

Supplementary Materials: Towards Practical Human Motion Prediction with LiDAR Point Clouds

Anonymous Authors

A OVERVIEW

In the supplementary materials, we first present visualizations of our method’s performance under challenging conditions, including occlusion and noise, to demonstrate its generalization capabilities and robustness. Next, we showcase the results of our diverse human motion prediction method, highlighting the unpredictability and diversity of the future human motions, and illustrating the significance of diverse predictions for more accurate decision-making. In addition, we also present the *real-time deployment and real-time results of our approach in the video of our supplementary materials*. Finally, we present the specifics of the pretrained LiDAR-based human parsing synthetic dataset, outlining the training methodologies and outcomes.

B MORE VISUALIZATION RESULTS

B.1 Occlusion and Noise Cases

In this section, we show the results of our method under occlusion and noise, as in Figure. 1 and Figure. 2. Our methods can still accurately forecast the future human motions even in such challenging cases, further highlighting the robustness and generalization ability of our approach.

B.2 Diverse Human Motion Prediction

Given the inherent spontaneity and unpredictability of human behavior, we extend our framework to diverse human motion predictions, covering up to four potential future motions. Here, we shown four different motion prediction results according to the same point cloud observations in Figure 3.

B.3 Real-world Applications

We show more real-world demos in the video of our supplementary materials to demonstrate the significance of our approach for real-world applications.

C LIDAR-BASED HUMAN PARSING

C.1 Synthetic Data

We train our LiDAR-based human parsing model on a synthetic dataset created using ray casting on human mesh models from the AMASS [1] dataset. Detailed steps of synthetic data generation are outlined below. We generated 700,000 frames for training and 300,000 frames for validation. To simulate LiDAR point clouds, human meshes are placed at distances ranging from 6 to 27 meters from a ray caster, using 2048 vertical scans (covering 360 degrees) and 128 LiDAR beams. Each beam’s emission direction is defined by a unit vector in the spherical coordinate system, $d = [\cos \varphi \sin \theta, \cos \varphi \cos \theta, \sin \varphi]$, where φ is the angle from the emission direction to the XY plane and θ represents the azimuth. The LiDAR center is $c = [0, 0, 2]$. The intersection point

$p = [p_x, p_y, p_z]$ is computed using:

$$p = c + d \frac{n^T (q - c)}{n^T d}, \quad (1)$$

where n is the normal vector of the corresponding mesh and q denotes any vertex point of the mesh. To bridge the gap between synthetic data and real LiDAR scans, random occlusions and noise are incorporated. The SMPL mesh vertices, which are known for their ordered and regular structure, normally provide 24 human body part labels. Due to the sparsity of LiDAR point clouds, we have simplified these to 9 primary categories: head, left arm, right arm, upper body, lower body, upper left leg, upper right leg, lower left leg, and lower right leg. Each LiDAR point is automatically labeled with the nearest vertex’s body part label, and randomly added noises are labeled as "noise".

C.2 Training Details and Results

The objective of the LiDAR-based human parsing task is to assign a label to each point, indicating its correspondence to a specific body part or noise, functioning as a form of segmentation. To achieve this, we implement the state-of-the-art object part segmentation method, PointNext [2], training it on 700,000 simulated frames of human LiDAR point clouds. We adhere to the training strategy recommended by PointNext, which includes data augmentation and training procedures. The experimental results of this method on a validation dataset of 300,000 simulated LiDAR frames is detailed in Table 1. Additionally, we also provide visualizations of our pretrained LiDAR-based human parsing on real LiDAR scans, illustrated in Figure 4.

REFERENCES

- [1] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5442–5451.
- [2] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems* 35 (2022), 23192–23204.

