

从序列到视觉：状态空间模型与视觉架构融合的全面综述与前沿展望

刘晓明, 刘静超

西京学院计算机学院, 陕西 西安

收稿日期: 2026年6月7日; 录用日期: 2026年7月8日; 发布日期: 2026年7月16日

摘要

状态空间模型(State Space Models, SSMs)作为控制理论和系统识别领域的经典方法, 近年来在深度学习的推动下焕发出新的生命力。特别是2023年提出的Mamba架构, 通过引入输入依赖的选择性机制和高效的并行扫描算法, 在自然语言处理领域取得了突破性进展, 挑战了Transformer的主导地位。这一成功自然地引发了计算机视觉领域的研究兴趣: 能否将这种高效的序列建模范式应用于具有空间结构的视觉数据? 本综述系统性地回顾了SSMs在计算机视觉领域的迁移、融合与应用进展。我们首先深入剖析了结构化SSMs (特别是Mamba)的核心原理与计算特性; 全面梳理了SSMs与视觉架构融合的三种主要范式: 纯SSM骨干网络、混合架构以及多模态融合模型; 最后, 我们探讨了这一领域当前面临的开放性问题, 并展望了未来可能的研究方向。本文旨在为研究者提供一个关于SSM在视觉任务中应用的全面视角, 推动更高效、更强大的视觉基础模型的发展。

关键词

状态空间模型, 选择性扫描, Mamba, 视觉Transformer, 卷积神经网络, 目标检测, 多模态学习, 高效计算

From Sequence to Vision: A Comprehensive Review and Frontier Prospects of the Integration of State Space Models and Vision Architectures

Xiaoming Liu, Jingchao Liu

School of Computer Science, Xijing University, Xi'an Shaanxi

Received: June 7, 2026; accepted: July 8, 2026; published: July 16, 2026

文章引用: 刘晓明, 刘静超. 从序列到视觉: 状态空间模型与视觉架构融合的全面综述与前沿展望[J]. 计算机科学与应用, 2026, 16(7): 59-70. DOI: 10.12677/csa.2026.167241

Abstract

State Space Models (SSMs), a classical methodology in control theory and system identification, have been revitalized by recent advances in deep learning. In particular, the Mamba architecture introduces input-dependent selective mechanisms and efficient parallel scan algorithms, achieving competitive long-sequence modeling performance while maintaining linear complexity. This review systematically surveys the migration, integration, and application of SSMs in computer vision. It analyzes the principles and computational characteristics of structured SSMs, especially Mamba; summarizes three major paradigms for integrating SSMs with vision architectures, including pure SSM backbones, hybrid architectures, and multimodal fusion models; and discusses open challenges and future research directions. The paper aims to provide researchers with a comprehensive perspective on applying SSMs to visual tasks and to promote the development of more efficient and powerful vision foundation models.

Keywords

State Space Model, Selective Scan, Mamba, Vision Transformer, Convolutional Neural Network, Object Detection, Multimodal Learning, Efficient Computation

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言：从序列建模到视觉理解

1.1. 计算机视觉架构的演变脉络

计算机视觉的发展经历了从手工特征到深度特征、从局部卷积到全局注意力、再到高效长序列建模的多次范式变化。卷积神经网络(CNNs)依靠局部连接、权重共享和平移等变性等二维归纳偏置,在图像分类、目标检测、语义分割等任务中长期占据主导地位。AlexNet 证明了深度卷积网络在大规模视觉识别中的潜力,ResNet 通过残差连接缓解深层网络退化问题,DenseNet 利用密集连接增强特征复用,EfficientNet 则通过复合缩放策略在模型深度、宽度和分辨率之间取得更优权衡[1]-[4]。

2020 年以后,Vision Transformer (ViT)将图像切分为 patch 序列,并使用标准自注意力进行全局建模,在大规模预训练条件下展示出强大表示能力[5]。随后,Swin Transformer 通过窗口注意力和层次化结构重新引入局部性,PVT 则通过金字塔式 token 表示增强密集预测任务适配性[6][7]。这些方法说明,视觉任务既需要全局依赖建模,也不能完全忽视二维空间结构。

然而,标准自注意力的时间和显存复杂度随 token 数量呈二次增长,高分辨率图像、视频、医学 3D 数据体和遥感大幅面影像会迅速放大这一瓶颈。因此,研究者开始重新关注能够以线性复杂度建模长距离依赖的序列模型。SSM,尤其是 Mamba,为视觉架构提供了 Transformer 之外的第三条路径:以递归状态和选择性扫描实现近似全局感受野,同时保留较好的计算可扩展性[8]-[10]。

1.2. 状态空间模型的复兴

在自然语言处理领域,状态空间模型(SSMs)的发展路径与视觉领域有着有趣的平行关系。S4 (结构化

状态空间序列模型)通过将连续时间系统离散化为深度序列模型,展示了在长序列建模任务上的卓越性能;其背后的 HiPPO 初始化与结构化矩阵思想进一步提升了长程记忆建模能力[8] [9]。

