Mamba-ND: Selective State Space Modeling for Multi-Dimensional Data

Appendix

7 Additional Discussions

7.1 Effective Receptive Field (ERF)

We visualize the effective receptive field (ERF) [40] of Mamba-ND in Fig. 6. ERF is computed by setting the gradient of the center patch to one and backpropagating through the network. In particular, we visualize the ERF of ImageNet-1K pretrained weights of Mamba-ND against the Bi-directional and Uni-directional baseline (Tab. 6, row 5 and Tab. 6, row 4). Sensitivity is measured by the whiteness of a pixel in the visualization. The Uni-Directional model exhibits a sharp cutoff and is insensitive to all patches after it in the flattened sequence. This is clearly undesirable in data where causal relations do not exist. The Bi-Directional model exhibits a global receptive field, but it is heavily biased in the horizontal direction. This is also undesirable because images can have vertical structures. The Multi-Directional model exhibits a more uniform pattern in the sensitivity visualization, which explains their superior performance.

7.2 Depths versus Widths

In Sec. 5.5, we showed that a complicated multi-directional design may not necessarily lead to improved performance over the alternating-directional baseline. We hypothesize that this result is related to the effective depth of the model.

Each of the block-level designs can be represented by a directed acyclic graph (DAG) of layers, where the direction of edges implies the computation order. In this interpretation, designs other than H+H-W+W-T+T- are trading tree depths for widths. Existing literature [56] shows that depth is more important than width in deep neural networks. Our results seem to reaffirm this conclusion.

8 Additional Ablation Studies

8.1 Additional ImageNet and HMDB-51 Results.

HMDB-51 and UCF-101 training requires ImageNet pretrained weights. Due to architectural differences, different HMDB-51 experiments may load different ImageNet pretrained weights. For completeness, we provide full results indicating which weights are loaded for each experiment. All models are trained with a learning rate of 1e-3 and a patch size of 16 for both pretraining and finetuning. We report the results in Tab. 8.

In addition to the results already referred to in the main paper, we also incorporate additional experiments on applying an alternating design to Bi-SSM.

In particular, the default Bi-SSM design processes each layer in $L+$ and $L-$ direction, which is equivalent to $H+$ and $H-$. We experiment with processing the input in $H+$ and $H-$ direction in odd layers and in $W+$ and $W-$ direction in even layers. This leads to a considerable improvement in both ImageNet-1K (+1.5) and HMDB-51 (+8.4), highlighting the effectiveness of alternating directional designs.

We also tried to train models from scratch on HMDB-51. These models all converge very slowly or diverge. We failed to obtain useful results.

Table 8: Detailed Ablation Results of various combinations of design choices. In the 2D Layer column, we list the layer-level design for ImageNet pretraining. In the 3D Layer column, we list the layer-level design for finetuning. (.) denotes layer-level grouping. For Bi-SSM, each layer incorporates both the forward direction and reverse direction, hence no +/− distinction is needed.

Kernel	2D Layer	IN1K↑	3D Layer	HMDB-51↑
1D-SSM	L+	76.4	L+	34.9
1D-SSM	(H+H-)(W+W-)	76.5	(H+H-)(W+W-)(T+T-)	49.8
1D-SSM	H+H-W+W-	**79.4**	(H+H-W+W-)(T+T-)	47.4
			H+H-W+W-T+T-	**59.0**
1D-SSM	(L+L-)	76.3	(L+L-)	46.3
Bi-SSM*	L	74.6	L	32.1
Bi-SSM*	HW	76.1	HWT	40.5
ND-SSM	-	77.2	-	46.7
Multi-Head	-	77.6	-	51.5

8.2 Video-Specific Ablations

8.3 Δ Initialization

When initializing new SSM layers in T+T- direction, we can adjust the scale of Δ to control the initial temporal receptive field. We show such results in Tab. 9. We find that 1.0 is the optimal value, which coincides with the prior conclusion from S4ND [44].

8.4 Weight Inflation

There are various alternatives for inflating the positional embedding from 2D to 3D. We considered two policies. Given a 3D sequence of shape $H \times W \times T$, the first option copies the original 2D embedding of shape $H \times W$ by T times and scales the values to $\frac{1}{T}$. The second option copies the 2D embedding once and places it at location $\frac{T}{2}$ on the temporal axis. All other embeddings are initialized to 0. We find no significant differences in performance between these two designs.

Table 9: Ablation Study on the Scaling Factor of Δ for HMDB-51 Experiments. We report the performance on HMDB-51 under different initializations of temporal layers (T+ and T-). $\Delta = 1.0$ is the optimal choice.

Δ	HMDB-51↑
0.1	58.0
0.2	57.8
1.0	59.0
5.0	55.3

8.5 Ablation on 3D Segmentation

Because we train 3D segmentation models from scratch instead of from 2D-pretrained initializations, it may offer better insight into the design choices in the 3D space. However, it may also be the case that segmentation is local in nature and may require less global information. For completeness, we compare the performance of different SSM designs in Tab. 10. Results show that our simple design of alternating the scan direction between layers is still the best performing one. However, 1D-SSM with no bidirectional design also exhibits strong performance.

