Mamba4D: Efficient 4D Point Cloud Video Understanding with Disentangled Spatial-Temporal State Space Models

Supplementary Material

1. Overview

The supplementary materials are structured as follows:

- We first analyze the reasons why our Mamba-based architecture can achieve better performance than its transformer-based counterparts in Section 2.
- We give more detailed illustrations about the dataset for 4D tasks in Section 3;
- More descriptions about the loss functions are provided in Section 4.
- Additional experiments about network settings and ablation studies are provided in Section 5.
- We analyze the prediction accuracy error bar in Section 6 and display more visualization results in Section 7.

2. Theoretical Analysis

Here, we give a theoretical analysis why our proposed Mamba4D can achieve such excellent performance in memory consumption, inference time, and numerical stability, compared to transformer-based counterparts.

a) Mamba eliminates quadratic attention storage ($\mathcal{O}(N^2d) \rightarrow \mathcal{O}(Nd) + \mathcal{O}(d^2)$), leading to 87.5% lower memory consumption by avoiding redundant attention maps. b) For Transformers, attention decays over distance. Mamba propagates global dependencies via a structured state-space recurrence $h_t = Ah_{t-1} + Bx_t$, enabling 5.36× faster inference. c) Numerical stability: Transformers suffer from softmax-induced vanishing gradients, while Mamba's state-space formulation maintains stable gradient flow, avoiding costly stabilization techniques. Stability is a core issue in the optimization view in scalability.

3. Datasets

MSR-Action3D. The MSR-Action3D dataset [4] is composed of 567 Kinect depth videos, including 20 action categories and 23K frames in total. We partition the train/test split following [2, 3], and sample 2048 points for each frame. Only point coordinates are available without point colors. Point cloud videos are partitioned into multiple equal-size clips. Video-level labels are directly used as clip-level labels when training. For testing, the mean of clip-level predicted probabilities is viewed as the video ones.

HOI4D. The HOI4D dataset [5] contains 2,971 training videos and 892 test videos for action segmentation. Each video sequence has 150 frames with each frame containing 2048 points. The dataset contains a total of 579K frames. All frames are annotated with 19 fine-grained action classes in the interactive scene.

Synthia 4D. The Synthia 4D [1] dataset is generated from the Synthia dataset [6], including 6 driving scene videos. Each video consists of 4 stereo RGB-D images captured from the top of the car. 3D point cloud videos are obtained from RGB and depth images. We follow [2] to split the training (19888 frames)/ validation (815 frames)/ test (1886 frames) sets. The evaluation metric is the mean Intersection over Union (mIoU).

4. Loss Functions

3D Action Recognition. In this task, the model is trained to classify a sequence of video frames into predefined action categories. The primary loss function employed is the Cross-Entropy Loss, which is defined as follows:

$$\mathcal{L} = -\Sigma_{i=1}^{N} y_i \log(p_i), \tag{1}$$

where N is the total number of videos, y_i is the true label for the i-th class and p_i is the predicted probability for the i-th class. The loss function supervises the model by providing a scalar value that quantifies the discrepancy between the predicted action probabilities and the true action labels. During the training process, the model parameters are optimized to minimize this loss value.

4D Action Segmentation. In the task of 4D action segmentation, the goal is to classify frames in a point cloud sequence into action categories. The primary loss function employed is the Cross-Entropy Loss, defined as:

$$\mathcal{L} = -\Sigma_{i=1}^{N} y_i \log(p_i), \tag{2}$$

where N represents the total number of frames in the point cloud sequence, y_i is the true action label for the i-th frame, and p_i is the predicted probability for the corresponding action class. This loss function measures the prediction error between the predicted and ground truth action labels across the point cloud sequence frames.

