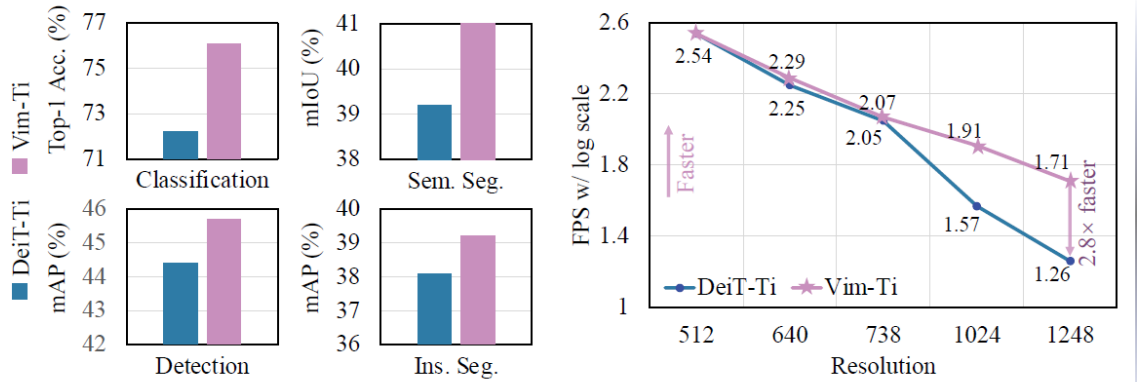
Multi-Scale VMamba: Hierarchy in Hierarchy Visual State Space Model

Yuheng Shi [1], Minjing Dong [1], Chang Xu [2]

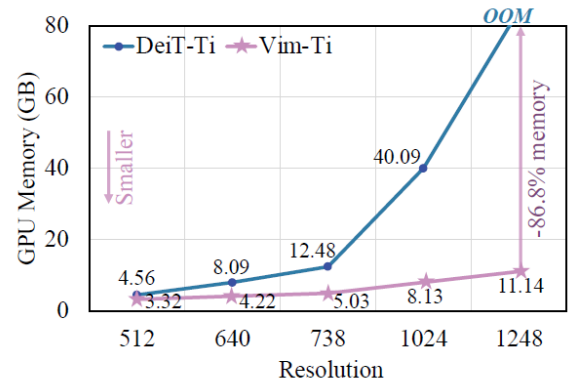
[1]City University of HongKong [2]University of Sydney

Background

- Vision Transformers show remarkable performance with global receptive field, but limited by quadratic complexity with respect to the token length
- Convolution Neural Networks show linear scaling complexity, but limited by the local receptive field
- Mamba have garnered attention for the ability to combine the best of both worlds: a global receptive field and linear scaling complexity



(a) Accuracy Comparison (b) Speed Comparison



(c) GPU Memory Comparison

Observation and Motivation

- Constraint on model size inherently limits the long-range modeling capabilities of SSMs in vision tasks. For ViM-Tiny, placing the class token in the middle of the sequence yields markedly better results than positioning it at the ends.

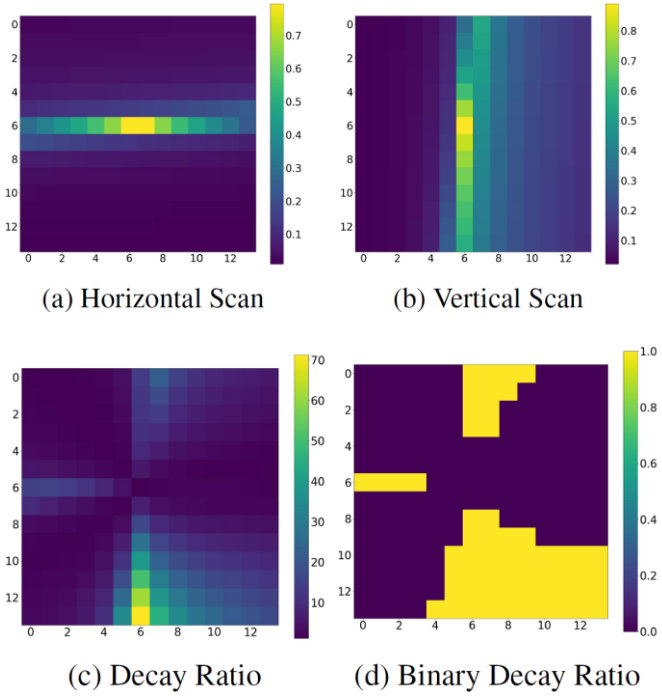
- The contribution of the m_{th} token to the construction of the n_{th} token decays significantly along their distance:

