
Transformer“一统江湖”
• 在各种理解、生成任务上性能领先
FocalNet-H (DINO)
Co-DETR: 65.900
Cascade Mask R-CNN (ResNet-50)
Box AP
2016
2018
2020
2022
2024
OpenAI
ChatGPT
OpenAI
Sora
Text-to-Video AI
图像理解
文本生成
视频生成
• why?
强大的建模能力
序列化处理、多模统一
全局视野、灵活位置编码
模型容量大、可规模化强
并行化强训练快
VALSE 2024 APR：面向大模型的新型高效率网络架构
2

Transformer的局限性

- 计算和存储复杂度为O(n^2)，随着序列长度n逞二次方增长
- 在多模态大模型、视频生成、具身基础模型等场景下都存在长序列的问题

黄仁勋和Transformer八子

"I think the world needs something better than the transformer," said Gomez. "I think all of us here hope it gets succeeded by something that will carry us to a new plateau of performance."

NLP: RWKV

- 核心思路：在Attention Free Transformer（AFT）的成对位置偏置矩阵中引入时序衰退机制，能够在训练时和Transformer一样并行，推理的时候可以等价为RNN。
- 关键方法：引入门控机制调制对历史信息的接收，引入Token shift 技术加强对相邻局部特征的融合。

$$\mathrm{AFT}(W,K,V)_t = \frac{\sum_{i=1}^{t} e^{w_{t,i}+k_i} \odot v_i}{\sum_{i=1}^{t} e^{w_{t,i}+k_i}}$$

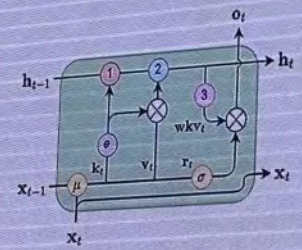
↓ 引入时序衰退机制

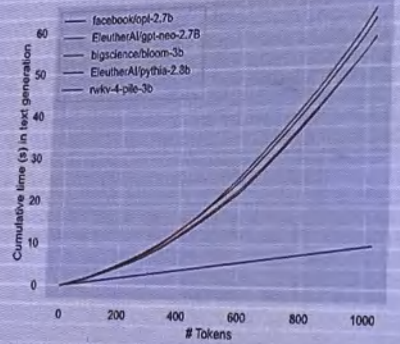
$$w_{t,i} = -(t-i)w$$

$$wkv_t = \frac{\sum_{i=1}^{t-1} e^{-(t-1-i)w+k_i} \odot v_i + e^{u+k_t} \odot v_t}{\sum_{i=1}^{t-1} e^{-(t-1-i)w+k_i} + e^{u+k_t}}$$

↓ 引入时序门控机制

RWKV: $o_t = W_o \cdot (\sigma(r_t) \odot wkv_t)$





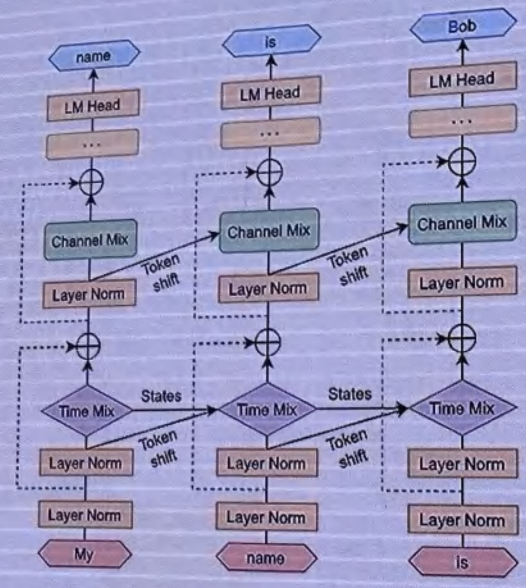
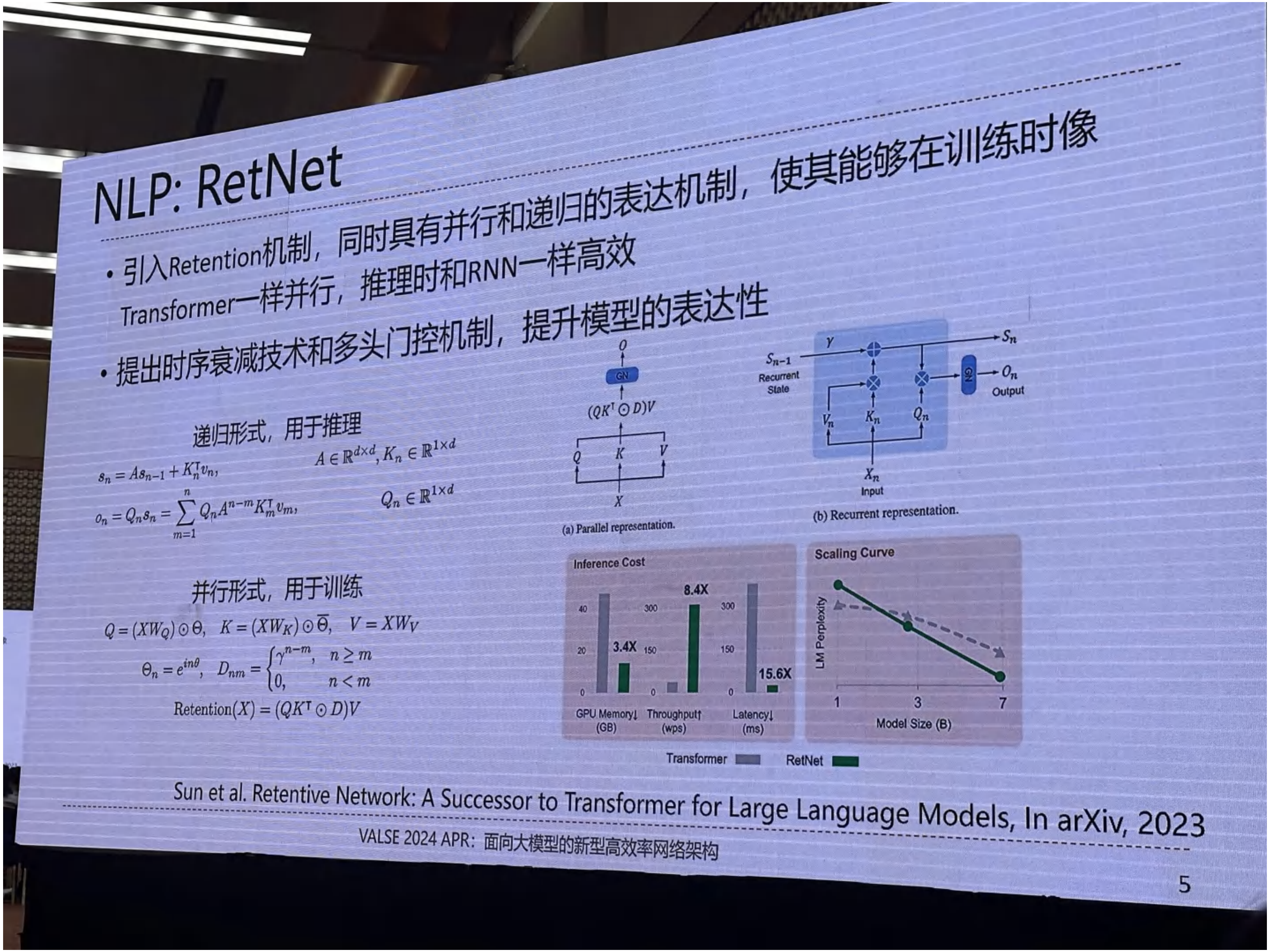


