MambaAD: Exploring State Space Models for Multi-class Unsupervised Anomaly Detection

Project Page: https://lewandofskee.github.io/projects/MambaAD/

Haoyang He[1*], Yuhu Bai[1*], Jiangning Zhang[2], Qingdong He[2], Hongxu Chen[1], Zhenye Gan[2], Chengjie Wang[2], Xiangtai Li[3], Guanzhong Tian[1], Lei Xie[1#]

[1]Zhejiang University [2]Youtu Lab, Tencent

[3]Nanyang Technological University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Comparison with CNN and Transformer based methods

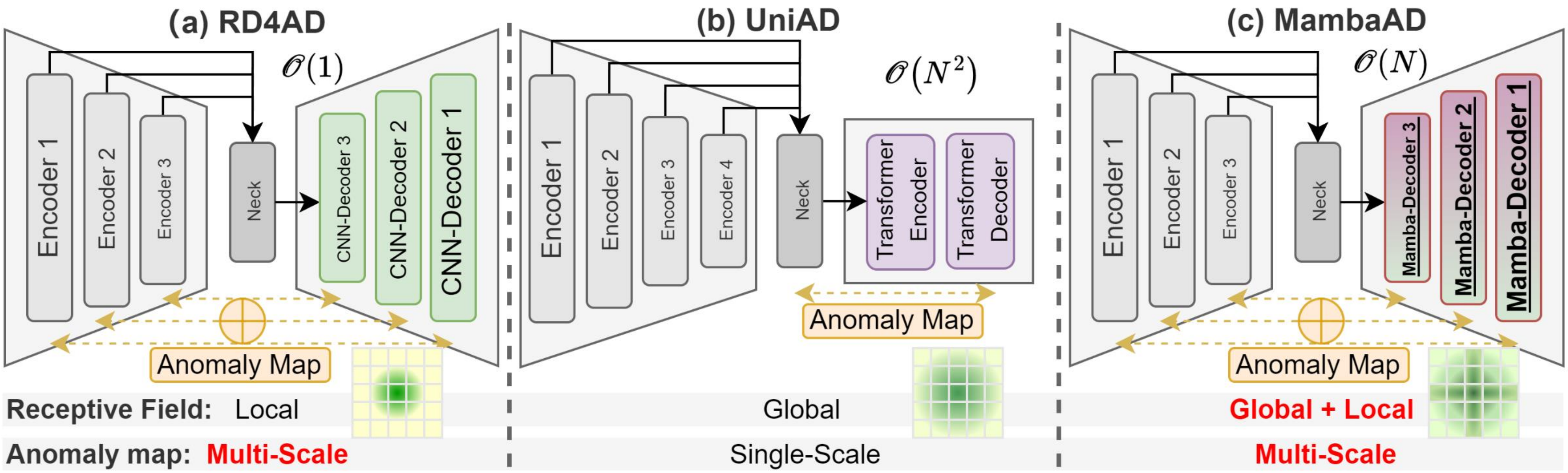


Figure 1: Compared with (a) local CNN-based RD4AD and (b) global Transformer-based UniAD, ours MambaAD with linear complexity is capable of integrating the advantages of both global and local modeling, and multi-scale features endow it with more refined prediction accuracy.