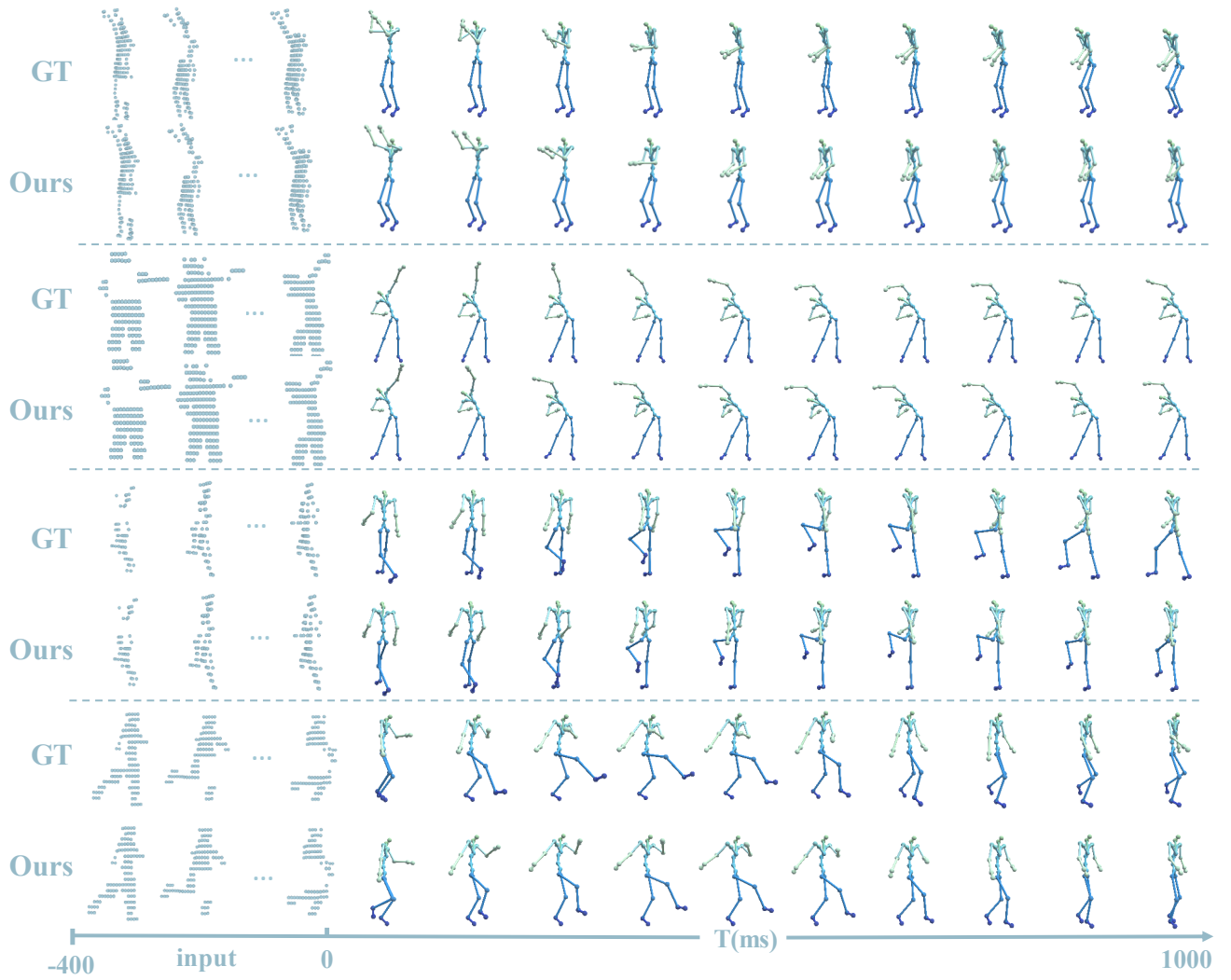


Figure 1: Visualisation of the results in the occlusion case demonstrates the robustness of our LiDAR-based human motion prediction approach even in scenarios with occlusions. “GT” denotes the future ground truth skeletons.

Table 1: LiDAR-based human parsing results on the validation set of our synthetic data.

head	left arm	right arm	upper body	lower body	upper left leg	lower left leg	upper right leg	lower right leg	noise	miou
94.10	90.77	90.25	82.66	91.25	90.67	95.27	90.66	95.48	97.83	91.89