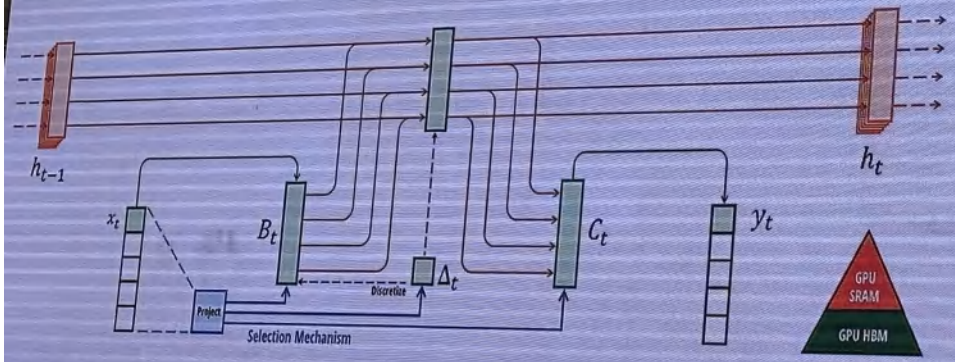
Figure 3: RWKV architecture for language modeling.

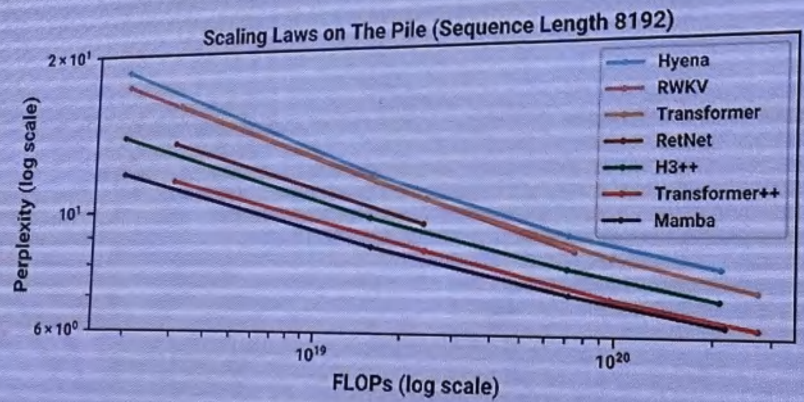
Peng et al. RWKV: Reinventing RNNs for the Transformer Era, In arXiv, 2023


NLP: RetNet
• 引入Retention机制，同时具有并行和递归的表达机制，使其能够在训练时像Transformer一样并行，推理时和RNN一样高效
• 提出时序衰减技术和多头门控机制，提升模型的表达性
递归形式，用于推理
$s_n = As_{n-1} + K_n^\intercal v_n,$
$A \in \mathbb{R}^{d\times d}, K_n \in \mathbb{R}^{1\times d}$
$o_n = Q_n s_n = \sum_{m=1}^{n} Q_n A^{n-m} K_m^\intercal v_m,$
$Q_n \in \mathbb{R}^{1\times d}$
并行形式，用于训练
$Q = (XW_Q) \odot \Theta, \quad K = (XW_K) \odot \overline{\Theta}, \quad V = XW_V$
$\Theta_n = e^{in\theta}, \quad D_{nm} = \begin{cases} \gamma^{n-m}, & n \ge m \\ 0, & n < m \end{cases}$
$\mathrm{Retention}(X) = (QK^\intercal \odot D)V$
O
GN
$(QK^\intercal \odot D)V$
Q K V
X
(a) Parallel representation.
S_{n-1}
Recurrent State
γ
S_n
O_n
Output
GN
V_n K_n Q_n
X_n
Input
(b) Recurrent representation.
Inference Cost
40 20 0
300 150 0
300 150 0
3.4X
8.4X
15.6X
GPU Memory↓ (GB)
Throughput↑ (wps)
Latency↓ (ms)
Scaling Curve
LM Perplexity
1 3 7
Model Size (B)
Transformer
RetNet
Sun et al. Retentive Network: A Successor to Transformer for Large Language Models, In arXiv, 2023
VALSE 2024 APR：面向大模型的新型高效率网络架构
5

NLP: Mamba

- 使用状态空间模型完成次二次方的序列建模任务
- 提出了 Mamba 模块：输入相关建模，次二次方复杂度，硬件感知算法
- 在 NLP 任务中取得了比 Transformer++ 更好的 efficiency 和 effectiveness





Gu et al. Mamba: Linear-time sequence modeling with selective state spaces, In arXiv, 2023

Transformer vs 新型高效网络架构

Architecture	Inference Time	Inference Memory	Parallel	Training Time	Training Memory
LSTM/LMU	$O(1)$	$O(1)$	✗	$O(N)$	$O(N)$
Transformer	$O(N)$	$O(N)^a$	✓	$O(N^2)$	$O(N)^b$
Linear Transformer	$O(1)$	$O(1)$	✓	$O(N)$	$O(N)$
H3/S4	$O(1)$	$O(1)$	✓	$O(N \log N)$	$O(N)$
Hyena	$O(N)$	$O(N)$	✓	$O(N \log N)$	$O(N)$
RWKV/Mamba/RetNet	$O(1)$	$O(1)$	✓	$O(N)$	$O(N)$

