PointNet4D: A Lightweight 4D Point Cloud Video Backbone for Online and Offline Perception in Robotic Applications

Supplementary Material

Implementation Details

In our approach, we adopted PointNet++ followed by five hybrid temporal fusion layers as the standard backbone for PointNet4D. In the action segmentation task, the network processes variable-length inputs per time step, from 1 up to 150 frames. In semantic segmentation, the network inputs the current frame along with the previous two frames to output dense semantic predictions for the current frame. For pretraining PointNet4D, we employed a 4-layer Transformer Decoder with a masked autoregressive approach for frame reconstruction. The pretrained encoder then initializes PointNet4D and is fine-tuned for downstream tasks. As mentioned in the method section, the only difference between PointNet4D++ and PointNet is that PointNet4D++ uses PointTransformer[1] as the feature extractor for per-frame point clouds. The pretraining strategy corresponding to PointNet4D++ is referred to as 4DMAP++. All experiments were conducted on 8 A800 GPUs. Pretraining was performed over 300 epochs, followed by 80 epochs of fine-tuning. Using a batch size of 32, we adopted a learning rate of 0.03. The other settings are consistent with P4Transformer[2] and PPTr[1].

Datasets Details for Other Offline Tasks

MSRAction-3D[3]: This dataset contains 567 videos across 20 daily action categories, with each video averaging around 40 frames. We followed the standard setup, using 270 videos for training and 297 for testing. We use the 24 frames as default. NTU-RGBD[4]: Comprising 56,880 videos in 60 fine-grained action categories, video lengths range from 30 to 300 frames. In the cross-subject setting, we split the dataset into 40,320 training and 16,560 testing videos. SHREC'17[5]: This dataset includes 2800 videos across 28 gesture classes. NvGesture[6]: Composed of 1532 videos covering 25 gesture classes, with 1050 videos for training and 482 for testing. Synthia 4D[7]: A synthetic outdoor driving dataset, Synthia4D generates 3D videos based on the Synthia dataset, capturing six driving scenarios with moving objects and cameras. We followed previous work, splitting the dataset into 19,888 training, 815 validation, and 1,886 testing frames.

Additional Results on Offline Tasks

3D Action Recognition on NTU-RGBD. We further validated the effectiveness of our approach for offline tasks on additional datasets. On NTU-RGBD, our PointNet4D achieved state-of-the-art results under supervised training and comparable results to M2PSC after pretraining with 4DMAP. Notably, M2PSC utilized human pose tracking for pretraining, whereas our pretraining did not incorporate any human priors. Beyond human action classification, our method also demonstrated competitive performance in gesture recognition tasks. These experiments further support the generality and effectiveness of our approach.

Methods	Acc.
3DV-Motion [8]	84.5
3DV-PointNet++ [8]	88.8
PSTNet [9]	90.5
PSTNet++ [10]	91.4
Kinet [11]	92.3
P4Transformer [2]	90.2
PST-Transformer [12]	91.0
PointNet4D [2]	90.5
PSTNet + PointCPSC[13] (50% Semi-supervised)	88.0
PSTNet + PointCMP[14] (50% Semi-supervised)	88.5
PSTNet + CPR[15] (End-to-end Fine-tuning)	91.0
P4Transformer + MaST-Pre[16] (50% Semi-supervised)	87.8
P4Transformer + MaST-Pre[16] (End-to-end Fine-tuning)	90.8
P4Transformer + M2PSC[17] (50% Semi-supervised)	88.7
P4Transformer + M2PSC[17] (End-to-end Fine-tuning)	91.3
4DMAP (50% Semi-supervised)	88.8
4DMAP (End-to-end Fine-tuning)	90.9

Table 1. Action recognition accuracy (%) on NTU-RGBD.

Offline 4D Action Recognition Tasks on MSRAction3D, SHREC'17 and NvGesture. We also conducted experiments on the 4D action recognition task on MSRAction3D. PointNet4D-PST refers to our model, which is built upon the PST-Transformer and incorporates our hybrid Mamba-Transformer layer. While M2PSC is a framework specifically designed for human point cloud video analysis, utilizing trajectory tracking of human points for pretraining, our method, as a general-purpose architecture, has already demonstrated competitive performance when compared to specialized methods tailored to the human domain. In addition, we have validated the effectiveness of PointNet4D and 4DMAP on the SHREC'17 [5] and NvGesture [6] gesture recognition tasks.

Offline Semantic Segmentation on Synthia 4D. We conducted experiments on an outdoor autonomous driving dataset to validate the effectiveness of our method. Our approach continues to yield significant performance improvements on this dataset, underscoring its versatility across various domains and data formats. This demonstrates the potential of our method as a robust general-purpose 4D backbone.