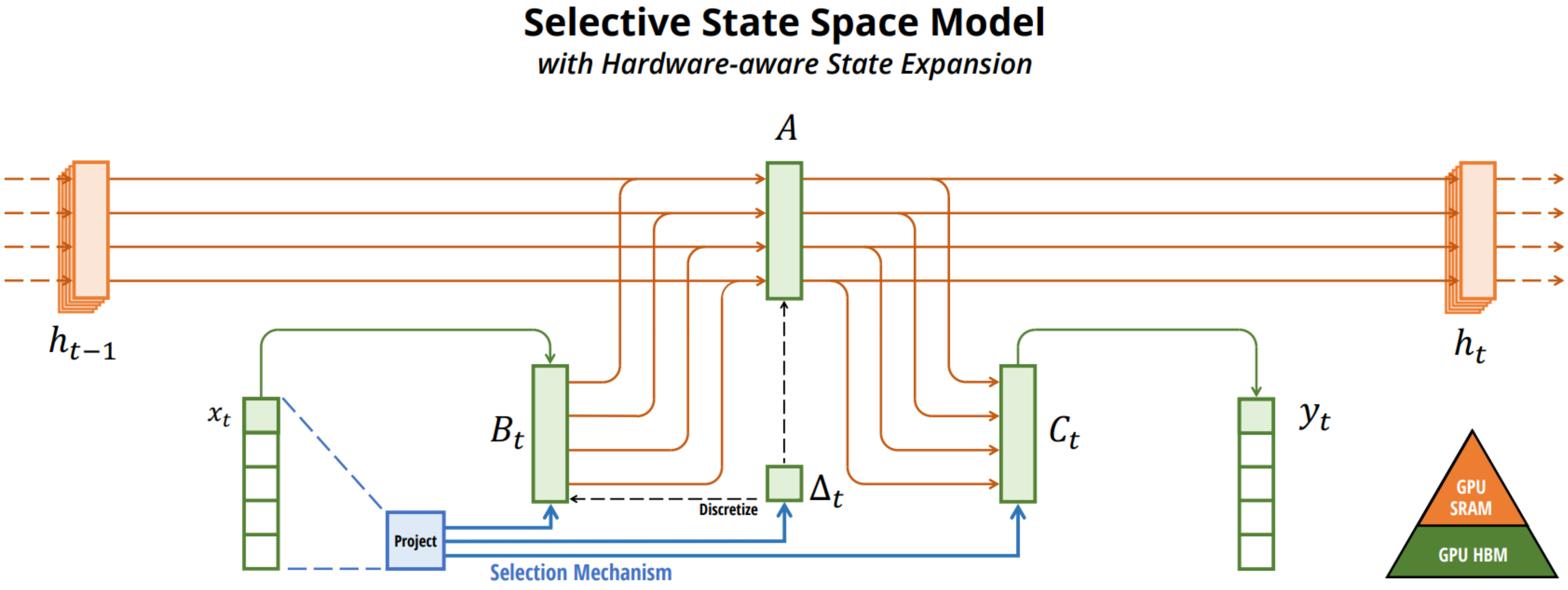
Introduction

Our Contributions

- We introduce **MambaAD**, which innovatively applies the Mamba framework to address multi-class unsupervised anomaly detection tasks. This approach enables **multi-scale training and inference** with **minimal model parameters and computational complexity.**
- We design a **Locality-Enhanced State Space** (LSS) module, comprising cascaded Mamba-based blocks and parallel multi-kernel convolutions, **extracts both global feature correlations and local information associations, achieving a unified model of global and local patterns**.
- We have explored a **Hybrid State Space** (HSS) block, **encompassing five methods and eight multi-directional scans**, to enhance the global modeling capabilities for complex anomaly detection images across various categories and morphologies.
- We demonstrate the superiority and efficiency of MambaAD in multi-class anomaly detection tasks, achieving **SoTA results on six distinct AD datasets with seven metrics** while maintaining remarkably low model parameters and computational complexity.

Method

Preliminaries



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

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \quad y_t = \mathbf{C}h_t.$$

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

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}\overline{\mathbf{B}}, \dots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \quad \mathbf{y} = \mathbf{x} * \overline{\mathbf{K}},$$

Figure Reference: Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.