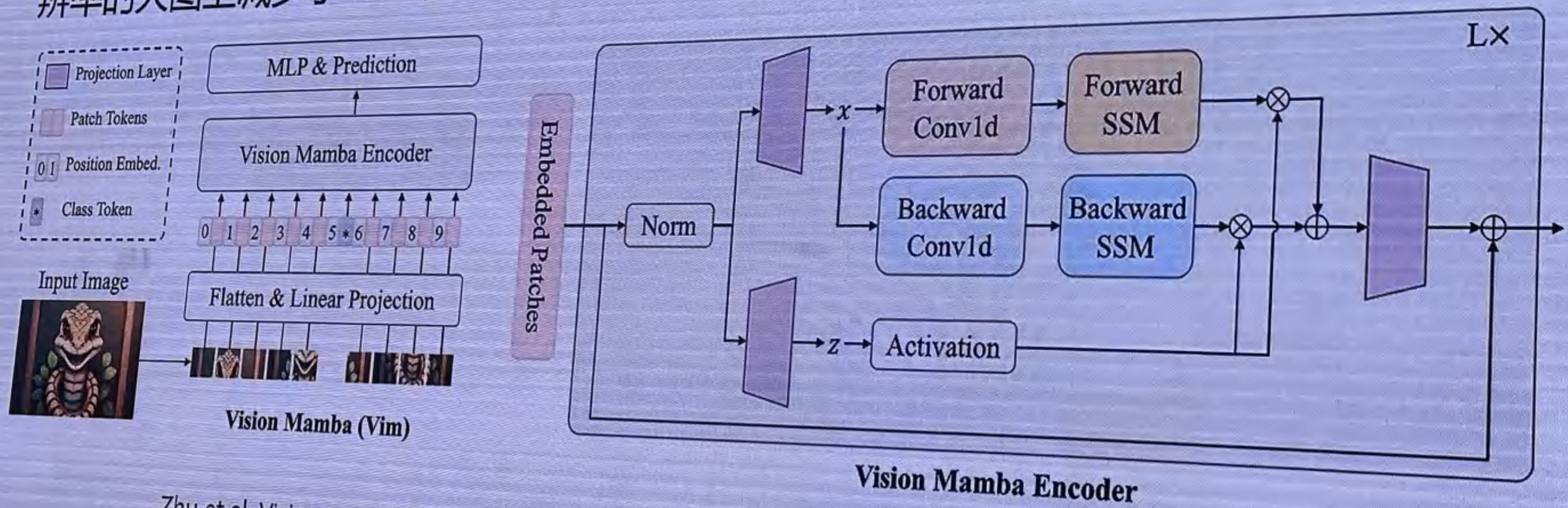
[a] $O(1)$ without KV cache; [b] With Flash Attention

RWKV-5 & -6 中关于Transformer和其它新型高效网络架构之间的推理、训练复杂度对比

Peng, Bo, et al. "Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence." *arXiv:2404.05892* (2024)

计算机视觉：Vision Mamba

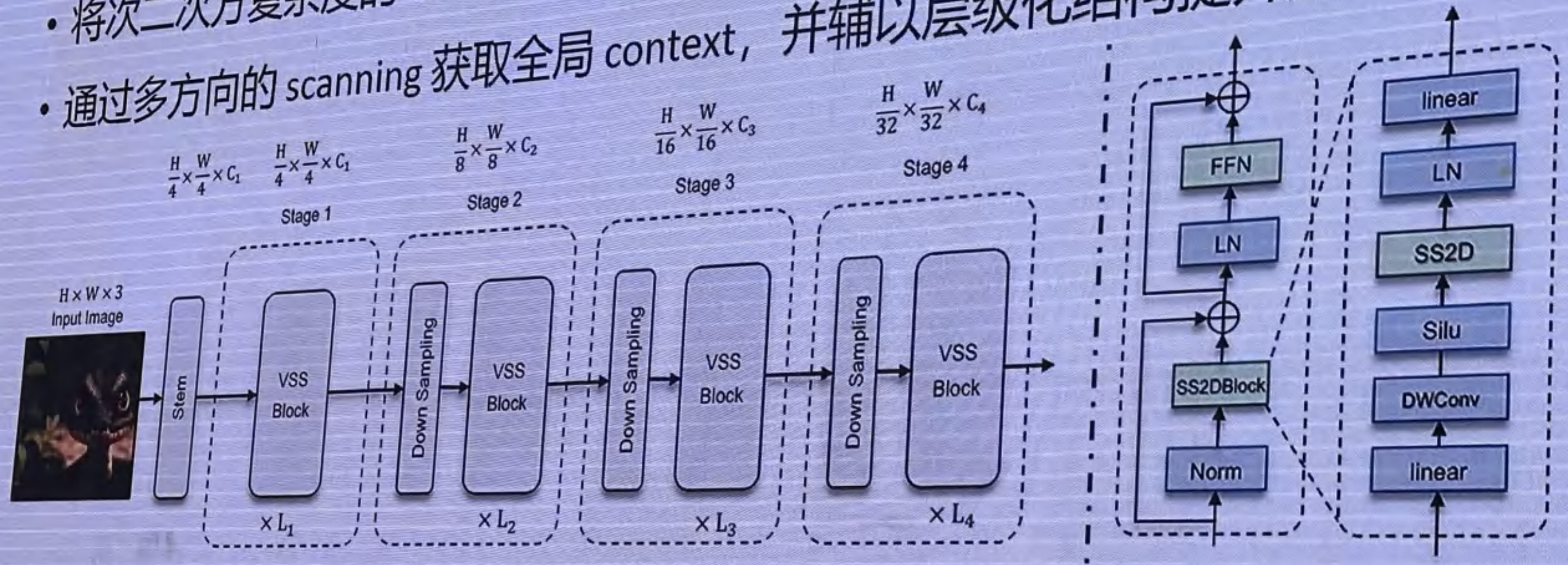
- 通过双向的结构设计使得 Mamba 能够拥有全局上下文感知
- 首次提出具有双向 Mamba 模块的通用视觉基础模型
- 在分类、检测、分割等视觉任务上获得了超越 ViT、DeiT 的结果，并且再 1248 分辨率的大图上减少了 86% 的显存占用，提速 2.8 倍



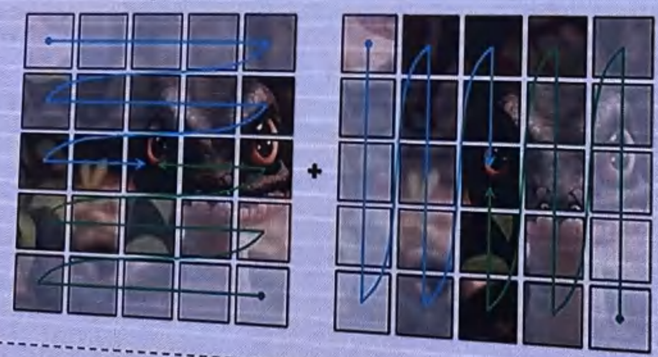
Zhu et al. Vision mamba: Efficient visual representation learning with bidirectional state space model, In ICML, 2024

