.963	0.956	0.964	0.976

利用S6 模块的因果预测能力，$\hat{f}_{i+1}$中的每个块可以逐步模拟$\hat{f}_i^{r\prime}$中对应块组的特征分布。为增强这一模拟过程，我们改变配对子序列的连接顺序以生成一个新的长序列，并将其输入到另一个S6 模块中。随后，从处理后的长序列中采样属于$\hat{f}_{i+1}$的块，并恢复其原始顺序，得到两个形状为$\mathbb{R}^{HW\times C}$的特征。将它们合并后通过一个线性层处理，生成一个形状为$\mathbb{R}^{HW\times 2C}$的新特征。为重新组织其特征分布，我们使用一个重排函数将其长度扩展为原来的四倍，并将通道数缩减为原来的四分之一，从而得到一个形状为$\mathbb{R}^{4HW\times C/2}$的上采样特征。最后，将原始的$f_i^{r\prime}$与上采样特征相加进行特征聚合。

3.5.3. VSS解码器层

我们基于VSS 模块实现了VSS 解码器层，旨在对来自CAU 的聚合特征进行解码。为探索通道间依赖关系，我们在SS2D 之后引入通道注意力机制（CAM）[22]，构成了我们的VSS 解码器层。这些层的处理流程主要遵循$LN \rightarrow Linear \rightarrow DWConv \rightarrow SS2D \rightarrow CAM \rightarrow LN \rightarrow Linear$的顺序，并带有残差连接。

4. 实验

4.1. 数据集和评价指标

对于RGB 显著目标检测（SOD），我们在五个常用基准数据集上评估了*Samba*，即DUTS [71]、DUT-O [77]、HKU-IS [34]、PASCAL-S [38] 和ECSSD [75]。对于RGB-D 显著目标检测，我们使用了五个基准数据集，包括NJUD [30]、NLPR [56]、SIP [9]、STERE [54] 和DUTLF-D [58]。关于RGB-T 显著目标检测，我们采用了三个基准数据集：VT821 [69]、VT1000 [67] 和VT5000 [66]。对于视频显著目标检测（VSOD），我们使用了五个广泛使用的基准数据集：DAVIS [57]、DAVSOD-easy [10]、FBMS [55]、SegV2 [33] 和VOS [37]。在RGB-D 视频显著目标检测方面，我们考虑了三个公开数据集，包括RDVS [53]、DVisal [36] 和ViDSOD [39]。我们采用三种显著性评估指标来衡量模型性能，即结构度量(S_m) [8]、最大F度量(F_m) [1] 和最大增强对齐度量(E_m) [5]。为了评估模型的计算复杂度和模型规模，我们还报告了乘加运算量（MACs）和参数数量（Params）。

4.2. 实现细节

我们的*Samba* 基于PyTorch 框架实现，并在NVIDIA 4090 GPU 上进行训练。参照先前研究，我们为

Table 3. 我们的*Samba* 与其他视频显著目标检测（VSOD）方法在五个基准数据集上的定量对比结果。

Method	Params (M)	MACs (G)	DAVIS [57]			DAVSOD-easy [10]			FBMS [55]			SegV2 [33]			VOS [37]		
			$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
CNN-based																	
MGAN [35]	91.51	123.57	0.913	0.894	0.965	0.740	0.611	0.778	0.909	0.903	0.946	0.902	0.869	0.950	0.797	0.725	0.829
PCSA [19]	-	-	0.900	0.877	0.960	0.725	0.590	0.759	0.872	0.844	0.917	0.886	0.848	0.938	0.802	0.699	0.816
FSNet [26]	83.41	35.32	0.922	0.909	0.972	0.760	0.637	0.796	0.875	0.867	0.918	0.849	0.773	0.920	0.678	0.621	0.755
DCFNet [80]	69.56	93.27	0.914	0.899	0.970	0.729	0.612	0.781	0.883	0.853	0.910	0.903	0.870	0.953	0.838	0.773	0.861
UGPL [59]	-	-	0.911	0.895	0.968	0.732	0.602	0.771	0.897	0.884	0.939	0.867	0.828	0.938	0.751	0.685	0.811
MMNet [85]	50.81	82.63	0.911	0.895	0.968	0.732	0.602	0.771	0.897	0.884	0.939	0.867	0.828	0.938	0.751	0.685	0.811
Transformer-based																	
MGTNet [52]	150.91	265.21	0.925	0.919	0.976	0.765	0.653	0.800	0.900	0.881	0.929	0.903	0.861	0.946	0.814	0.727	0.819
UFO [21]	55.92	248.80	0.918	0.906	0.978	0.747	0.626	0.799	0.858	0.868	0.911	0.888	0.850	0.951	-	-	-
CoSTFormer [41]	-	-	0.923	0.906	0.978	0.779	0.667	0.819	0.869	0.861	0.913	0.874	0.813	0.943	0.791	0.708	0.811
VSCode-S [47]	74.72	93.76	0.936	0.922	0.973	0.800	0.710	0.835	0.905	0.902	0.939	0.946	0.937	0.984	-	-	-
Samba	54.94	71.64	0.943	0.936	0.985	0.813	0.734	0.856	0.925	0.922	0.954	0.943	0.938	0.987	0.870	0.820	0.898