Method

Overview

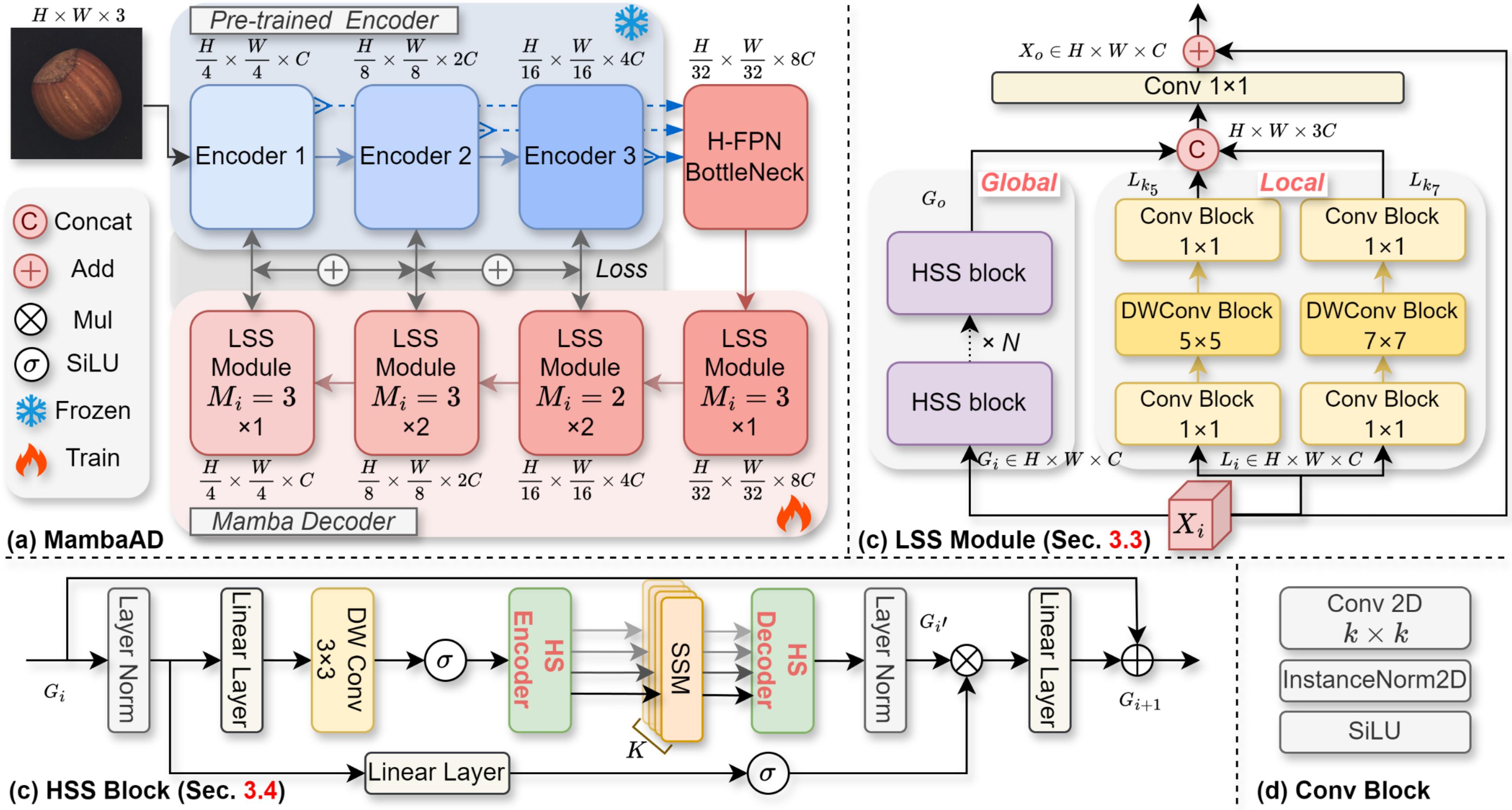


Figure 2: Overview of the proposed MambaAD, which employs pyramidal auto-encoder framework to reconstruct multi-scale features by the proposed efficient and effective Locality-Enhanced State Space (LSS) module. Specifically, each LSS consists of: 1) cascaded Hybrid State Space (HSS) blocks to capture global interaction; and 2) parallel multi-kernel convolution operations to replenish local information. Aggregated multi-scale reconstruction error serves as the anomaly map for inference.

