MambaVision: A Hybrid Mamba-Transformer Vision Backbone

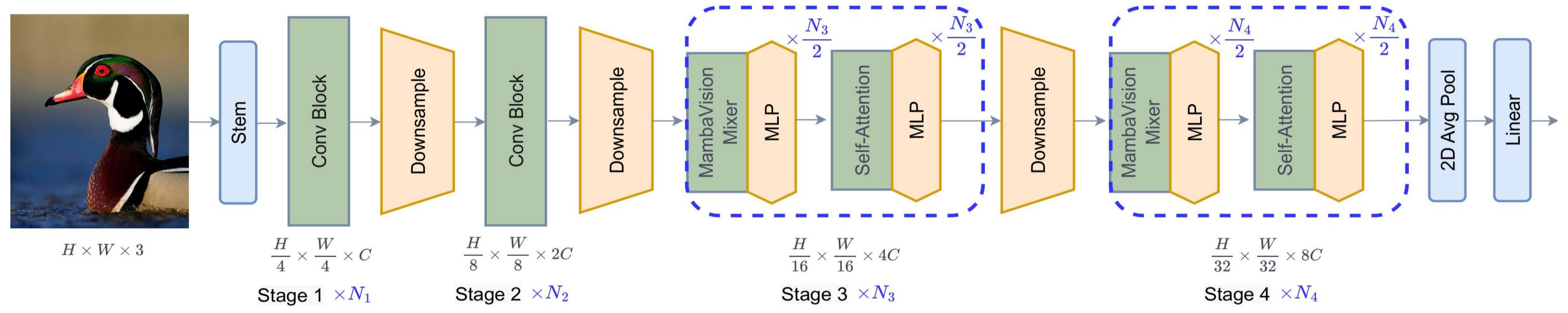
Ali Hatamizadeh, Jan Kautz

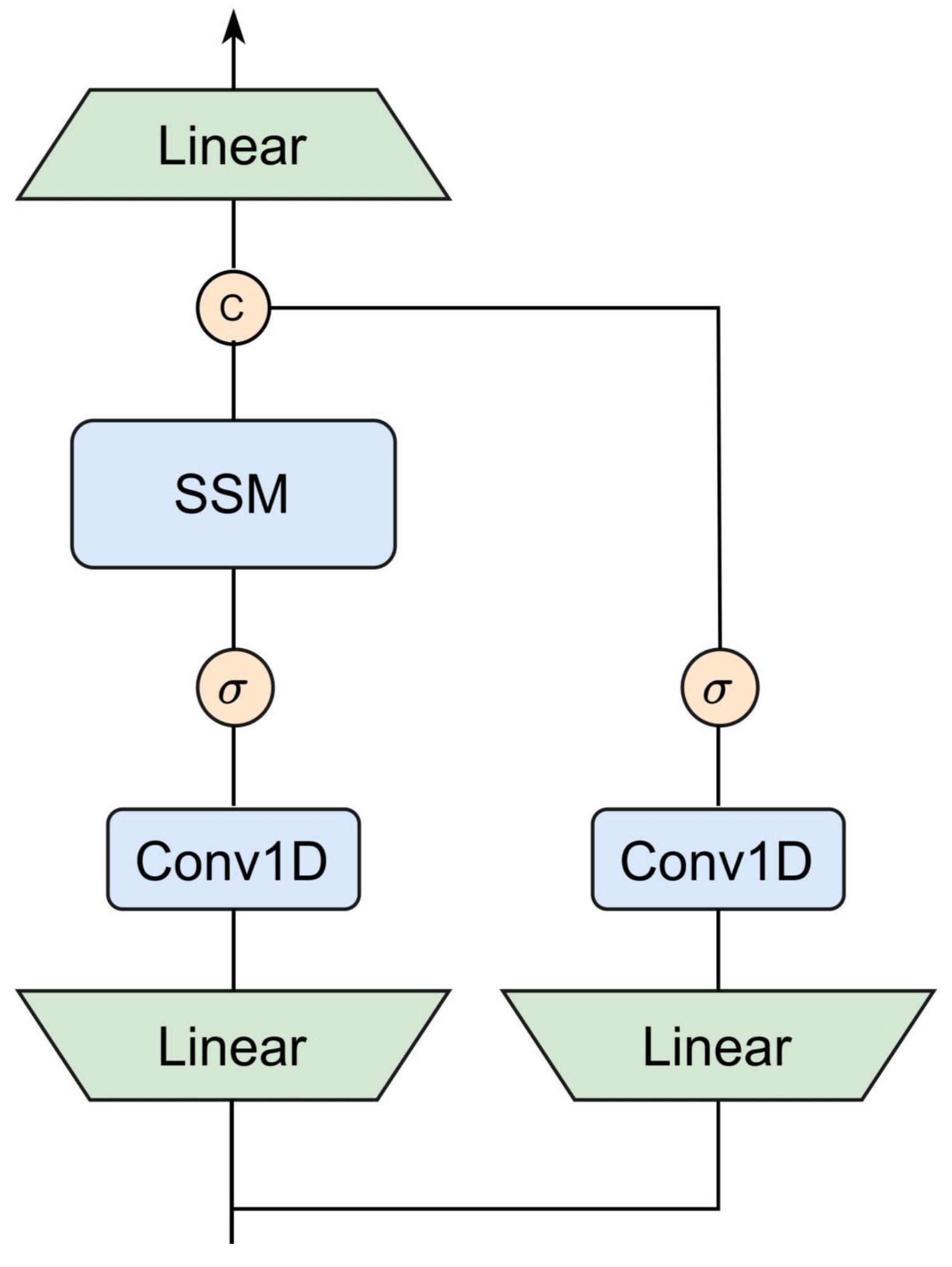
Introduction

- Vision Transformers excel at capturing global context but incur quadratic computational costs, while Mamba-based models run in linear time yet struggle with long-range spatial dependencies. **MambaVision** bridges this gap by redesigning the Mamba mixer to use non-causal, symmetric convolution branches and by adding self-attention in the final layers.