Table 4. 我们的*Samba* 与其他先进RGB-T 显著目标检测（SOD）方法在三个基准数据集上的定量对比结果。

Method		CNN-based								Transformer-based		
		MIDD [65]	ECFFNet [89]	CGFNet [70]	CSRNet [25]	MGAI [62]	TNet [4]	CGMDR [2]	SPNet [82]	SwinNet-B [45]	VSCode-S [47]	*Samba*
Params (M)		52.43	-	69.92	1.01	87.09	87.04	-	104.03	199.18	74.72	54.94
MACs (G)		217.13	-	382.63	5.76	78.37	54.90	-	67.59	122.20	93.76	71.64
VT821 [69]	$S_m\uparrow$	0.871	0.877	0.881	0.885	0.891	0.899	0.894	0.913	0.904	0.926	0.934
	$F_m\uparrow$	0.847	0.835	0.866	0.855	0.870	0.885	0.872	0.900	0.877	0.910	0.927
	$E_m\uparrow$	0.916	0.911	0.920	0.920	0.933	0.936	0.932	0.949	0.937	0.954	0.965
VT1000 [67]	$S_m\uparrow$	0.916	0.924	0.923	0.919	0.929	0.929	0.931	0.941	0.938	0.952	0.953
	$F_m\uparrow$	0.904	0.919	0.923	0.901	0.921	0.921	0.927	0.943	0.933	0.947	0.956
	$E_m\uparrow$	0.956	0.959	0.959	0.952	0.965	0.965	0.966	0.975	0.974	0.981	0.983
VT5000 [66]	$S_m\uparrow$	0.868	0.876	0.883	0.868	0.884	0.895	0.896	0.914	0.912	0.925	0.928
	$F_m\uparrow$	0.834	0.850	0.852	0.821	0.846	0.864	0.877	0.905	0.885	0.900	0.919
	$E_m\uparrow$	0.919	0.922	0.926	0.912	0.930	0.936	0.939	0.954	0.944	0.959	0.963

Table 5. 我们的*Samba* 与其他先进RGB-D 视频显著目标检测（VSOD）方法在三个公开数据集上的定量对比结果。

Method		CNN-based			
		DCTNet+ [53]	DVSOD [36]	ATFNet [39]	*Samba*
Params (M)		90.69	97.34	124.07	60.28
MACs (G)		117.94	276.46	54.36	96.60
RDVS [53]	$S_m\uparrow$	0.869	0.689	0.741	0.883
	$F_m\uparrow$	0.814	0.574	0.592	0.834
	$E_m\uparrow$	0.914	0.733	0.785	0.936
DVisal [36]	$S_m\uparrow$	0.814	0.729	0.723	0.847
	$F_m\uparrow$	0.807	0.648	0.659	0.825
	$E_m\uparrow$	0.909	0.813	0.809	0.914
ViDSOD [39]	$S_m\uparrow$	0.877	0.770	0.875	0.923
	$F_m\uparrow$	0.820	0.687	0.832	0.895
	$E_m\uparrow$	0.901	0.846	0.911	0.944