计算机视觉：VMamba

- 将次二次方复杂度的 Mamba 方法引入视觉任务
- 通过多方向的 scanning 获取全局 context，并辅以层级化结构提升效果



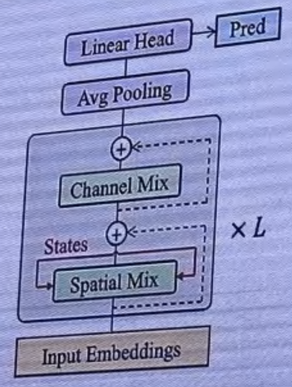
- 在视觉任务上取得了超越多层级 Transformer (Swin) 的结果。



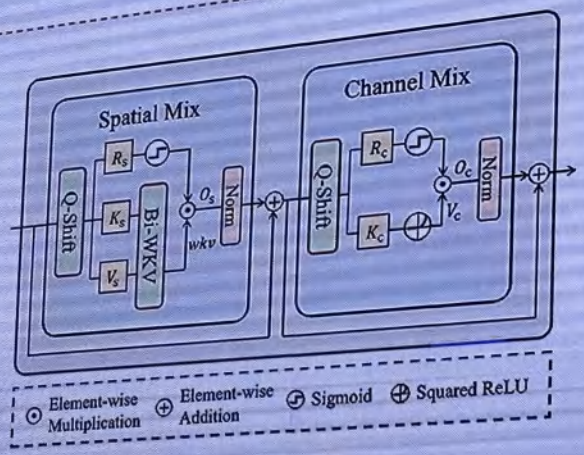
Liu et al. Vmamba: Visual state space model, In Arxiv, 2024.