MambaVision Architecture

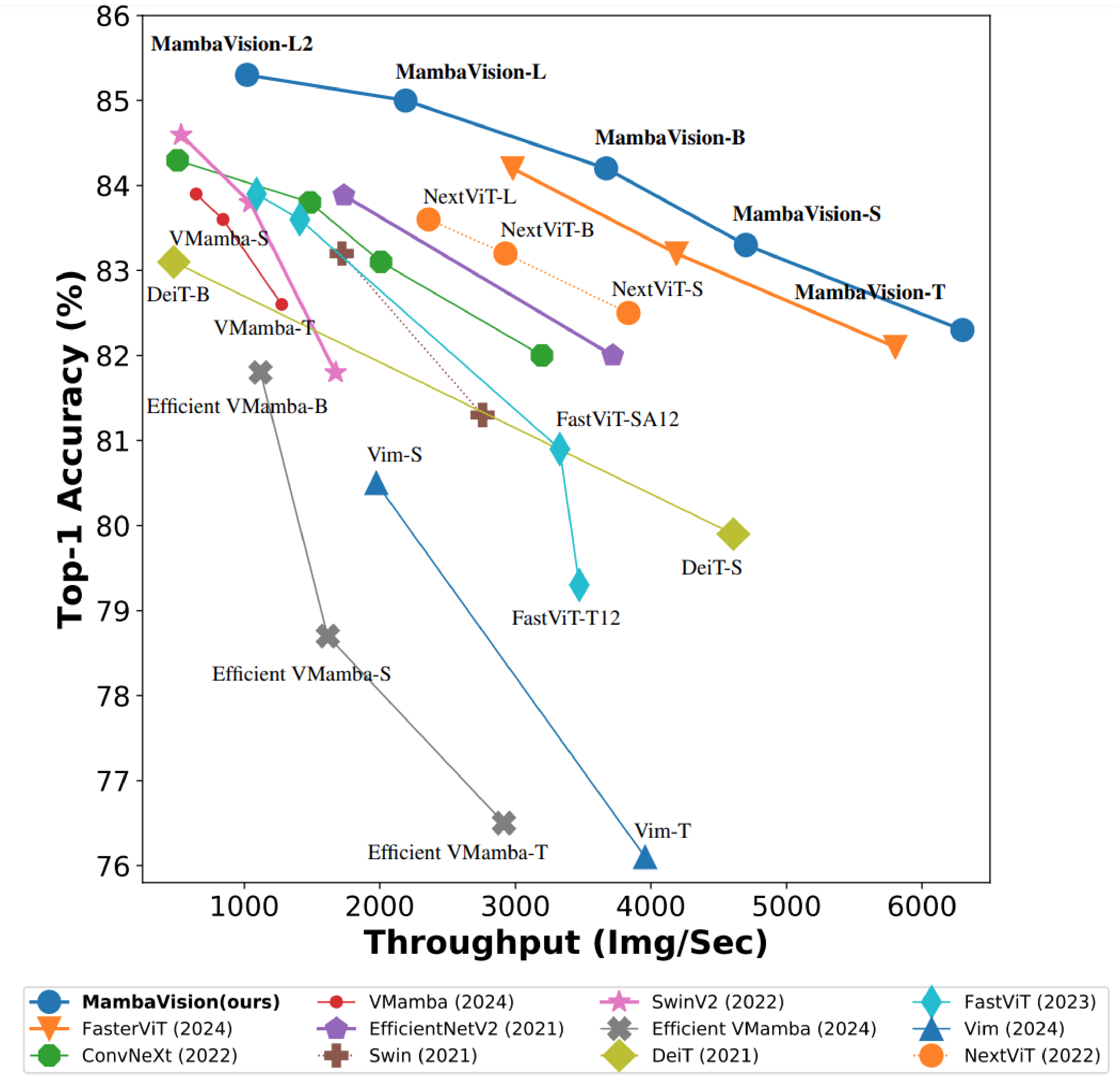
- **Hierarchical Four-Stage Backbone:** Stages 1–2 use lightweight CNN residual blocks and strided downsamplers for fast feature extraction.
- **Hybrid Token Mixing**: Within each of the last two stages, the first half of layers use the MambaVision mixer and the second half use multi-head self-attention to recover global context.