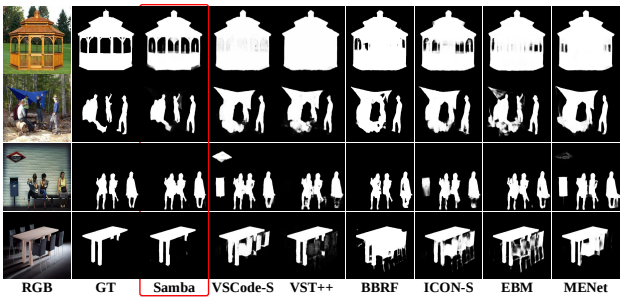

Figure 7. 与先进RGB 显著目标检测（SOD）方法的可视化对比结果。

各任务配置的训练集如下：RGB 显著目标检测（SOD）使用DUTS 训练集；RGB-D 显著目标检测使用NJUD、NLPR 和DUTLF-D 训练集；RGB-T 显著目标检测使用VT5000 训练集；视频显著目标检测（VSOD）使用DAVIS 和DAVSOD 训练集。由于RGB-D 视频显著目标检测缺乏基准训练集，我们在RDVS、DVisal 和ViDSOD 数据集上分别进行模型的训练与测试。训练过程中，我们采用AdamW 优化器，初始学习率为$1e-4$，批量大小设置为2。所有输入图像在训练和测试时均统一调整为448×448 尺寸，且在训练阶段采用随机翻转、随机裁剪和随机旋转等多种数据增强策略。模型在30 个训练周期后收敛。

4.3. 方法比较

定量评估。由于*Samba* 是用于处理通用显著目标检测（SOD）任务的统一模型，我们在五个SOD 任务上开展了*Samba* 与现有先进方法的对比实验，包括10 个RGB SOD 模型、10 个RGB-D SOD 模型、10 个VSOD 模型、10 个RGB-T SOD 模型和3 个RGB-D VSOD 模型，结果如表～1、2、3、4 和5 所示。综合实验结果表明，*Samba* 在21 个数据集上超越了现有的基于CNN 和Transformer 的先进SOD 模型，且参数数量相当、计算量（MACs）相对较低，彰显了其优越性能。具体而言，在RGB SOD 任务中，*Samba* 的参数数量和MACs 低于除VST-S++[40] 外的Transformer 类方法，同时在给定数据集上取得最优

Table 6. *Samba* 在三个RGB 显著目标检测（SOD）、三个RGB-D 显著目标检测和一个RGB-D 视频显著目标检测（VSOD）数据集上的消融实验结果如下（加粗结果为最优）。

Settings	DUTS[71]			ECSSD[75]			DUT-O[77]			NJUD[30]			NLPR[56]			DUTLF-D[58]			RDVS[53]		
	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
Ours	**0.932**	**0.930**	**0.966**	**0.889**	**0.859**	**0.922**	**0.953**	**0.965**	**0.978**	**0.949**	**0.956**	**0.975**	**0.947**	**0.941**	**0.976**	**0.956**	**0.964**	**0.976**	**0.883**	**0.834**	**0.936**
A1	0.922	0.916	0.950	0.881	0.843	0.909	0.946	0.954	0.966	0.940	0.941	0.961	0.942	0.932	0.965	0.947	0.952	0.963	0.876	0.822	0.929
A2	0.926	0.919	0.952	0.882	0.846	0.914	0.946	0.952	0.964	0.941	0.945	0.966	0.945	0.933	0.962	0.949	0.957	0.968	0.878	0.827	0.930
A3	0.929	0.924	0.961	0.884	0.852	0.917	0.950	0.961	0.973	0.945	0.949	0.970	0.944	0.937	0.974	0.949	0.959	0.971	0.880	0.829	0.931
A4	0.928	0.927	0.963	0.886	0.853	0.917	0.947	0.960	0.971	0.943	0.948	0.965	0.943	0.938	0.972	0.947	0.956	0.972	0.879	0.831	0.932
A5	0.930	0.928	0.963	0.886	0.857	0.918	0.951	0.962	0.973	0.946	0.952	0.972	0.945	0.940	0.973	0.949	0.961	0.972	0.881	0.831	0.933
A6	0.928	0.926	0.963	0.886	0.856	0.918	0.948	0.960	0.971	0.947	0.954	0.973	0.945	0.940	0.972	0.952	0.962	0.975	0.880	0.830	0.933
B1	0.927	0.923	0.960	0.887	0.856	0.912	0.952	0.961	0.976	0.941	0.952	0.966	0.942	0.939	0.972	0.953	0.957	0.969	0.879	0.828	0.932
B2	0.921	0.918	0.952	0.877	0.848	0.905	0.949	0.955	0.970	0.936	0.944	0.962	0.939	0.933	0.965	0.947	0.954	0.961	0.874	0.819	0.928
B3	0.923	0.924	0.956	0.884	0.855	0.913	0.946	0.959	0.968	0.937	0.946	0.966	0.938	0.933	0.963	0.950	0.955	0.966	0.878	0.826	0.933
B4	0.931	0.928	0.964	0.887	0.857	0.921	0.951	0.965	0.974	0.947	0.955	0.972	0.946	0.939	0.971	0.955	0.963	0.974	0.882	0.834	0.934
B5	0.929	0.926	0.964	0.888	0.854	0.920	0.950	0.964	0.976	0.948	0.953	0.971	0.944	0.936	0.973	0.955	0.962	0.971	0.879	0.832	0.931
B6	0.926	0.922	0.963	0.888	0.853	0.916	0.949	0.961	0.974	0.945	0.952	0.971	0.940	0.934	0.972	0.952	0.959	0.972	0.877	0.833	0.928
C1	0.932	0.930	0.966	0.889	0.859	0.922	0.953	0.965	0.978	0.944	0.954	0.970	0.945	0.936	0.973	0.952	0.961	0.969	0.878	0.831	0.926
C2	0.929	0.926	0.965	0.886	0.857	0.918	0.951	0.962	0.973	0.946	0.953	0.973	0.944	0.938	0.972	0.953	0.963	0.972	0.881	0.833	0.932