Methods		Accuracy (%)
Supervised Learning	MeteorNet [18]	88.50
	PSTNet [9]	91.20
	PSTNet++ [10]	92.68
	Kinet [11]	93.27
	PPTr [1]	92.33
	P4Transformer [2]	90.94
	PST-Transformer [12]	93.73
	Mamba4D	92.68
	PointNet4D	91.61
	PointNet4D-PST	93.75
End-to-end Fine-tuning	PSTNet + PointCPSC [13]	92.68
	PSTNet + CPR [15]	93.03
	PSTNet + PointCMP [14]	93.27
	P4Transformer + MaST-Pre [16]	91.29
	PST-Transformer + MaST-Pre [16]	94.08
	P4Transformer + M2PSC [17]	93.03
	PST-Transformer + M2PSC [17]	94.84
	4DMAP	92.65
	4DMAP-PST	94.76

Table 2. Action recognition accuracy on MSRAction-3D.

Methods	NvG	SHR
FlickerNet [19]	86.3	-
PLSTM [20]	85.9	87.6
PLSTM-PSS [20]	87.3	93.1
Kinet [11]	89.1	95.2
P4Transformer [2] (30 Epochs)	84.8	87.5
P4Transformer [2] (50 Epochs)	87.7	91.2
PointNet4D (30 Epochs)	85.0	87.7
PointNet4D (50 Epochs)	87.7	91.0
P4Transformer + MaST-Pre[16] (30 Epochs)	87.6	90.2
P4Transformer + MaST-Pre[16] (50 Epochs)	89.3	92.4
P4Transformer + M2PSC[17] (30 Epochs)	88.0	90.9
P4Transformer + M2PSC[17] (50 Epochs)	89.6	92.8
4DMAP (30 Epochs)	87.8	90.5
4DMAP (50 Epochs)	89.6	92.6

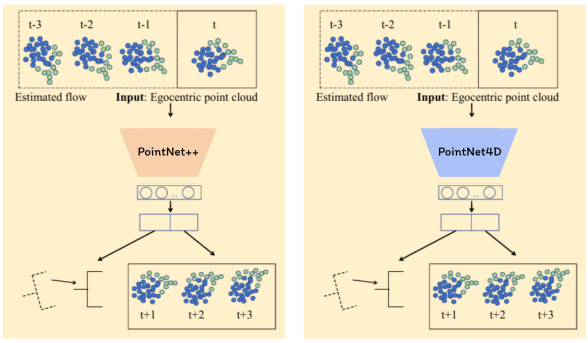
Table 3. Gesture recognition accuracy (%) on NvG and SHR.

Offline Setting	Clip Length	Bldn	Road	Sdwlk	Fence	Vegittn	Pole	Car	T.Sign	Pedstrn	Bicycl	Lane	T.Light	mIoU
3D MinkNet14 [21]	1	89.39	97.68	69.43	86.52	98.11	97.26	93.50	79.45	92.27	0.00	44.61	66.69	76.24
4D MinkNet14 [21]	3	90.13	98.26	73.47	87.19	99.10	97.50	94.01	79.04	92.62	0.00	50.01	68.14	77.24
PointNet++ [22]	1	96.88	97.72	86.20	92.75	97.12	97.09	90.85	66.87	78.64	0.00	72.93	75.17	79.35
MeteorNet-m [21]	2	98.22	97.79	90.98	93.18	98.31	97.45	94.30	76.35	81.05	0.00	74.09	75.92	81.47
MeteorNet-l [21]	3	98.10	97.72	88.65	94.00	97.98	97.65	93.83	84.07	80.90	0.00	71.14	77.60	81.80
P4Transformer [23]	1	96.76	98.23	92.11	95.23	98.62	97.77	95.46	80.75	85.48	0.00	74.28	74.22	82.41
P4Transformer [23]	3	96.73	98.35	94.03	95.23	98.28	98.01	95.60	81.54	85.18	0.00	75.95	79.07	83.16
PointNet4D	3	97.23	98.35	94.73	95.93	99.08	98.11	97.60	81.04	87.18	0.00	76.90	77.97	83.67
4DMAP	3	97.30	98.31	95.13	96.79	99.35	98.16	97.91	80.88	89.60	0.00	78.01	77.69	84.10

Table 4. Evaluation for semantic segmentation on Synthia 4D.