Table 10: Ablation Study on Layer Designs. We report top-1 accuracy on the ImageNet-1K validation set and HMDB-51 split 1. We report Dice score on BTCV dataset. The Alt-Directional design is the top-performing one.

		IN1K↑	HMDB-51 ↑	BTCV ↑
Alt-Directional	Block Level	79.4	59.0	81.5
Multi-Head	Layer-Level	77.6	51.5	81.4
ND-SSM	Layer-Level	77.2	46.7	80.2
1D-SSM	-	76.4	34.9	80.7
Bi-SSM	Layer-Level	74.6	32.1	77.9

8.6 Patch Size

One crucial element of Mamba-ND's success depends on its capability to achieve a global context at linear complexity. This allows it to use a lower patch size or higher resolution input given fixed computation capacity when compared with self attention based methods. We highlight this in figure Fig. 7. On ImageNet-1K, we explore the effect of different patch sizes. We report the results in Tab. 11.

8.7 Ordering

In the alternating directional setup, there is still room for exploration in the ordering of axes. For example, there is no reason to assume TWH+ is better

Table 11: Ablation Study on patch size. We report the performance on ImageNet-1K. We find that lower patch size achieves better result.

	patch size	IN1K↑
ViT-B	16	77.9
DeiT-S	16	79.8
DeiT-S	8	75.2
Swin-T	4	81.3
Mamba-2D-S	8	81.7
Mamba-2D-S	16	79.4

than WTH+ for H+. There is also no intuitive reason to prefer H+ over W+ as the direction of the first layer. This means the design space incorporates $2^3 \cdot 3! = 48$ possible choices (there are 2 options for each of H, W, T, and there are 3! orderings of the first three layers). Due to computational constraints, we are unable to explore all options. We attempted two incremental changes from the base setup on BTCV dataset: 1) Changing the order of the first three layers to WHT and TWH. 2) Swapping the order of TWH+ to WTH+. We found no meaningful difference between the results. We hypothesize that the residual connection is sufficient to mitigate the effect of different orderings.

9 Implementation Details

9.1 Model Size

Mamba-2D-S has 24 layers and a hidden size of 384. Mamba-2D-B has 24 layers and a hidden size of 768. Mamba-3D-S has 32 layers and a hidden size of 384. Mamba-3D-S+ has 32 layers and a hidden size of 512. Mamba-3D-B has 32 layers and a hidden size of 768. For ERA5 weather forecast, we use a 12 layer version of Mamba-3D-B. In terms of the parameter count, two Mamba Layers is roughly equivalent to one ViT block of the same dimension.

9.2 Additional Hyperparameters

On HMDB-51, we use a learning rate of 6.0e-4. On UCF-101, we use a learning rate of 1.0e-3. We set RandAug to $(9, 4)$ for both experiments. We do not use label smoothing or cutmix. We set the dropout and drop path rate both to 0.1. On ImageNet-1K, we use a learning rate of 1.0e-3. We set RandAug to (2, 10). We set label smoothing to 0.1, Mixup to 0.8 and CutMix to 1.0. We set the dropout and drop path rate both to 0.1. For all image and video ablations, we fix the learning rate to be 1e-3 and use a patch size of 16. On BTCV, we use a learning rate of 1.0e-4 and a drop rate of 0.1. On ERA5, we use a learning rate of 5.0e-4 and a drop rate of 0.1.

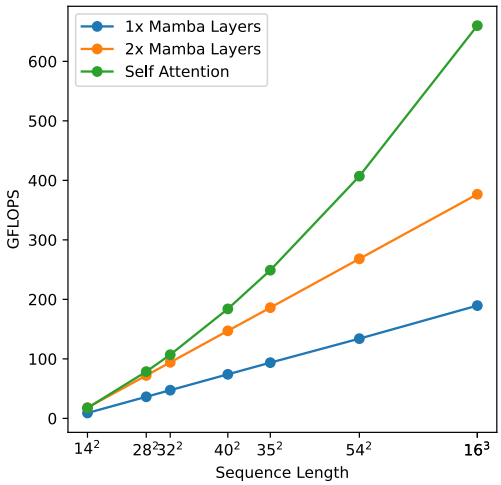

Fig. 7: Mamba-ND Computational Complexity with Respect to Sequence Length. We plot the computational cost of the standard 12-layer Vision Transformer (ViT-B), and comparable Mamba-2D with 12 and 24 Mamba layers. In terms of parameter count, 2 Mamba layers are roughly equivalent to 1 ViT layer, which includes a Multi-head Attention (MHA) module and a Feed Forward Network (FFN). 14^2 reflects the standard 224×224 image with a patch size of 16. Sequence length increase can either reflect a lower patch size or a higher input resolution.

10 Limitations

While we have provided a relatively comprehensive set of experiments to explore the design space of multi-dimensional selective state-space models, there remain unexplored areas due to the exponential number of possible orderings. In this paper, we primarily explored various combinations of row-major ordering. However, there may be other reasonable choices such as diagonal or zig-zag patterns. We leave these possibilities for future work.

