Appendix for *VSSD: Vision Mamba with Non-Causal State Space Duality*

Supplementary Material

Image/Batch Size	ConvNeXt-T	Swin-T	VMamba-T	VSSD-T
224/128				
GFLOPs	4.5	4.5	4.9	5.0
Params. (M)	29	28	30	28
Memory (MB)	1670	2402	3204	1946
Acc. (%)	82.1	81.3	82.6	**83.8**
384/128				
GFLOPs	13.1	14.0	14.4	15.4
Memory (MB)	4654	8118	9169	5537
Acc. (%)	81.0	80.7	82.4	**83.5**
512/128				
GFLOPs	23.3	26.6	25.4	27.4
Memory (MB)	8181	19731	16204	9751
Acc. (%)	78.0	79.0	80.9	**81.9**
640/64				
GFLOPs	36.5	45.0	39.6	42.8
Memory (MB)	6417	20893	12710	7645
Acc. (%)	74.3	76.6	78.6	**79.4**
768/64				
GFLOPs	52.5	70.7	57.1	61.7
Memory (MB)	9189	OOM	18262	10954
Acc. (%)	69.5	73.1	74.7	**75.9**
1024/32				
GFLOPs	93.3	152.5	101.5	109.6
Memory (MB)	8182	OOM	16276	9750
Acc. (%)	55.4	61.9	62.3	**65.0**

Table 1. Performance comparison of VSSD-T against widely used vision models on ImageNet-1K across different image resolutions on an RTX 4090 GPU. OOM indicates out-of-memory errors.

1. Analyzing Generalization Ability Across Increasing Input Resolutions.

Following VMamba [3], we also present detailed comparison on ImageNet-1K with increasing image resolutions with CNN-based ConvNext [5], attention-based Swin [4] and SSM-based VMamba [3]. The detailed results are presented in Table 1. At the standard 224×224 resolution, VSSD-T achieves 83.8% top-1 accuracy, outperforming ConvNeXt-T (82.1%), Swin-T (81.3%), and VMamba-T (82.6%) while maintaining a competitive parameter count of 28M and GFLOPs count of 5.0. The performance advantage of VSSD-T becomes more pronounced at higher resolutions. At 384×384, our model achieves 83.5% accuracy, surpassing VMamba-T by 1.1 percentage points. This trend continues through 512×512 (81.9%), 640×640 (79.4%), and 768×768 (75.9%) resolutions, where VSSD-T consistently outperforms all competitors. Notably, at the 1024×1024 resolution, VSSD-T achieves 65.0% accuracy, significantly outperforming ConvNeXt-T (55.4%), Swin-T (61.9%), and VMamba-T (62.3%). Our VSSD also demonstrates significantly better memory efficiency than both Swin-T and VMamba-T. For instance, at 512×512 resolution and batch size of 128, VSSD-T consumes only 9751MB of memory compared to VMamba-T's 16204MB and Swin-T's 19731MB. At higher resolutions (768×768 and 1024×1024), Swin-T encounters out-of-memory errors, while VSSD-T continues to operate efficiently. These results highlight VSSD-T's exceptional balance between accuracy, computational efficiency, and memory usage, making it particularly well-suited for high-resolution image analysis tasks.

2. Additional Comparison.

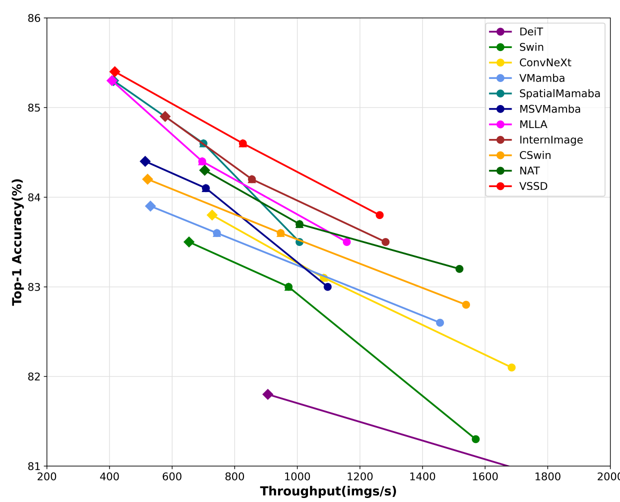

Figure 1. Efficiency comparison with more SOTA works with the same setting as the main paper.