More Details of 4D Imitation Learning(4DIL)

We also validated the significant potential of PointNet4D within an imitation learning framework, using the HandoverSim[24] benchmark for evaluation. All settings were kept consistent with the default configuration of HandoverSim. The Sequential setting represents the scenario where the robot performs the grasping action after the human hand has remained stationary. This setup does not require 4D information, as the hand and object states do not change once they are stationary, relying solely on 3D perception.



(a) 4D Imitation Learning used in GenH2R (b) 4D Imitation Learning based on PointNet4D

Figure 1. Comparison of our 4DIL with GenH2R [25].

The Simultaneous setting, on the other hand, involves both the human hand and the object in motion during the robot's grasping action, making this setup significantly more challenging than the Sequential one. We reproduced the results of GenH2R in the Simultaneous setting and confirmed them with the authors. To ensure a fair comparison, we used the same codebase and replaced PointNet++ with PointNet4D in the GenH2R model, using a 3-frame time window. We report the results of our method in the more challenging Simultaneous setting.

References

[1] Hao Wen, Yunze Liu, Jingwei Huang, Bo Duan, and Li Yi. Point primitive transformer for long-term 4d point cloud video understanding. In *European Conference on Computer Vision*, pages 19–35. Springer, 2022.

[2] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4D transformer networks for spatio-temporal modeling in point cloud videos. In *CVPR*, 2021.

[3] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010.

[4] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[5] Quentin De Smedt, Hazem Wannous, Jean-Philippe Vandeborre, Joris Guerry, Bertrand Le Saux, and David Filliat. SHREC'17 Track: 3D hand gesture recognition using a depth and skeletal dataset. In *3DOR*, 2017.

[6] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network. In *CVPR*, 2016.

[7] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[8] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 3DV: 3D dynamic voxel for action recognition in depth video. In *CVPR*, 2020.

[9] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. PSTNet: Point spatio-temporal convolution on point cloud sequences. In *ICLR*, 2021.

[10] Hehe Fan, Xin Yu, Yi Yang, and Mohan Kankanhalli. Deep hierarchical representation of point cloud videos via spatio-temporal decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9918–9930, 2021.

[11] Jia-Xing Zhong, Kaichen Zhou, Qingyong Hu, Bing Wang, Niki Trigoni, and Andrew Markham. No pain, big gain: classify dynamic point cloud sequences with static models by fitting feature-level space-time surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8510–8520, 2022.

[12] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point spatio-temporal transformer networks for point cloud video modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2181–2192, 2022.

[13] Xiaoxiao Sheng, Zhiqiang Shen, Gang Xiao, Longguang Wang, Yulan Guo, and Hehe Fan. Point contrastive prediction with semantic clustering for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16515–16524, 2023.

[14] Zhiqiang Shen, Xiaoxiao Sheng, Longguang Wang, Yulan Guo, Qiong Liu, and Xi Zhou. Pointcmp: Contrastive mask prediction for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1212–1222, 2023.

[15] Xiaoxiao Sheng, Zhiqiang Shen, and Gang Xiao. Contrastive predictive autoencoders for dynamic point cloud self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9802–9810, 2023.

[16] Zhiqiang Shen, Xiaoxiao Sheng, Hehe Fan, Longguang Wang, Yulan Guo, Qiong Liu, Hao Wen, and Xi Zhou. Masked spatio-temporal structure prediction for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16580–16589, 2023.

[17] Yuehui Han, Can Xu, Rui Xu, Jianjun Qian, and Jin Xie. Masked motion prediction with semantic contrast for point cloud sequence learning.

[18] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. MeteorNet: Deep learning on dynamic 3D point cloud sequences. In *ICCV*, 2019.

[19] Yuecong Min, Xiujuan Chai, Lei Zhao, and Xilin Chen. FlickerNet: Adaptive 3D gesture recognition from sparse point clouds. In *BMVC*, 2019.

[20] Yuecong Min, Yanxiao Zhang, Xiujuan Chai, and Xilin Chen. An efficient PointLSTM for point clouds based gesture recognition. In *CVPR*, 2020.

[21] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019.

[22] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017.

[23] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14204–14213, 2021.

[24] Yu-Wei Chao, Chris Paxton, Yu Xiang, Wei Yang, Balakumar Sundaralingam, Tao Chen, Adithyavairavan Murali, Maya Cakmak, and Dieter Fox. Handoversim: A simulation framework and benchmark for human-to-robot object handovers. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6941–6947. IEEE, 2022.

[25] Zifan Wang, Junyu Chen, Ziqing Chen, Pengwei Xie, Rui Chen, and Li Yi. Genh2r: Learning generalizable human-to-robot handover via scalable simulation demonstration and imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16362–16372, 2024.