Table 1: Detailed throughput comparison of Swin [2], ConvNeXt [3], and VMamba [1] models tested on a RTX 4090 GPU with a batch size of 128 at an image resolution of 224.

Model	**Top-1 Acc**(%)	**#Params**	**FLOPs** (G)	**Thru.** (imgs/sec)	**Train Thru.** (imgs/sec)	**Memory** (MB)
Swin-T [2]	81.3	28 M	4.5	1577	503	2906
ConvNeXt-T [3]	82.1	29 M	4.5	1800	570	1670
VMamba-T [1]	82.2	23 M	5.6	603	151	6639
MSVMamba-T	82.8	33 M	4.6	1092	331	4532
Swin-S [2]	83.0	50 M	8.7	970	309	2987
ConvNeXt-S [3]	83.1	50 M	8.7	1088	352	1753
VMamba-S [1]	83.5	44 M	11.2	425	106	6882
MSVMamba-S	84.1	51 M	9.2	665	232	4910
Swin-B [2]	83.5	88 M	15.5	654	219	3928
ConvNeXt-B [3]	83.8	89 M	15.4	729	240	2380
VMamba-B [1]	83.7	76 M	18.0	314	77	8853
MSVMamba-B	84.3	91 M	16.3	476	127	6347

Table 2: Efficiency comparison of models at increased input resolutions, tested on a RTX 4090 GPU with a batch size of 32.

Model	**Resolution**	**#Params**	**FLOPs** (G)	**Thru.** (imgs/sec)	**Memory** (MB)
Swin-T	768 / 640 / 512	28 M	70.7 / 45.0 / 26.6	61 / 108 / 213	20006 / 13051 / 5123
ConvNeXt-T	768 / 640 / 512	29 M	52.5 / 36.5 / 23.3	150 / 217 / 341	4733 / 3348 / 2214
VMamba-T	768 / 640 / 512	23 M	66.1 / 45.9 / 29.4	64 / 95 / 166	19293 / 13428 / 8628
MSVMamba-T	768 / 640 / 512	33 M	53.9 / 37.4 / 23.9	89 / 129 / 206	13051 / 9128 / 5919

Table 3: Ablation with more baselines. The CrossScan is utlized in VMamba, while the Uni-Scan and Bi-Scan denote Uni-directional Scan and Bi-directional Scan respectively.

Model-Name	**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**

Table 4: Detailed ablation results complementing fine-grained object detection and instance segmentation using Mask R-CNN on the COCO dataset.

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	$75.1_{\uparrow 5.5}$	$41.4_{\uparrow 3.3}$	$37.9_{\uparrow 2.3}$

References

[1] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.