To establish a more comprehensive analysis for our VSSD model, we present detailed comparisons with more advanced architectures specially designed for vision perception tasks, including ConvFormer [10], SG-Former [7], SMT [2], MaxViT [8], BiFormer [11], CAFormer [10], EfficientVMamba [6] and Groot-VL [9]. As shown in Table 2, we categorize the comparison across three model scales: tiny, small, and base according to parameter and FLOPs counts. Our VSSD consistently demonstrates superior performance across all scales when evaluated on the ImageNet-1K dataset. In the tiny model category, VSSD achieves 83.8% top-1 accuracy, matching BiFormer while outperforming other attention-based models like SG-Former (83.2%) and SSM-based models like GrootVL (83.4%). The performance advantage of VSSD extends to the base model category, where it reaches 85.4% accuracy, surpassing CAFormer (85.2%) and GrootVL (84.8%). This comprehensive comparison validates the effectiveness of our proposed VSSD as a powerful alternative to existing paradigms in vision model architecture. Additionally, we also provide efficiency comparison with more SOTA works listed in the main paper in Fig .1.

Method	Type	#Param.	FLOPs	Top-1 Acc(%)
Tiny Models				
ConvFormer [10]	Conv	27M	3.9G	83.0
SG-Former [7]	Attn	23M	4.8G	83.2
MaxViT [8]	Attn	31M	5.6G	83.6
BiFormer [11]	Attn	26M	4.5G	83.8
SMT-T [2]	Conv+Attn	20M	4.8G	83.7
CAFormer [10]	Attn	26M	4.1G	83.6
EffVMamba [6]	Conv+SSM	33M	4.0G	81.8
GrootVL [9]	SSM	30M	4.8G	83.4
VSSD	SSD	28M	5.0G	**83.8**
Small Models				
ConvFormer [10]	Conv	40M	7.6G	84.1
SG-Former [7]	Attn	39M	7.5G	84.1
MaxViT [8]	Attn	69M	11.7G	84.5
CAFormer [10]	Attn	39M	8.0G	84.5
BiFormer [11]	Attn	57M	9.8G	84.3
SMT-T [2]	Conv+Attn	32M	7.7G	84.3
GrootVL [9]	SSM	51M	8.5G	84.2
VSSD	SSD	50M	8.1G	**84.6**
Base Models				
ConvFormer [10]	Conv	57M	12.8G	84.5
SG-Former [7]	Attn	78M	15.6G	84.7
CAFormer [10]	Attn	56M	13.2G	85.2
MaxViT [8]	Attn	120M	23.4G	85.0
GrootVL [9]	SSM	91M	15.1G	84.8
VSSD	SSD	89M	16.1G	**85.4**

Table 2. **Additional Comparison across More Advanced Models on ImageNet-1K.**

3. More Detailed information of VSSD

More Details of the Proposed VSSD Model. The VSSD model initiates with a series of overlapping convolutions serving as the stem, followed by four progressive stages of processing. First three stages are equipped with VSSD Block, comprising a NC-SSD block and a FFN. We provide illustration in Fig. 2 for clarity. Besides, the detailed setting of VSSD variants are shown in the Tab. 3.

More Detailed Configuration of ImageNet-1K Training. Our experiments are conducted using the ImageNet-1K dataset [1]. Each model undergoes training for 300 epochs, which includes a 20-epoch warm-up phase. We employ the AdamW optimizer, setting the betas to (0.9, 0.999) and the momentum to 0.9. A cosine decay scheduler manages the learning rate, complemented by a weight decay rate of 0.05. The batch sizes and peak learning rates are set to 1024/1e-3 for the Tiny and Small models, and 2048/1.2e-3 for Base model, respectively. To enhance model accuracy and generalization, we incorporate exponential moving average (EMA) techniques and apply label smoothing with a coefficient of 0.1. The stochastic depth drop rates for our Tiny, Small, and Base models are set at 0.2, 0.4, and 0.6, respectively. Further details are provided in Tab. 4.