结果。尽管部分CNN 方法（如ICON-R[93]、EDN[73]和CSF-R2[15]）计算效率高于*Samba*，但其性能显著低于*Samba*。在RGB-D/T SOD、VSOD和RGB-D VSOD任务中，除BBSNet[11]、VST-S++[40]、FSNet[26]、SPNet[82]、TNet[4]、CSRNet[25]，大多数方法（无论基于CNN 还是Transformer）的计算复杂度均高于*Samba*，但*Samba* 在给定数据集上仍优于这些方法。

需要注意的是，VSCode-S 旨在实现通用提示方案，因此接受了所有SOD 子任务的联合训练数据，而*Samba* 仅对单个SOD 子任务进行独立训练。从这一角度看，VSCode-S 的训练数据量多于*Samba*。

质量评估。为清晰展示*Samba* 的优越性，我们在图7中展示了性能领先的RGB 显著目标检测（SOD）模型的可视化对比结果。第一行展示了一个带有中空部分的大型目标。与其他模型相比，*Samba* 能够准确检测到中空区域并生成更可靠的结果。第二行和第三行呈现了两个包含多个显著目标的场景，对比结果显示，*Samba* 能够有效定位所有显著目标，并且分割精度显著高于其他模型。最后一行展示了背景复杂且存在目标遮挡的场景，在此场景中，*Samba* 成功检测到显著目标，而其他模型则误将非显著区域识别为目标。

4.4. 消融实验

为验证*Samba* 中不同组件的相对贡献，我们通过从模型中移除或替换组件的方式开展了全面的消融实验。如表6 所示，我们在三个RGB 显著目标检测数据集（DUTS、ECSSD、DUT-O）、三个RGB-D 显著目标检测数据集（NJUD、NLPR、DUTLF-D）以及一个RGB-D 视频显著目标检测数据集（RDVS）上进行了一系列实验。

SGMB 的有效性。为验证SGMB 的有效性，我们首先删除SGMB 模块（在表～6 中记为变体“A1”），然后使用SS2D 模块替换SGMB 中的SG-SS2D 模块（记为变体“A2”）。对比结果表明，完整模型性能显著优于“A1” 和“A2”，这说明SGMB 对检测性能的提升具有重要贡献。由于SGMB 的核心思想在于用于保持显著块空间连续性的SNS 算法，为评估其作用，我们对显著区域采用三种其他扫描方式（即图～1 中的(a)、(b)和(c)），这些方式无法保留显著块的空间连续性，相应评估记为“A3”“A4” 和“A5”。与这些变体相比，我们的模型在所有评估数据集上始终表现更优，凸显了保留显著块空间连续性的重要性。在SNS 算法中，我们通过改变方向生成三种扫描路径变体以增强鲁棒性，为验证多方向路径的作用，我们使用三份初始扫描路径副本替换变体路径（记为“A6”），“Ours” 与“A6” 的对比结果证明了使用多方向路径的有效性。