计算机视觉: Vision RWKV

- 将NLP里的线性算子RWKV引入到视觉领域，并证明了其具有良好的可规模化能力

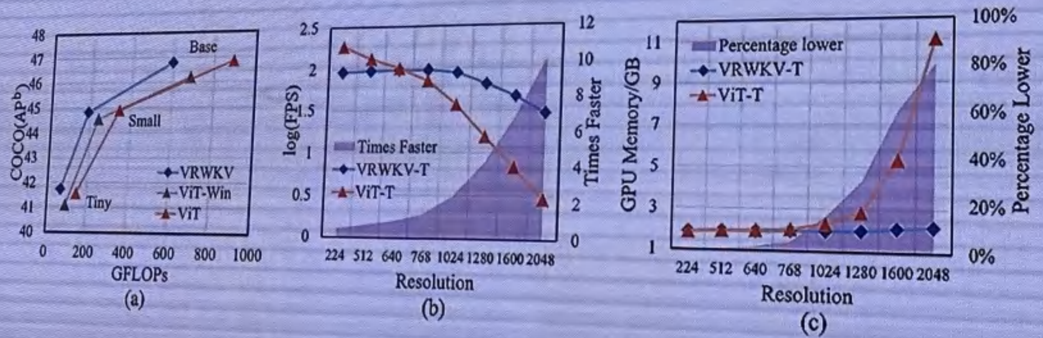


(a) Vision-RWKV Architecture

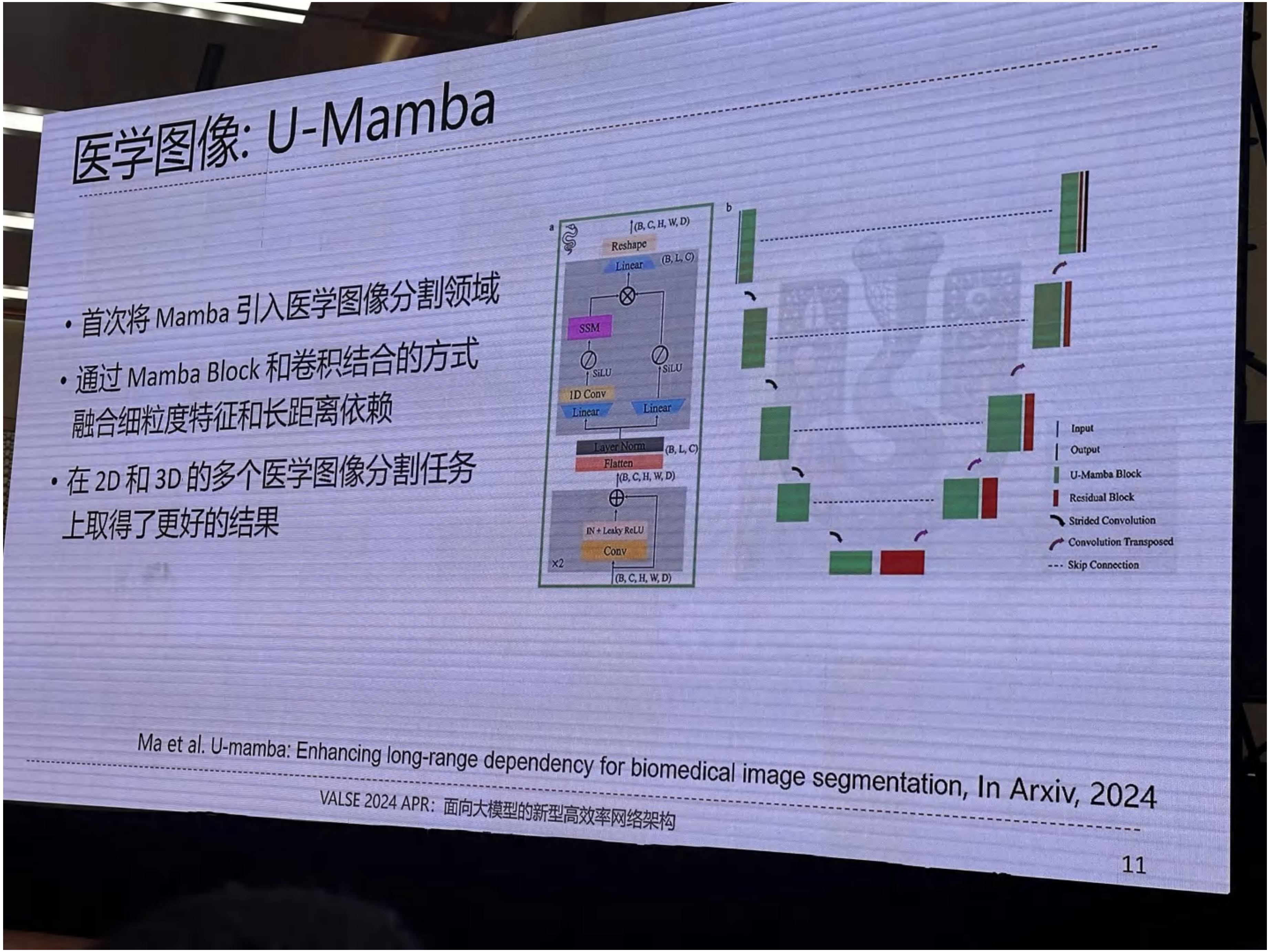


(b) Vision-RWKV Encoder Layer

- 引入双向扫描机制，和上下左右四个方向的Token shift机制，极大地增强了RWKV处理2D视觉信号的能力

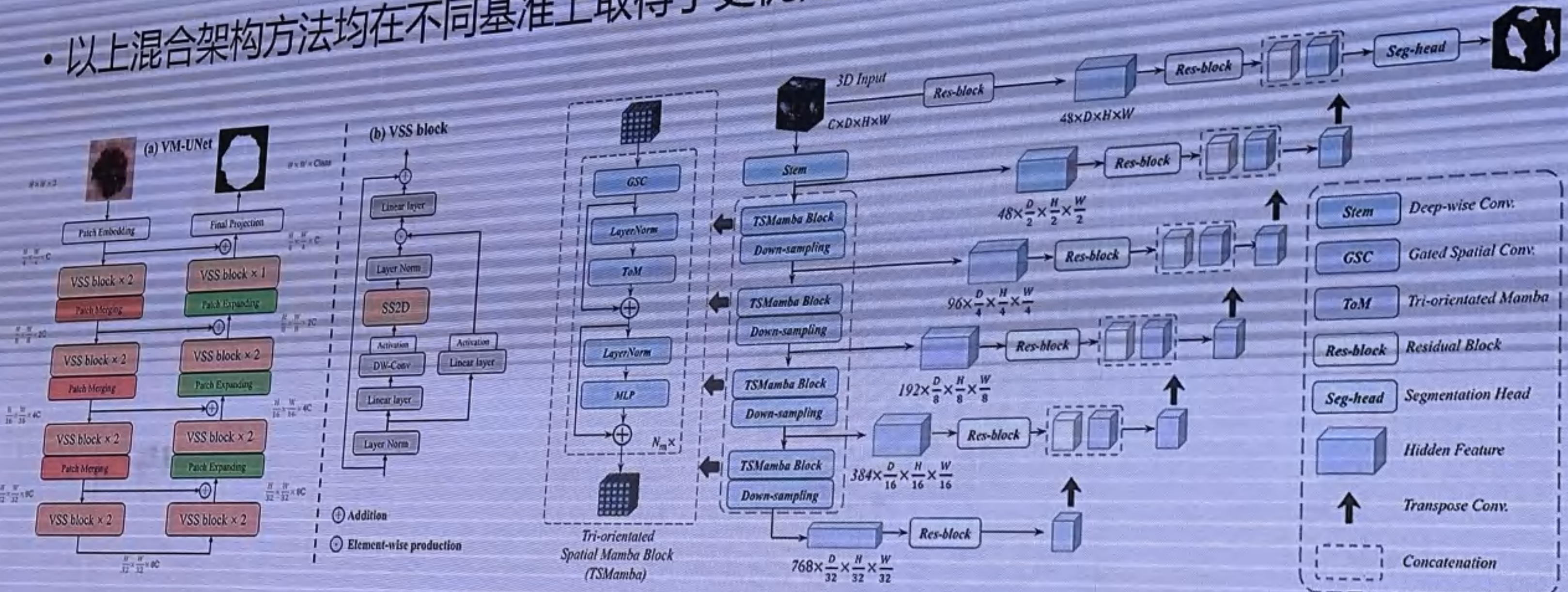


Duan et al. Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures, In arXiv, 2024


医学图像: U-Mamba
• 首次将 Mamba 引入医学图像分割领域
• 通过 Mamba Block 和卷积结合的方式融合细粒度特征和长距离依赖
• 在 2D 和 3D 的多个医学图像分割任务上取得了更好的结果
a
(B, C, H, W, D)
Reshape
Linear
(B, L, C)
SSM
SiLU
SiLU
1D Conv
Linear
Linear
Layer Norm
(B, L, C)
Flatten
(B, C, H, W, D)
IN + Leaky ReLU
Conv
×2
(B, C, H, W, D)
b
Input
Output
U-Mamba Block
Residual Block
Strided Convolution
Convolution Transposed
Skip Connection
Ma et al. U-mamba: Enhancing long-range dependency for biomedical image segmentation, In Arxiv, 2024
VALSE 2024 APR：面向大模型的新型高效率网络架构
11

医学图像

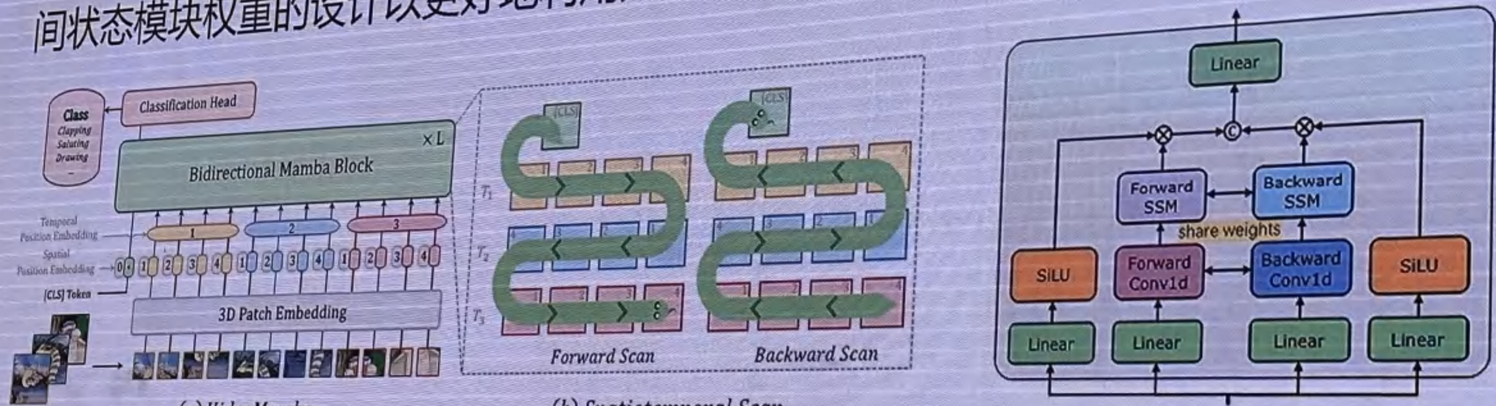
- VM-Unet 和 SegMamba 将 Mamba 和 Unet 结合以处理不同的医学视觉任务
- 以上混合架构方法均在不同基准上取得了更优的结果



Ruan et al. VM-UNet: Vision Mamba UNet for Medical Image Segmentation, In Arxiv, 2024
Xing et al. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation, In Arxiv, 2024

视频理解

- VideoMamba / Video Mamba Suite 均使用了 Vim 构建视频理解网络
- VideoMamba 使用了多方向的 scanning 策略， Video Mamba Suite 使用了共享空间状态模块权重的设计以更好地利用方向偏置