2023 年, Mamba 的出现标志着 SSMs 的进一步成熟[10]。其核心创新在于:

- 1) 选择性机制: 使 SSM 参数(B, C, Δ)依赖于输入, 实现基于内容的动态状态转移;
- 2) 硬件感知设计: 采用并行扫描算法, 将递归计算转化为适合 GPU 并行的高效实现;
- 3) 简化架构: 比 S4 更简单的参数化和更稳定的训练特性。

Mamba 在语言建模、DNA 序列分析和音频处理等任务中展示了与 Transformer 相当甚至更好的性能, 同时保持了序列长度的线性复杂度。

1.3. 融合的动机与意义

将 SSMs 引入计算机视觉的根本动机源于一个核心观察: 视觉信号本质上是具有强烈空间相关性的高维数据。处理这种数据需要:

- 1) 局部性: 相邻像素通常高度相关;
- 2) 层次性: 从边缘到纹理, 从部件到物体的多尺度表示;
- 3) 全局上下文: 场景理解需要整合图像全局信息;
- 4) 计算效率: 实际应用(特别是移动端和实时系统)对延迟和功耗的严格要求。

现有的架构在这些需求之间做出了不同的权衡:

CNNs: 擅长局部性和层次性, 但全局上下文建模能力有限;

ViTs: 强大的全局建模, 但计算成本高且缺乏固有的局部归纳偏置;

高效 ViT 变体: 试图平衡效率与性能, 但常引入复杂的工程实现。

SSMs, 特别是 Mamba, 提供了一个有前景的第三条道路: 线性复杂度的全局建模能力。如果能够将 SSMs 与视觉架构成功融合, 我们可能获得: 比 CNNs 更强的长程依赖建模; 比 ViTs 更高的计算效率; 比混合架构更简单的设计。

本综述的后续部分将详细探讨这一融合的现状、方法和前景。

2. 理论基础: 从连续系统到深度序列模型

2.1. 连续时间状态空间模型

经典的状态空间模型描述了一个线性时不变系统:

连续时间表示:

$$\dot{h}(t) = Ah(t) + Bx(t)$$

$$y(t) = Ch(t) + Dx(t)$$

其中:

$x(t) \in \mathbb{R}^d$: 输入信号;

$h(t) \in \mathbb{R}^n$: 隐状态;

$y(t) \in \mathbb{R}^d$: 输出信号;

$A \in \mathbb{R}^{n \times n}$: 状态转移矩阵;

$B \in \mathbb{R}^{n \times d}$: 输入矩阵;

$C \in \mathbb{R}^{d \times n}$: 输出矩阵;

$D \in \mathbb{R}^{d \times d}$: 跳跃连接(通常可选)。

2.2. 离散化与深度学习适配

为了适应数字计算和深度学习框架, 需要对连续系统进行离散化。使用零阶保持(ZOH)方法, 选取离散化步长 Δ :

离散时间表示:

$$\begin{aligned} h_k &= \bar{A}h_{k-1} + \bar{B}x_k \\ y_k &= Ch_k + Dx_k \end{aligned}$$

其中:

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\exp(\Delta A) - I)A^{-1}B \approx (\Delta A)^{-1}(\exp(\Delta A) - I)\Delta B$$

在实际实现中, 常采用更稳定的离散化方法, 如双线性变换或欧拉方法。

2.3. 选择性状态空间模型——Mamba 的核心

Mamba 的关键创新是引入选择性机制, 使参数 B 、 C 、 Δ 成为输入 x 的函数:

$$\begin{aligned} B_k &= \text{Linear}_B(x_k) \\ C_k &= \text{Linear}_C(x_k) \\ \Delta_k &= \text{Softplus}(\text{Linear}_\Delta(x_k) + \tau) \end{aligned}$$

这种选择性使得模型能够根据当前输入动态调整状态转移过程, 实现了比传统线性时不变 SSM 更强的表达能力。选择性 SSM 可以理解作为一种内容感知的循环网络, 能够在不同时间步关注不同的输入特征。