CAU 的有效性。所提出的CAU 模块通过建模层级特征间的上下文依赖关系实现可学习上采样，从而促进特征对齐与聚合。为评估其效果，我们将其与三种其他上采样方法进行对比：首先使用最近邻插值替换CAU 中的上采样操作（记为变体“B1”），此外还研究了两种可学习上采样方法DUpsampling[64] 和DySample[42] 并分别用于特征上采样（记为“B2” 和“B3”）。将“B1”“B2”“B3” 与完整模型对比可知，CAU 性能显著优于所有替代方法。在CAU 模块中，新型块配对与排序方案是上采样过程的核心，为验证该设计的有效性，我们将原始配对序列偏移1 个块、3 个块和5 个块，得到变体“B4”“B5”“B6”，这些变体的性能随偏移量增加而明显下降，表明所提块配对与排序方案的有效性。

MFM 的有效性。为验证MFM 模块的作用，我们将其替换为简单拼接后接卷积的结构（记为变体“C1”），此外还使用文献[53] 中提出的RFM 模块重新实现特征转换器（记为变体“C2”）。对比结果清晰表明，MFM 在促进多模态信息交互方面具有显著优势。

5. 结论

本文首次将状态空间模型应用于显著目标检测（SOD）任务，提出基于纯Mamba架构的统一框架——显著性Mamba（*Samba*），可灵活处理通

用SOD任务。我们关注扫描序列中显著块的空间连续性，设计了显著性引导的Mamba模块（SGMB），其核心为空间邻域扫描（SNS）算法，通过动态调整扫描方向保持显著块的空间连续性。此外，提出上下文感知上采样（CAU）模块，通过建模层级特征间的上下文依赖关系促进特征对齐与聚合。大量实验表明，*Samba*在5类SOD任务、21个数据集上超越了现有先进的CNN和Transformer基模型，且计算成本更低。

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604. IEEE, 2009. 5

[2] Gang Chen, Feng Shao, Xiongli Chai, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. Cgmdrnet: Cross-guided modality difference reduction network for rgb-t salient object detection. *IEEE TCSVT*, 32(9):6308–6323, 2022. 2, 6

[3] Qian Chen, Zhenxi Zhang, Yanye Lu, Keren Fu, and Qijun Zhao. 3-d convolutional neural networks for rgb-d salient object detection and beyond. *IEEE TNNLS*, 35(3):4309–4323, 2024. 2

[4] Runmin Cong, Kepu Zhang, Chen Zhang, Feng Zheng, Yao Zhao, Qingming Huang, and Sam Kwong. Does thermal really always matter for rgb-t salient object detection? *IEEE TMM*, 25:6971–6982, 2022. 1, 2, 6, 7

[5] Yang Cao Bo Ren Ming-Ming Cheng Ali Borji Deng-Ping Fan, Cheng Gong. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018. 5

[6] Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Xuhui Liu, Juan Zhang, Guodong Guo, and Baochang Zhang. Fusion-mamba for cross-modality object detection. *arXiv preprint arXiv:2404.09146*, 2024. 1, 2

[7] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018. 3

[8] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017. 5

[9] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE TNNLS*, 32(5):2075–2089, 2020. 5

[10] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 5, 6

[11] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *ECCV*, pages 275–292. Springer, 2020. 5, 7

[12] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, pages 3052–3062, 2020. 2, 5

[13] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. *IEEE TPAMI*, 44(9):5541–5559, 2022. 2

[14] Keren Fu, Yao Jiang, Ge-Peng Ji, Tao Zhou, Qijun Zhao, and Deng-Ping Fan. Light field salient object detection: A review and benchmark. *Computational Visual Media*, 8(4):509–534, 2022. 1

[15] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In *ECCV*, pages 702–721. Springer, 2020. 5, 7

[16] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 3

[17] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2

[18] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *NeurIPS*, volume 34, pages 572–585, 2021. 2

[19] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, volume 34, pages 10869–10876, 2020. 2, 6

[20] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, pages 222–241. Springer, 2024. 1

[21] Ruohao Guo, Xianghua Ying, Yanyu Qi, and Liao Qu. Unitr: A unified transformer-based framework for co-object and multi-modal saliency detection. *IEEE TMM*, 2024. 1, 2, 6