$$C_n \prod_{i=m}^{n} \bar{A}_i \bar{B}_m = C_n \bar{A}_{(m \to n)} \bar{B}_m,$$

$$where\ \bar{A}_{(m \to n)} = e^{\sum_{i=m}^{n} \Delta_i A}$$

- Multi-scan strategy reduces the decay of influence, but also introduce redundancy.

Classification strategy	ImageNet top-1 acc.
Mean pool	73.9
Max pool	73.4
Head class token	75.2
Double class token	74.3
Middle class token	76.1



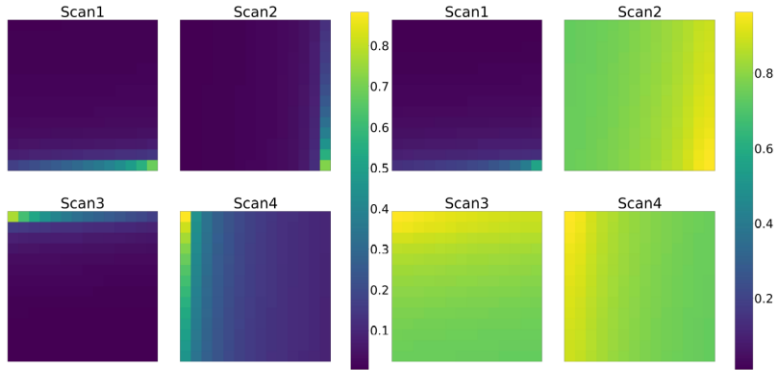
(a) Horizontal Scan (b) Vertical Scan

(c) Decay Ratio (d) Binary Decay Ratio

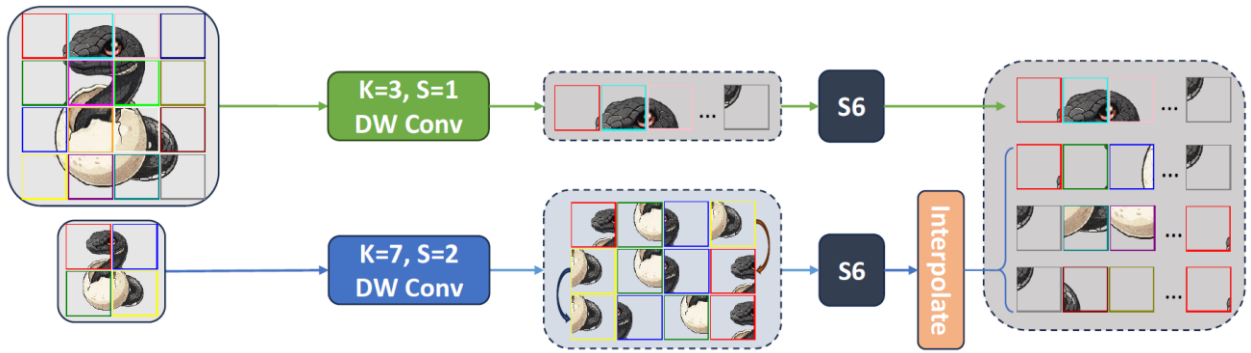
Method

- The most effective and direct way to alleviate the long-range forgetting problem is to reduce the number of tokens.