2.4. 并行扫描算法: 从递归到并行

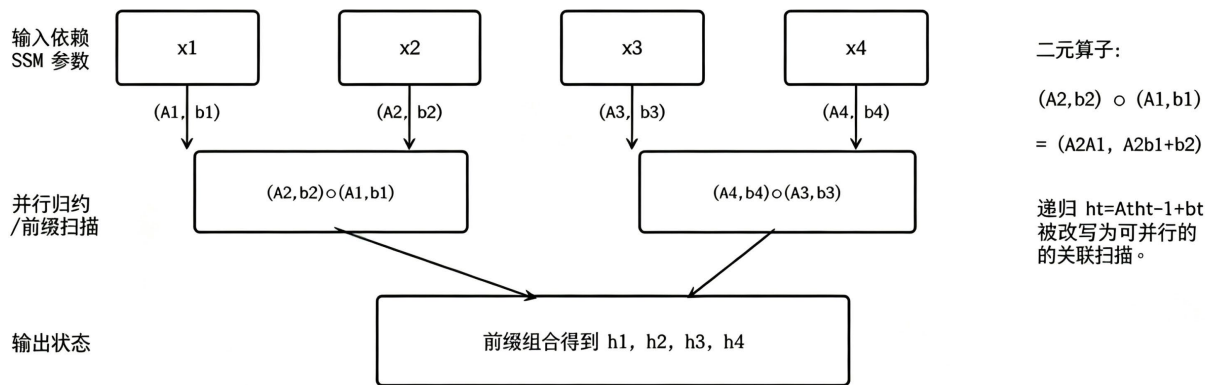


Figure 1. Simplified workflow of the parallel scan algorithm: rewriting recursive state updates as associative affine-prefix combinations

图 1. 并行扫描算法的简化流程: 将递归状态更新重写为可结合的仿射变换前缀组合

传统递归计算:

$$h_k = \bar{A}_k h_{k-1} + \bar{B}_k x_k$$

在 GPU 上顺序执行效率低下。Mamba 采用并行扫描算法(也称为前缀和算法), 将递归计算转化为适合并行硬件的操作:

对于关联扫描操作, 定义二元运算符:

$$(a_i, b_i) \circ (a_j, b_j) = (a_j a_i, a_j b_i + b_j)$$

那么状态计算可以并行化为:

$$h_k = \left(\prod_{i=1}^k \bar{A}_i \right) h_0 + \sum_{i=1}^k \left(\prod_{j=i+1}^k \bar{A}_j \right) \bar{B}_i x_i$$

这个变换正是扩展实现的核心, 使 SSM 能够在 GPU 上高效并行计算。其整体流程见图 1。

2.5. 与自注意力的理论对比

如表 1 所示, 这一理论对比解释了为什么 SSM 在长序列任务上可能比注意力机制更具优势, 特别是在需要处理极长序列或内存受限的场景中。

Table 1. Comparison between SSM and self-attention mechanisms

表 1. SSM 与自注意力机制对比

特性	自注意力(标准)	选择性 SSM (Mamba)
计算复杂度	$O(L^2 d)$	$O(L d^2)$
内存复杂度	$O(L^2)$	$O(L d)$
长程依赖	显式全连接	隐式递归传递
归纳偏置	全局交互, 置换不变	因果/序列依赖
并行性	完全并行	需要扫描算法

3. SSM 与视觉架构的融合范式

3.1. 融合策略分类

3.1.1. 纯 SSM 视觉骨干网络

纯 SSM 视觉骨干网络试图尽量减少或替代自注意力, 把图像 patch 序列主要交给 SSM 模块处理。其核心挑战是如何在一维扫描中保留二维位置关系。代表性工作包括 Vim、VMamba 与 LocalMamba。

Vim (Vision Mamba) 将图像切分为 patch 序列, 引入位置编码, 并使用双向 Mamba block 同时建模前向和后向视觉上下文。其设计动机是缓解单向序列建模对图像位置敏感性的不足, 并在 ImageNet、COCO、ADE20K 等任务上验证了高分辨率场景下的效率优势[11]。

VMamba 提出 Visual State-Space (VSS) block 与 SS2D (2D Selective Scan) 模块, 通过四个方向的扫描路径连接二维图像与一维选择性扫描。它强调全局有效感受野和输入缩放效率, 是视觉 Mamba 路线中最具代表性的层次化骨干之一[12]。

LocalMamba 认为直接展平图像会拉长相邻 patch 的序列距离, 因此引入窗口化局部扫描, 并允许不同层搜索不同扫描方向。该方法本质上是在 SSM 中重新注入局部二维依赖, 说明视觉 SSM 不能只追求全局扫描, 也需要保留局部结构[13]。

3.1.2. 混合架构: 渐进式融合策略

混合架构并不把 SSM 视为 CNN 或 Transformer 的完全替代品, 而是把它作为长程依赖建模模块嵌入成熟视觉框架中。这类方法通常更稳健, 也更容易迁移到检测、分割、医学影像等任务。

MambaVision 将 Mamba 与 Transformer 组织为层次化混合骨干, 并重新设计 Mamba mixer, 使其更适合视觉特征建模。该路线体现了“局部/层次特征 + 全局 token mixer”的折中思路, 在 ImageNet、COCO

