MAP: Unleashing Hybrid Mamba-Transformer Vision Backbone's Potential with Masked Autoregressive Pretraining

Supplementary Material

Additional Experiments on 3D Tasks

To verify that our method not only performs well on 2D tasks but is also effective for 3D tasks, we conducted experiments on additional 3D tasks. In addition to the ModelNet40[26] classification experiment, we also tested on the more challenging ScanObjectNN[21] dataset. Unlike ModelNet40, which is a synthetic dataset, ScanObjectNN is a real-world object dataset and is commonly evaluated under three settings: OBJ_BG, OBJ_ONLY, and PB_T50_RS. Among these, the PB_T50_RS setting is the most challenging. Comparing the results of HybridNet and Mamba3D under the Supervised Learning Only setting reveals that HybridNet performs only slightly better than Mamba3D. However, both HybridNet and Mamba3D achieve significant performance improvements after MAP pretraining. This further validates that the MAP pretraining strategy is not only effective for hybrid frameworks but also enhances the pure Mamba framework. Comparing the results of Mamba3D under Point-BERT, Point-MAE, and MAP, it is evident that MAP demonstrates a significant performance advantage. This proves that even within the pure Mamba framework, MAP can surpass the performance of BERT-style and MAE-style pertaining. We also conducted experiments on the few-shot learning task of ModelNet40 to validate the effectiveness of MAP. After MAP pretraining, both HybridNet and Mamba3D achieved significant performance improvements. On the more fine-grained task of ShapeNetPart[28] part segmentation, we also demonstrated that MAP can bring significant performance improvements to both hybrid frameworks and the pure Mamba framework.

Method	5-way		10-way	
	10-shot ↑	20-shot ↑	10-shot ↑	20-shot ↑
Supervised Learning Only				
DGCNN [24]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
Transformer [22]	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
Mamba3D [7]	92.6 ± 3.7	96.9 ± 2.4	88.1 ± 5.3	93.1 ± 3.6
HybridNet	92.8 ± 3.2	97.0 ± 1.8	88.4 ± 4.3	93.1 ± 3.8
with Self-supervised pretraining				
DGCNN+*OcCo*[23]	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
OcCo [23]	94.0 ± 3.6	95.9 ± 2.7	89.4 ± 5.1	92.4 ±4.6
PointMamba [9]	95.0 ± 2.3	97.3 ± 1.8	91.4 ± 4.4	92.8 ± 4.0
MaskPoint [11]	95.0 ± 3.7	97.2 ± 1.7	91.4 ± 4.0	93.4 ± 3.5
Point-BERT [29]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
Point-MAE [14]	96.3 ± 2.5	97.8± 1.8	92.6 ±4.1	95.0 ± 3.0
Mamba3d+*P-B* [29]	95.8 ± 2.7	97.9 ± 1.4	91.3 ± 4.7	94.5 ± 3.3
Mamba3d+*P-M* [14]	96.4 ± 2.2	98.2 ±1.2	92.4 ± 4.1	95.2 ± 2.9
Mamba3d+*MAP*	97.1 ± 3.1	98.7 ±1.3	92.8 ± 2.1	95.8 ± 3.1
HybridNet+*MAP*	97.3 ±2.8	98.7 ±0.8	93.0 ± 3.6	96.0 ± 2.7

Table 1. Few-shot classification on ModelNet40 dataset. Overall accuracy (%) without voting is reported. *P-B* and *P-M* represent Point-BERT and Point-MAE strategy, respectively.

Method	PT	#P ↓	#F ↓	ScanObjectNN		
				OBJ_BG ↑	OBJ_ONLY ↑	PB_T50_RS ↑
Supervised Learning Only: Dedicated Architectures						
PointNet[16]	×	3.5	0.5	73.3	79.2	68.0
PointNet++[17]	×	1.5	1.7	82.3	84.3	77.9
DGCNN[24]	×	1.8	2.4	82.8	86.2	78.1
PointCNN[8]	×	0.6	-	86.1	85.5	78.5
DRNet [19]	×	-	-	-	-	80.3
SimpleView[4]	×	-	-	-	-	80.5±0.3
GBNet[20]	×	8.8	-	-	-	81.0
PRA-Ne[3]	×	-	2.3	-	-	81.0
MVTN[6]	×	11.2	43.7	92.6	92.3	82.8
PointMLP[13]	×	12.6	31.4	-	-	85.4±0.3
PointNeXt[18]	×	1.4	3.6	-	-	87.7±0.4
P2P-HorNet[25]	✓	-	34.6	-	-	89.3
DeLA[2]	×	5.3	1.5	-	-	90.4
Supervised Learning Only: Transformer or Mamba-based Models						
Transformer	×	22.1	4.8	79.86	80.55	77.24
PCT[5]	×	2.9	2.3	-	-	-
PointMamba[10]	×	12.3	3.6	88.30	87.78	82.48
PCM[30]	×	34.2	45.0	-	-	88.10±0.3
SPoTr[15]	×	1.7	10.8	-	-	88.60
PointConT[12]	×	-	-	-	-	90.30
Mamba3d w/o vot.[7]	×	16.9	3.9	92.94	92.08	91.81
Mamba3d w/ vot.[7]	×	16.9	3.9	94.49	92.43	92.64
HybridNet w/o vot.	×	19.3	4.4	92.81	92.28	91.97
HybridNet w/ vot.	×	19.3	4.4	94.50	92.58	92.66
With Self-supervised pretraining						
Transformer	*OcCo*	22.1	4.8	84.85	85.54	78.79
Point-BERT	*IDPT*	22.1+1.7†	4.8	88.12	88.30	83.69
MaskPoint	*MaskPoint*	22.1	4.8	89.30	88.10	84.30
PointMamba	*Point-MAE*	12.3	3.6	90.71	88.47	84.87
Point-MAE	*IDPT*	22.1+1.7†	4.8	91.22	90.02	84.94
Point-M2AE	*Point-M2AE*	15.3	3.6	91.22	88.81	86.43
Mamba3d w/o vot.	*Point-BERT*	16.9	3.9	92.25	91.05	90.11
Point-MAE	*Point-MAE*	22.1	4.8	90.02	88.29	85.18
Mamba3d w/o vot.	*Point-MAE*	16.9	3.9	93.12	92.08	92.05
Mamba3d w/ vot.	*Point-MAE*	16.9	3.9	95.18	94.15	93.05
Mamba3d w/o vot.	*MAP*	16.9	3.9	93.62	92.75	92.65
Mamba3d w/ vot.	*MAP*	16.9	3.9	95.64	94.87	93.76
HybridNet w/o vot.	*MAP*	19.3	4.4	93.88	93.03	92.95
HybridNet w/ vot.	*MAP*	19.3	4.4	95.84	94.97	93.87