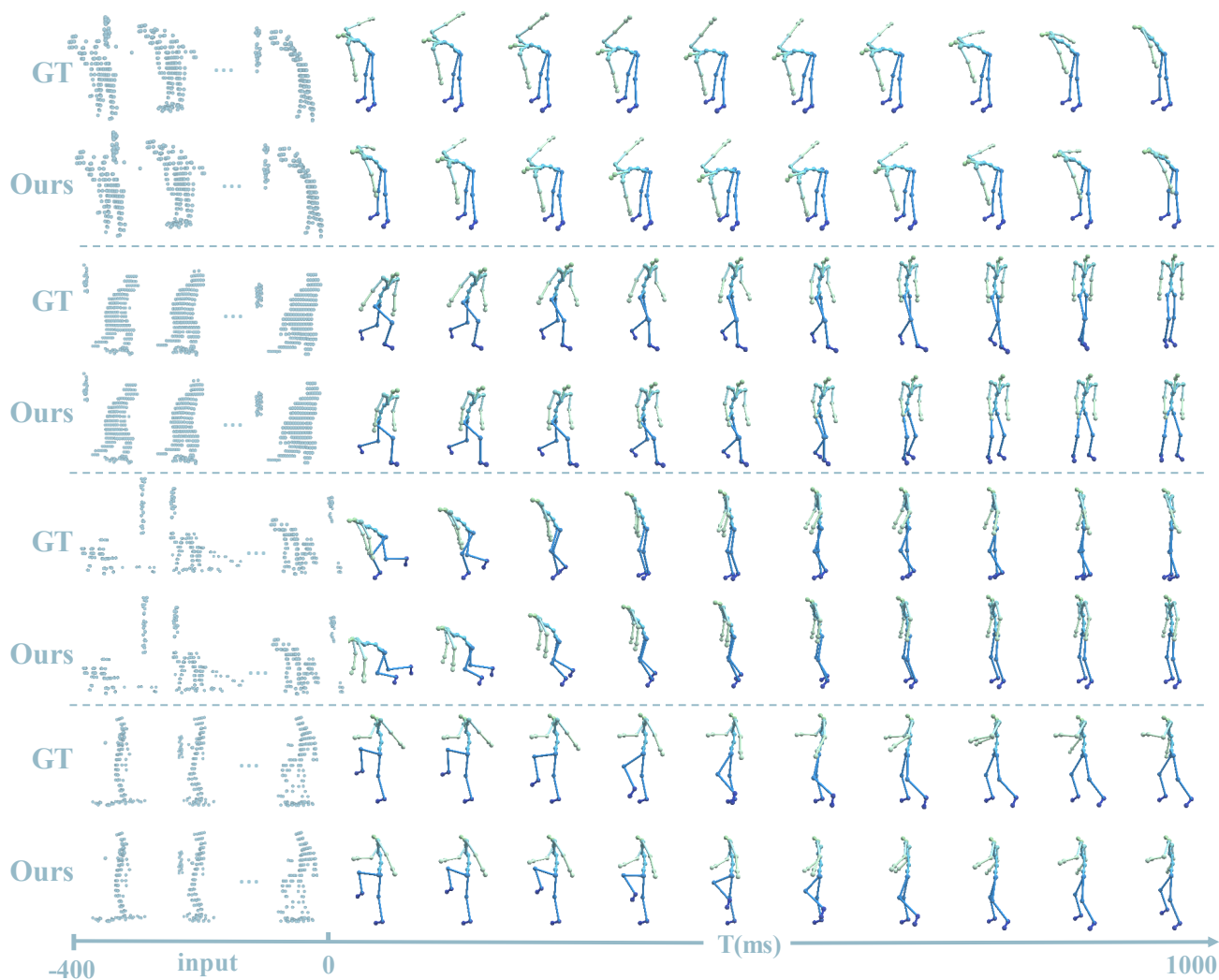


Figure 2: Visualisation of the results in the noise case demonstrates the robustness of our LiDAR-based human motion prediction approach even in noisy environments. “GT” indicates the future ground truth skeletons.