检测和 ADE20K 分割中给出了较完整的结果[14]。

U-Mamba 面向生物医学图像分割, 设计 CNN-SSM 混合模块: 卷积负责局部纹理和边界, SSM 负责远距离依赖。对于 CT、MRI、内镜和显微图像等任务, 这种设计比单纯替换成序列模型更符合医学图像的结构特点[15]。

部分遥感分割研究通常采用 CNN 主分支与 VSS/SSM 辅助分支的双分支结构, 让卷积保留空间细节, 让 SSM 提供大范围上下文。这对遥感图像中尺度变化大、背景复杂、目标分布稀疏的问题较有针对性。

3.1.3. 多模态融合架构

多模态融合架构代表了 SSM 在视觉领域应用的最前沿方向——构建能够统一处理视觉、语言、音频等多种模态信息的模型。这一方向的核心愿景是创建真正统一的多模态理解系统。

3.1.4. 代表性模型性能对比

表 2 汇总了部分代表性视觉 SSM/Mamba 模型在公开资料中报告的主流基准结果。由于不同论文的训练策略、预训练数据、输入分辨率、吞吐测试硬件和下游框架并不完全一致, 该表更适合用于横向了解设计趋势, 而不应被理解为严格公平的排行榜。

Table 2. Performance and design comparison of representative vision SSM/Mamba models

表 2. 代表性视觉 SSM/Mamba 模型性能与设计对比

模型 Model	范式与核心设计 Paradigm & design	主要结果 Key results	参数/FLOPs Params/FLOPs	主要结论 Conclusion
Vim-S	纯 SSM; 双向 Mamba + 位置编码	ImageNet-1K Top-1 80.3%; 论文报告在高分辨率特征提取中优于 DeiT 类基线, 并给出 COCO/ADE20K 结果	26 M/5.1 G	高分辨率场景下显存与速度优势明显[11]
VMamba-S	层次化 SSM; VSS block + SS2D 四向扫描	ImageNet-1K Top-1 83.5%; 用于 COCO 检测和 ADE20K 分割	44 M/11.2 G	四向扫描弥补 1D SSM 与 2D 视觉结构之间的差距[12]
Local-Mamba-S	局部扫描 SSM; 窗口选择性扫描 + 层级方向搜索	ImageNet-1K Top-1 83.7%; 提供检测与分割代码	50 M/11.4 G	窗口扫描能改善直接展平导致的邻域破坏[13]
MambaVision-T	混合 Mamba-Transformer; 层次化骨干 + 改造 Mamba mixer	ImageNet-1K Top-1 82.3%; box mAP 51.1, mask mAP 44.3; mIoU 46.0	31.8 M/4.4 G	速度 - 精度 Pareto 更稳健[14]
MambaVision-B	混合 Mamba-Transformer; 更大规模层次化混合骨干	ImageNet-1K Top-1 84.2%; box mAP 52.8, mask mAP 45.7; mIoU 49.1	97.7 M/15.0 G	大模型下兼顾分类、检测和分割表现[14]
U-Mamba	CNN-SSM 混合; 局部卷积 + Mamba 长程依赖	非主分类模型; 在多类医学图像分割任务中优于多类 CNN/Transformer 基线	随配置变化	医学图像中局部边界与全局依赖互补[15]

3.2. 关键设计考量

3.2.1. 空间到序列的转换策略

将二维视觉特征转换为适合 SSM 处理的一维序列是 SSM 视觉应用的核心挑战。不同的转换策略在

计算效率、空间信息保留和实现复杂度方面有着显著差异。

选择空间到序列转换策略时需要考虑多个因素:

- 1) 任务需求: 分类任务可能对空间局部性要求较低, 而分割和检测任务需要精确的空间关系保持。
- 2) 计算预算: 复杂的扫描策略(如希尔伯特曲线)可能带来可观的额外计算开销。
- 3) 实现复杂度: 简单的扫描策略更容易集成到现有框架中, 也更容易进行硬件优化。
- 4) 扩展性: 策略是否容易扩展到三维(视频)或更高维数据。

3.2.2. 空间到序列扫描策略的比较

扫描策略不是无关紧要的工程细节, 而是视觉 SSM 是否有效的关键结构假设, 如表 3 所示。光栅扫描最简单但可能破坏二维邻域, 双向扫描改善因果性但仍受路径限制, 四向扫描增强方向覆盖, 窗口扫描则显式保护局部结构, 如图 2 所示。未来更有潜力的方向可能是“动态扫描”: 根据图像内容、层级深度或任务需求自适应选择扫描路径, 而不是固定一种全局路径。