Method

Hybrid Scanning Methods and Directions

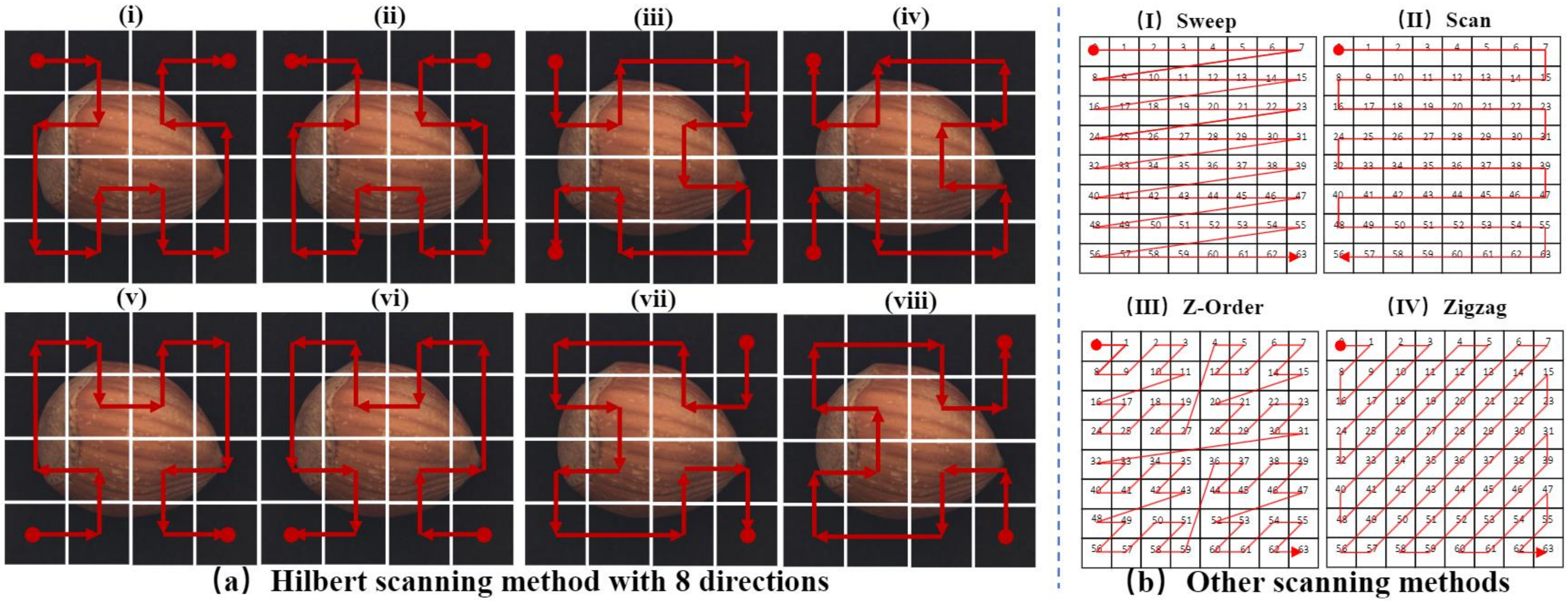


Figure 3: Hybrid Scanning directions and methods. (a) The Hilbert scanning method with 8 scanning directions is used for HS Encoder and Decoder. (b) The other four scanning methods for comparison.