(a) VideoMamba (b) Spatiotemporal Scan

- 以上两方法均取得了包括视频理解、跨模态交互等多项任务的最好结果

Chen et al. Video Mamba Suite: State Space Model as a Versatile Alternative for Video Understanding, In Arxiv, 2024
Li et al. VideoMamba: State Space Model for Efficient Video Understanding, In Arxiv, 2024

点云理解: PointMamba

- 将NLP里的Mamba架构引入到3D点云处理
- 提出点云序列重排序方法，使得Mamba中的SSM模块在处理点云序列时具有更为合理的扫描顺序

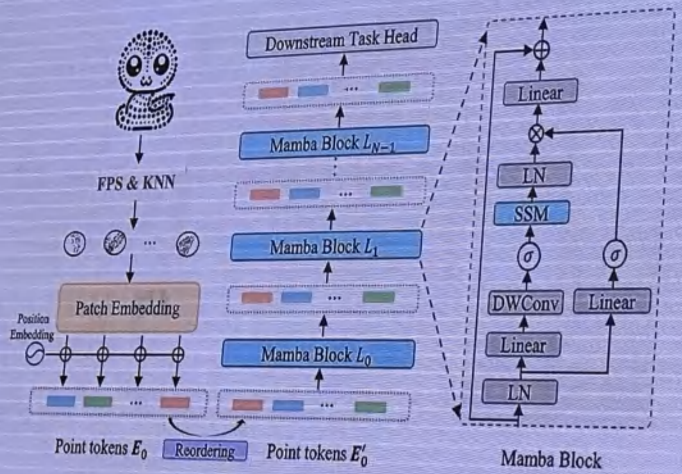


Table 3: Object classification on ScanObjectNN [50]. We evaluate PointMamba on three variants, with PB-T50-RS being the most challenging. Accuracy (%) is reported. * denotes reproduced results, † indicates that using simple rotational augmentation [9] for training.

Methods	Backbone	Param. (M) ↓	FLOPs (G) ↓	OBJ-BG ↑	OBJ-ONLY ↑	PB-T50-RS ↑
PointNet [40]	-	3.5	0.5	73.3	79.2	68.0
PointNet++ [41]	-	1.5	1.7	82.3	84.3	77.9
PointCNN [26]	-	0.6	0.9	86.1	85.5	78.5
DGCNN [56]	-	1.8	2.4	82.8	86.2	78.1
PRANet [6]	-	-	-	-	-	81.0
MVTN [20]	-	11.2	43.7	-	-	82.8
PointNeXt [43]	-	1.4	1.6	-	-	87.7
PointMLP [34]	-	13.2	31.4	-	-	85.4
RepSurf-U [46]	-	1.5	0.8	-	-	84.3
ADS [23]	-	-	-	-	-	87.5
Training from scratch						
Transformer [64]	Transformer-based	22.1	4.8	79.86	80.55	77.24
Point-MAE* [37]	Transformer-based	22.1	4.8	86.75	86.92	80.78
PointMamba (ours)	Mamba-based	**12.3**	**3.6**	**88.30**	**87.78**	**82.48**
Training from pre-training						
Point-BERT [64]	Transformer-based	22.1	4.8	87.43	88.12	83.07
Point-MAE [37]	Transformer-based	22.1	4.8	90.02	88.29	85.18
Point-MAE† [37]	Transformer-based	22.1	4.8	92.77	91.22	**89.04**
PointMamba (ours)	Mamba-based	**12.3**	**3.6**	90.71	88.47	84.87
PointMamba† (ours)	Mamba-based	**12.3**	**3.6**	**93.29**	**91.91**	88.17

Liang et al. PointMamba: A Simple State Space Model for Point Cloud Analysis, In arXiv, 2024

底层视觉: MambaIR

- 将NLP中的Mamba引入到图像复原任务
- 设计了残差空间状态模块，通过额外引入卷积和channel attention来缓解Mamba在处理low-level视觉中出现的局部像素遗忘和通道冗余问题

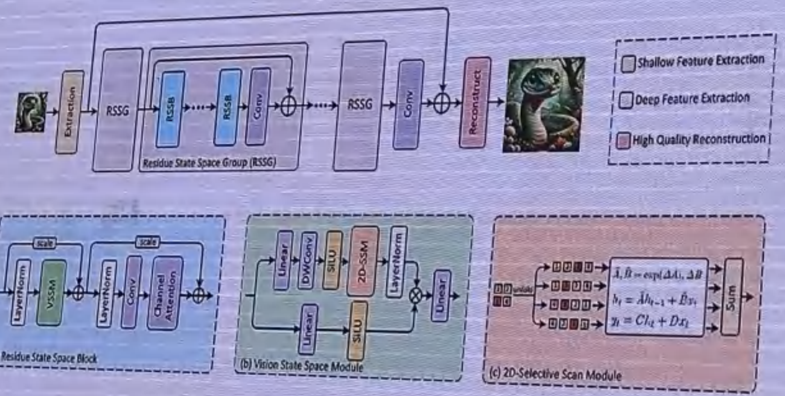


Table 3: Quantitative comparison on classic image super-resolution with state-of-the-art methods. The best and the second best results are in red and blue.

Method	scale	Set5		Set14		BSDS100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [37]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [76]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [12]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
HAN [51]	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
IGNN [79]	×2	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
CSNLN [48]	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
NLSA [47]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
ELAN [75]	×2	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
IPT [8]	×2	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [36]	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
SRFormer [80]	×2	38.51	0.9627	34.44	0.9253	32.57	0.9046	34.09	0.9449	40.07	0.9802
MambaIR	×2	38.57	0.9627	34.67	0.9261	32.58	0.9048	34.15	0.9446	40.28	0.9806
MambaIR+	×2	38.60	0.9628	34.69	0.9260	32.60	0.9048	34.17	0.9443	40.33	0.9806
EDSR [37]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN [77]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [12]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
HAN [51]	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
IGNN [79]	×3	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
CSNLN [48]	×3	34.74	0.9300	30.66	0.8482	29.33	0.8105	29.13	0.8712	34.45	0.9502
NLSA [47]	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
ELAN [75]	×3	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517
IPT [8]	×3	34.81	-	30.85	-	29.38	-	29.49	-	-	-
SwinIR [36]	×3	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
SRformer [80]	×3	35.02	0.9323	30.94	0.8540	29.48	0.8156	30.04	0.8865	35.26	0.9543
MambaIR	×3	35.08	0.9323	30.99	0.8536	29.51	0.8157	29.93	0.8841	35.43	0.9546
MambaIR+	×3	35.13	0.9326	31.06	0.8541	29.53	0.8162	29.98	0.8838	35.55	0.9549
EDSR [37]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [76]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [12]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
HAN [51]	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IGNN [79]	×4	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
CSNLN [48]	×4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201
NLSA [47]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
ELAN [75]	×4	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
IPT [8]	×4	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [36]	×4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
SRFormer [80]	×4	32.93	0.9041	29.08	0.7953	27.94	0.7502	27.68	0.8311	32.21	0.9271
MambaIR	×4	33.03	0.9046	29.20	0.7961	27.98	0.7503	27.68	0.8287	32.32	0.9272
MambaIR+	×4	31.13	0.9054	29.25	0.7971	28.01	0.7510	27.80	0.8303	32.48	0.9281