Table 3. Comparison of space-to-sequence scanning strategies

表 3. 空间到序列扫描策略比较

扫描策略	基本思想	优点	局限	代表性关联工作
光栅/展平扫描	按行或按列把 $H \times W$ patch 展成 1D 序列	实现简单、内存连续、硬件友好	跨行邻居在序列中可能距离较远, 二维局部性被破坏	早期 ViT 展平思想、Vim 的 patch 序列处理 [5] [11]
双向扫描	同时进行正向和反向序列建模	缓解单向因果偏置, 获得前后文信息	仍主要沿单一展平路径, 二维方向覆盖有限	Vim 的 bidirectional Mamba block [11]
四向/交叉扫描	沿左-右、右-左、上-下、下-上多路径扫描	更好覆盖二维上下文, 增强全局有效感受野	多路径增加实现和融合成本	VMamba 的 SS2D/Cross-Scan [12]
窗口局部扫描	先划分窗口, 在窗口内扫描, 再跨窗口融合	保留邻域关系, 适合局部纹理和边界	窗口大小和跨窗口交互需额外设计	LocalMamba 的 windowed selective scan [13]
空间填充曲线扫描	用 Hilbert/Z-order 等曲线尽量保持邻近关系	理论上更好保持空间邻近性	实现复杂, GPU 连续访问和工程优化较难	可作为未来视觉 SSM 扫描策略探索方向

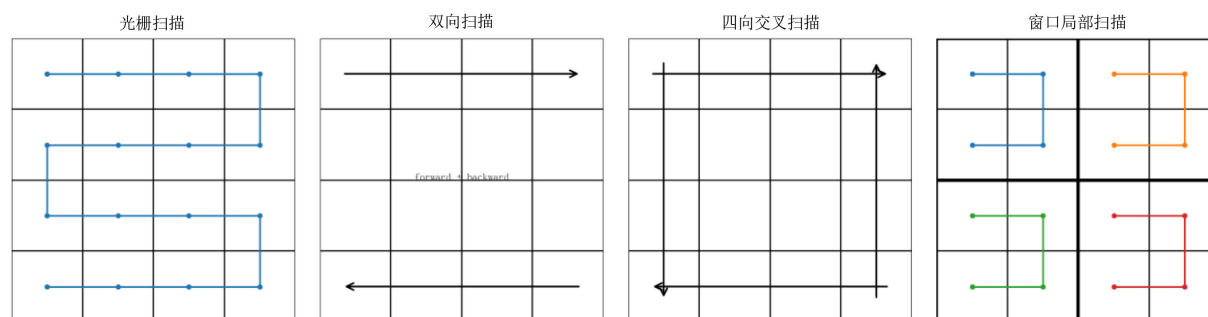


Figure 2. Schematic illustration of common space-to-sequence scanning strategies: different paths determine how SSMs “read” two-dimensional images

图 2. 常见空间到序列扫描策略示意图: 不同路径决定了 SSM 如何“阅读”二维图像

3.2.3. 选择性机制在视觉任务中的特化设计

选择性是 Mamba 架构的核心创新,但在视觉任务中需要重新思考其设计和实现方式,以更好地适应图像数据的特性。

选择性机制在视觉任务中的设计是一个丰富而复杂的研究领域,需要在表达能力和计算效率之间、在通用性和任务特异性之间、在理论优雅性和工程实用性之间找到恰当的平衡。随着研究的深入,我们预期将看到更多针对视觉数据特性优化的选择性设计,进一步释放 SSM 在视觉任务中的潜力。

4. 挑战与限制

4.1. 技术实现层面的核心障碍

4.1.1. 计算图优化的理论空白

选择性 SSM 在视觉任务中的高效实现面临基础性理论挑战。与 CNN 的规则计算模式相比,SSM 的选择性机制创建了数据依赖的计算图,这种动态性导致:

1) 静态优化与动态执行的矛盾:现代深度学习编译器(如 TVM、XLA)依赖静态计算图进行优化。SSM 的选择性路径在每次前向传播时都可能变化,破坏了传统编译优化的前提假设。这使得自动算子融合、内存重用等关键技术难以应用。

2) 内存访问模式的随机性:基于输入内容的扫描顺序导致不可预测的内存访问模式。在现代 GPU 架构中,合并内存访问对性能至关重要。随机访问不仅降低缓存命中率,还可能触发代价高昂的全局内存访问,严重抵消线性复杂度的理论优势。

3) 并行扫描的细粒度调度:虽然并行扫描算法提供了理论上的并行性,但实际实现中的调度开销可能很大。特别是在处理非均匀选择性的情况下,工作负载难以均衡分配,导致 GPU 流处理器利用率低下。

4.1.2. 二维递归的硬件不友好性

将一维 SSM 扩展到二维视觉空间在硬件实现上面临固有困难:

1) 数据局部性的双重损失:二维图像同时具有行局部性和列局部性,任何扫描策略都会至少损失一个方向的局部性。硬件设计(如缓存行大小、内存预取策略)通常假设访问模式具有一定连续性,二维扫描破坏了这种连续性。