Table 2. Results on 3D classification tasks. Our results are highlighted in blue. PT: pre-training strategy.

Method	mIoU$_C$ (%) ↑	mIoU$_I$ (%) ↑	#P ↓	#F ↓
Supervised Learning Only				
PointNet [16]	80.4	83.7	3.6	4.9
PointNet++ [17]	81.9	85.1	1.0	4.9
DGCNN [24]	82.3	85.2	1.3	12.4
Transformer [22]	83.4	85.1	27.1	15.5
Mamba3D[7]	83.7	85.7	23.0	11.8
HybridNet	83.5	85.6	25.1	12.9
with Self-supervised pretraining				
OcCo [23]	83.4	84.7	27.1	-
PointContrast [27]	-	85.1	37.9	-
CrossPoint [1]	-	85.5	-	-
Point-MAE [14]	84.2	86.1	27.1	15.5
PointMamba [9]	84.4	86.0	17.4	14.3
Point-BERT [29]	84.1	85.6	27.1	10.6
Mamba3d+*P-B* [29]	84.1	85.7	21.9	9.5
Mamba3d+*P-M* [14]	84.3	85.8	23.0	11.8
Mamba3d+*MAP*	84.5	86.0	23.0	11.8
HybridNet+*MAP*	84.7	86.3	25.1	12.9

Table 3. Part segmentation on ShapeNetPart dataset. Our results are highlighted in blue. The class mIoU (mIoU$_C$) and the instance mIoU (mIoU$_I$) are reported, with model parameters #P (M) and FLOPs #F (G).

References

[1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022.

[2] Binjie Chen, Yunzhou Xia, Yu Zang, Cheng Wang, and Jonathan Li. Decoupled local aggregation for point cloud learning. *arXiv preprint arXiv:2308.16532*, 2023.

[3] Silin Cheng, Xiwu Chen, Xinwei He, Zhe Liu, and Xiang Bai. Pra-net: Point relation-aware network for 3d point cloud analysis. *IEEE Trans. Image Process. (TIP)*, 30:4436–4448, 2021.

[4] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 3809–3820. PMLR, 2021.

[5] Meng-Hao Guo, Junxiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. PCT: point cloud transformer. *Comput. Vis. Media*, 7(2):187–199, 2021.

[6] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. MVTN: multi-view transformation network for 3d shape recognition. In *Int. Conf. Comput. Vis. (ICCV)*, pages 1–11. IEEE, 2021.

[7] Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. *arXiv preprint arXiv:2404.14966*, 2024.

[8] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 828–838, 2018.

[9] Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024.

[10] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024.

[11] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[12] Yahui Liu, Bin Tian, Yisheng Lv, Lingxi Li, and Fei-Yue Wang. Point cloud classification using content-based transformer via clustering in feature space. *IEEE/CAA Journal of Automatica Sinica*, 2023.

[13] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. In *Int. Conf. Learn. Represent. (ICLR)*. OpenReview.net, 2022.

[14] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[15] Jinyoung Park, Sanghyeok Lee, Sihyeon Kim, Yunyang Xiong, and Hyunwoo J Kim. Self-positioning point-based transformer for point cloud understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 21814–21823, 2023.

[16] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 77–85, 2017.

[17] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 5099–5108, 2017.

[18] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022.

[19] Shi Qiu, Saeed Anwar, and Nick Barnes. Dense-resolution network for point cloud classification and segmentation. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 3812–3821, 2021.

[20] Shi Qiu, Saeed Anwar, and Nick Barnes. Geometric back-projection network for point cloud classification. *IEEE Trans. Multimedia (TMM)*, 24:1943–1955, 2022.

[21] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1588–1597, 2019.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 5998–6008, 2017.

[23] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Int. Conf. Comput. Vis. (ICCV)*, pages 9782–9792, 2021.

[24] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph.*, 38(5):146:1–146:12, 2019.

[25] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2P: tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022.

[26] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[27] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 574–591. Springer, 2020.

[28] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

[29] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022.

[30] Tao Zhang, Xiangtai Li, Haobo Yuan, Shunping Ji, and Shuicheng Yan. Point could mamba: Point cloud learning via state space model. *arXiv preprint arXiv:2403.00762*, 2024.