Guo et al. MambaIR: A Simple Baseline for Image Restoration with State-Space Model, In arXiv, 2024

遥感图像: RS-Mamba

- 将Mamba应用到遥感图像分析领域，用于处理非常高分辨率 (VHR) 遥感图像的密集预测任务
- 设计了全方向扫描机制，在水平、垂直、对角和反对角方向上进行选择性扫描，增强了在多个方向上对上下文信息的全局建模能力

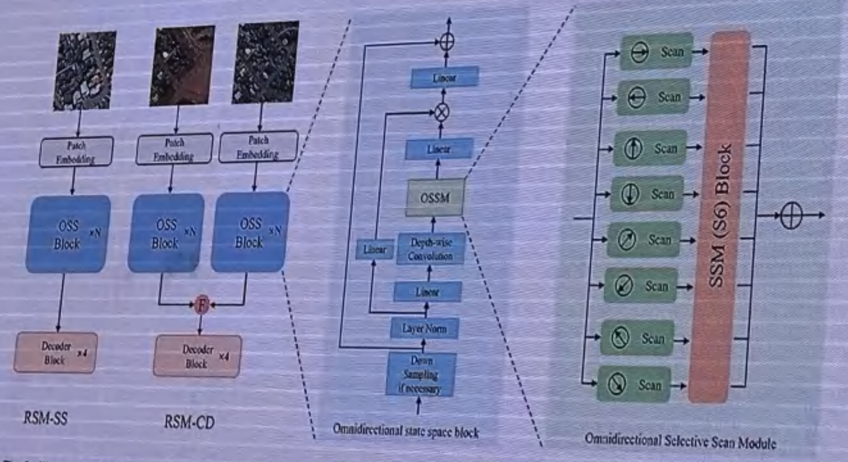


Fig. 2. Illustration of the Overall structure of RSM-SS and RSM-CD. RSM-SS and RSM-CD can globally model the context of images in multiple directions with linear complexity using the omnidirectional selective scan.

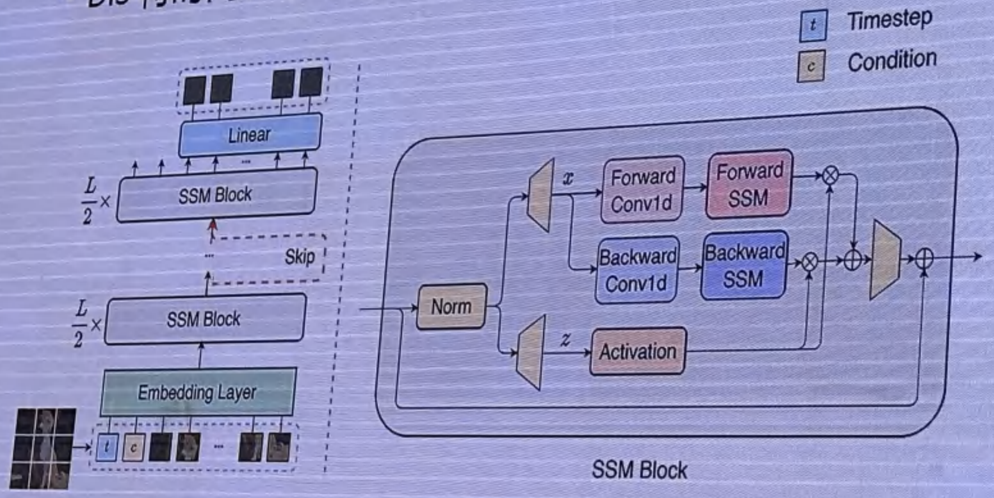
TABLE III
ACCURACY COMPARISON ON THE MASSACHUSETTS ROAD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

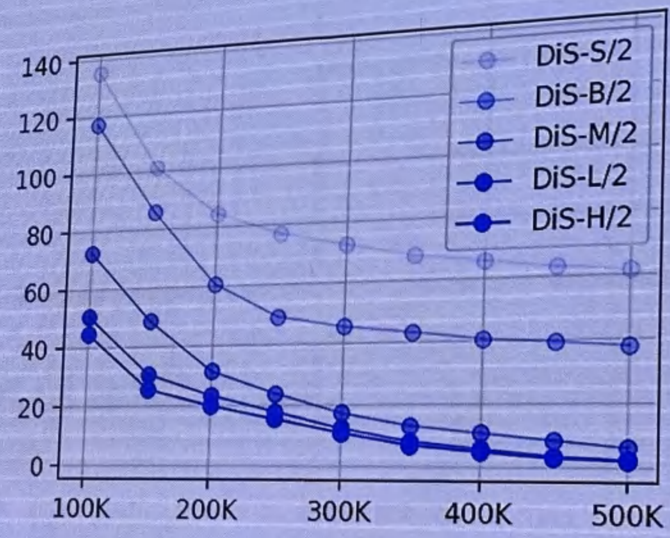
Methods	P (%)	R (%)	F1 (%)	IoU (%)
SegNet [44]	76.09	78.23	77.15	62.79
U-Net [17]	77.53	77.82	77.67	63.50
ResUNet [23]	78.77	77.45	78.10	64.07
D-LinkNet [51]	78.34	77.91	78.12	64.10
HRNetv2 [46]	79.01	78.20	78.60	64.75
Deeplabv3+ [48]	75.14	72.56	73.83	58.51
SIINet [52]	85.36	74.13	79.35	65.77
RoadFormer [8]	80.54	**78.90**	79.71	66.27
BDTNet [54]	82.99	76.37	79.54	66.03
RSM-SS	**86.52**	75.24	**80.49**	**67.35**

Chen et al. RSMamba: Remote Sensing Image Classification with State Space Model, In arXiv, 2024