- Setting all scans on a downsampled feature map will ignore fine-grained features and result in unavoidable information loss. Scanning along the full-resolution feature map is also essential.



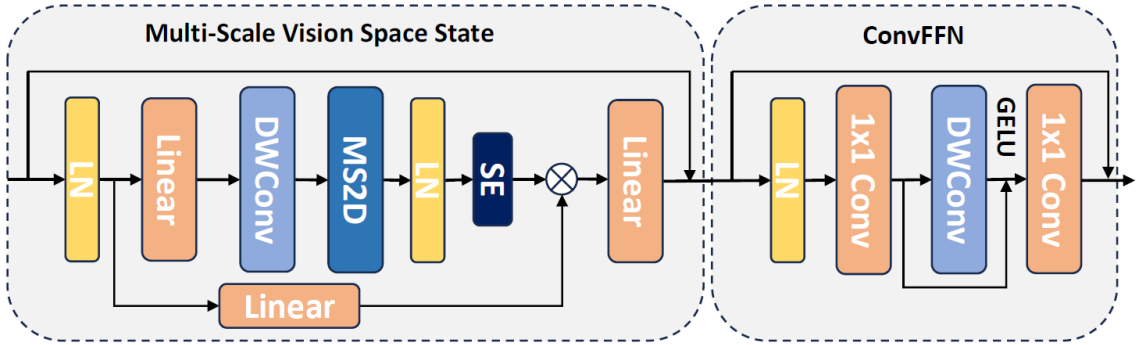
(a) SS2D (b) MS2D



Method

- Our design introduces a channel mixer to augment the flow of information across different channels

- a Squeeze-Excitation (SE) block is integrated subsequent to the MS2D, as informed by LocalMamba



Model	Blocks	Channels	*ssm ratio*	*FFN ratio*	#param.(M)	GFLOPs
Nano	[1, 2, 5, 2]	[48, 96, 192, 384]	2	2	7	0.9
Micro	[1, 2, 5, 2]	[64, 128, 256, 512]	2	2	12	1.5
Tiny	[2, 2, 9, 2]	[96, 192, 384, 768]	1	4	32	5.1
Small	[2, 3, 20, 2]	[96, 192, 384, 768]	1	4	50	8.8
Base	[2, 3, 20, 2]	[128, 256, 512, 1024]	1	4	88	15.5

Experiments

- Replacing SSD in VMamba with our MS2D bring improvement in both accuracy, GPU memory cost and efficiency.

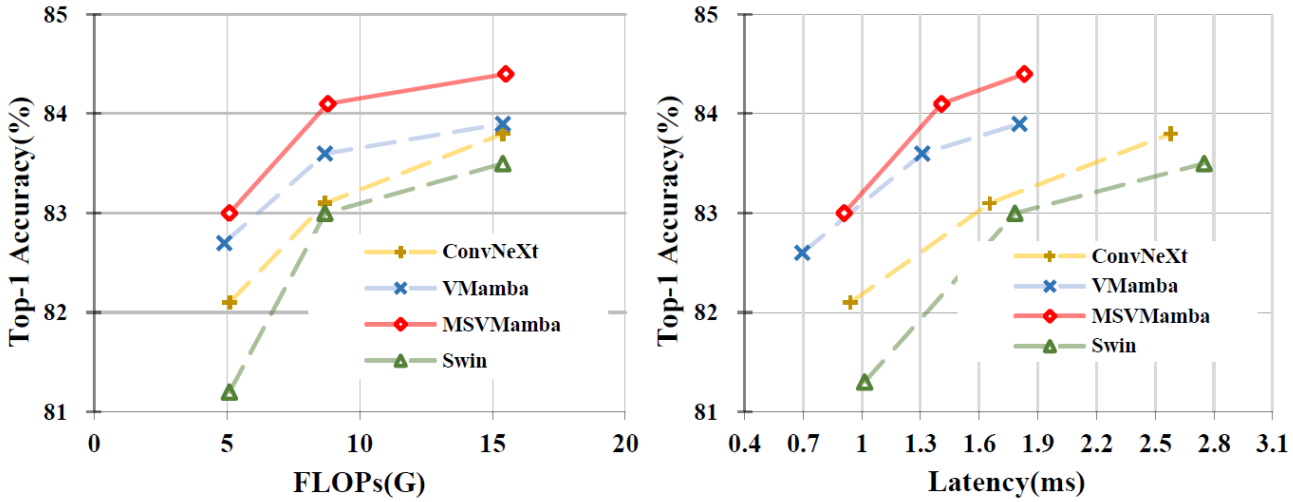
- Compared to other scanning strategy, our MS2D continuous to lead.