2) 状态管理的维度灾难:真正的二维递归需要维护二维状态网格,而不是一维状态向量。这不仅需要平方级的内存增长,还使状态更新模式复杂化,难以利用现有的张量核心优化。

3) 跨处理单元的同步开销:在分布式训练或大模型推理中,SSM 的有状态特性需要频繁的状态同步。这引入了额外的通信开销,特别是在处理高分辨率图像时可能成为瓶颈。

最新的空间 SSM 变体(如 Vim、Vision Mamba)尝试通过分解方法近似二维递归,但这只是权宜之计。长远来看,可能需要全新的硬件架构支持,例如设计专门的 SSM 处理单元,或开发支持高效二维状态传递的内存系统。

4.2. 训练动态的深层次难题

4.2.1. 选择性机制的优化曲面复杂性

SSM 的选择性参数创建了高度非凸的优化曲面:

1) 参数耦合的强度效应: Δ (离散化步长)、 B (输入矩阵)、 C (输出矩阵)的选择性相互耦合,形成复杂的优化地形。小的参数扰动可能导致状态转移行为的剧烈变化,这类似于混沌系统中的敏感依赖现象。

2) 模式坍塌风险:在训练过程中,选择性机制可能坍塌到简单模式,如对所有输入采用相同的选择

策略。这虽然稳定但失去了选择性的优势。防止模式坍塌需要精心设计的正则化和初始化策略。

3) 长尾分布的优化困难: 真实世界视觉数据的分布通常具有长尾特性。SSM 可能过度适应头部类别, 而在尾部类别上表现不佳, 因为选择性机制可能倾向于将有限的计算资源分配给频繁出现的模式。

理论创新需求: 需要发展新的优化理论来处理选择性系统的独特特性。控制理论中的自适应系统理论可能提供启示, 但也需要针对深度学习的特定环境进行调整。元学习或课程学习框架可能帮助 SSM 学习在不同数据分布下的选择性策略。

4.2.2. 梯度流动的路径依赖性

SSM 的梯度计算具有独特的时间依赖性:

1) 梯度传播的时序敏感度: 由于 SSM 的状态传递是时序的, 早期时间步的梯度信息必须通过长链传播。这增加了数值不稳定的风险, 特别是在处理长序列时。

2) 选择性的梯度截断效应: 选择性机制本质上是一个门控系统, 可能导致梯度传播路径的动态变化。某些路径可能在某些时间步被完全阻断, 导致对应参数的梯度为零, 即使它们在长期任务中可能很重要。

3) 二阶优化器的适应困难: Adam 等自适应优化器基于梯度的一阶和二阶矩估计调整学习率。SSM 动态变化的梯度统计使这些估计变得不可靠, 可能导致学习率调整不当。

算法改进前景: 需要设计专门针对 SSM 的优化算法。时间感知的梯度裁剪、选择性路径的梯度补偿、以及考虑时间依赖性的学习率调度都是值得探索的方向。此外, 从强化学习借鉴的 credit assignment 技术可能帮助改善梯度传播。

4.3. 模型鲁棒性与安全性隐患

4.3.1. 对抗性脆弱性的结构性根源

SSM 的对抗鲁棒性问题有深刻的架构根源。对抗攻击者可能通过微调像素序列顺序而不改变像素值, 就能显著改变 SSM 的输出。这种顺序敏感性在卷积模型中不存在, 因此针对卷积设计的防御策略对 SSM 可能无效; 由于选择性参数直接依赖于输入, 攻击者可能精心构造输入以操纵选择过程。例如, 使模型在关键区域应用弱选择性, 从而忽略重要特征;

SSM 维护的内部状态可能被精心设计的输入污染, 这种污染可能持续影响后续多个时间步的预测, 即使在恶意输入之后恢复正常输入。

需要专门研究 SSM 对抗攻击的新形式和新防御方法。动态检测机制、状态净化技术、以及对抗性训练的特殊变体都是必要的安全措施。在部署 SSM 系统前, 必须进行全面的审计。

4.3.2. 分布外泛化的理论不确定性

SSM 在新领域数据上的行为难以预测。

1) 选择策略的领域特异性: 在一个领域学习的选择性策略可能在另一个领域表现不佳。例如, 在自然图像上学习的基于纹理的选择性策略, 在医学图像上可能完全失效。

2) 状态空间的域间迁移: SSM 的状态表示在不同领域之间可能不具备可迁移性。这限制了少样本学习和领域适应的效果。

3) 极端值处理的脆弱性: 由于选择性机制依赖于输入统计, 遇到训练分布之外的极端值时, 可能产生不可预测的行为。

4) 泛化理论缺口: 需要建立 SSM 的泛化理论, 特别是理解选择性机制如何影响泛化边界。领域自适应、测试时适应、以及元学习都可能为改善 SSM 的分布性能提供工具。