MambaVision Mixer

ImageNet-1K Pareto Front

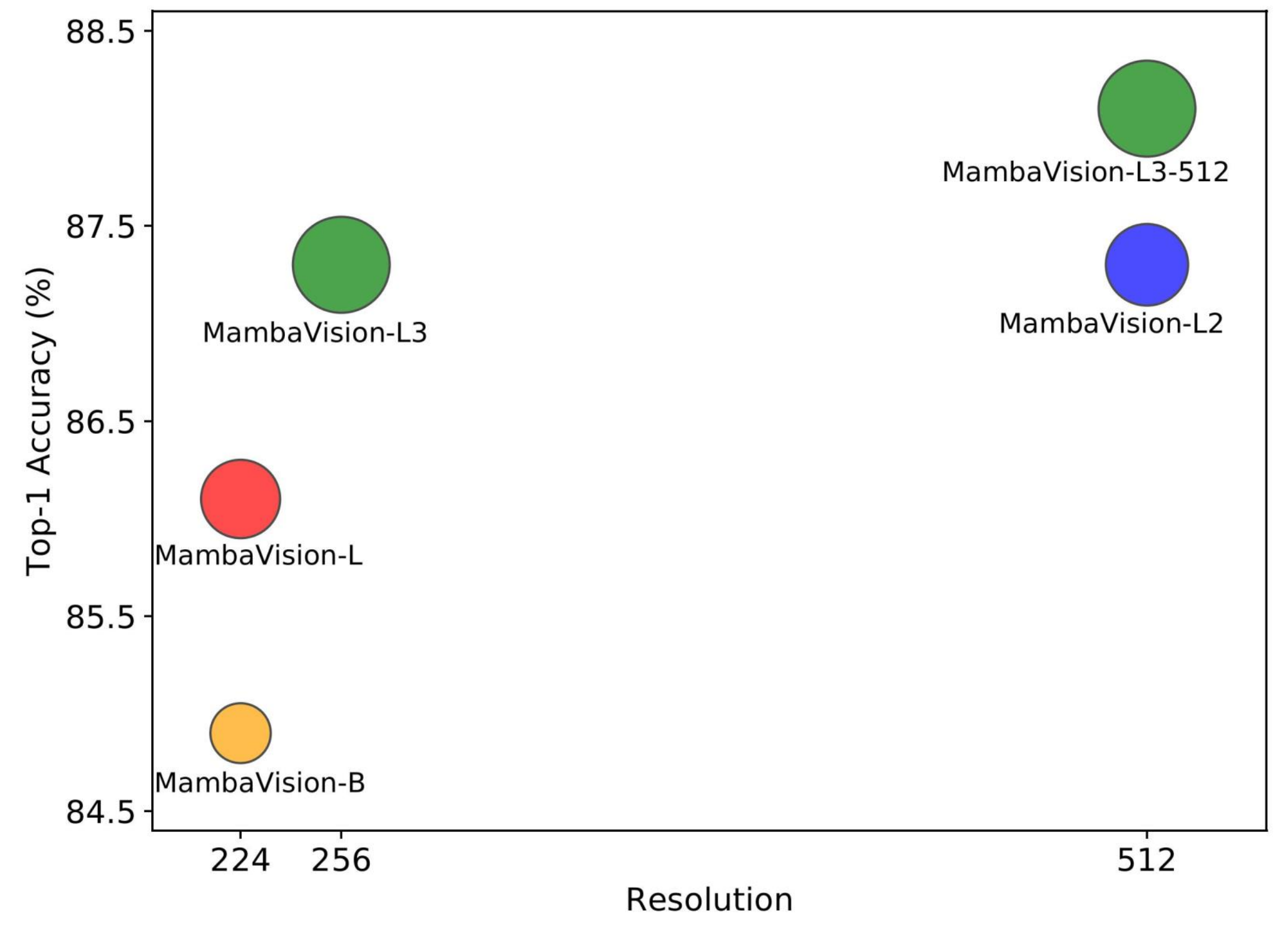


Detection & Segmentation (COCO)