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 5

[23] Xihang Hu, Fuming Sun, Jing Sun, Fasheng Wang, and Haojie Li. Cross-modal fusion and progressive decoding network for rgb-d salient object detection. *IJCV*, pages 1–19, 2024. 1, 2, 5

[24] Nianchang Huang, Yang Yang, Dingwen Zhang, Qiang Zhang, and Jungong Han. Employing bilinear fusion and saliency prior information for rgb-d salient object detection. *IEEE TMM*, 24:1651–1664, 2021. 2

[25] Fushuo Huo, Xuegui Zhu, Lei Zhang, Qifeng Liu, and Yu Shu. Efficient context-guided stacked refinement network for rgb-t salient object detection. *IEEE TCSVT*, 32(5):3111–3124, 2021. 2, 6, 7

[26] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *ICCV*, pages 4922–4933, 2021. 2, 6, 7

[27] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *CVPR*, pages 9471–9481, 2021. 5

[28] Yao Jiang, Xin Li, Keren Fu, and Qijun Zhao. Transformer-based light field salient object detection and its application to autofocus. *IEEE TIP*, 33:6647–6659, 2024. 1

[29] Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1):17, 2024. 1

[30] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan

Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119. IEEE, 2014. 5, 7

[31] Yun Yi Ke and Takahiro Tsubono. Recursive contour-saliency blending network for accurate salient object detection. In *WACV*, pages 2940–2950, 2022. 2

[32] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *ECCV*, pages 630–647. Springer, 2022. 2, 5

[33] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013. 5, 6

[34] Guanbin Li and Yizhou Yu. Visual saliency based on multi-scale deep features. In *CVPR*, pages 5455–5463, 2015. 5

[35] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019. 2, 6

[36] Jingjing Li, Wei Ji, Size Wang, Wenbo Li, et al. Dvsod: Rgb-d video salient object detection. In *NeurIPS*, volume 36, 2024. 1, 2, 5, 6

[37] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE TIP*, 27(1):349–364, 2017. 5, 6

[38] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 5

[39] Junhao Lin, Lei Zhu, Jiaxing Shen, Huazhu Fu, Qing Zhang, and Liansheng Wang. Vidsod-100: A new dataset and a baseline model for rgb-d video salient object detection. *IJCV*, pages 1–19, 2024. 1, 2, 5, 6

[40] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE TPAMI*, 2024. 1, 2, 5, 6, 7

[41] Nian Liu, Kepan Nan, Wangbo Zhao, Xiwen Yao, and Junwei Han. Learning complementary spatial–temporal transformer for video salient object detection. *IEEE TNNLS*, 2023. 2, 6

[42] Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *ICCV*, pages 6027–6037, 2023. 2, 7

[43] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1, 2

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[45] Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE TCSVT*, 32(7):4486–4497, 2021. 5, 6

[46] Yukang Lu, Dingyao Min, Keren Fu, and Qijun Zhao. Depth-cooperated trimodal network for video salient object detection. In *ICIP*, pages 116–120. IEEE, 2022. 2

[47] Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In *CVPR*, pages 17169–17180, 2024. 5, 6

[48] Mingcan Ma, Changqun Xia, Chenxi Xie, Xiaowu Chen, and Jia Li. Boosting broader receptive fields for salient object detection. *IEEE TIP*, 32:1026–1038, 2023. 2, 5

[49] Xianping Ma, Xiaokang Zhang, and Man-On Pun. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE GRSL*, 2024. 1

[50] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *ICLR*, 2022. 1

[51] S Mahdi H Miangoleh, Zoya Bylinskii, Eric Kee, Eli Shechtman, and Yağiz Aksoy. Realistic saliency guided image enhancement. In *CVPR*, pages 186–194, 2023. 1

[52] Dingyao Min, Chao Zhang, Yukang Lu, Keren Fu, and Qijun Zhao. Mutual-guidance transformer-embedding network for video salient object detection. *IEEE SPL*, 29:1674–1678, 2022. 6

[53] Ao Mou, Yukang Lu, Jiahao He, Dingyao Min, Keren Fu, and Qijun Zhao. Salient object detection in rgb-d videos. *IEEE TIP*, 33:6660–6675, 2024. 1, 2, 5, 6, 7