4D Semantic Segmentation. In the task of 4D semantic segmentation, the goal is to classify each point in a sequence of point clouds into semantic categories. The primary loss function employed in this task is the weighted Cross-Entropy Loss, which is defined as follows:

$$\mathcal{L} = -\Sigma_{i=1}^{N} w_i y_i \log(p_i), \tag{3}$$

where N is represents the total number of points in the point cloud, w_i is the weight for the i-th class to handle class imbalance, y_i is the true label for the i-th point, and the p_i is

Table 1. Efficiency gains by replacing CNN or Transformer in P4Transformer [2] with Mamba. ✓means replacing with Mamba.

Intra	Inter	Acc (%)	Speed (ms)	GPU (G)
CNN	Trans	90.94	154	2.1
✓	Trans	91.36	202	2.3
CNN	✓	91.67	**56.7**	**1.5**
✓	✓	**92.68**	102	1.8

Table 2. Comparison between disentangling or unifying spatial and temporal modeling. Both two methods have the same accuracy, but the disentangling one has higher efficiency.

Modeling	Acc (%)	Speed (ms)	Tokens
Disentangling	**92.68**	**102.4**	**768**
Unifying	**92.68**	154.0	1536

Table 3. Ablation studies on different anchor strategies. For short-term modeling, smaller stride is better for capturing rapid movements.

Method	Acc (%)	Method	Acc (%)
Fixed (stride=2)	**92.68%**	Fixed (stride=4)	89.19%
Fixed (srtide=8)	88.50%	Multi-scale (stride=2,4,8)	90.24%

the predicted probability for the i-th class. The loss function supervises the 4D semantic segmentation task by providing a measure of the prediction error between the predicted semantic labels and the true labels for each point in the point cloud.

5. Additional Experiments

Efficiency Quantification. We quantify individual efficiency gains by replacing each component with Mamba in Table 1. Both replacements largely increase accuracy. However, replacing CNN with Mamba would degrade efficiency, which is partially mitigated by replacing Transformer with Mamba. The final model is a trade-off between accuracy and efficiency.

Ablation Studies in Spatial-Temporal Modeling. We follow [3] to disentangle the spatial and temporal dimensions. We conduct ablation studies by comparing different spatio-temporal modeling methods: disentangling or unifying. From Table 2, both methods have the same recognition accuracy, but the disentangling one has a higher efficiency.

Ablation Studies in Fixed or Adaptive Anchors. We add ablation studies by extended anchor intervals or replacing fixed anchors by adaptive multi-scale anchors in Table 3. For the effective short-term modeling, smaller stride is better for capturing rapid movements.

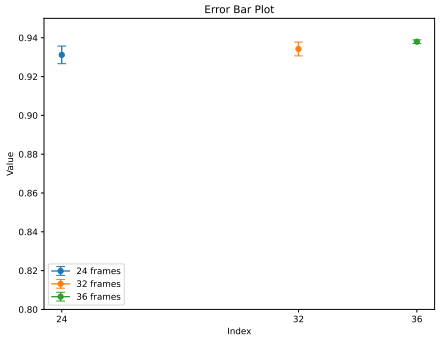

Figure 1. **Error bars of our estimated action recognition accuracy for 24, 32, and 36 frames as inputs.** We can see a stable performance with a small fluctuation around the mean accuracy.

6. Error Bar Analysis

We plot the error bar on the action recognition accuracy for 24, 32, and 36 frames as input in Fig. 1. From the figure, we can see stable performance with a small fluctuation around the mean accuracy.

7. Visualization

We show more visualization results in Fig. 2 and Fig. 3 respectively for the 4D action segmentation and semantic segmentation. In Fig. 3, all the predicted segmentation labels are highly overlapped with the Ground Truth, which shows the perfect accuracy of our method.