4. Limitations

This paper lacks experiments involving larger models and more extensive datasets, such as those using the ImageNet-22K benchmark [1]. Consequently, the scalability of the proposed VSSD model remains an area ripe for further exploration.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[2] Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. Scale-aware modulation meet transformer. In *ICCV*, 2023. 1, 2

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

[4] 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. 1

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 1

[6] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. 1, 2

[7] Sucheng Ren, Xingyi Yang, Songhua Liu, and Xinchao Wang. Sg-former: Self-guided transformer with evolving token reallocation. In *ICCV*, 2023. 1, 2

[8] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022. 1, 2

[9] Yicheng Xiao, Lin Song, Shaoli Huang, Jiangshan Wang, Siyu Song, Yixiao Ge, Xiu Li, and Ying Shan. Grootvl: Tree topology is all you need in state space model. In *NeurIPS*, 2024. 1, 2

[10] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

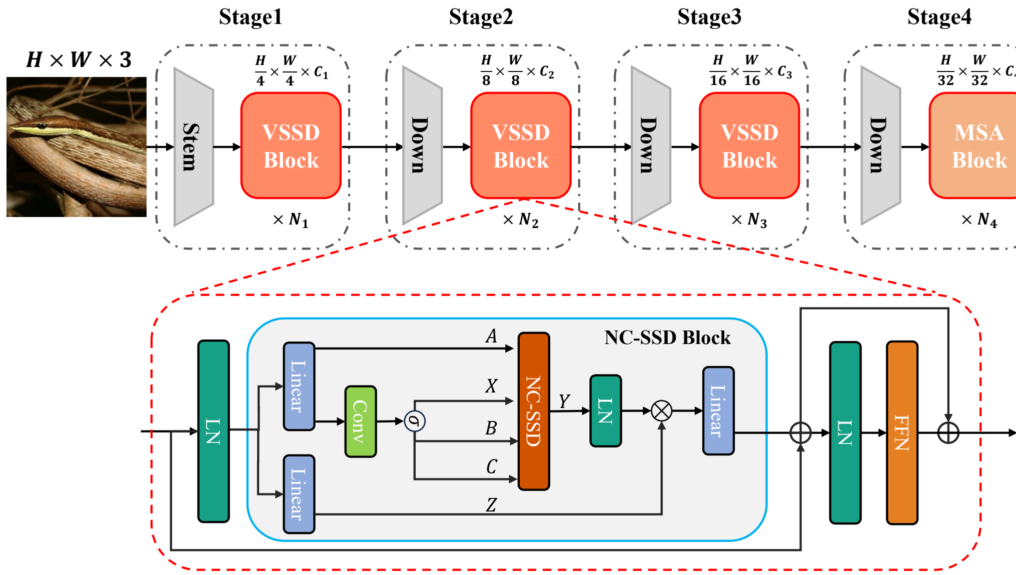

Figure 2. **Overall Architecture of the Proposed VSSD Model**. Local Perception Units (LPU) are omitted in this visualization for brevity.

Model	Blocks	Channels	Heads	SSD Ratio	#Param	FLOPs
VSSD-Tiny	[2, 2, 8, 2]	[96, 192, 384, 768]	[4, 4, 8, 16]	1	28M	5.0G
VSSD-Small	[2, 4, 15, 4]	[96, 192, 384, 768]	[4, 4, 8, 16]	1	51M	8.1G
VSSD-Base	[3, 4, 18, 5]	[96, 192, 384, 768]	[3, 6, 12, 24]	2	89M	16.1G

Table 3. **Model Specifications of VSSD varints.**

[11] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson Lau. Biformer: Vision transformer with bi-level routing attention. In *CVPR*, 2023. 1, 2

Settings	**Tiny**	**Small**	**Base**
Input resolution		224^2	
Epochs		300	
Batch size	1024	1024	2048
Optimizer		AdamW	
Adam ϵ		1e-8	
Adam (β_1, β_2)		(0.9, 0.999)	
Learning rate	1e-3	1e-3	1.2e-3
Learning rate decay		Cosine	
Warmup epochs		20	
Weight decay		0.05	
Rand Augment		rand-m9-mstd0.5-inc1	
Cutmix		1.0	
Mixup		0.8	
Cutmix-Mixup switch prob		0.5	
Random erasing prob		0.25	
Label smoothing		0.1	
Stochastic depth rate	0.2	0.4	0.6
Random erasing prob		0.25	
EMA decay rate		0.9999	

Table 4. **Detailed Configuration Parameters for ImageNet-1K Training.**