[54] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461. IEEE, 2012. 5

[55] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2013. 5, 6

[56] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *ECCV*, pages 92–109. Springer, 2014. 5, 7

[57] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 5, 6

[58] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 5, 7

[59] Yongri Piao, Chenyang Lu, Miao Zhang, and Huchuan Lu. Semi-supervised video salient object detection based on uncertainty-guided pseudo labels. In *NeurIPS*, volume 35, pages 5614–5627, 2022. 2, 6

[60] Yuan Shi, Bin Xia, Xiaoyu Jin, Xing Wang, Tianyu Zhao, Xin Xia, Xuefeng Xiao, and Wenming Yang. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, 2024. 1

[61] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 2

[62] Kechen Song, Liming Huang, Aojun Gong, and Yunhui Yan. Multiple graph affinity interactive network and a variable illumination dataset for rgbt image salient object detection. *IEEE TCSVT*, 33(7):3104–3118, 2022. 1, 6

[63] Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection. *IEEE TMM*, 2023. 1, 2, 5

[64] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In *CVPR*, pages 3126–3135, 2019. 2, 7

[65] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient

object detection. *IEEE TIP*, 30:5678–5691, 2021. 2, 6

[66] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE TMM*, 25:4163–4176, 2022. 5, 6

[67] Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang. Rgb-t image saliency detection via collaborative graph learning. *IEEE TMM*, 22(1):160–173, 2019. 5, 6

[68] Zifu Wan, Yuhao Wang, Silong Yong, Pingping Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256*, 2024. 1, 2

[69] Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *IGTA*, pages 359–369. Springer, 2018. 5, 6

[70] Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. Cgfnet: Cross-guided fusion network for rgb-t salient object detection. *IEEE TCSVT*, 32(5):2949–2961, 2021. 6

[71] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 5, 7

[72] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *CVPR*, pages 10031–10040, 2023. 1, 2, 5

[73] Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE TIP*, 31:3125–3136, 2022. 2, 5, 7

[74] Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1):37, 2024. 1

[75] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013. 5, 7

[76] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024. 1, 2

[77] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. 5, 7

[78] Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. *arXiv preprint arXiv:2403.17839*, 2024. 1

[79] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. In *NeurIPS*, volume 34, pages 15448–15463, 2021. 5

[80] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, pages 1553–1563, 2021. 1, 2, 6

[81] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. Revisiting feature fusion for rgb-t salient object detection. *IEEE TCSVT*, 31(5):1804–1818, 2020. 2

[82] Zihao Zhang, Jie Wang, and Yahong Han. Saliency prototype for rgb-d and rgb-t salient object detection. In *ACM MM*, pages 3696–3705, 2023. 2, 6, 7

[83] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019. 2

[84] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Rs-mamba for large remote sensing image dense prediction. *arXiv preprint arXiv:2404.02668*, 2024. 1, 2

[85] Xing Zhao, Haoran Liang, Peipei Li, Guodao Sun, Dongdong Zhao, Ronghua Liang, and Xiaofei He. Motion-aware memory network for fast video salient object detection. *IEEE TIP*, 2024. 1, 2, 6

[86] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51. Springer, 2020. 2, 5

[87] Minghang Zhou, Tianyu Li, Chaofan Qiao, Dongyu Xie, Guoqing Wang, Ningjuan Ruan, Lin Mei, and Yang Yang. Dmm: Disparity-guided multispectral mamba for oriented object detection in remote sensing. *arXiv preprint arXiv:2407.08132*, 2024. 1, 2

[88] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *ICCV*, pages 4681–4691, 2021. 2, 5

[89] Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. *IEEE TCSVT*, 32(3):1224–1235, 2021. 6

[90] Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, and Zhenyu He. Saliency-associated object tracking. In *ICCV*, pages 9866–9875, 2021. 1

[91] Enze Zhu, Zhan Chen, Dingkai Wang, Hanru Shi, Xiaoxuan Liu, and Lei Wang. Unetmamba: Efficient unet-like mamba for semantic segmentation of high-resolution remote sensing images. *arXiv preprint arXiv:2408.11545*, 2024. 1, 2

[92] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *ICML*, volume 235, pages 62429–62442. PMLR, 2024. 1, 2

[93] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE TPAMI*, 45(3):3738–3752, 2022. 2, 5, 7