References

[1] 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. 1

[2] 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. 1, 2

[3] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. In *International Conference on Learning Representations*, 2021. 1, 2

[4] 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. 1

[5] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 21013–21022, 2022. 1

[6] 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. 1

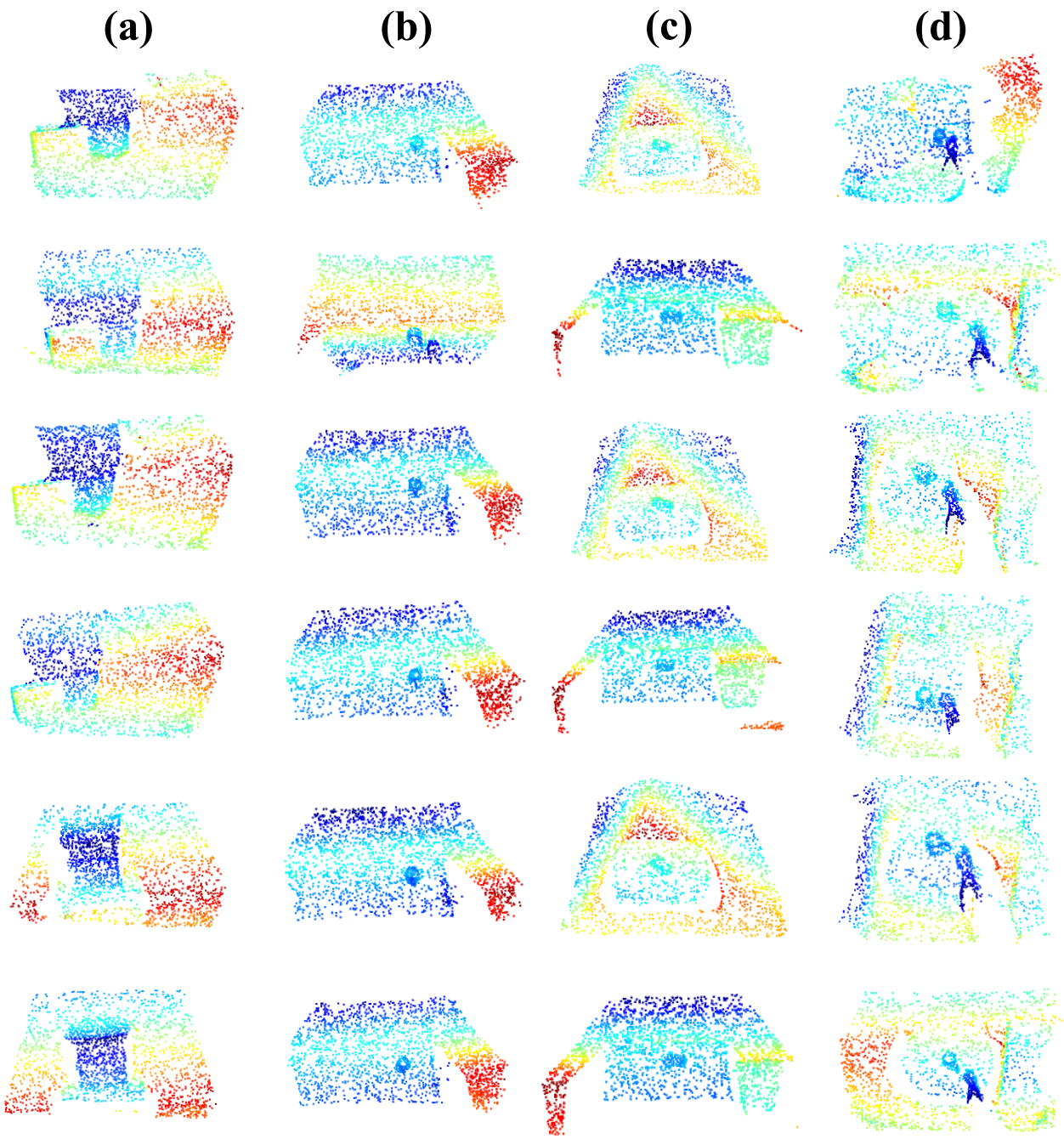

Figure 2. **More visualization samples of the 4D action segmentation.**

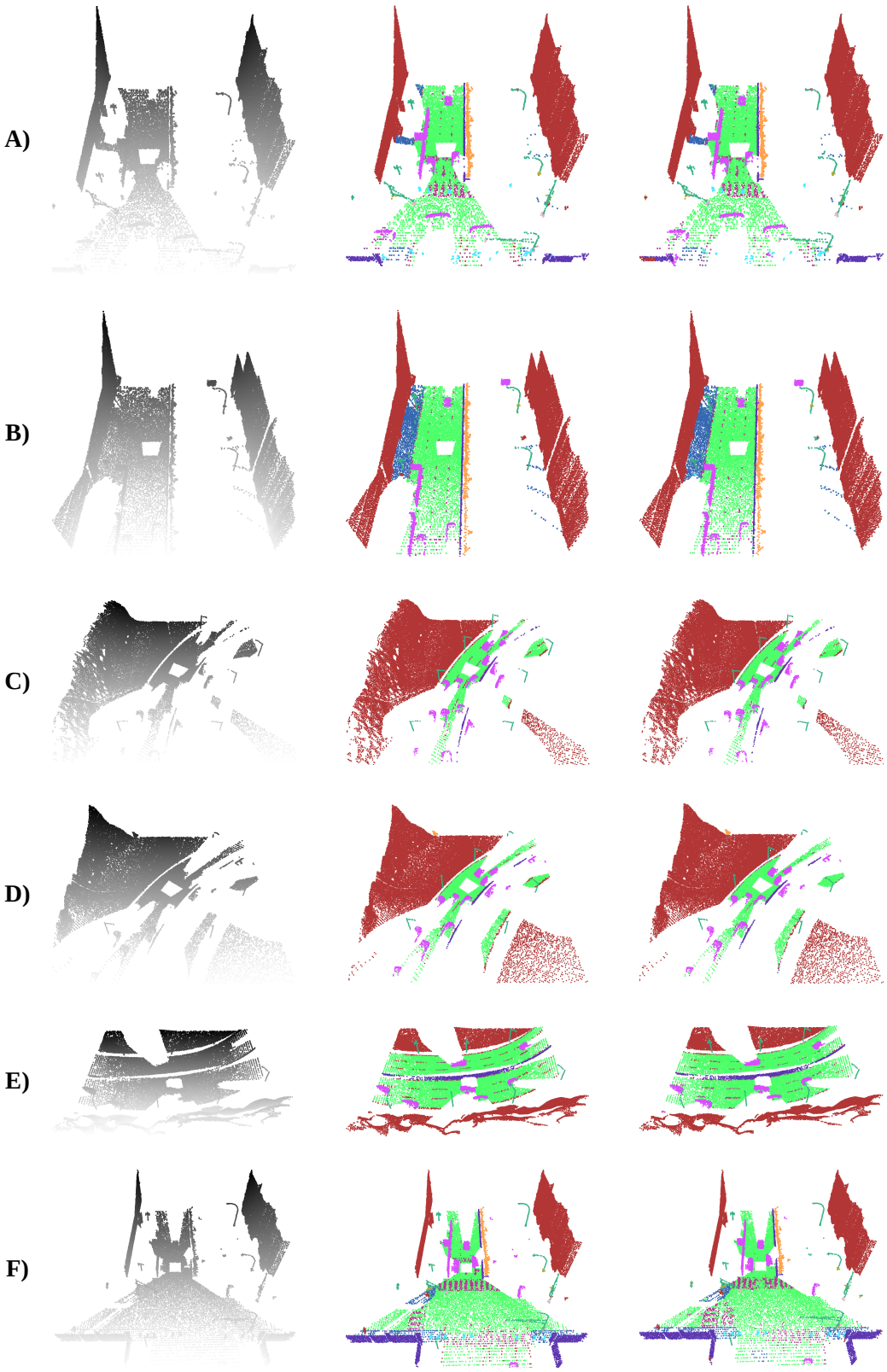

Figure 3. **More visualization samples of the 4D semantic segmentation.**