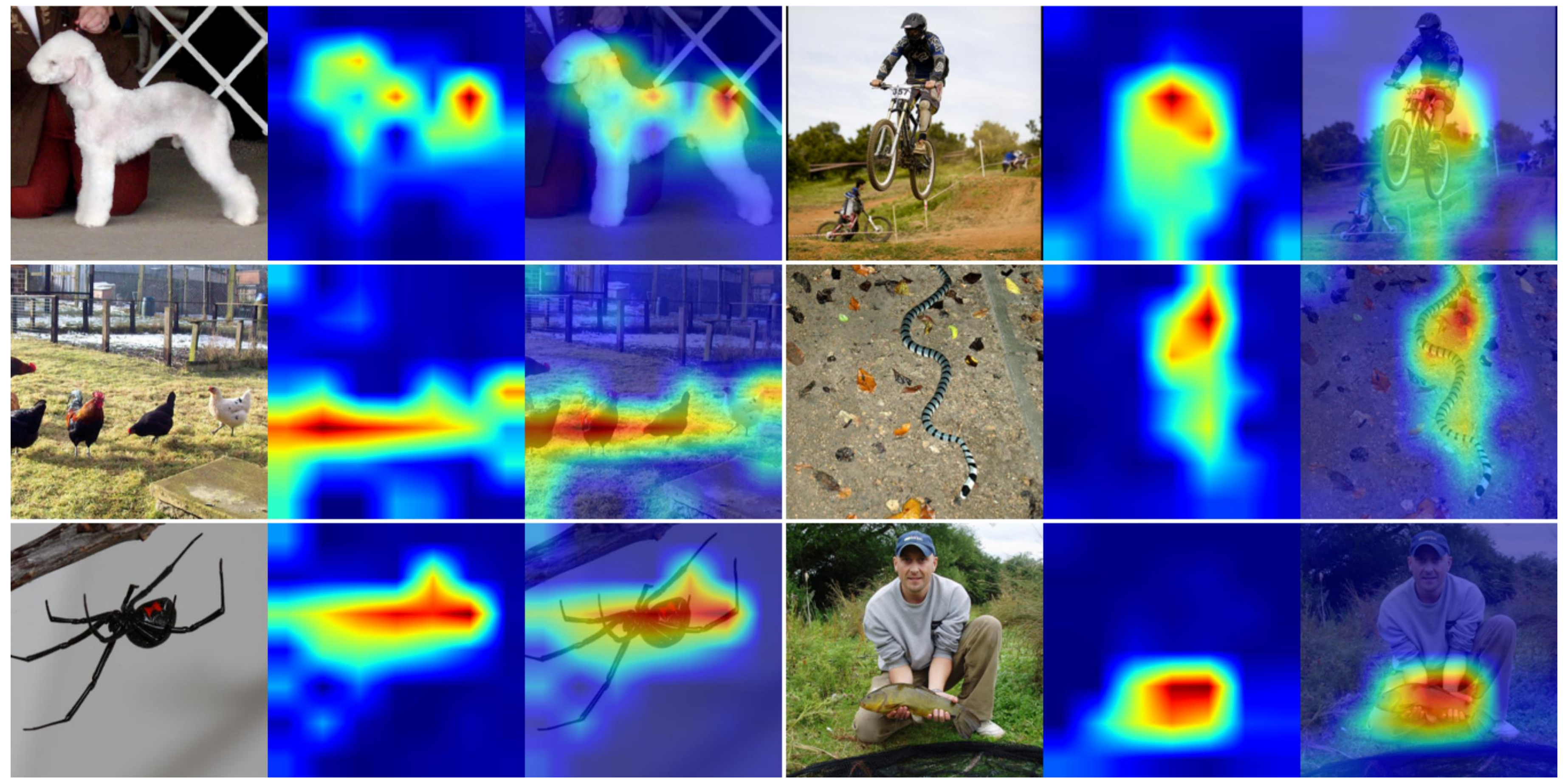
Backbone	Params (M)	FLOPs (G)	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
DeiT-Small/16 [28]	80	889	48.0	67.2	51.7	41.4	64.2	44.3
ResNet-50 [12]	82	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T [21]	86	745	50.4	69.2	54.7	43.7	66.6	47.3
ConvNeXt-T [23]	86	741	50.4	69.1	54.8	43.7	66.5	47.3
MambaVision-T	86	740	**51.1**	70.0	55.6	**44.3**	67.3	47.9
X101-32 [35]	101	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S [21]	107	838	51.9	70.7	56.3	45.0	68.2	48.8
ConvNeXt-S [23]	108	827	51.9	70.8	56.5	45.0	68.4	49.1
MambaVision-S	108	828	**52.3**	71.1	56.7	**45.2**	68.5	48.9
X101-64 [35]	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B [21]	145	982	51.9	70.5	56.4	45.0	68.1	48.9
ConvNeXt-B [23]	146	964	52.7	71.3	57.2	45.6	68.9	49.5
MambaVision-B	145	964	**52.8**	71.3	57.2	**45.7**	68.7	49.4

Scalability

- Pretrained on ImageNet-21K, achieving up to **88.1%** accuracy



Interpretability



Conclusion

- MambaVision combines SSM mixers with self-attention for fast, global context modeling, surpassing CNNs, Transformers, and Mamba backbones.
- It sets new benchmarks on major datasets and easily scales to larger data, higher resolutions.
- Code: https://github.com/NVlabs/MambaVision