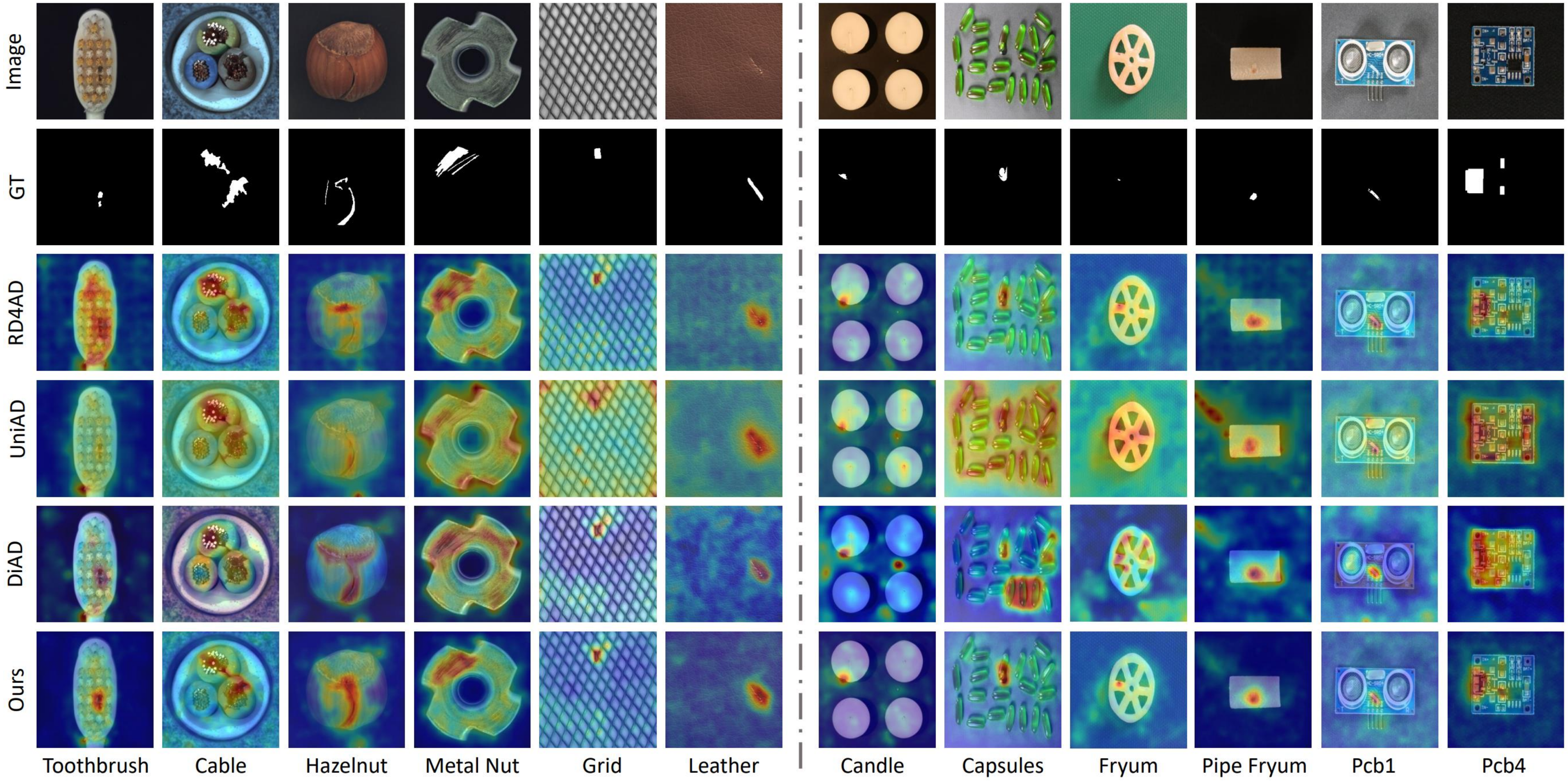
Comparison with SoTAs

◻ Quantitative Results on Three Datasets, more in Appendix

Dateset	Method	Image-level			Pixel-level				mAD
		AU-ROC	AP	F1_max	AU-ROC	AP	F1_max	AU-PRO	
MVTec-AD [2]	RD4AD [8]	94.6	96.5	95.2	96.1	48.6	53.8	91.1	82.3
	UniAD [44]	96.5	98.8	96.2	96.8	43.4	49.5	90.7	81.7
	SimpleNet [26]	95.3	98.4	95.8	96.9	45.9	49.7	86.5	81.2
	DeSTSeg [50]	89.2	95.5	91.6	93.1	54.3	50.9	64.8	77.1
	DiAD [14]	97.2	99.0	96.5	96.8	52.6	55.5	90.7	84.0
	MambaAD (Ours)	**98.6**	**99.6**	**97.8**	**97.7**	**56.3**	**59.2**	**93.1**	**86.0**
VisA [53]	RD4AD [8]	92.4	92.4	**89.6**	98.1	38.0	42.6	**91.8**	77.8
	UniAD [44]	88.8	90.8	85.8	98.3	33.7	39.0	85.5	74.6
	SimpleNet [26]	87.2	87.0	81.8	96.8	34.7	37.8	81.4	72.4
	DeSTSeg [50]	88.9	89.0	85.2	96.1	**39.6**	43.4	67.4	72.8
	DiAD [14]	86.8	88.3	85.1	96.0	26.1	33.0	75.2	70.1
	MambaAD (Ours)	**94.3**	**94.5**	89.4	**98.5**	39.4	**44.0**	91.0	**78.7**
Real-IAD [39]	RD4AD [8]	82.4	79.0	73.9	97.3	25.0	32.7	89.6	68.6
	UniAD [44]	83.0	80.9	74.3	97.3	21.1	29.2	86.7	67.5
	SimpleNet [26]	57.2	53.4	61.5	75.7	2.8	6.5	39.0	42.3
	DeSTSeg [50]	82.3	79.2	73.2	94.6	**37.9**	**41.7**	40.6	64.2
	DiAD [14]	75.6	66.4	69.9	88.0	2.9	7.1	58.1	52.6
	MambaAD (Ours)	**86.3**	**84.6**	**77.0**	**98.5**	33.0	38.7	**90.5**	**72.7**

Comparison with SoTAs

Qualitative Results on MVTec-AD and VisA Datasets



Ablation Studies

Qualitative Results on MVTec-AD and VisA Datasets

Table 2: Incremental Ablations.

Basic Mamba	LSS	HSS	MVTec-AD	VisA
✓			82.1	72.9
✓	✓		84.9	78.0
✓	✓	✓	**86.0**	**78.9**

Table 3: Ablation Study on the LSS Module.