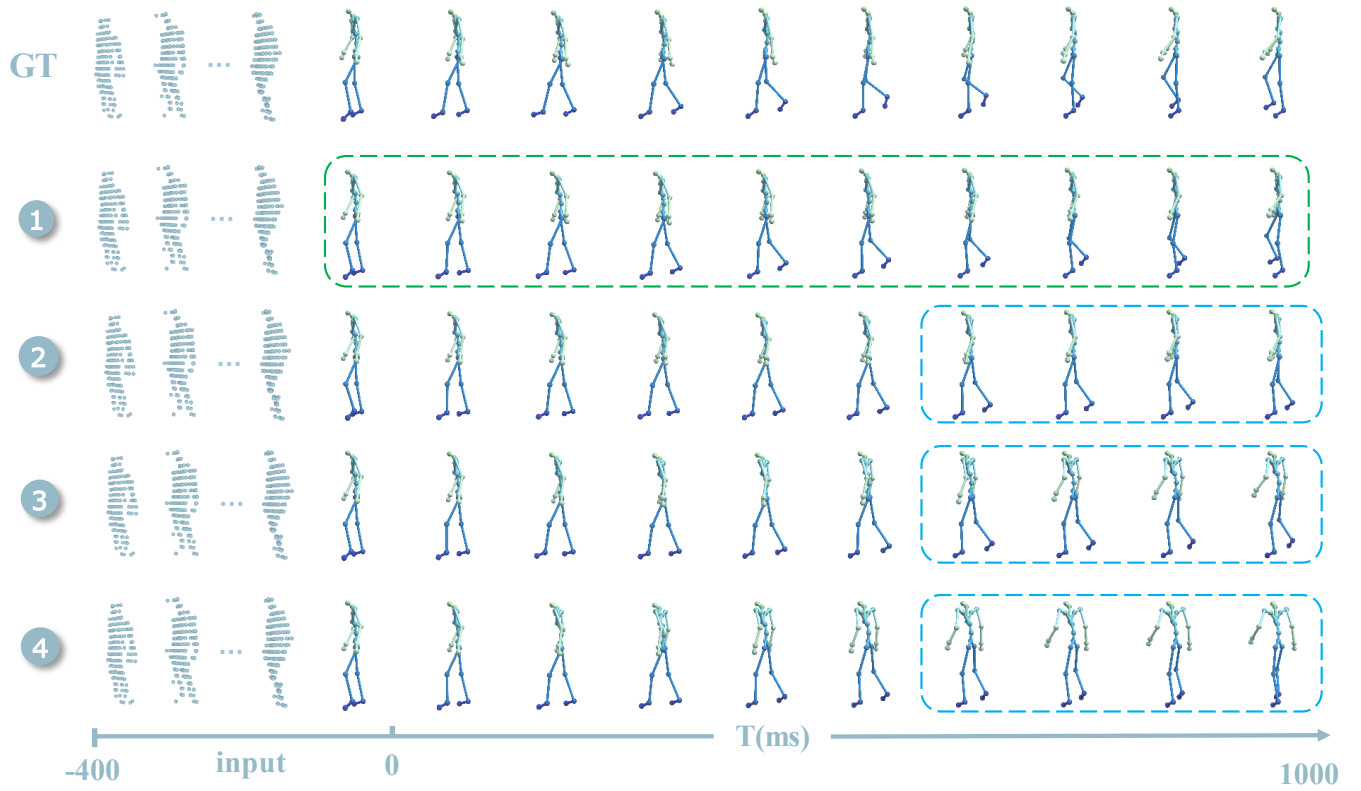


Figure 3: Visualization results of our extended diverse human motion prediction on LIPD. “GT” refers to the future ground truth skeletons. Using the same historical input, we display four different future motions predicted by our network simultaneously. The predictions in the green dashed box are closest to the ground truth, while the other three in the blue dashed box, though varying from the ground truth, depict plausible alternatives where a person might move forward or turn at different speeds.

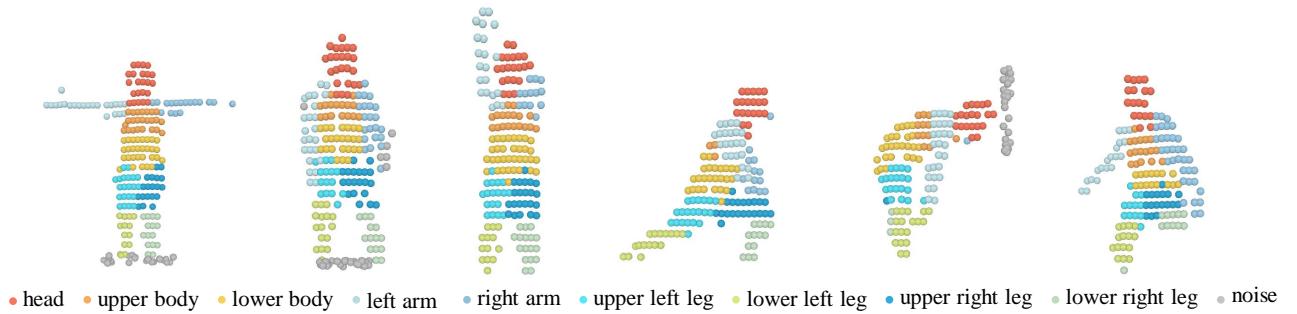


Figure 4: Visualisation of our pretrained LiDAR-based human parsing network applied to real LiDAR scans.