4.4. SSM 应用于空间数据的固有理论局限性

这是视觉 SSM 最需要平衡讨论的问题。SSM 本质上是序列模型，它对“顺序”非常敏感；而静态图像本质上是二维拓扑结构，邻近关系并不只存在于一条路径上。把二维图像展平为一维序列，不可避免地引入人为顺序：某些原本相邻的 patch 在序列中可能相距很远，某些在二维空间并不相邻的 patch 反而可能在扫描路径上相邻。

与之相比，CNN 的二维归纳偏置更直接。卷积核天然在局部二维邻域内共享权重，平移等变性使模型不需要从数据中重新学习“相邻像素更相关”这一常识；池化和层次结构又逐步扩大感受野。也就是说，CNN 把视觉空间结构写进了模型结构本身，而 SSM 需要通过扫描路径、位置编码、多方向融合或局部窗口来间接恢复这些结构。

Transformer 的自注意力虽然缺少 CNN 那样强的局部归纳偏置，但它能显式建立任意两个 token 的交互关系；SSM 则通过有限维隐状态压缩历史信息，长程依赖是隐式传递的。这意味着 SSM 在效率上更优，但可能在细粒度空间对应、目标边界定位和多实例关系建模上不如显式注意力直观。

因此，视觉 SSM 的理论局限可概括为三点：第一，一维序列偏置与二维空间拓扑之间存在结构错配；第二，有限状态维度可能形成信息瓶颈，难以无损保留所有空间关系；第三，扫描路径选择本身是一种人为归纳偏置，不同任务可能需要不同路径。更合理的定位是：SSM 并不是 CNN/Transformer 的简单替代，而是适合长序列、高分辨率和需要全局上下文的补充模块。

4.5. 可解释性问题与潜在研究方向

视觉 SSM 的可解释性目前仍明显弱于 CNN 的特征图可视化和 Transformer 的注意力图分析。CNN 可以通过卷积核、激活图、Grad-CAM 等方式解释局部响应；Transformer 至少可以通过注意力权重近似观察 token 之间的显式交互。SSM 的状态传播则更隐式：信息被压缩进隐状态，并经过输入依赖的选择性门控连续更新，很难直接回答“模型在某一步到底记住了什么、遗忘了什么”。

当前困难主要包括：第一，隐状态维度没有明确语义，不同通道可能混合位置、纹理、类别和上下文信息；第二，选择性参数 B 、 C 、 Δ 随输入变化，解释对象从静态权重变为动态轨迹；第三，多方向扫描和窗口扫描会产生多条状态路径，路径间融合后的贡献难以分解；第四，下游任务中 SSM 常与 CNN、MLP、注意力或解码器混合，单独归因更加困难。

潜在研究方向包括：构建状态轨迹可视化方法，显示不同空间位置对隐状态的写入、保留和读取强度；设计 scan attribution，将最终预测分解到不同扫描方向或窗口；分析 Δ 、 B 、 C 的分布与图像语义区域之间的关系；将可解释性约束纳入训练，例如鼓励边界区域、小目标区域或文本相关区域产生可解释的选择性响应。对于医学、遥感和自动驾驶等高风险任务，可解释性不是锦上添花，而是模型可信部署的重要前提。

5. 结论与展望

5.1. 技术演进的宏观图景

回顾计算机视觉的发展历程，我们经历了从手工特征到学习表示、从局部处理到全局理解、从静态分析到动态推理的演变。在这一宏观图景中，SSM 代表了一个重要的转折点。

传统视觉系统主要关注空间维度，时间被视为独立的维度。SSM 提供了一个统一的框架，将空间和时间都视为序列维度，这种视角的转变可能从根本上改变我们处理视觉信息的方式。

CNN 使用静态的卷积核，Transformer 使用静态的注意力模式，而 SSM 引入基于输入的动态选择性。

这种从静态到动态的转变,使模型能够根据具体情境调整计算策略,更接近生物智能的灵活性。

传统深度学习模型通常是无状态的,每次推理都是独立的。SSM 的有状态特性使其能够保持历史信息,支持连续的、交互式的智能行为。

5.2. SSM 的技术定位再思考

SSM 不应被视为 CNN 或 Transformer 的替代品,而是视觉架构工具箱中的重要补充。

1) 与 CNN 的互补关系: CNN 擅长捕获局部模式和空间结构,SSM 擅长建模长程依赖和时间动态。最优的视觉系统可能需要同时包含两者, CNN 提供基础的感知能力,SSM 提供高层次的推理能力。

2) 与 Transformer 的对比优势: 在处理极长序列时,SSM 的计算效率优势明显;在处理需要精确位置信息时,Transformer 的注意力机制可能更合适。任务特性和资源约束决定哪种架构更合适。