Method	Params(M)	FLOPs(G)	MVTec-AD	VisA
Local	13.0	5.0	81.7	72.5
Global	22.5	7.5	82.1	72.9
Local + Global	25.7	8.3	**86.0**	**78.9**

Table 6: Efficiency comparison of SoTA methods.

Method	Params(M)	FLOPs(G)	mAD
RD4AD[12]	80.6	28.4	82.3
UniAD [47]	**24.5**	**3.6**	81.7
DeSTSeg [55]	35.2	122.7	81.2
SimpleNet [30]	72.8	16.1	77.1
DiAD [18]	1331.3	451.5	84.0
MambaAD (Ours)	25.7	8.3	**86.0**

Table 4: Ablation studies on the pre-trained backbone and Mamba decoder depth.

Backbone	Decoder Depth	Image-level			Pixel-level				Params(M)	FLOPs(G)
		AU-ROC	AP	F1_max	AU-ROC	AP	F1_max	AU-PRO		
ResNet18	[2,2,2,2]	96.7	98.6	95.8	95.7	47.9	52.4	89.1	14.6	4.3
	[3,4,6,3]	96.6	98.8	96.4	96.8	53.2	56.2	91.8	20.3	6.2
ResNet34	[2,2,2,2]	98.0	99.3	97.0	97.6	55.4	58.2	92.7	20.0	6.5
	[2,9,2,2]	97.6	99.3	97.3	97.7	56.4	59.0	93.2	26.1	7.9
	[3,4,6,3]	**98.6**	**99.6**	97.8	97.7	56.3	59.2	93.1	25.7	8.3
ResNet50	[3,4,6,3]	98.4	99.4	97.7	97.7	54.2	57.0	92.3	251.0	60.3
WideResNet50	[3,4,6,3]	**98.6**	99.5	**98.0**	**98.0**	**57.9**	**60.3**	**93.8**	268.0	68.1

Table 5: Ablations on different scanning methods and directions.

Index	HS Methods with Different Directions					Image-level			Pixel-level			
	Sweep	Scan	Zorder	Zigzag	Hilbert	AU-ROC	AP	F1_max	AU-ROC	AP	F1_max	AU-PRO
1	8	-	-	-	-	98.1	99.4	97.2	97.5	56.8	58.8	92.9
2	-	8	-	-	-	98.0	99.4	97.2	97.6	56.6	59.0	**93.4**
3	-	-	8	-	-	98.1	99.4	97.4	97.6	56.6	59.0	93.0
4	-	-	-	8	-	98.2	99.4	97.6	97.6	56.3	58.8	93.1
5	-	-	-	-	2	97.9	99.3	97.1	**97.7**	56.5	**59.2**	93.1
6	-	-	-	-	4	98.0	99.4	97.0	**97.7**	**56.9**	59.1	93.2
7	-	-	-	-	8	**98.6**	**99.6**	**97.8**	**97.7**	56.3	**59.2**	93.1
8	-	-	-	4	4	96.8	99.0	97.0	97.4	54.4	57.0	92.8
9	-	-	4	-	4	97.5	99.2	97.4	97.5	55.0	57.4	93.1
10	-	4	-	-	4	97.4	99.1	96.8	97.5	55.5	57.9	93.3
11	4	-	-	-	4	98.0	99.3	97.4	97.6	56.2	58.5	93.3
12	-	2	2	2	2	97.5	99.2	97.1	97.5	55.4	57.9	92.9

Conclusion

Qualitative Results on MVTec-AD and VisA Datasets

- This paper introduces MambaAD, the first application of the Mamba framework to AD. MambaAD consists of a pre-trained encoder and a Mamba decoder, with a novel LSS module employed at different scales and depths. The LSS module, composed of sequential HSS modules and parallel multi-core convolutional networks, combines Mamba's global modeling prowess with CNN-based local feature correlation. The HSS module employs HS encoders to encode input features into five scanning patterns and eight directions, which facilitate the modeling of feature sequences in industrial products at their central positions. Extensive experiments on six diverse AD datasets and seven evaluation metrics demonstrate the effectiveness of our approach in achieving SoTA performance.
- Limitations, Broader Impact and Social Impact. The model is not efficient enough and more lightweight models need to be designed. This study marks our initial attempt to apply Mamba in AD, laying a foundation for future research. We hope it can inspire lightweight designs in AD. MambaAD exhibits significant practical implications in enhancing industrial production efficiency.

Thanks