Model	MS2D	SE	ConvFFN	$N=1$	#param.	FLOPs	Top-1 Acc(%).	AP_b	AP_m
VMamba-Nano					4.4M	0.87G	69.6	38.1	35.6
MSVMamba-Nano	✓				4.8M	0.89G	$71.9_{\uparrow 2.3}$	$39.1_{\uparrow 1.0}$	$36.3_{\uparrow 0.7}$
	✓	✓			5.3M	0.89G	$72.4_{\uparrow 2.8}$	$39.5_{\uparrow 1.4}$	$36.5_{\uparrow 0.9}$
	✓	✓	✓		6.6M	0.94G	$74.4_{\uparrow 4.8}$	$41.0_{\uparrow 2.9}$	$37.8_{\uparrow 2.2}$
	✓	✓	✓	✓	6.9M	0.88G	$75.1_{\uparrow 5.5}$	$41.4_{\uparrow 3.3}$	$37.9_{\uparrow 2.3}$

Model	MS2D	SE	ConvFFN	$N=1$	#param.	FLOPs	Top-1 Acc(%).	FPS	Memory (MB)
VMambav1-Tiny					23M	5.6G	80.3	603	6639
MSVMamba-Tiny	✓				24M	4.8G	$80.9_{\uparrow 0.6}$	866	4780
	✓	✓	✓	✓	33M	4.6G	$81.4_{\uparrow 1.1}$	1092	4533
MSVMamba-Tiny†	✓	✓	✓	✓	32M	5.1G	$81.7_{\uparrow 1.4}$	1097	2413

Setting	**#param.(M)**	**GFLOPs**	**Accuracy (%)**
Uni-Scan	4.4	0.87	68.9
Bi-Scan	4.4	0.87	69.5
CrossScan	4.4	0.87	69.6
MS2D	4.8	0.89	**71.9**

Experiments



Model	Top-1 Acc(%)	#Params	FLOPs (G)	Thru. (imgs/sec)	Memory (MB)
Swin-T [32]	81.3	28 M	4.5	986	2402
ConvNeXt-T [33]	82.1	29 M	4.5	1062	1670
VMambav1-T [30]	82.2	23 M	5.6	603	6639
VMambav3-T [30]	82.6	30 M	4.9	1456	3204
MSVMamba-T	83.0	32 M	5.1	1097	2413
Swin-S [32]	83.0	50 M	8.7	561	2596
ConvNeXt-S [33]	83.1	50 M	8.7	605	1753
VMambav1-S [30]	83.5	44 M	11.2	425	6882
VMambav3-S [30]	83.6	50 M	8.7	764	5780
MSVMamba-S	84.1	50 M	8.8	708	2545
Swin-B [32]	83.5	88 M	15.5	363	3362
ConvNeXt-B [33]	83.8	89 M	15.4	387	2380
VMambav1-B [30]	83.7	76 M	18.0	314	8853
VMambav3-B [30]	83.9	89 M	15.4	555	7826
MSVMamba-B	84.3	88 M	15.5	545	3358

Our MSVMamba also demonstrates good scalability and gets better in accuracy-efficiency curve along model size.

香港城市大學
City University of Hong Kong

THE UNIVERSITY OF SYDNEY

Thanks for watching!