3) 混合智能的必然性: 未来的视觉系统很可能是多种架构的智能混合,根据输入特性和任务需求动态选择合适的计算范式。SSM 为这种混合智能提供了新的组件选项。

SSM 不仅提供了新解决方案,也提出了新问题: SSM 的动态计算图对传统的计算理论提出了挑战,需要发展处理动态性和状态的新理论框架;SSM 揭示了现有深度学习硬件在支持动态计算方面的不足,推动下一代 AI 芯片设计;SSM 的选择性学习机制提出了新的机器学习理论问题,如动态架构的学习理论、状态保持系统的泛化理论等。

5.3. 研究范式的转变

SSM 可能推动计算机视觉研究范式的几个重要转变。

1) 从架构设计到架构学习: 传统上,神经网络架构主要由人类设计。SSM 的选择性机制使部分架构决策可以学习得到,这可能导致从架构设计向架构学习的范式转变。

2) 从静态优化到动态适应: 传统模型优化主要关注静态性能指标。SSM 的动态特性要求我们同时考虑计算效率和适应性,优化问题变得更加复杂和动态。

3) 从孤立任务到连续智能: 大多数视觉研究关注孤立任务(如单张图像分类)。SSM 的状态保持能力鼓励我们研究连续智能系统,能够持续学习、适应和交互。

5.4. 结语

SSM 在视觉领域的旅程充满了挑战——从理论基础的薄弱到实现困难的艰巨,从训练不稳定的困扰到部署复杂性的障碍。然而,正是这些挑战定义了研究的边界,激发了创新的动力。

每一次技术范式的转变都会经历类似的艰难时期。卷积神经网络在 21 世纪初被视为不切实际,Transformer 在 2017 年刚提出时也面临诸多质疑。今天,它们都已成为深度学习的基础组件。SSM 可能正在经历类似的早期发展阶段。

我们站在一个充满可能性的历史节点上。SSM 为我们提供了重新思考视觉智能本质的机会,挑战我们对计算、学习和理解的基本假设。无论 SSM 最终成为主流技术还是被更新的架构超越,这一探索过程都将丰富我们的认知,推动整个领域向前发展。

最终,技术的历史不是线性进步的历史,而是多样探索的历史。SSM 代表了多样性中的重要一支——它提醒我们,在追求视觉智能的道路上,还有无数未被探索的可能性和未被想象的解决方案。

让我们以开放的心态拥抱这种多样性,以严谨的态度面对挑战,以创新的精神探索未知。在 SSM 和其他新兴技术的共同推动下,我们可能正迈向一个视觉智能的新时代——一个更加高效、更加灵活、更加理解我们世界的时代。

参考文献

- [1] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90.
- [2] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 27-30 June 2016, 770-778.
<https://doi.org/10.1109/cvpr.2016.90>
- [3] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, 21-26 July 2017, 2261-2269.
<https://doi.org/10.1109/cvpr.2017.243>
- [4] Tan, M. and Le, Q.V. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, 9-15 June 2019, 6105-6114.
- [5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.
- [6] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, 10-17 October 2021, 9992-10002. <https://doi.org/10.1109/iccv48922.2021.00986>
- [7] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., *et al.* (2021) Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, 10-17 October 2021, 548-558. <https://doi.org/10.1109/iccv48922.2021.00061>
- [8] Gu, A., Dao, T., Ermon, S., Rudra, A. and Ré, C. (2020) HiPPO: Recurrent Memory with Optimal Polynomial Projections. *Advances in Neural Information Processing Systems*, **33**, 1474-1487.
- [9] Gu, A., Goel, K. and Ré, C. (2022) Efficiently Modeling Long Sequences with Structured State Spaces. arXiv: 2111.00396.
- [10] Gu, A. and Dao, T. (2023) Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv: 2312.00752.
- [11] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W. and Wang, X. (2024) Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vienna, 21-27 July 2024, 62429-62442.
- [12] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., *et al.* (2024) VMamba: Visual State Space Model. *Advances in Neural Information Processing Systems* 37, Vancouver, 10-15 December 2024, 103031-103063.
<https://doi.org/10.52202/079017-3273>
- [13] Huang, T., Pei, X., You, S., Wang, F., Qian, C. and Xu, C. (2025) LocalMamba: Visual State Space Model with Windowed Selective Scan. In: Del Bue, A., Canton, C., Pont-Tuset, J. and Tommasi, T., Eds., *Computer Vision—ECCV 2024 Workshops*, Springer, 12-22. https://doi.org/10.1007/978-3-031-91979-4_2
- [14] Hatamizadeh, A. and Kautz, J. (2025) Mambavision: A Hybrid Mamba-Transformer Vision Backbone. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, 10-17 June 2025, 25261-25270.
<https://doi.org/10.1109/cvpr52734.2025.02352>
- [15] Ma, J., Li, F. and Wang, B. (2024) U-Mamba: Enhancing Long-Range Dependency for Biomedical Image Segmentation. arXiv: 2401.04722.