11 Concurrent Works

There have been recent works such as VisionMamba [63], VMamba [35], and U-Mamba [41] that focus on applying Mamba to specific application fields such as classification or segmentation. Contrary to these works, which have a limited design space, Mamba-ND explores the generic recipe to adopt Mamba-Kernel to multi-dimensional data, which is under-explored. For example, VisionMamba can be considered as a variant of Bi-SSM with additional designs, and VMamba can be considered as hierarchical ND-SSM with additional designs.

References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems **34**, 24206–24221 (2021)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Baron, E., Zimerman, I., Wolf, L.: 2-d ssm: A general spatial layer for visual transformers. arXiv preprint arXiv:2306.06635 (2023)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
5. Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. arXiv preprint arXiv:2211.02556 (2022)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5659–5667 (2017)
8. Christoph, R., Pinz, F.A.: Spatiotemporal residual networks for video action recognition. Advances in neural information processing systems **2** (2016)
9. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in neural information processing systems **34**, 3965–3977 (2021)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848
11. Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., Yuan, L.: Davit: Dual attention vision transformers. In: European Conference on Computer Vision. pp. 74–92. Springer (2022)
12. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134 (2022)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
15. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)

16. Gu, A., Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces (Dec 2023). https://doi.org/10.48550/arXiv.2312.00752
17. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: The International Conference on Learning Representations (ICLR) (2022)
18. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)
19. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al.: Era5 hourly data on single levels from 1979 to present. Copernicus climate change service (c3s) climate data store (cds) **10**(10.24381) (2018)
22. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
23. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
24. Islam, M.M., Bertasius, G.: Long movie clip classification with state-space video models. arXiv preprint arXiv:2204.01692 (2022)
25. Islam, M.M., Hasan, M., Athrey, K.S., Braskich, T., Bertasius, G.: Efficient movie scene detection using state-space transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18749–18758 (2023)
26. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
28. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 780–787 (2014)
29. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
30. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
31. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten Digit Recognition with a Back-Propagation Network. In: Advances in Neural Information Processing Systems. vol. 2. Morgan-Kaufmann (1989)
32. Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z.: Attention mechanism-based cnn for facial expression recognition. Neurocomputing **411**, 340–350 (2020)

33. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)
34. Lin, X., Petroni, F., Bertasius, G., Rohrbach, M., Chang, S.F., Torresani, L.: Learning to recognize procedural activities with distant supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13853–13863 (2022)
35. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
36. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
37. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
38. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
39. Lu, J., Mottaghi, R., Kembhavi, A., et al.: Container: Context aggregation networks. Advances in neural information processing systems **34**, 19160–19171 (2021)
40. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems **29** (2016)
41. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
42. Mangalam, K., Fan, H., Li, Y., Wu, C.Y., Xiong, B., Feichtenhofer, C., Malik, J.: Reversible Vision Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10830–10840 (2022)
43. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3163–3172 (2021)
44. Nguyen, E., Goel, K., Gu, A., Downs, G.W., Shah, P., Dao, T., Baccus, S.A., Ré, C.: S4nd: Modeling images and videos as multidimensional signals using state spaces. Advances in Neural Information Processing Systems **35** (2022)
45. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., Grover, A.: Climax: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343 (2023)
46. Nguyen, T., Jewik, J., Bansal, H., Sharma, P., Grover, A.: Climatelearn: Benchmarking machine learning for weather and climate modeling. arXiv preprint arXiv:2307.01909 (2023)
47. Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Madireddy, S., Maulik, R., Kotamarthi, V., Foster, I., Grover, A.: Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. arXiv preprint arXiv:2312.03876 (2023)
48. Poli, M., Massaroli, S., Nguyen, E., Fu, D.Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., Ré, C.: Hyena hierarchy: Towards larger convolutional language models. arXiv preprint arXiv:2302.10866 (2023)
49. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

50. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
51. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
52. Tian, Y., Xie, L., Wang, Z., Wei, L., Zhang, X., Jiao, J., Wang, Y., Tian, Q., Ye, Q.: Integrally pre-trained transformer pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18610–18620 (2023)
53. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
54. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
55. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. pp. 1747–1756. PMLR (2016)
56. Vardi, G., Yehudai, G., Shamir, O.: Width is less important than depth in relu neural networks. In: Conference on Learning Theory. pp. 1249–1281. PMLR (2022)
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
58. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
59. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
60. Zhang, B., Wang, L., Wang, Z., Qiao, Y., Wang, H.: Real-time action recognition with enhanced motion vector cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2718–2726 (2016)
61. Zhang, B., Wang, L., Wang, Z., Qiao, Y., Wang, H.: Real-time action recognition with deeply transferred motion vector cnns. IEEE Transactions on Image Processing **27**(5), 2326–2339 (2018)
62. Zhang, X., Tian, Y., Xie, L., Huang, W., Dai, Q., Ye, Q., Tian, Q.: Hivit: A simpler and more efficient design of hierarchical vision transformer. In: The Eleventh International Conference on Learning Representations (2022)
63. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model (2024)