扩散模型: DiS

- DiS 首次将 Vim 模块引入 Diffusion 生成领域
- DiS 将时间、条件和噪声图像块视作 token，并通过跳跃连接贯通浅层和深层



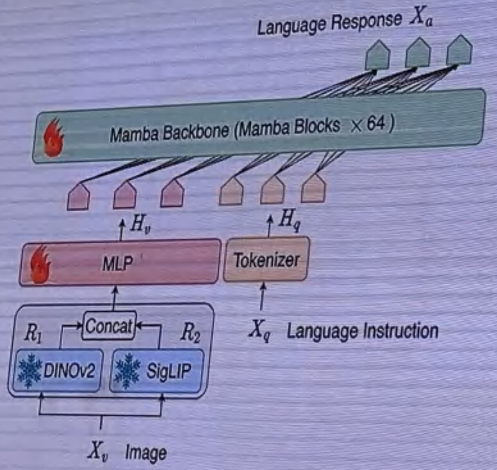


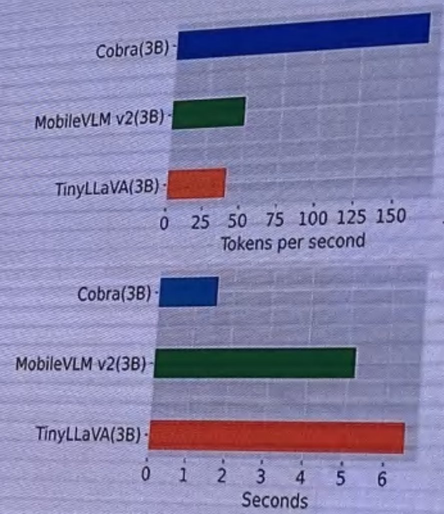
- DiS 展现了较强的可扩展性并在多个图片生成基准上展现了卓越的结果

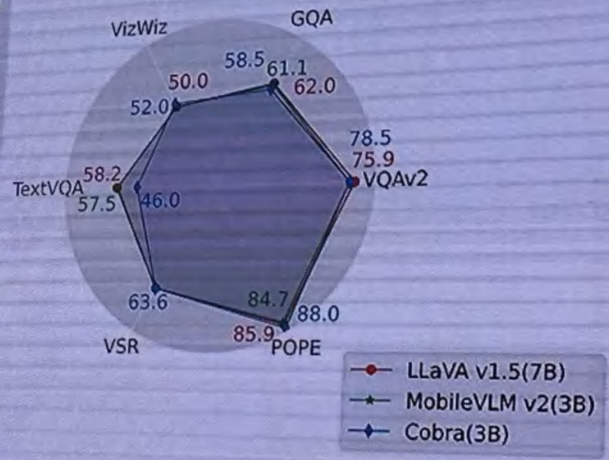
Fei et al. Scalable Diffusion Models with State Space Backbone, In Arxiv, 2024

多模态模型: Cobra

- Cobra 首次将 Mamba 引入多模态大语言模型领域
- Cobra 通过线性层将视觉基础模型的输出映射为 Mamba 可接受的向量





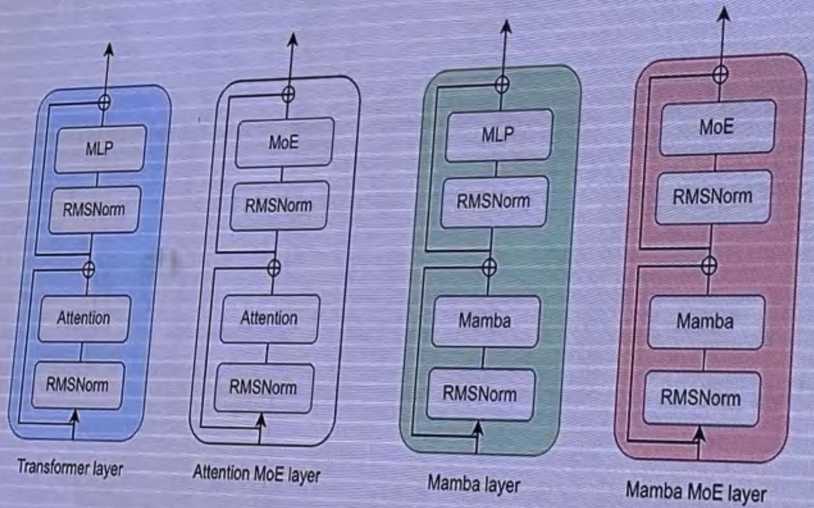


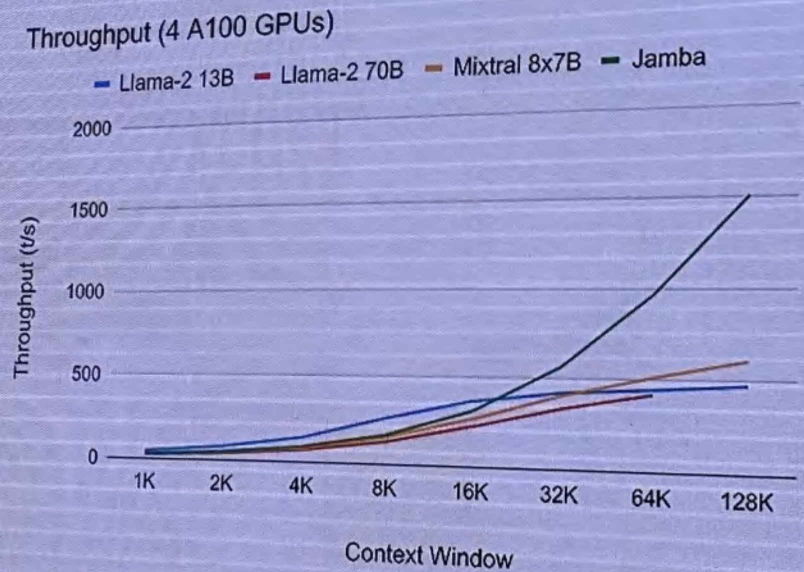
- Cobra 在多个基准取得了更好的结果，以及更高的效率

Zhao et al. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference, In Arxiv, 2024

混合架构: Jamba

- Jamba 希望同时结合 Transformer (高效用) 和 Mamba (高效率) 的优点
- 通过交错排列Transformer和Mamba层的块来实现，同时在某些层中加入MoE以增加模型容量，并控制被激活的参数量
- 仅采用1个80G现存的GPU，实现了大规模数据集上的训练





Lieber et al. Jamba: A Hybrid Transformer-Mamba Language Model, In Arxiv, 2024

总结和展望

- 长序列的表征学习在Sora视频生成、多模态大模型领域非常关键，探索类似Mamba的低复杂度序列建模网络是一个非常重要的问题。
- Mamba等模型尚存在的挑战：（1）缺乏超大规模预训练上的验证；（2）缺乏灵活的多模态自监督训练方法；（3）串行模型的并行训练与硬件实现息息相关，更高效率的并行训练方法需要进一步挖掘。
- 值得进一步探索的方向：（1）混合架构，有机的结合transformer、mamba等方法的优点；（2）算子优化，新型架构和异构硬件的协同优化；（3）视频生成：充分发挥新型架构在长序列建模方面的低复杂度优势；（4）多模态建模：构建类似于fuyu模型那种原生的多模态模型。